The University of Sheffield

Learning Attention Mechanisms and Context: An Investigation into Vision and Emotion

Md Asif Jalal

Supervisor: Prof. Roger K Moore

A report submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science

in the

Department of Computer Science

April 30, 2021

Declaration

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: ______________________________

Signature: ______________________________

Date: ______________________________

Acknowledgement

Alhamdulillah! for everything in my life. I have been fortunate for having a wonderful family, great teachers, fantastic colleagues and some precious friends. As I am on the verge of finishing my PhD thesis, I look back at these exciting years and feel immensely grateful for the contribution of these people in my life.

I wish to thank my parents and sister for their unconditional love and selfless criticism to teach me the value of continuous learning and self-betterment. My father first introduced me to the deconstruction of ideas into cogent pieces and built upon those pieces. My mother taught me to be perseverant to accomplish goals. I am eternally grateful. To my little sister: you make me a better person in ways you would not realise, and I can not imagine a better sister or a happier childhood.

I want to thank my supervisor Prof. Roger K Moore, who granted me the research opportunity with his careful supervision. But most importantly, I am immensely grateful because he taught me to be a researcher and thinker. Furthermore, he made me an independent thinker by encouraging me to push my limits and providing me with the luxury of trying and failing. I will always cherish our discussions. I want to thank Prof. Thomas Hain for his guidance, who taught me and encouraged me to question every minute details and explore persistently to build a better understanding of machine learning. I am grateful for getting the research opportunity to collaborate with him. I want to thank Prof. Lyudmila Mihaylova for her kind guidance and support. I want to thank Prof. Eleni Vasilaki and Prof. Guy J Brown for their thoughtful advice throughout my PhD study. I am grateful to all my reviewers for reviewing my research. I want to thank Dr Heidi Christensen and Prof. Bjoern W. Schuller for reviewing my thesis.

I would not be able to complete my PhD journey without Dr Erfan Loweimi and Dr Rosanna Milner. I want to thank Erfan, the friend, philosopher, guide and brother, for continually encouraging me and showing me the direction when I lose motivation. I want to thank Rosanna for being so supportive and inspirational, and for teaching me to be organised and thorough in my research. I want to thank Ms Anna Ollerenshaw, a wonderful human being, friend, labmate and colleague, for her constant positive encouragement and help. I am also grateful to her for reviewing my research with thoughtful comments. I wish to thank Dr Rabab Algadhy, Dr Hardik Sailor and Dr Mauro Nicolao for their precious advice and thoughtful discussions throughout my PhD study. I am grateful to all my SPandH and MINI group mates for being so supportive and kind.

I am lucky to be surrounded by some wonderful childhood friends who are with me through thick and thin. I want to thank my dearest friends Dr Sounak Bepari, Mr Aniruddha Roy, Mr Sandip Mahato, Dr Dibyayan Chakraborty, Dr Arka Mallick, and Mr Arghya Biswas for all these exciting years of fun, adventures, healthy criticisms, support and genuine friendship. I want to thank my two remarkable cousins Mr Md Kader and Ms Jinat Aktar, for their selfless guidance and love.

Finally, I want to thank Saba, the love of my life, and a source of motivation, for tolerating all of my frivolity and nonsense. I am grateful for your grace and simplicity.

Abstract

Attention mechanisms for context modelling are becoming ubiquitous in neural architectures in machine learning. The attention mechanism is a technique that filters out information that is irrelevant to a given task and focuses on learning task-dependent fixation points or regions. Furthermore, attention mechanisms suggest a question about a given task, i.e. 'what' to learn and 'where/how' to learn for task-specific context modelling. The context is the conditional variables instrumental in deciding the categorical distribution for the given data. Also, why is learning task-specific context necessary? In order to answer these questions, context modelling with attention in the vision and emotion domains is explored in this thesis using attention mechanisms with different hierarchical structures. The three main goals of this thesis are building superior classifiers using attention-based deep neural networks (DNNs), investigating the role of context modelling in the given tasks, and developing a framework for interpreting hierarchies and attention in deep attention networks. In the vision domain, gesture and posture recognition tasks in diverse environments, are chosen. In emotion, visual and speech emotion recognition tasks are chosen. These tasks are selected for their sequential properties for modelling a spatiotemporal context. One of the key challenges from a machine learning standpoint is to extract patterns which bear maximum correlation with the information encoded in its signal while being as insensitive as possible to other types of information carried by the signal. A possible way to overcome this problem is to learn task-dependent representations. In order to achieve that, novel spatiotemporal context modelling networks and the mixture of multi-view attention (MOMA) networks are proposed using bidirectional long-short-term memory network (BLSTM), convolutional neural network (CNN), Capsule and attention networks. A framework has been proposed to interpret the internal attention states with respect to the given task. The results of the classifiers in the assigned tasks are compared with the *state-of-the-art* DNNs, and the proposed classifiers achieve superior results. The context in speech emotion recognition is explored deeply with the attention interpretation framework, and it shows that the proposed model can assign word importance based on acoustic context. Furthermore, it has been observed that the internal states of the attention bear correlation with human perception of acoustic cues for speech emotion recognition. Overall, the results demonstrate superior classifiers and context learning models with interpretable frameworks. The findings are very important for speech emotion recognition systems. In this thesis, not only better models are produced, but also the interpretability of those models are explored, and their internal states are analysed. The phones and words are aligned with the attention vectors, and it is seen that the vowel sounds are more important for defining emotion acoustic cues than the consonants, and the model can assign word importance based on acoustic context. Also, how these approaches for emotion recognition using word importance for predicting emotions are demonstrated by the attention weight visualisation over the words. In a broader perspective, the findings from the thesis about gesture, posture and emotion recognition may be helpful in tasks like human-robot interaction (HRI) and conversational artificial agents (such as Siri, Alexa). The communication is grounded with the symbolic and sub-symbolic cues of intent either from visual, audio or haptics. The understanding of intent is much dependent on the reasoning about the situational context. Emotion, i.e. speech and visual emotion, provides context to a situation, and it is a deciding factor in the response generation. Emotional intelligence and information from vision, audio and other modalities are essential for making human-human and human-robot communication more natural and feedback-driven.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"The human brain, then, is the most complicated organisation of matter that we know."* - Isaac Asimov

The complex phenomenon of the human mind, built on the physical organisation of the brain, is governed by thoughts (cognition) and feelings (emotion). The mind can accumulate information from a multitude of sources with multiple modalities and conduct the information in different ways to learn knowledge and achieve different goals flexibly, which is a very intricate process, and to this very day, much is unknown. The accumulated information is entangled and too complicated to be processed quickly, and therefore needs some form of filtering and selection mechanism to process the information efficiently. The brain uses the attention mechanism to filter goal-oriented inference and bias in order to learn contextual cues dynamically. Attention is a standalone mechanism, which is used on spatial information, time, sensory modalities and goals (Posner & Petersen (1990*a*), Posner (2012)).

The biological neural basis for attention has been an inspiration for the computational neural network modelling of attention, and since then, attention has become increasingly popular in machine learning. The computational attention models try to mimic brain attention mechanisms by focusing on fixation points as 'glimpses' and combining those 'glimpses' to learn meaningful task-specific context cues (Larochelle & Hinton (2010), Mnih et al. (2014)). In machine learning, learning a specific context for a given task or goal is a big undertaking due to the sub-symbolic abstract nature of the neural networks. The context cues are the conditional variables instrumental in deciding the categorical or dimensional distribution for a given task and the given data. The primary motivation of this research is to investigate the role of context and attention learning and to be able to interpret the sub-symbolic abstract representations of these learned attributes empirically. Vision and emotion information is used in this research to study the role of context and attention mechanisms in computational models.

"All learning has an emotional base." - Plato

The astounding revelation from Plato indicated the relation between human learning and emotional response two thousand years ago. Since then, thousands of interdisciplinary research and experiments have been conducted to investigate this phenomenon. Sartre (1962) referred to emotion as this natural 'magical being', that spontaneously affects the conscious reasoning of our worldview as a 'magical' transformation of the situation. By ipso facto, the influence of emotions in learning among humans is crucial. Emotion understanding is a multimodal and interdisciplinary topic. Emotional intelligence is essential for understanding meaning due to the dependency of meaning on the context, and this turns it into a complicated signal; encoding a large amount of information that can be categorised into lingual content, speaker-dependent attributes and environmental clues. Emotion is among the speaker-related information and plays an essential role in human-human communication.

The most natural way of communication between humans is language. A language is a symbolic form of communication and the union of sounds, words, sentences and gestures. Language captures an infinitely huge possibility of meanings. Language can be perceived and understood with acoustic cues and visual cues. Visual cues are natural and have some unique advantages over the other communication mediums (Fong et al. (2003)). Visual cues can convey a universal language that can serve the fundamental purpose of communication. Gestures and postures construct universal visual language, and task-specific gestures/postures can also be designed without the help of any particular hardware (sign language sensor gloves) or any special spoken language training (Kalpagam Ganesan et al. (2018)). Language is not perceived solely by a stream of sounds or gestures, and before understanding a language, one needs to ground the language. The communication is grounded with the symbolic and sub-symbolic cues of intent either from visual, audio or haptics. The understanding of intent is greatly dependent on the reasoning about the situational context. The meaning and intent can be propositional or inferential. The situational context can come from both spatiotemporal information and emotional reasoning about the spatiotemporal information. Furthermore, emotion classification raises the investigation about 'what' is said and 'how' / 'where' it is said. Therefore learning context with spatial and temporal dependencies is very important. In this research, attention mechanisms are used to learn the spatiotemporal context for recognising visual and emotional cues.

The ground truth is hugely dependant on the observers (individual perception), who associate the acoustic cue patterns with discrete emotion states. However, emotion boundaries cannot be clearly defined. Emotion perception also varies among different persons and cultures. Therefore, some questions arise about these acoustic cues concerning the acoustic phone boundaries and the lengths of these cues. This research investigates these question with attention models and an attention alignment framework for speech emotion recognition.

Although emotion recognition and context-based learning are a very sophisticated issue, the limitations and the protocols should be understood for applying these in practice. Picard (2000) said,

> *"Without emotion, computers are not likely to attain creative and intelligent behaviour, but with too much emotion, we, the maker, may be eliminated by our creation."*

In this era of machine intelligence, we are surrounded by physical and virtual robots or agents where human-robot / human-virtual agent interaction is becoming prevalent as human-human interaction. Robots are being used in various social scenarios such as: personal use (Jones & Schmidlin (2011)), education (Saerbeck et al. (2010)), medical and assistive agents (Robins et al. (2009)). One of the main goals of human-robot spoken language interaction is to make robots intelligent enough to communicate and work with humans in a human-like manner. Robots and artificial agents have been envisioned as social companions for humans. Overall, it is crucial to understand how emotion recognition may be used in human-robot interaction. Therefore, future research can move into the direction of building an artificial social companion that mimics the essential traits of human communication, using emotional feedback to help humans in different tasks and aid the human caregiver.

As the use of speech-driven user interfaces such as *Siri*, *Alexa*, *Google assistant*, *Bixby* increases rapidly in daily life, a lack of emotional intelligence is becoming more evident. To this end, among others, these virtual agents should become capable of distinguishing emotions. It is highly desirable to add this critical dimension to these virtual agents. This research presents the scope of emotion recognition and emotion perception with interpretable speech-based computational attention models. The thesis also presents vision-based gesture/posture recognition and vision-based emotion recognition models proposed in this research using deep neural networks and attention networks.

When the computational models learn to recognise visual and emotional cues, it may reduce the perceptual difference between human and artificial agents for human-robot interaction (HRI). I have briefly discussed how the research about vision and emotion may help HRI. If we do not perceive the meaning of language based on situations, we might not understand the situation appropriately. Emotion, i.e. speech and visual emotion, provides context to a situation, and it is a deciding factor in the response generation. Emotional intelligence and information from vision, audio and other modalities are essential for making human-human and human-robot communication more natural and feedback-driven.

1.1 Research Questions

The excerpt from the previous discussion states that learning and communicating employ context-dependent attention mechanisms. The attention mechanisms focus on contextual cues, i.e. visual cues, speech acoustic cues. However, the nature of these contextual cues is not very clear computationally. For example, acoustic phones and sounds comprise speech acoustic cues, but it is unclear how they are decided or the boundaries and lengths chosen for these phones/ sounds. Furthermore, it is also challenging to interpret and visualise these attention context cues computationally. Computationally, gesture and posture recognition systems face many challenges regarding spatial and temporal modelling and dependencies between themselves. Some research questions arise from these premises.

1. How visual cue learning with attention can overcome the challenges of gesture and posture modelling? Does hierarchy in the network play a role in learning

attention? In order to discuss that, the hierarchical capsule context and attention learning are presented in Chapter 3 and 4.

2. How context learning for speech emotion can make better speech emotion classifiers and how the nature of context representation differs over different neural network models? Is there some form of universal representation of emotion context among different variations of speech utterances in the same language? These are discussed, and the experiments are presented in Chapter 5, 6 and 7 with the speech emotion recognition models and crosscorpus training. In crosscorpus training, elicited, acted, and natural speech emotion data is used to determine if there is a common representation of emotion context over these various corpora.

3. How the acoustic cue length for context representation determines speech emotion recognition? What are the nature and phonemic boundaries of these speech emotion context cues? These questions are explored with computational modelling and visualisation in Chapter 7.

4. How acoustic emotion context may aid an artificial agent for human-robot interaction? These are briefly discussed in Chapter 8.

1.2 Contributions

In order to explore the above-presented questions, I have made some contributions during this research. These are

- A sign language framework using capsules and adaptive pooling has been proposed for recognising American sign language (Presented in Chapter 3, Section 3.2). This framework uses state-of-the-art capsule networks with adaptive pooling for training variable resolution images with quick convergence training and better noise robustness.

- A vehicle logo recognition model has been proposed, and the model is challenged with distortion, noise, rotation challenges, and empirically, it is proved to be robust. (Presented in Chapter 3, Section 3.3)

- Dual learning and dual fusion techniques for action recognition have been investigated, and different stages of attention fusion are performed to build an action recognition framework (Presented in Chapter 4, Section 4.2). Dual learning is discussed in Chapter 2, and it states the interdependency of spatial and temporal dependency. The research also used data dependent modelling using data augmentation in the training process.

- Temporal representation learning in capsules for speech emotion recognition is proposed with a novel hybrid temporal capsule network. The intermediate capsule representations have been analysed (Presented in Chapter 5, Section 5.3). This research shows the temporal clusters over different hierarchies in capsule networks and proposes a temporal capsule based speech emotion recognition model.

- A spatiotemporal context modelling framework is proposed to model bias and context in the samples for speech emotion recognition (Presented in Chapter 5, Section 5.3).

- A hierarchical residual mixture of multi-view attention model (MOMA) has been proposed for speech emotion classification (Presented in Chapter 6, Section 6.2). This research extensively investigates the hierarchical representation of attention for speech emotion recognition by mapping attention weights with acoustic phone units, which gives better interpretability of the model.

- The intermediate attention representations and clusters in hierarchical models have been analysed (Presented in Chapter 6,7, Section 6.2,7.2).

- The effects of linguistic units to build the perception of attention models for speech emotion recognition are explored with a novel empirical framework (Presented in Chapter 7). The vowel-consonant phoneme boundaries have been analysed and visualised for speech emotion context learning.

- The universality of emotions across different speech emotion corpora in the English language (also between acted, elicited and natural speech emotion) is discussed. (Presented in Chapter 6, Section 6.4)

The details of the publications during this period are given below (the last one is currently on review.)

- **Jalal, M. A.**, Chen, R., Moore, R. K., & Mihaylova, L. (2018, July). American sign language posture understanding with deep neural networks. In 2018 21st International Conference on Information Fusion (FUSION) (pp. 573-579). IEEE.

- Chen, R., **Jalal, M. A.**, Mihaylova, L., & Moore, R. K. (2018, July). Learning capsules for vehicle logo recognition. In 2018 21st International Conference on Information Fusion (FUSION) (pp. 565-572). IEEE.

- **M. A. Jalal**, W. Aftab, R. K. Moore and L. Mihaylova, ”Dual Stream Spatio-Temporal Motion Fusion With Self-Attention For Action Recognition,” 2019 22nd International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2019, pp. 1-7.

- **M. A. Jalal**, L. Mihaylova and R. K. Moore, “An End-to-End Deep Neural Network for Facial Emotion Classification,” 2019 22nd International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2019, pp. 1-7.

- **Jalal, M. A.**, Loweimi, E., Moore, R. K., & Hain, T. (2019, September). “Learning temporal clusters using capsule routing for speech emotion recognition”. In Proc. Interspeech (Vol. 2019, pp. 1701-1705).

- **M. A. Jalal**, R. K. Moore and T. Hain, “Spatio-Temporal Context Modelling for Speech Emotion Classification,” 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 853-859

- R. Milner, **M. A. Jalal**, R. W. M. Ng and T. Hain, “A Cross-Corpus Study on Speech Emotion Recognition,” 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 304-311.

- H. B. Sailor, S. Deena, **M. A. Jalal**, R. Lileikyte and T. Hain, ”Unsupervised Adaptation of Acoustic Models for ASR Using Utterance-Level Embeddings from Squeeze and Excitation Networks,” 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 980-987

- **Jalal, M. A.**, Milner, R., Hain, T, & Moore, R. K. “Removing Bias with Residual Mixture of Multi-View Attention for Speech Emotion Recognition”. In Interspeech 2020.

- **Jalal, M. A.**, Milner, R., & Hain, T, “Empirical Interpretation of Speech Emotion Perception with Attention Based Model for Speech Emotion Recognition”. In Interspeech 2020.

- Broughton, S. J., **Jalal, M. A.**, & Moore, R. K. “Investigating Deep Neural Structures and their Interpretability in the Domain of Voice Conversion”. Interspeech 2021 (in review)

1.3 Structure of the Thesis

At the beginning of each chapter, I describe the specific domain and the main research problems that the chapter will address. The structure of the chapter is presented by describing the sections and the individual research goal for each section. Each section serves separate research questions, and they are addressed using proper background discussion, modelling, experiments, results analysis and discussion. Each section summarises the findings in the discussion. The rest of the Thesis structures is as follows. Chapter 2 discusses the cognitive and psychological theories behind learning. As the motivation of this research is to template biological learning mechanisms to model the computational frameworks and HRI, Chapter 2 constructs the motivation of this research by critically comparing the pragmatism and plausibility of different theories and architectures. Chapter 3 presents the models and paradigms for visual language cue recognition for posture and static images with capsule learning. Chapter 4 presents the models and paradigms for visual language cue recognition for gestures and visual emotion through attention mechanisms. Chapter 5 proposes the investigation and models for speech emotion recognition. It briefly discourses the history and models for computational speech recognition. After that, Chapter 5 proposes novel frameworks for speech emotion recognition. Models are proposed for representation and context learning through capsule networks and convolutional self-attention networks. The capsule representations are analysed with t-SNE plots to show that distinct context representations are being learned in these capsules. Chapter 6 proposes a novel mixture of a multi-view attention (MOMA) model to remove task-specific bias with a residual hierarchical attention model. The attention weights are aligned with the input segments to show the context of learning at different hierarchies of the network. After that, the universality of speech emotion context is discussed with an attention model and four different types of emotion corpora. Chapter 7 presents two different neural network models with attention to investigate the roles of acoustic cue and vowel-consonant boundaries for speech emotion perception in attention models. A framework is proposed for interpreting the speech emotion classifier models with attention alignment. This chapter suggests that vowel-consonant boundaries and acoustic cues create a computational perception of linguistic units for speech emotion recognition. The importance and implications of emotion recognition for learning and behaviour modelling in HRI are discussed in chapter 8. Finally, Chapter 9 concludes and proposes a future research path.

Chapter 2

Learning and Attention

Learning is the acquisition of knowledge with continuous interaction with the physical and social environment. In most biological organisms, learning is essential to its cognitive development about abstract concepts or schemas (Section 2.1.3) as well as for motor skill development. There is evidence that learning starts in human beings since fetal state and they learn their mother's voice. The theories about learning often revolve around child knowledge and skill acquisition. Language development is one of the most complicated processes that depend upon both abstract concept learning and motor skill acquisition. We cannot indeed say the learning is a simple cognitive function with some internal and external deciding factors. However, some argue that each person's genome is crucial in some aspects of learning such as colour perception, interpretation of the physical world, emotion interpretation and regulation. Different theories have been proposed to explain the initial and native philosophy of knowledge acquisition. There have been fundamental differences between these theories, and they have their strengths and weaknesses. The most advanced machine learning techniques are envisioned from the biological and cognitive learning mechanisms. The computational models tend to follow the principles of learning symbolic and sub-symbolic representation in biological organisms. In this chapter, the cognitive theories of language, knowledge and skill acquisition are discussed. The machine learning techniques used in this research are also discussed with their inspirations from cognitive science, psychology and computational neuroscience.

The rest of this chapter is organised as follows: Section 2.1.1 to 2.1.5 present learning theories in infants. Section 2.2.1 to 2.2.4 discuss the cognitive and neurosceience perspectives of different computational deep neural networks. Section 2.3 presents the corpora used in this research. Finally, Section 2.5 summarises this chapter.

2.1 Cognitive Theories

2.1.1 Nativism

The existence of having an innate cognitive and psychological foundation is a way of understanding learning, language acquisition and knowledge acquisition in humans. The nativists say that the universal rules for language acquisition are innately encoded in our biological foundation. They cite language acquisition as pre-programmed natural phenomena. Chomsky (1959) proposed that the process of language acquisition has to be guided by pre-defined knowledge, which may be perceived as innate knowledge. Children are born with innate, domain-specific knowledge, which infers the direct influence of genes in cognitive development. Pinker (2009) suggested that by understanding and knowing some simple sentences and word formation, the infant can construct syntactic trees. They derive the grammar of a language from these syntactic trees. Eventually, more syntactic trees help the child to develop more complex grammatical rules. It was claimed by Fisher et al. (1994) and Gleitman et al. (1988) that they have already acquired the pieces of syntactic information which can generate the rest of the syntax tree. Leslie (1994) has hypothesised that children are born with a theory of mind. Pinker (2000) has emphasised that the theory of mind is an innate body of knowledge. Thus irrespective of language, the cognitive process of a child can extract the generalisation properties and rules of that language. Gleitman & Newport (1995), Nazzi et al. (1998), and Nazzi et al. (2000) showed in their research that newborn babies could distinguish their native language from a foreign language, indicating an innate structure in mind.

For example, one of the fundamental hurdles in language acquisition is to understand and generalise basic abstract algebraic rules. Wynn & Chiang (1998) hypothesised that children have innate math knowledge. Seven-month-old infants can learn the abstract algebraic generalisations. According to Samuels (1998), the mind is like a library, and some of the shelves are pre-stocked. The retrieval and use of information depend on the domain. A language-oriented domain-general cognitive process recruits innate knowledge for knowledge acquisition.

However, emotion acquisition and emotion regulation have few pieces of evidence that show the existence of an innate mind. Field et al. (1986) found that four-month-old infants become more distressed when they see their mothers' still face than being separated briefly from their mothers. This suggests that children can distinguish primary emotions in their early days and the state of emotional availability in their parents (Bornstein et al. (2012)). Emotional availability constructs the first paradigm of communication between a mother and a child. The mother and the child not only recognise each other's emotion but also respond to it as a reciprocal process (Biringen & Robinson (1991)). In the early phases of learning, parents' emotional expressions work as a stimulus to engage or withdraw the child from any physical, social and conceptual exploration (Sorce et al. (1985), Sorce & Emde (1981)). This research suggests that in the early stages of a child's development, there is clear evidence of emotion recognition and emotion regulation. Izard (1977), Campos & Barrett (1984), Oster et al. (1992), Izard (1994) stated that children are born with an innate state of mind which helps them regulate emotions though it is not as good as adults. However, it is adaptive, and emotion regulation becomes fine-grained over time. Nevertheless, the nativist theories do not explain how it is structured or what is the general structure of emotional intelligence. Clearly, the understanding and recognition of emotion differ culturally. Needham & Baillargeon (2000) showed that seven and half months' infants assume that a keyring and keys move separately, but eight and half months' infants assume that the keyring and keys move together. Therefore with experience, infants learn the association between objects. This evidence of learning through experience provides support for the empiricist theories.

2.1.2 Empiricism

The empiricist theories emphasise sensory experiences such as stress, social and cultural environments in order to acquire knowledge and language. The empiricists believe that development is a continuous process of interaction between genes and the socio-cultural environment. Thelen & Smith (1994) concluded that it is impossible that the source of the knowledge is solely based on genes. According to Smith & Thelen (2003)

> 'Development is seen as the emergent product of many decentralised and local interactions that occur in real-time. We examine how studying the multi causality of real-time processes could be the key to understanding change over developmental time.' –Smith & Thelen (2003)

Bates et al. (1995) studied and showed the individual differences in early language development varies with their individuality and environment. Bruner (1987) emphasised social interaction and interpersonal communication for knowledge acquisition. According to Bruner (1961), there are three modes of representation in the memory, i.e. enactive, iconic and symbolic. The symbolic representation is language. Worgan & Moore (2010) proposed that perception and the meaning of language is a factor of ‘social affordances’, i.e. reciprocal conversation can provide a sense of meaning based on the speakers and situation. This theory added flexibility to the concept of meaning based on social learning and sensory adaptation rather than being restricted to fixed object categories. Meltzoff & Borton (1979) experimented with four-week-old babies by giving them dummies of different types to suck on. When he gave them the dummies again, he observed that the baby preferred the dummy, which the baby had sucked on before.

It has been observed that newborns distinguish between old and new stimuli better than one month infants. There is documentation about the fact that newborn infants differentiate their mother’s face from a similar-looking woman. This evidence can empower nativist claim. Nevertheless, infants or newborns can not predict object movement.

2.1.3 Constructivism

Although constructivism advocates that learning frequently occurs from surroundings, it differs from empiricism in learning motivation. According to the constructivist theories, learning happens as a part of exploring the world in order to achieve goals and build a world model.

Piagets Theory

Piaget, one of the most influential in this field, combined the ‘nature and nurture’ mechanisms. He named the building blocks of intelligence schema. According to him, the schema is

> *"a cohesive and repeatable action sequence which possess tightly interconnected component actions that are governed by a core meaning."*

He proposed four stages of development

1. **Sensorimotor Stage (0-2 years old):** Acquisition of basic motor function
2. **Pre-operational stage (2-7 years old):** Acquisition of symbolic representations of objects and actions.
3. **Concrete operational stage (7-11 years old):** Starts adopting other people's perspectives and starts thinking logically.
4. **Formal Operational stage (11+ Years):** Starts abstract thinking and complex problem-solving.

According to Piaget, these stages are influenced genetically. The children have a structure of mind (Piaget (1952)), and with those schemata, they have the primary learning block. He mentioned three processes for learning.

1. **Assimilation:** The process to incorporate new information from the learned schema and extending the knowledge.
2. **Accommodation:** The process to change an existing schema into a new schema to accommodate new information and experiences.
3. **Equilibration:** When new information and experiences occur in a child's mind, then it causes disequilibrium and its necessary to accommodate new information and knowledge. The process to handle this is called equilibration.

Piaget's theory has certain limitations.

- Piaget assumed the child has no innate biases.

- Piaget's theory does not focus on learning.
- He ignored the role of social learning. In which his contemporary, Vygotsky, (Vygotsky (1978)) succeeded.
- Piaget studied children from the same socioeconomic background. So it would not be fair to generalise his findings.

The Piagetian perspective dictates that emotions are present in children at the time of birth, and the development of emotion is parallel with cognition, which only becomes complex over time in adulthood. In this context, Vygotsky's sociocultural theory is relevant (Vygotsky (1978)). According to this theory, the role of peers and adults is crucial for the child to evolve. Vygotsky emphasised how culture affects children's cognitive development. Tomasello & Rakoczy (2003) emphasised constant interaction and emergent development is necessary for constructing the language competence of a child. According to Tomasello (2001), children attempt to replicate adult utterances via usage-based syntactic operations, and they try to modify the schema. Slightly deviating from the core constructivist view, Hoemann et al. (2019) suggested that there might be interdependence between the emotion words of parents describing an infant's actions and emotional instances of the motivation of the infant. The infants learn to inflict importance on tasks based on those emotion motivations to explore knowledge and skill exploration tasks. Thus these emotion words become part of the child's social environment that teach the child cultural variants of emotion dependent concepts.

2.1.4 Statistical Learning

Statistical learning is proposed as a mechanism by which people learn patterns from the environment. In statistical learning theory, it is suggested that infants do knowledge acquisition from the statistical properties of sensory inputs. They discover the abstract structures from these statistical properties. Maye et al. (2002) designed a model to describe the phenomenon of why children learn to discriminate between native and non-native speakers. Studies have shown that Japanese children can distinguish between English and their native language Kuhl (2007). Maye et al. (2008) investigated the phenomenon and stated that in order to discriminate a language from his/her native language, the language must fit its discrimination profile, and this phonetic contrast may happen differently at different stages of life. It can indicate that the development of an infant may have different stages, as Piaget has stated. The linguistic feature extraction and sensitivity to phonetic variation can be different at different stages of life. The sensitivity to phonetic variation plays a vital role in building the perception of emotional context cues.

Kim et al. (2009) concluded that statistical learning is long-term learning with implicit learning mechanisms. Baldwin et al. (2008) stated that statistical learning could play a pivotal part in action segmentation which would help to model dynamic human action via a statistical structure. From the literature review, we can conclude two things

1. Statistical learning is a mechanism to choose the most likely and optimised sequence from previous experiences.

2. Learning mechanisms are different at different stages of life, and the difference is how they extract features from linguistic inputs and how they model it.

Statistical learning is long term and implicit process. However, some aspects are still not clear. Such as

- Are the learning mechanisms innate or not?

- Does the surrounding environment increase and decreases the amount of learning process?

2.1.5 Dual Learning

Processing both spatial and temporal information is a significant component in cognitive development. The neurobiology evidence for dual streams in visual, auditory and haptic information processing was found in multiple studies (Mishkin et al. (1983), Romanski (2007), Rauschecker (1998), Rauschecker & Tian (2000), Reed et al. (2005)). The research on auditory neurobiology proposed that auditory cortical processing happened in dual pathways. It was hypothesised that one of the pathways serve spatial processing, and the other pathway was responsible for temporal processing (Rauschecker (1998), Rauschecker & Tian (2000), Rauschecker & Scott (2009)). Further research showed evidence that not only do dual pathways exist, but they interact with each other (Cloutman (2013)). This research prompted a deep investigation into the 'what' and 'where' question (discussed in Chapter 1, 3, 5, 7) that helps us understand the nature of connections between these pathways.

In Chapter 4, a dual-stream architecture depicting the computational neurobiology model of dual-stream learning for action classification is presented. The interpretation of how the dual pathways will interact with each other is proposed in a computational model. The first stream deals with spatial nature, and the other learns the temporal dynamics of the speech. They distil the information further by contextualising the spatial representation with temporal dynamics and classifying the extracted representations. The goal is to build a deep spatiotemporal model of sequence through leveraging the relative context encoded by the interaction of dual pathways.

Rauschecker et al. first suggested that the dual pathways, i.e. 'dorsal auditory pathway' and 'ventral auditory pathway', play separate but complementary roles in auditory processing. The dorsal subserves the localisation of sounds or phonemes, and the ventral subserves the identification of ‘auditory objects’ and higher-order object representation (Rauschecker & Tian (2000), Goldman-Rakic et al. (1996)). Therefore, the ventral auditory pathway subserves the perception (‘what’), and the dorsal pathway subserves the action (‘where’/‘how’) (Goodale & Milner (1992)). In this thesis, the interaction between a ventral and dorsal network has been shown with a self-attention mechanism. The two stream’s interaction with each other and the degree of self-attention are decided using a parameter learned by the network through backpropagation.

2.2 Deep Neural Networks

2.2.1 Multi-Layer Perceptron and Feed Forward Neural Network

Multilayer perceptrons or feedforward neural networks are classical bio-inspired deep learning models. The feedforward paradigm is a homeostasis process that adjusts the internal states of a model in a continuative way. The intuition of feedforward neural networks can be traced back to the theoretical connectionist frameworks, which were rooted in the mid-1940s by McCulloch & Pitts (1943). Hebb (1949) postulated a theory for associative learning which suggests that an increase in synaptic efficacy is a result of the repetition of reverberatory activities between neighbouring neurons. This connectionist regime was further expanded by the research of Fahlman & Hinton (1987), Hinton (1989), Tryon (1993), Singer (1995) etc.

Singer (1995) examined neural mechanisms in the mammal cerebral cortex and described the neural activity in terms of feedforward and reciprocal (feedback) processes.

According to Singer (1995), these processes use some specific input recombination mechanism, and they are correlated as proposed by Hubel & Wiesel (1962). Singer (1995) further added that the input stimuli or signal are processed by a group of neurons or neuron cells. Each of these individual cells produces an ambiguous response to the input, but jointly they represent one feature or a group of features. Thus, the internal representational units are distributed and processed in parallel by a knowledge-sharing paradigm. A single unit can represent different overlapping features and feature constellations. However, due to dynamic grouping, these units are reassociated to form meaningful and context-dependent feature space. Fahlman & Hinton (1987), and Hinton (1989) supported this hypothesis. Feldman & Ballard (1982) differed with the idea that a single unit can represent different overlapping feature constellations. However, they all agreed that these neural mechanisms occur in a connectionist paradigm. Singer (1995) claimed that the underlying neural feedforward architecture does not require initial training. Monkeys develop selectivity about characteristics in the neurons before birth, and a similar phenomenon happens in other mammals (Hubel & Wiesel (1963)). These pieces of evidence support the nativist claims of an innate structure.

Initially, the connectionist paradigms assumed that neurons have two states. These states are distinguished by specific neural thresholds, and the neuron's activation states are influenced by the weighted connections from the neighbouring neurons (McCulloch & Pitts (1943), Hebb (1949)). Fahlman & Hinton (1987), Hinton (1989) and Tryon (1993) suggested some properties of these connectionist models. Such as

- They consist of a set of units and work as massively distributive parallel processing systems.
- The units have an activation state and an input function.
- The connections have a transfer function, and they are weighted.

- The models have a specific environment and a learning rule.

In this section, we will follow the similarities between the cognitive feedforward model and the computational feedforward model. Computationally, the goal of a feedforward neural network is function approximation i.e. for a classifier, $y = f^* (x)$ maps the input x to a categorical distribution y. The feedforward network is denoted by $y = f^* (x, \theta)$ where θ is denoting the learnable parameters learned by either backpropagation or any similar learning algorithm (Goodfellow et al. (2016)). The forward propagation happens in two stages, preactivation and activation. In the i layers, in the preactivation stage, the linear transformation of the weights of the incoming connections from previous layers are aggregated (Hinton (1989)). Each of these units will be

$$h_i = b_i + \sum \mathbf{W_i} y_{i-1}, \tag{2.1}$$

Where W_i denotes the weight matrices in the incoming connections from layer $(i - 1)$ to layer i. The extra parameter b_i is the threshold 'bias' and a learnable parameter. The weighted sums are passed through different activation functions to introduce non-linearity to the network. According to Hinton (1989), the learning occurs by changing the connection weights or by adding or removing connections and temporary states of the units represent short-term memory. Composition of such functions are performed to make the network deeper and have a richer representation i.e. $f (x) = f_n (f_{n-1} ... (f_1 (x, \theta)))$. Where n is the number of layers. These compositions form a directed acyclic graph (DAG) as there are no feedback connections. The number of layers (depth) and the size of each layer (width) are very crucial. Otherwise, the network will memorise (overfit) the data rather than generalise. The layers are optimised purely empirically as we do not have a theoretical relation between the amount of data and the network size.

However, a crucial factor often overlooked is the initialisation of the parameters. Poor initialisation can lead to disastrous performance. Some research reflects a nativist claim by showing that parameter initialisation with a clear variance and distribution leads to better results (Hanin & Rolnick (2018)). Furthermore, the feedforward models have a specific cost function, an optimisation procedure and a learning rule based on the model family and the task (Goodfellow et al. (2016), Strang (2019)). Deeper models generally tend to learn better representations. Hinton (1989), Hinton & Anderson (2014) found that these representations are distributed among hierarchies, and each unit represents part of the object or concept as local representation. By scaling up in a distributed manner, they can represent different things.

2.2.2 Convolutional Neural Network

Convolutional neural networks (CNN) have a grid-like topology with overlapping common weight shared layers. Though it was made popular in the late 1900s with the seminal paper by Lecun (LeCun et al. (1998)), the ideas about convolution and weight sharing among neurons are far older. Hubel & Wiesel (1959) and Hubel & Wiesel (1962) found that the neurons in the receptive fields of a cat's visual cortex were divided into different mutually antagonistic excitatory and inhibitory regions, which were activated by different region stimuli of the retina. They used simultaneous and specific stimuli over the regions and found the phenomena of summation and antagonism between the neighbouring regions. These phenomena were claimed as the basis of learning the specific nature of shape, size and orientation. Also, it suggests a hierarchical and grid-like structure such as simple cells, complex cells, lower-order hypercomplex cells, and higher-order hypercomplex cells. The cells in the higher hierarchy have bigger receptive fields than those in the lower hierarchy. The first computational model of these structures was proposed as the *Neurocognition* model by Fukushima (1980). The *Neurocognition* model proposed a hierarchical overlapping ‘C-plane’ and ‘S-plane’ comprising several ‘C-cell’s and ‘S-cell’s to perform self-organisation and learn from stimulus ‘without any teacher’. Kar (1986) suggested that there is a similarity between the computational convolution-correlation process and the distributed process of perception-memory in the brain. It also suggested an associative nature or memory and cognition along with ontological hierarchies for semantic representations.

The term ‘convolution’ in convolutional neural network comes from the mathematical operation convolution, which is a mathematical operation operated on two functions to produce a new function that expresses the amount of overlap of one function shifted over another function. If $f(x)$ and $g(x)$ are two functions over a continuous variable x, the convolution over an infinite interval would be

$$f(\mathbf{x}) * g(\mathbf{x}) = \int_{-\infty}^{+\infty} f(\Gamma) \times g(\mathbf{x} - \Gamma) d\Gamma, \tag{2.2}$$

where Γ is continuous time step, $*$ is the convolution operator and $\times$ is ordinary multiplication. If f and g are functions over discrete variables, with k being a discrete time, then convolution of g over f will be defined as

$$y[\mathbf{x}] = f[\mathbf{x}] * g[\mathbf{x}] = \sum_{k=-\infty}^{+\infty} f[\mathbf{k}] \times g[\mathbf{x} - \mathbf{k}], \tag{2.3}$$

where $y[x]$ is the resulting output, $f[x]$ is the input and $g[x]$ is the filter. In signal processing, $g[x]$ is the impulse function over $f[x]$. For functions over two discrete variables x and y the convolution would be

$$y[\mathbf{x},\mathbf{y}] = f[\mathbf{x},\mathbf{y}] * g[\mathbf{x},\mathbf{y}] = \left(\sum_{n_1=-\infty}^{+\infty}\sum_{n_2=-\infty}^{+\infty} f[\mathbf{n_1},\mathbf{n_2}] \times g[\mathbf{x}-\mathbf{n_1},\mathbf{y}-\mathbf{n_2}]\right), \tag{2.4}$$

where n_1 and n_2 are discrete timestamps. In a linear time invariant system the output $y[x]$ can be seen as the combination of convolution operations of kernel $g[x]$ on input $f[x]$. In image processing, the interval is finite. So, the equation (2.4) will be converted to

$$y[\mathbf{x_c},\mathbf{y_c}] = f[\mathbf{x_a},\mathbf{y_a}] * g[\mathbf{x_b},\mathbf{y_b}] = \left(\sum_{n_1=0}^{x_a-1}\sum_{n_2=0}^{y_a-1} f[\mathbf{n_1},\mathbf{n_2}].g[\mathbf{x}-\mathbf{n_1},\mathbf{y}-\mathbf{n_2}]\right), \tag{2.5}$$

where $0 < c < a+b-1$. The $g[x_b, y_b]$ can be perceived as the local receptive field, sharing the same set of weights, working on different parts of the input image $f[x_a, y_a]$, in which the neurons extract visual features (edges, corners or more abstract features) and combine the set of outputs to form feature maps. If kernels of size $[x \times y \times N]$ ([height $\times$ width $\times$ depth], and $n = 1, 2, \cdots, N$) are used, the n^{th} convolutional feature map can be denoted as:

$$y_n = f\left(\sum_j g_n * x_j\right), \tag{2.6}$$

where g_n is the n^{th} kernel and x_j ($j = 1, 2, \cdots, J$) is the j^{th} input feature map of size $[A \times B]$ and $f(\cdot)$ is a nonlinear activation function adaptable to a more complex problem space because the output of convolution is linear. The Rectified Linear Unit (ReLU) is applied with convolution layers (Nair & Hinton (2010), Xu, Wang, Chen & Li (2015)). ReLU is formally defined as:

$$f(x) = \begin{cases} x, & \text{if } x >= 0. \\ 0, & \text{if } x < 0. \end{cases} \tag{2.7}$$

Generally, CNNs contain several of these convolution layers along with spatial or temporal sub-sampling (LeCun et al. (1998)). As mentioned in the previous cognitive and neuroscience studies, this corresponds to shared regions and sparsed interactions in different regions and in separate hierarchies. CNNs offer these important functionalities in the way of sparse interactions, parameter sharing with kernels and equivariant representations (Goodfellow et al. (2016)) as well as allowing variable length input data. In CNNs the receptive fields or the kernels have smaller sizes than the input. The receptive fields do sparse interactions by moving over the data with or without overlapping regions. Each parameter in the kernel is used repeatedly over the whole input data in overlapping or non-overlapping regions and by this parameter sharing, the convolution process becomes more efficient than the standard feedforward neural nets. These mechanisms make the CNNs equivariance to translation. However, CNNs are not equivariant to rotation, scaling and pixel positions, and they are prone to adversarial attacks (Goodfellow et al. (2016)).

2.2.3 Recurrent Neural Networks

Neural mechanisms in the mammal cerebral cortex are described in terms of feedforward and reciprocal (feedback) processes of excitatory and inhibitory cells, which use some specific input recombination mechanisms (Hubel & Wiesel (1959, 1962), Singer (1995)). Wilson & Cowan (1972) proposed that the temporal dynamics between the neurons can be explained with multiple hysteresis loops and limit cycles within the subpopulation of the neurons. Sequence and context learning are important aspects of cognition.

Computationally the sequence learning can be constructed by modelling set of stimulli and the corresponding response over time. If the input vector X^T for a time sequence $1, 2, ...t$ be $x_1, x_2, ...x_t$ with the corresponding output labels $y_1, y_2, ..., y_t$ and $y \in 1, 2, ..., k$, we need to compute the posterior probability (P)

$$y_t^k = P(c_1, c_2, ..c_t | X^T), \tag{2.8}$$

here k is the total number of classes, and $c_1, c_2, .., c_t$ are conditional probabilities.

For using sequence data for training a classifier, there are mainly two approaches. One approach, sets the input data into overlapping time windows to provide the context of a particular data frame in that time instance. The second approach uses recurrent neural networks directly. In a regular feedforward neural network, it is not easy to optimally set the overlapping time frame and, in most of the cases, it is data dependent (Waibel et al. 1989, Graves & Schmidhuber 2005). Recurrent neural networks can store the previous and also the next context (bi-directional RNNs) in the hidden layers and correlate among the data points in the neighbouring sequence. If x_t is the input at time step t , s_{t-1} and s_t are hidden states at time steps $t-1$ and t then output o_t at time t would be

$$o_t = softmax((f(Wx.x_t + Ws.s_{t-1} + b)), \tag{2.9}$$

Where f is an activation function (generally ReLU or tanh activation function), b is the bias vector and Wx and Ws are weight matrices. Unfortunately, RNNs lose the gradient value in long time sequences and are not learned properly Hochreiter (1998). Long Short-Term Memory network (LSTM) (Hochreiter & Urgen Schmidhuber 1997) is proposed to overcome this problem. LSTM is an RNN with a special memory block containing information about the previous temporal states in the sequence. They outperform the normal RNNs to learn sequenced data (Sundermeyer et al. 2012). LSTM applies those past contexts of the memory blocks in the decoding process for prediction. Further studies show that not only the past sequence contexts but also the future contexts are useful for context-sensitive sequence modelling (Graves & Schmidhuber 2005, Schuster & Paliwal 1997). Bidirectional RNNs train the input sequence forwards and backwards in two separate RNNs (Graves & Schmidhuber 2005, Schuster & Paliwal 1997). Those RNNs feed forward to the same output layer. The output o_t of a Bi-directional RNN will be

$$o_t = (W_{g_t o} g_t + W_{h_t o} h_t + b_o), \tag{2.10}$$

where $W_{g_t o}$, $W_{h_t o}$ are the weight matrices between the hidden layers g, h and output layer o, g_t is the forward hidden sequence layer propagating from time $t = 1$ to T and h_t is the backward sequence layer propagating from time $t = T$ to 1.

$$g_t = f(W_{xg} x_t + W_{gg} g_{t+1} + b_g), \tag{2.11}$$

$$h_t = f(W_{xh} x_t + W_{hh} h_{t-1} + b_h), \tag{2.12}$$

where $W_{xg}, W_{gg}, W_{xh}, W_{hh}$ are weight matrices between the corresponding layers.

2.2.4 Attention Mechanism

Attention is a widely employed ability in cognition for flexibly using and controlling perceptual and cognitive information in order to achieve salience, awareness, arousal, learning and task-oriented reasoning. The concept of attention was popularised in psychology by (Posner & Boies (1971), Posner (1980)) as a covert and overt orientation mechanism in the brain for detecting visual stimuli. Posner (1980) found a relationship between attention and the brain systems that control perception and motion. Attention was initially perceived only as an orientation mechanism and eye control in the visual field as attention shifts. However, later studies found that attention is prominent in orienting and filtering (focus) other sensory information as well (Posner (1980), Posner & Petersen (1990*b*)). Further research found evidence for both top-down and bottom-up attention found in the visual pathways with spatial attention and feature-based attention (Noudoost et al. (2010), Linsley et al. (2018), Bichot et al. (2015)). Spatial and feature-based attention perform normalisation in the visual cortex, and they are additive (Reynolds & Heeger (2009), Hayden & Gallant (2009)). Similar attention mechanism was reported for auditory attention with spatial and non-spatial information (Gomes et al. 2000, Shomstein & Yantis 2006, Spence & Santangelo 2010). Heinke & Humphreys (2003*a*) proposed the attention mechanism as a dynamic routing process. The attention cues are also subject to other sensory modalities and cross-modal sensory processing by combining perceptual information with past knowledge and perform efficient task selection and execution (Malinowski (2013), Spence & Driver (2004), Lindsay (2020)).

The biological attention mechanism inspires the computational models of attention for machine learning. However, the computational models do not always follow the findings from the biological attention phenomena (Lindsay (2020)). Larochelle & Hinton (2010) introduced a 'glimpse' selection and combination mechanism (attention) for a restricted Boltzmann machine (RBM). Attention became popular when it was used in the *' sequence to sequence'* models by providing attention context vectors to the hidden states (Bahdanau et al. (2014), Cho et al. (2015)). Basically, these attention cues provide a context based alignment operation. Different scoring functions are used to calculate attention scoring such as content/location-based, additive, dot-product, normalised dot product. (Bahdanau et al. (2014), Luong et al. (2015), Vaswani et al.

(2017)). These scoring mechanisms put attention alignment weights either on the whole input or on special regions of the input. The self-attention mechanism was introduced to relate different regions of the same sequence to generate a positional dependency based representation of the same sequence (Zhang et al. (2018)). Vaswani et al. (2017) introduced the transformer model which uses transformations of input and previous output as key, query & values to get the next output using scaled dot product attention. The transformer model lacks positional encoding, which is an essential factor for preserving positional relationships in input information. The positional encoding problem was addressed by adding temporal convolutions with self-attention to model spatiotemporal context dependency (Mishra et al. (2017)).

2.3 Data

In order to evaluate the proposed methods for this research purpose, I have reused several publicly available emotion corpora in the experiments. These corpora are briefly discussed in this section.

2.3.1 AffectNet

AffectNet is introduced by Mollahosseini et al. (2017*b*). It is an image database for detecting and recognising facial expressions of emotions in the wild. It consists of both the categorical (joy, anger, sad, happy, etc.) and the dimensional labelling (valence, arousal) of emotions in the wild. Having more than 140000 image samples without any controlled environment (e.g. lighting condition, alignment, rotation, head posture etc.), makes this database very challenging to use for emotion recognition. The facial landmarks (OpenCV) of the images are also provided with the dataset. The average image resolution is 425 × 425 with STD of 349 × 349 pixels. Professional annotators labelled the images.

There are eight emotion categories in the AffectNet database. The emotions are neutral (80276), happy (146198), sad (29487), surprise (16288), fear (8191), disgust (5264), anger (28130), contempt (5135) with numbers being the numbers of annotated samples for each category.

2.3.2 FAU-AiBO

FAU-Aibo by Schuller et al. (2009), Steidl (2009), Batliner et al. (2008) is a speech emotion corpus that has 9 hours of children speech recording. Both male and female children participated. The children's are communicating with Sony's pet robot Aibo. Children from two different schools (Mont, Ohm) have been recorded. The recordings are segmented manually into meaningful small chunks using prosodic criteria. The files are 16kHz, 16bit, mono channel .wav file. FAU-Aibo consists of 5 emotional classes: anger, emphatic, neutral, positive and rest (other categories). FAU consists of children speech recordings with Sony's pet robot Aibo, so the emotions are natural and spontaneous.

FAU consists of two sets, namely Ohm and Mont, which cover 55% and 45% of the whole data, respectively, with totally disjoint speakers. The Ohm and Mont sets are used for speaker disjoint training and testing. In one scenario, Mont is used as a training set, and Ohm is used as a test set and vice versa. In this research, another train/test set has been used for an overlapping speaker scenario. For training the system, 75% of the data (randomly chosen) was employed, and the remaining 25% were used for testing. Since it consists of two subsets, i.e. Ohm and Mont, one was used for train and the other one to test. These two sets are disjoint in terms of speakers, which makes the test conditions more challenging than a 75/25% case and provides a better platform for evaluating the robustness of the system. The downside, however, is that a lower amount of data becomes available for training.

2.3.3 RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) by Livingstone & Russo (2018) is an audio-visual corpus. It contains recordings from 24 professional (12 male and 12 female) actors with a north American accent. The speech recordings include emotions calm, happy, sad, angry, fear, surprise, and disgust. The song recordings contain calm, happy, sad, angry, and fearful emotions. Only its speech part is utilised in this research which covers eight acted emotional expressions: *neutral*, *calm*, *happy*, *sad*, *angry*, *fearful*, *surprise* and *disgust*. It is also annotated with two classes, i.e. strong emotion and normal emotion intensity. The RAVDESS speaker-independent scenario is performed by using the first 19 speakers for training and four remaining speakers for testing. The same train/test set is used for all the systems mentioned in this research. For the other scenario, 75% of the samples are used as training, and the remaining 25% samples are used for testing purpose. The same train/test set has been used for all the systems mentioned in this research. Each of these conditions is available in three modality formats: Audio-only, Audio-Video, Video-only (no sound). In this research, only the audio modality has been used with speech recordings.

2.3.4 IEMOCAP

The IEMOCAP corpus by Busso et al. (2008) is used for validating the proposed framework. The corpus contains utterances from ten speakers (five male and five female) for 12 hours of recording. The sessions are dyadic (between two speakers) and either scripted or improvised for eliciting emotions. Four sessions, containing a total of eight speakers, are used for training. The remaining session, which contains

Table 2.1: *Number of emotion utterances in IEMOCAP (IEM4).*

Emotion	IEM4
happy + excitement	595 + 1041
sad	1084
anger	1103
neutral	1708

two speakers, is used for testing. In the literature it is common for IEMOCAP to be evaluated with four classes: *happy*, *sad*, *anger* and *neutral* (where *happy* is combined with *excitement*) Li et al. (2018*b*). The utterances are split into a training set of 4290 samples (Sessions 1-4) and a test set of 1241 samples (Session 5). This is referred to as IEM4 in this research. IEMOCAP with 'big-six' emotions Ekman (1992), which are *happy*, *sad*, *anger*, *surprise*, *disgust* and *fear* is denoted by IEM in Table 5.6 and Section 5.4.

2.3.5 eNTERFACE

The eNTERFACE (ENT) corpus has the big-six emotion classes (*happy*, *sad*, *anger*, *surprise*, *disgust* and *fear*) and contains 1 hour of acted segments (Martin et al. (2006)). There are 44 speakers (8 female) and each speaker has 5 recordings of each emotion. After inspection of the data only 43 speakers (Spkr6's recordings have not been segmented) and 1287 segments (Spkr23 is missing three *happy* segments) are used. The speakers are from 14 different nations with different English accents. The training set has 38 speakers (Spkr1 to Spkr39 without Spkr6), and the test set has the last five speakers (Spkr40 to Spkr44).

2.3.6 MOSEI

The MOSEI (MOS) corpus is created from YouTube videos and annotated for the big-six emotions using Amazon Mechanical Turk. The utterances have been segmented, and it contains more than 1000 speaker's utterances with various accents (Zadeh et al. (2018)). The emotional speech is described as natural though the speakers recorded themselves consciously in from of the camera. Thus, whether MOSEI is natural or not is debatable. The official training, validation and test set splits for the ACL 2018 conference have been used, where the training and validation sets are combined for training.

The speech emotion corpora that have been discussed in this section (Section 2.3) are from three categories such as: natural (FAU-AiBo, MOSEI), elicited (IEMOCAP) and acted (RAVDESS, eNTERFACE). Each of the corpora have participants from different gender, accent, age and sex. Although there was some research that investigated crosscorpus impact, those studies focused on the cross-lingual aspect of emotion. However, it is necessary to know if the acted, elicited, and natural corpora within the same language can benefit each other. In other words, if there is a common latent representation and understanding of emotion across different corpora in a language. The crosscorpus study Milner et al. (2019) presents whether the emotion representation learning from one corpus can be used for speech emotion recognition in other corpora (presented in Chapter 5). [1]

[1] The crosscorpus SER work is published as R. Milner, M. A. Jalal, R. W. M. Ng and T. Hain,"A Cross-Corpus Study on Speech Emotion Recognition," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 304-311. The thesis author contributed to the implementation of the attention model used in the cross-corpus experiment and the presentation of the work.

2.3.7 UCF-101

The UCF-101 dataset (Soomro et al. (2012*a*)) consists of 101 actions annotated for 13320 Youtube video clips. This dataset is among the biggest (in terms of the number of classes and videos) publicly available and annotated datasets to date. The videos have been uploaded by non-professionals which include additional challenges such as shaking cameras, inconsistent viewpoints and changing resolutions. Moreover, there are groups of classes that are quite similar to each other such as violin and cello playing.

2.3.8 ASL

The 'Kaggle' American Sign Language Letter database of hand gestures has 24 classes of letters. The database contains all the letters except 'J' and 'Z' because they do not have static postures. The training set consists of 27455 items, and the test set has 7172 items. The dataset contains greyscale 28 × 28 pixel images. According to the dataset description, the images have been modified with at least 50+ variations. For example, 5% random pixelation, +/- 15% brightness/contrast, three degrees rotation etc. These modifications also altered the image resolution.

Vehicle Logo

The largest publicly available VLR dataset by Huang et al. (2015) has ten categories, and each category has 1000 training images, 150 test images, and each image size is [70 × 70]. For simulating real scenarios and making the test more challenging, three more variants of the test set have been introduced, i.e. *Dataset 1*, *Dataset 2* and *Dataset 3*. *Dataset 1* contains test images with noise. Zero mean Gaussian noise with 0.1 variance has been applied to the original test images. *Dataset 2* contains test images with random rotation in the range $[-25^\circ, 25^\circ]$. A random combination of noise, rotation, occlusion (maximum 20%) has been applied in *Dataset 3*.

2.4 Evaluation

Unweighted accuracy (UA) and the weighted accuracy (WA) are used to evaluate the results. The UA calculates accuracy in terms of the total correct predictions divided by total samples, which gives equal weight to each class. As IEM4 is imbalanced across the emotion classes, the WA is calculated as well, which weighs each class according to the number of samples in that class:

$$UA = \frac{TP + TN}{P + N}, \qquad WA = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}) \tag{2.13}$$

where P is the number of correct positive instances (equivalent to $TP + FN$) and N is the number of correct negative instances (equivalent to $TN + FP$).

2.5 Summary

In Sections 2.1.1, 2.1.2 , 2.1.3 and 2.1.4, four main classes of cognitive theories have been presented. Nativism proposes that infants are equipped with all their perceptual abilities by birth, but empiricism claims that cognitive functions develop through experience, learning and genetics. However, statistical learning does model the world view through experience and the probabilities of one event happening after another. Both nativists and empiricists agree on an innate mind. Empiricists agree that the mind has an innate, underlying structure to cope with the physical world. By statistical learning, one can use any of the nativist or empiricist assumptions and build a world model. Therefore, both agree that cognitive development follows after birth, but there is a contrast on how much prior knowledge it needs. The postulates from constructivism have a similar goal as empiricism, but they are more pragmatic for building a world model with statistical learning.

Attention (Section 2.2.4) plays a vital role in selecting what abstract sub-symbolic representation enters into memory, which basically constructs learning. This process manages multiple sensory information in the neural pathways, i.e. communication between spatial and temporal representation. Dual learning described in Section 2.1.5, shows the communication in the ventral and dorsal hierarchy, and it is explored with a computational model in Section 4.1. Cross-modal perception can be observed in infants. It is the procedure of learning something via one sense and transfer it to others, in a bigger picture that is a generalisation. Section 2.2.1 to 2.2.4 discuss the deep neural network used in this research from a cognitive and neuroscience perspective. The purpose of these sections is to discuss how deep neural networks are inspired by the learning mechanisms in neurons.

Section 2.3.1 to 2.3.6 briefly presents the emotion corpora used in this research and the motivations behind using them.

In conclusion, a combination of the theories of learning is perhaps the most efficient strategy to take. For this purpose, this thesis combines the strong points of learning theories and deep neural networks for statistical learning and emotion modelling.

Chapter 3

Learning Visual Cues with Capsules

In human visual processing, visual information may contain many objects and many points of interest, but few saliency regions are taken into consideration to attain a particular goal at any given time. The saliency regions are chosen with a visual cue based selection method (Posner (1980)). This selection method in human visual perception can be location-coded, stimuli oriented or task-specific. A selection mechanism is a form of attention that can choose a single saliency region or multiple grouped regions. Hinton (1981), Hinton & Anderson (2014) suggested the view-invariant object recognition model. Olshausen et al. (1995) proposed the dynamic routing circuit model based on a top-down attention mechanism for visual perception learning. SIAM (Heinke & Humphreys (2003*b*)) combined these previous approaches and proposed a dynamic circuit routing selective attention model, which combines both top-down and bottom-up attention. The dynamic routing and task/stimuli specific region grouping were further explored with the capsule networks (Sabour et al. (2017)). The multiple saliency regions, which are grouped together to denote similar characteristics or stimuli, are called capsules in capsule networks (Sabour et al. (2017), Jalal et al. (2018)). In this chapter, capsule-based visual cues are further explored in a posture recognition and vehicle logo recognition framework. Visual language or sign language is a visually oriented, natural, non-verbal medium for communication. However, posture sign language is a static (fixed fingerspelling postures) mode of communication between people. Static sign language understanding has scope for simple non-verbal communication in Human-Robot Interaction (HRI) as it would take less computational time due to the limited number of postures.

In this chapter, two capsule-based models have been proposed, and they have been demonstrated with two separate corpora. Challenging scenarios, such as noise and distortion, are introduced to test the robustness of these models. Firstly, a capsule-based deep neural network sign posture translator for an American Sign Language (ASL) fingerspelling (posture) has been presented in Section 3.1. The performance validation shows that the approach can successfully identify sign language, with unweighted accuracy (UA) of approximately 99%. Unlike previous neural network approaches, which mainly used fine-tuning and transfer learning from pre-trained models, the proposed capsule network architecture does not require a pre-trained model. The framework uses a capsule network with adaptive statistical pooling which is the key to its high accuracy. Secondly, a capsule learning framework is proposed for vehicle logo recognition in Section 3.2. Vehicle logo recognition is a crucial part of vehicle recognition in intelligent transportation systems, and the data is suitable for this research. Because the vehicle logo corpus has images added with distortion, rotation and noise. Finally, Section 3.3 summarises the chapter and draws conclusion.

3.1 Posture Recognition

Sign language is a visually oriented, natural, non-verbal communication medium, which is used by millions of hearing-impaired people around the globe as their first language. According to the British Deaf Association, there are 151,000 people who use sign language (Association (2018)). The main two components of sign language are fingerspelling (postures) and dynamic hand movement (gestures) (Lahamy & Lichti (2012)). Hearing-impaired people often find it quite challenging to communicate with non-hearing impaired people because most do not know sign language. Therefore, an artificial sign language translator would be useful in reducing the communication barrier.

Although there has been extensive research on ASL detection, it still remains a relevant research field due to the unavailability of an accurate method. Two different approaches are mainly used: using gloves with sensors to detect joint movements and using vision (Mehdi Y. N. (2002) , Sidney Fels & Hinton (1993), Pigou et al. (2015)).

Sensor-based methods with Bayesian networks and neural network classifiers were popular in the early 2000s (Mehdi Y. N. (2002), Singh et al. (2017), Chuan et al. (2014), Vutinuntakasame et al. (2011), Sidney Fels & Hinton (1993)). Wearable sensor gloves are used for getting the relative motion of the fingers and hands to get the kinematic parameters to predict the sign language. Cheap wearable technologies are proposed in Singh et al. (2017), Chuan et al. (2014), Vutinuntakasame et al. (2011).

Linear classifiers are also widely used for detecting postures and gestures because they are relatively simple models compared to Bayesian models and they get high recognition accuracies (96%) (Singha & Das (2013), Aryanie & Heryadi (2015), Sharma et al. (2013)). Hidden Markov models and Bayesian models achieved higher accuracies (Starner & Pentland (1996), Jebali et al. (2013)). The problems with those approaches arise from hand-coded feature extraction with heavy pre-processing and the constrained experimental environments (Garcia & Viesca (2016)).

Neural networks have an advantage over these networks because they learn essential features to classify the data (Admasu & Raimond (2010)). Feedforward neural networks also need image processing and hand-coded feature extraction. Convolutional Neural Networks (CNNs) have been very useful for recognising gestures and events (Sze et al. (2017)). Garcia & Viesca (2016), and Pigou et al. (2015) used CNNs for recognising sign language. Convolutional layers work as filters and do not require hand-coded feature inputs. However, CNNs have some fundamental defects. They work as filters but do not preserve all the spatial-temporal relationships of an image due to the use of pooling operations.

Sign language understanding may also be seen as a gesture classification task. Bheda & Radpour (2017) used Deep Convolutional Neural Network (DCNN) to explore this using a gesture classification approach. The depth and colour of an image are also used in some research. Ameen & Vadera (2017) used a CNN to classify American sign language (ASL) using the depth and colour of images, and they report 82% precision and 80% recall in their experiments. The major weakness in these models is that they do not preserve the spatial relations among the regions of the image. The spatial relations among the image regions construct the context. The dynamic routing by Sabour et al. (2017) models spatial relationships forming the context in a hierarchical routing technique using capsules. The proposed model in this section uses this technique to learn visual context information for American sign language recognition. The model contains three convolutional layers, two spatial adaptive pooling layers, one 6D convolutional capsule layer and one fully connected capsule layer.

This section aims to present an efficient framework for sign language posture understanding. Using adaptive layer pooling allows the network to be trained with variable sized images as it produces fixed-length matrices. Networks trained with multiple sized images could enhance scale-invariance. Those feature matrices are fed into the capsule layer. Capsule layers provide transitional invariance and robustness to the framework.

The images in the dataset are rotated slightly, making it quite challenging [1].

The rest of the section is organised as follows. In Section 3.1.1, the deep learning approaches, which are used in the proposed framework, are discussed. Section 3.1.2 explains the proposed deep framework. Simulation results and discussions are presented in Section 3.1.3, 3.1.4 and Section 3.1.5 summarises the work.

3.1.1 Approach

Convolutional Neural Network

A convolution is a mathematical operation operated on two functions to produce a new function that expresses the amount of overlap of one function shifted over another function. A detailed explanation could be found in Section 2.2.2. LeCun et al. (1998) proposed LeNet with several of these convolution layers along with spatial or temporal sub-sampling. Krizhevsky et al. (2012*a*) was the first popular deep convolutional neural network which achieved the best performance on ImageNet in 2012. By using convolutional kernels, the CNNs do weight sharing. Thus, the CNNs use fewer weight parameters compared to their feedforward neural network counterparts. The convolution process preserves the spatial information of the image, and the feature vector represents richer representations than simple feedforward DNNs. A smaller kernel size has been used to save more pixel-level information (Zeiler & Fergus (2014)).

Adaptive Pooling

Pooling or subsampling plays a significant role in CNNs. Although the feature map loses some of its spatial features after a pooling operation, it has been proven to be

[1] This section is published as Jalal, M. A., Chen, R., Moore, R. K., & Mihaylova, L. (2018, July). American sign language posture understanding with deep neural networks. In 2018 21st International Conference on Information Fusion (FUSION) (pp. 573-579). IEEE.

useful for CNNs (Wu & Gu (2015), Krizhevsky et al. (2012*b*)). The pooled vector is given by $y = f(y_1, y_2, ..y_n)$, where $y_i \in \mathbb{R}^d (i = 1, 2, ..., N)$, d is the dimensionality and n is the number of feature descriptors. The $f(\cdot)$ function is the pooling operation. Spatial pooling preserves the spatial features by pooling vectors in local spatial pooling bins. The size of the spatial bins is proportional to the input image size. In this experiment, a spatial adaptive max-pooling method is used (He et al. (2015)). The output of each filter from the previous convolution layer is stored in each of the spatial bins using a max pooling method. If there are n filters in the last convolution layer and the number of spatial bins is m, the output of the pooling layer will be n, m dimensional vectors. Adaptive spatial pooling is efficient in image classification, object detection problems with higher accuracies and keeps more spatial properties of an image (Liu et al. (2016), Tsai et al. (2015)). Also, it provides the network with the ability to be trained with multiple sized images.

The Capsule Neural Network

Unlike the connections between adjacent convolutional layers (which are through neurons in a CNN) adjacent capsule layers are connected by capsules in a capsule network. For each capsule (represented as a vector), the output of the capsule j can be denoted as:

$$\mathbf{v}_j = g(\mathbf{s}_j), \tag{3.1}$$

where v_j is the output vector of capsule j, $\mathbf{s}_j$ is the input to the capsule j and $g(\cdot)$ is a squashing function given by:

$$g(\mathbf{x}) = \frac{||\mathbf{x}||^2}{1 + ||\mathbf{x}||^2} \frac{\mathbf{x}}{||\mathbf{x}||}, \tag{3.2}$$

where $\mathbf{x}$ is the input vector of the function.

The squashing function makes the length of a short vector shrink close to 0, and long vectors shrink close to 1. The length of a capsule is used to represent the existence probability of the corresponding entity or part of an entity. Parameters in each capsule represent various properties such as position, scale and orientation of a particular entity (Sabour et al. (2017)).

Apart from the capsules in the primary capsule layer, the total input of the capsule $\mathbf{s}_j$ can be calculated by:

$$\mathbf{s}_j = \sum_i c_{ij} \mathbf{o}_{j|i}, \tag{3.3}$$

where $\mathbf{o}_{j|i}$ is the predicted output of the capsule j (in the current capsule layer) made by the capsule i from the previous capsule layer. The coefficients c_{ij} are coupling coefficients determined by a routing algorithm

$$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})}, \tag{3.4}$$

where b_{ij} denotes that the log prior probabilities of capsule i are coupled with capsule j from the previous capsule layer to the current capsule layer. The term k is an index through all capsules in the current layer. The coefficients b_{ij} are initialized to zero and their values are updated by a routing algorithm. In the routing algorithm, b_{ij} is updated by:

$$b_{ij}^{(r+1)} = b_{ij}^{(r)} + \mathbf{v}_j \cdot \mathbf{o}_{j|i}, \tag{3.5}$$

where r is the iteration index. The scalar product of $\mathbf{v}_j$ and $\mathbf{o}_{j|i}$ is the cosine similarity measurement. The term $j|i$ means from i^{th} capsule to j^{th} capsule. If the two vectors are similar, the b_{ij} will be updated with a large step. On the contrary, if the two vectors are different, b_{ij} will be updated with a small step. Intuitively, a lower level capsule i predicts the output of capsule j. The probability that the capsule i and the capsule j

are coupled is determined by the product of the prediction and the actual output.

In equations (3.3) and (3.5), the predictions $\mathbf{o}_{j|i}$ are given by:

$$\mathbf{o}_{j|i} = \mathbf{W}_{ij}\mathbf{u}_i, \tag{3.6}$$

where $\mathbf{u}_i$ is the output of capsule i in the previous capsule layer and $\mathbf{W}_{ij}$ are transformation matrices connecting capsules between the previous capsule and the current capsule layer. The final capsule layer connects the classification labels directly. There are D (D is the number of classes) capsules, where the length of each capsule represents the prior probability of the classification object (Chen et al. (2018)).

A margin loss is applied in capsule network and the loss L_d for the class $d = 1, 2, \cdots, D$ is given by

$$\begin{aligned} L_d = T_d\ max(0, m^+ - ||\mathbf{v}_d||)^2 \\ + \lambda(1 - T_d)\ max(0, ||\mathbf{v}_d|| - m^-)^2, \end{aligned} \tag{3.7}$$

where $T_d = 1$ if and only if a digit of class d exists and $||\mathbf{v}_d||$ represents the length of vector $\mathbf{v}_d$. For the correct class d, loss starts to accumulate when the length of the $\mathbf{v}_d$ is under m^+. Meanwhile, for the unrelated class d, loss begins to accumulate when the length of the $\mathbf{v}_d$ is beyond m^-. The λ is a controlling parameter and the total loss $L_{total} = \sum_d L_d$, which simply sums the losses from all the final layer capsules.

In the capsule network, the back-prorogation is applied to update the weights in the convolutional kernels and the transformation matrices. The routing phase is applied to update the weights for the coupling process coefficients c, and the log prior probabilities b (Chen et al. (2018)). The vector to vector transformation could potentially extract more robust features than scaler to scale transformation in a CNN (Sabour et al.

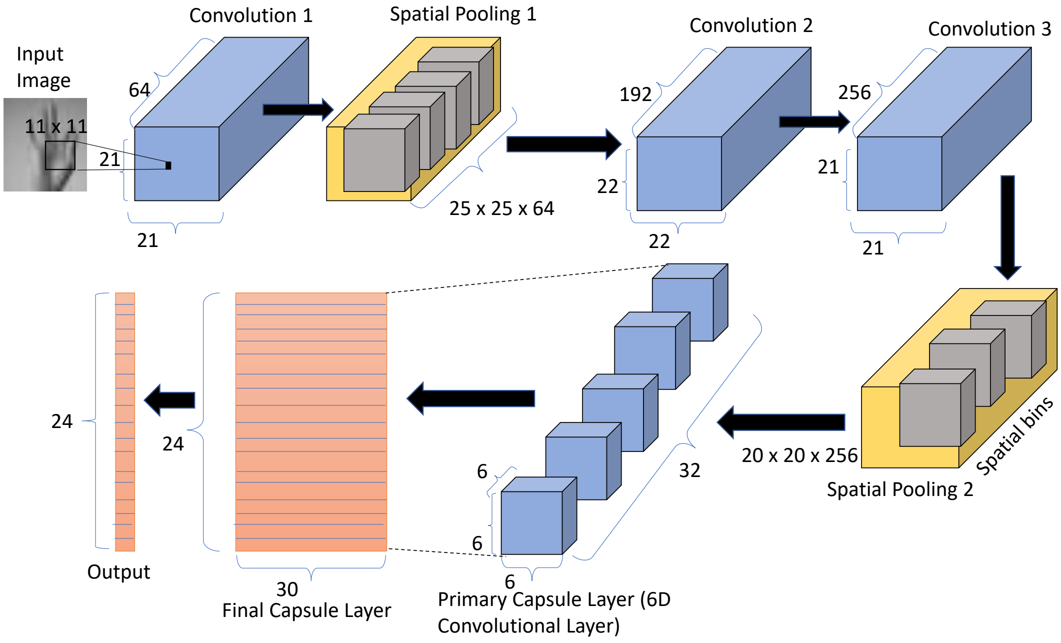


Figure 3.1: *The proposed framework consisting of a convolutional layer and spatial adaptive pooling with capsule network routing.*

(2017)).

3.1.2 The proposed framework

The general architecture of the proposed framework is given in Figure 3.1. The first convolution layer has 64, [11×11] sized ([height × width]) convolution kernels. The convolution process is applied with a stride of 1, padding 2 and ReLU activation function. A spatial adaptive max-pooling layer follows this layer. Neural networks with fixed filter size extract differently sized features for different sized images. Fully connected layers need fixed-length input. If the feature scale is different due to different sized images, it would be a problem. Hence, the advantage of using this adaptive pooling layer is, it can take an input of variable size, arbitrary aspect ratio, scales and produce fixed-sized output. Therefore, this method allows flexibility in the network. The first adaptive pooling layer produces 64 [25 × 25] neurons. After this, two convolution layers are stacked together having 192, [5 × 5] and 256, [3 × 3] ([height × width]) kernels respectively with padding 1 and ReLU activation function. An adaptive pooling layer follows, producing 256 channels of fixed output neurons [20 × 20].

On the top of the second pooling layer, the primary capsule layer has thirty-two channels of six convolutional units with the kernel size of [9×9] and stride 2. Thirty-two channels with a block size of [6 × 6 × 6] are generated. Each channel contains 6×6=36 capsules, with each capsule containing a vector of length six. The detailed process of capsule generation can be found in our work Chen et al. (2018). Each capsule in the convolutional capsule layer shares its weights with each other.

Transformation matrices connect adjacent capsule layers. The convolutional capsule layer has [32 × 6 × 6] sized capsule outputs. A routing process is applied in the capsule layer according to the work of Sabour et al. (2017). The final capsule layer gives 24 output classes with 30 length vectors per class. Those are passed through a softmax function, and the class probabilities are computed. PyTorch has been used to implement the model and the 'AdaptiveMaxPooling' library function for the pooling.

The probabilities are used to compute the loss at each training case. They are also used as a mask to nullify the unwanted output vector classes. The remaining output vector is used to reconstruct the image through another decoder network with feedforward neural networks. The reconstruction loss is also used as a regularising expression to add more robustness in the network.

3.1.3 Experimental Setup

The capsule network from Sabour et al. (2017) is used as the benchmark for this work. A deeper version of that capsule network has also been designed with four capsule layers to compare with the proposed framework. The deep capsule net contains one simple convolution layer for feature map extraction and one convolutional capsule layer, and

Table 3.1: *Summary of experiment results for ASL*

Method	Maximum Test Accuracy (UA%)		Minimum Test error (%)	
	With Reconstruction	Without Reconstruction	With Reconstruction	Without Reconstruction
BaseLine	97.65	99.70	0.652	0.633
Deep Capsulenet	98.87	98.54	0.639	0.634
ConvNet	-	99.26	-	-
Proposed model	99.74	99.60	0.641	0.632

three fully connected capsule layers. When capsule layers are stacked together, they are one of the most computationally costly frameworks. After using different numbers of routing iterations, it has been observed that three rotations are efficient enough to get quick convergence with reasonable computational complexity. All of these results presented in this section of work are based on three routing iterations.

A CNN model has also been used to compare the performance of the proposed framework. Alexnet by Krizhevsky et al. (2012*b*) is an efficient CNN architecture. Model parameters have been changed to accommodate it with dataset (image size 28×28 pixels). It is called *ConvNet* in Figure (3.3) and in the discussion of the results.

The performance evaluation of the networks is conducted in Python using the PyTorch library. The machine on which the experiments have been performed has an Intel core i5-3210M CPU @ 2.50GHz × 4 , 8 GB Primary memory and one Nvidia GeForce GTX 1070 (extended GPU). The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

Dataset

The 'Kaggle' American Sign Language Letter database of hand gestures, described in Section 2.3.8, has been used to evaluate the framework.

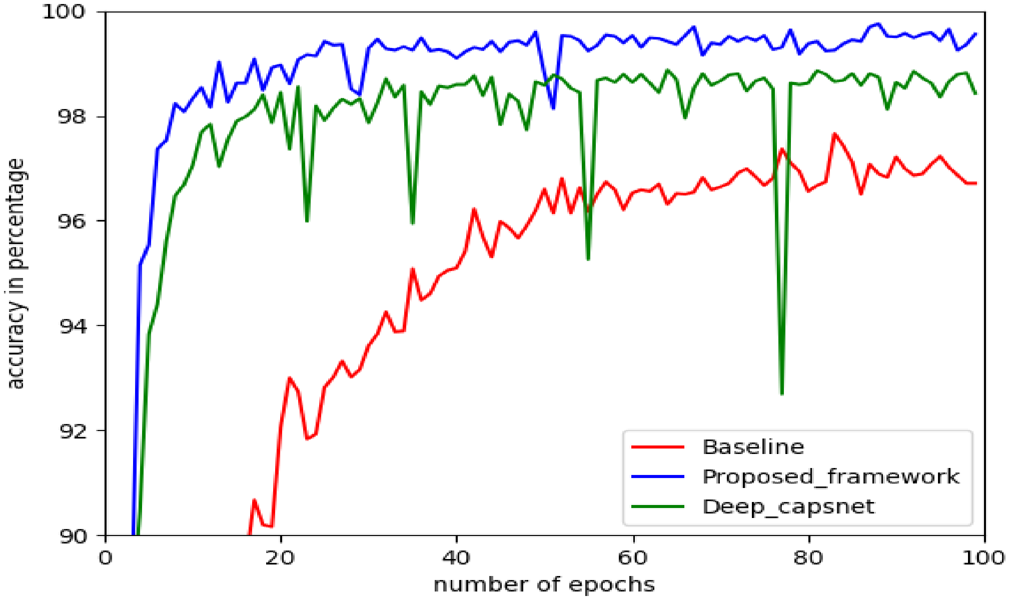


Figure 3.2: *The accuracy (UA%) of the proposed model, baseline, and deep capsulenet on the testing data while considering reconstruction loss.*

3.1.4 Results

Result of the proposed capsule network with reconstruction loss

According to (Sabour et al. (2017)), the reconstruction loss is considered while calculating the total loss in each training batch. Figure 3.2 illustrates the performance of the proposed method in terms of testing accuracies. The testing accuracies are evaluated in each training epoch with the test data.

The performance of the baseline and the deep capsulenet are shown in Figure 3.2. Some observations can be made from the figures. The rate of convergence is higher in the proposed model than the baseline two-layer capsulenet. The 4-layer deep capsulenet is closer to the proposed network in terms of convergence rate, but it is less robust. The best accuracy of the baseline framework is 97.65%. The best accuracy of the proposed framework is 99.74%.

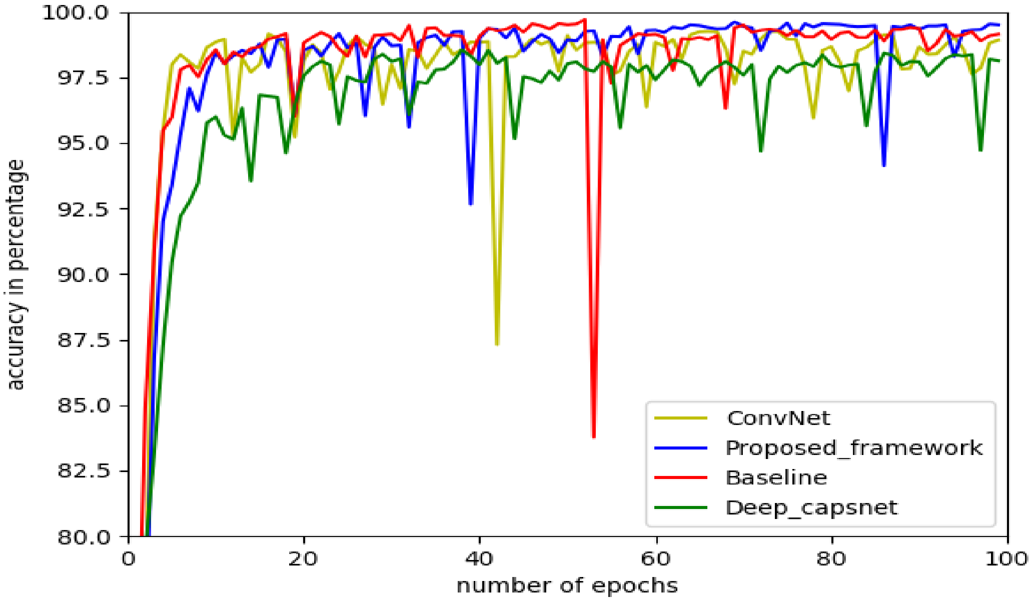
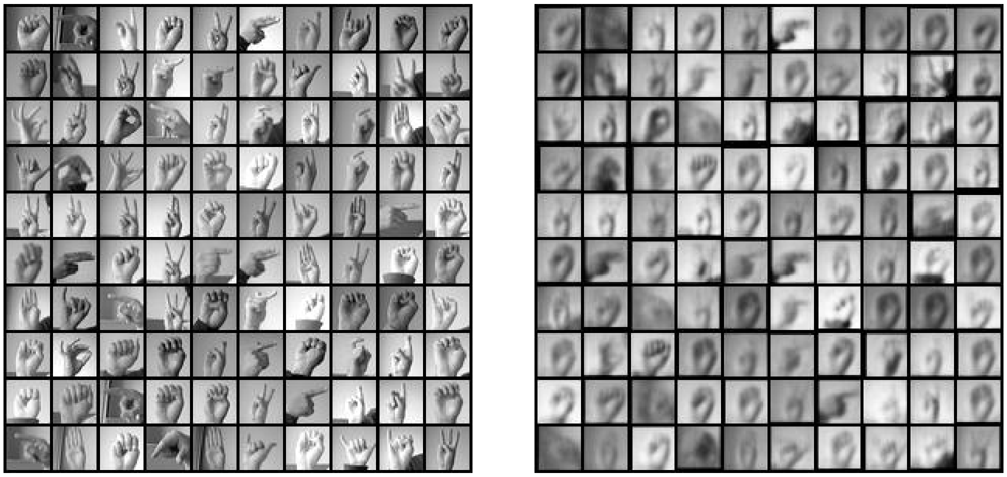


Figure 3.3: *The accuracy (UA%) of the proposed model, deep capsulenet, the baseline method and the convnet on testing data without considering reconstruction loss.*

Result of the proposed capsule network without reconstruction loss

In this section, the training is completed without the image reconstruction loss (regulariser). Figure 3.3 represents the performance of different frameworks in each training epoch, up to 100 epochs. The baseline method has the higher best accuracy, but the stability and consistency (test accuracy over epochs in Figure 3.3) in the proposed method are slightly better. The 4-layer deep Capsulenet is stable and robust, but the test accuracies dropped significantly (between epochs 50 to 60). One of the reasons could be the absence of the regularising term while calculating the loss (for example, reconstruction loss). There are fewer capsules in the proposed method, which also could be the reason.

Table 3.1 gives a summary of all the experiments. Clearly, it can be seen that, for this dataset, the proposed method is an improvement of the 4-layer capsule net and



(a) *Ground truth* **(b)** *Reconstructed*

Figure 3.4: *Ground truth vs. Reconstructed Images*

baseline 2-layer capsulenet.

3.1.5 Discussion

In this section, a posture learning framework has been proposed for sign language recognition. Although the image quality of the dataset is not very high, the framework shows good results. This framework is robust for rotation and varying image quality. The concept of capsules and pooling are used simultaneously in the network. Originally, Capsules Network's routing by agreement came as an alternative to pooling methods (Sabour et al. (2017)). This research confirms that using both pooling and capsules routing on the same network can improve the networks accuracy and convergence speed. The adaptive pooling used in this framework allows the network to train with multiple sized images, which adds scale-invariance and prevents the network from overfitting.

3.2 Vehicle Logo Recognition

Vehicle Logo Recognition (VLR) is a popular research topic as it facilitates an intelligent transport system as well as traffic monitoring (Psyllos et al. (2010)). For example, fraudulent number plates can be detected if the number does not match the license plate database. VLR can also help to improve an intelligent parking system as well (Fernández-Isabel & Fuentes-Fernández (2015)). VLR faces some practical challenges due to change in visual alignment, rotation and noise, which makes it an ideal scenario for exploring positional dependency-based context cues. Traditionally, handcrafted features are used to represent the image, and the features can be either global or local features. Global features consider all pixels for generating a feature vector over the whole image like histogram oriented gradients (HOG) features (Dalal & Triggs (2005)). Nevertheless, this process makes the feature sensitive to shift, rotation, distortion and scaling. Local features such as SIFT and SURF considers small distinguishable areas and index them accordingly (Lowe (2004), Bay et al. (2008)). These local features are much better than the HOG features, but they are much more computationally complex. Local and global types of features have been used for the VLR task (Llorca et al. (2013), Chen et al. (2016)). However, with the advent of deep neural nets (DNNs), the learning of feature representations from raw images are done by the deep generative and discriminative neural networks (Krizhevsky et al. (2012*a*), Deng et al. (2009)). They performed much better and more efficiently. The DNN based end-to-end image classification became popular with the Alexnet, and different deep neural architectures have been used in the VLR tasks as well (Huang et al. (2015, 2017)). Though the popular DNN based architectures such as CNNs have proven effective in image classification tasks, they fail in some adversarial attacks such as pixel value variations (Moosavi-Dezfooli et al. (2016), Su et al. (2019)). Sabour et al. (2017) proposed a capsule-based hierarchical attention modelling network which deals with these limitations of CNNs.

In this section, a novel VLR classification framework is proposed based on the capsule network. The proposed system performs better than the state-of-the-art CNNs with and without image changes such as rotation and occlusion, and image degradations, including blurring and the noise effects [2]. The rest of this section is organised as follows. In Section 3.2.1, methods based on CNNs capsule networks are introduced. Section 3.2.2 explains the proposed VLR classification framework based on the capsule network. Experimental setup and data are discussed in Section 3.2.3. The results are presented in Section 3.2.4 and Section 3.2.5 discusses the results and summaries the work.

3.2.1 Approach

In this experiment the approach used for building the VLR model architecture is similar to 3.1.1 except it does not have an adaptive pooling operation because all the images in the corpora has same dimension.

3.2.2 The Proposed Framework

This section demonstrates the capsule VLR framework shown in Figure 3.5. The input image is converted to a feature vector using two convolutional layers. The first convolution layer has kernel size $[21 \times 21 \times 128]$ ([height×width×depth]) with stride 2. ReLU nonlinearity has been applied here. The second convolutional layer forms the primary capsules shown in Figure 3.5. Ten groups of convolutional kernels having size $[12 \times 12 \times 10]$ is applied with a stride of 2. Ten convolutional units are generated of the

[2] This section is published as Chen, R., Jalal, M. A., Mihaylova, L., & Moore, R. K. (2018, July). Learning capsules for vehicle logo recognition. In 2018 21st International Conference on Information Fusion (FUSION) (pp. 565-572). IEEE. The thesis author formulated the problem, designed and implemented the model and did the experiments.

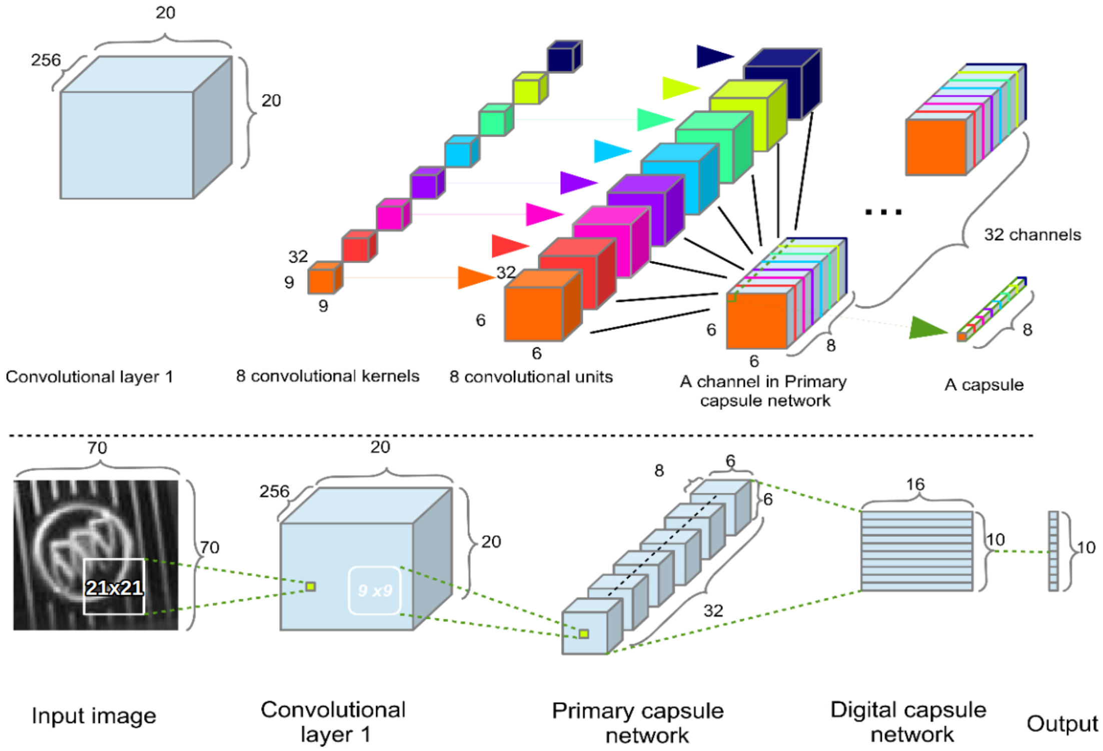


Figure 3.5: *The capsule generation process in the proposed primary capsule layer (Chen et al. (2018)).*

size $[7 \times 7 \times 10]$. These units are re-grouped into ten channels, including one layer from each convolutional units, i.e. each channel is made up from $7 \times 7 = 49$ capsules amidst individual capsule being a vector with ten entries. The primary capsule layer and the final capsule layer are connected through transformation matrices. The reconstruction loss is considered for regularisation and prevent the network from overfitting. The reconstruction process is a decoder connecting the last capsule layer with 2 hidden layers, and each hidden layer has 2048 neurons. The routing process occurs with the primary capsule layer and the final capsule layer. The primary capsule layer has $7 \times 7 \times 10 = 490$ capsules, and the final capsule layer has ten capsules, which requires 4900 transformation matrices sized $[10 \times 30]$. The network weights are learned through backpropagation.

3.2.3 Experimental Setup

The CNN baseline mentioned in Chen et al. (2018) is used as the baseline. The simulation hardware specification is Intel I5 (3210M CPU 2.5GHz 4), 8G RAM and an Nvidia GTX 1070 (extended GPU). The learning rate has been set to be 0.0001. The performance of each method is measured in terms of accuracy (percentage of correctly classified testing images). The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

Dataset

The largest publicly available VLR corpus by Huang et al. (2015) is used to evaluate the proposed framework. The corpus, described in Section 2.3.8, has three sets, and each of these sets (dataset1, dataset2, dataset3) has various challenging scenarios.

Table 3.2: *Summary of experiment results with original test set and challenging VLR test sets: Dataset 1 (added noise), Dataset 2 (added rotation), Dataset 3 (added combination of rotation, noise, occlusion).*

Model	Accuracy (UA%)			
	Original Testset	Dataset 1	Dataset 2	Dataset 3
VLR_caps	**100.0**	**98.5**	**94.8**	**66.0**
CNN	99.4	85.6	89.0	56.5

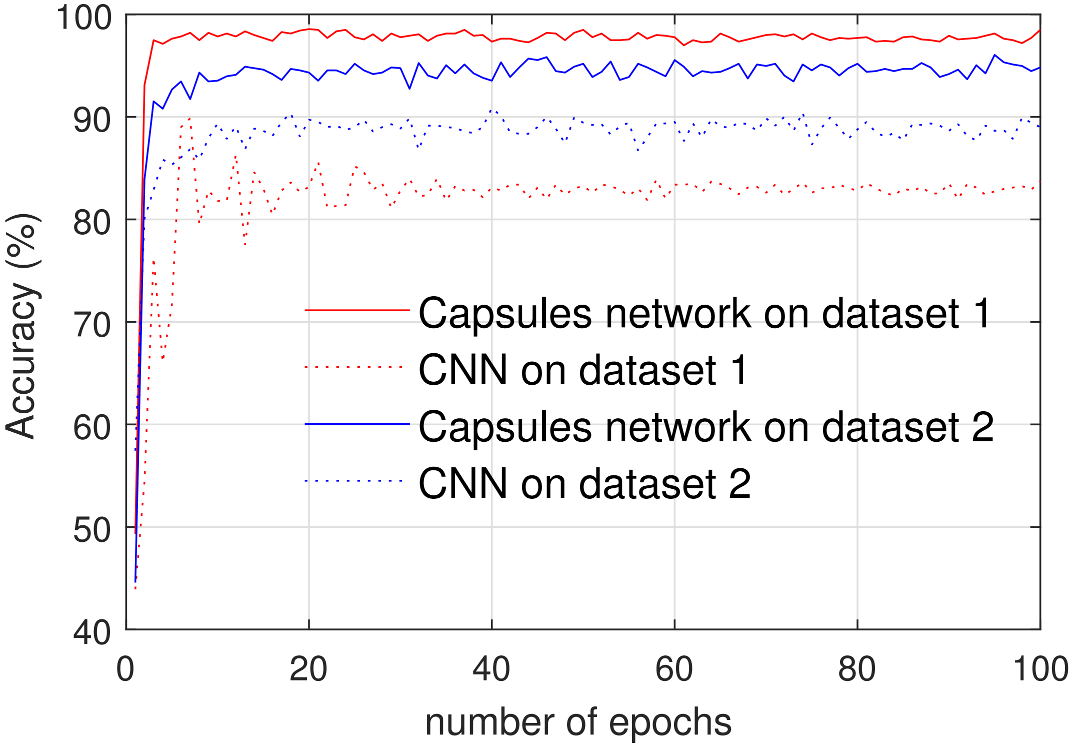
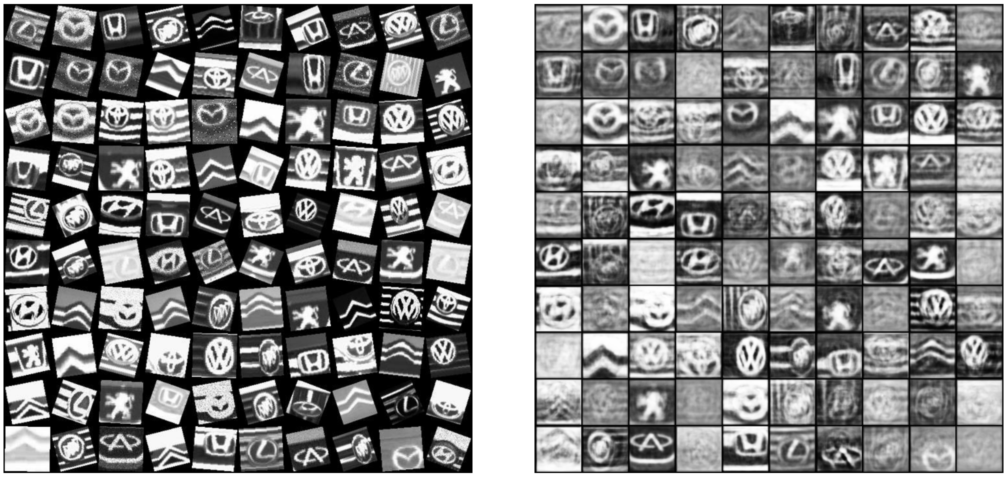


Figure 3.6: *Test accuracy over Dataset 1 and Dataset 2.*



(a) *Ground truth* **(b)** *Reconstructed*

Figure 3.7: *Ground truth vs. Reconstructed Images on Dataset 2*

3.2.4 Results

The results are shown in Table 3.2. Although in the original VLR test set the proposed VLR Capsules network and the CNN have similar performance, the VLR Capsules network reached to that level of accuracy in fewer epochs compared to the CNN. The difficulty level in the original test set is also lower compare to the other testing conditions. The VLR Capsules network produce superior unweighted accuracy in all the scenarios (Figure 3.6). However, it is noticeable that VLR Capsule network is robust with noise and rotation, but when everything is combined along with oscillation and pixel blurring, then the performance drops significantly.

3.2.5 Discussion

Capsule network is advantageous despite changes in image rotation, occlusion, noise and other forms of image degradation. In all the scenarios, it performs better than CNNs, which is also confirmed by these experiments. Although the number of parameters is significantly higher, the hierarchical architecture is different. The routing process is a hierarchical relational attention mechanism. That is why the capsules are learning task-specific individual representations. This will be explored further in the next chapter. Overall, the vehicle logo recognition problem has similar challenges as static sign language recognition, such as rotation, scaling and distortion. Both these experiments show that capsules and hierarchical attentions are beneficial and highly effective to deal with these type of scenarios.

3.3 Summary

Hand postures and gestures can be a great medium of visual communication as they carry universal language and intent information. Section 3.1 proposes a framework that recognises American sign language in static postures. One of the big challenges for visual language recognition are that visual image is subject to constant changes in rotation, scale and other distortions. Thus task-specific context should be learned based on spatial relationship context. Therefore the models will be more invariant to rotation, scaling, and robust to noises and distortions. The proposed models show their robustness to distortions and noise and invariance to rotation and scaling.

Chapter 4

Learning Visual Cues with Attention

A visual language consists of a set of gestures, postures and facial expressions. The visual languages share similar linguistic properties with their respective spoken languages. Human action recognition in diverse and realistic environments is a challenging task. As discussed earlier in Chapter 1, Section 2.1.5 and Section 2.2.4, visual perception is obtained by obtaining visual cues with spatiotemporal cross-modal communication and combining other multi-sensory information for context representation. Visual emotion may denote intent and the situational context. The dorsal and ventral (spatial and temporal) channels are represented as dual channel neural network, and the communication between them as attention fusion at different hierarchies has been studied.

Facial emotion is also a nonverbal communication medium in human-human communication. Facial expression recognition (FER) is a significantly challenging task in computer vision. With the advent of deep neural networks, facial expression recognition corpora have transitioned from lab-controlled settings to more natural environments. Although the new FER corpora have a huge number of samples, they have a high imbalance in samples per target class distribution. Thus deep neural networks (DNNs) suffer from overfitting the data and biases towards specific categories. An end-to-end convolutional-self attention framework is proposed with data augmentation for classifying facial emotions as discrete emotional states. The research is presented in Section 4.2. The AffectNet database is used to validate the framework. The AffectNet database has a large number of image samples in-the-wild (natural and without any lab-restricted environment) settings, which makes this database very challenging. Recognition of action and gestures, along with emotion has a significant impact on the intent and meaning. Due to the prevalence of complex real-world problems, it is non-trivial to produce a rich representation of actions and to produce an effective categorical distribution of large action classes.

In this chapter, two frameworks have been proposed, and they are demonstrated with two experimental setups and two separate corpora. Firstly, a dual-stream spatio-temporal fusion architecture for human action classification is proposed in Section 4.1. The spatial and temporal representations are fused using an attention mechanism. Two fusion techniques have been investigated, and it has been shown that the proposed architecture achieves accurate results with much fewer parameters as compared to the traditional deep neural networks. Secondly, a convolutional self-attention facial emotion recognition model is proposed in Section 4.2. Finally, Section 4.3 summarises the chapter and draws conclusion.

4.1 Gesture Recognition

Human activity recognition in a real-world environment is gaining popularity for its various applications in day to day life. It aims to classify human actions by a series of observations of human actions at a given period. There are numerous applications of action recognition, such as human-robot interaction, wearable technologies, surveillance, multimedia content annotation and measuring similarity. The temporal, motion and contextual aspect of a video makes it different from standard image classification. The spatiotemporal feature representation and generalisation are non-trivial due to the real-world obstacles, such as jitter, lighting conditions, camera viewpoint changes and camera motion.

Many studies have been published on this problem so far. Gaussian mixture models, SVM models and probabilistic models were proposed using hand-crafted features (Laptev & Lindeberg (2003), Laptev et al. (2008), Vrigkas et al. (2014), Zhu et al. (2014), Fathi & Mori (2008)). However, deep neural models become very popular because they can generate high-level features from low level features (Bengio & Lecun (2007), Bengio (2009), Hinton & Salakhutdinov (2006)). Initially, the deep neural architectures do not perform exceedingly compare to the traditional hand-crafted feature based methods (Peng et al. (2016)). Deep convolutional neural architectures are also introduced for vision-based action recognition problems (Simonyan & Zisserman (2014*a*), Karpathy et al. (2014*a*), Tran et al. (2014)). Since then different variants of convolutional neural networks (CNN) were introduced exploring spatial and temporal modelling (Ji et al. (2013), Wang et al. (2016)).

This huge revolution in action recognition research also evolved the experimental data. From stationary camera and controlled environment-oriented (Schuldt et al. (2004)) action database, the research community has moved towards more in the wild and real-world oriented database (Soomro et al. (2012*b*), Karpathy et al. (2014*b*), Abu-El-Haija et al. (2016)).

In this section, the spatiotemporal relationship between the sequences in a video for action recognition has been investigated. An attention mechanism Wang et al. (2018)

has been adopted in the proposed framework [1]. The contribution is two-fold

i. A deep neural net framework that performs with high accuracy (state-of-the-art with UCF-101) using fewer parameters compared to the current state-of-the-art architectures has been proposed.

ii. The fusion between the spatial and temporal channels has been investigated.

This section is organised as follows. Section 4.1.1 discusses the previous work related to this research, Section 4.1.2 explains the approach and the basic building blocks of the proposed frameworks, Section 4.1.3 describes the proposed models, Section 4.1.4 describes the experimental scenarios and interprets the results. Finally, Section 4.1.5 concludes this experiment and proposes future research path.

4.1.1 Related Works

The CNN based methods that are applied for image and video processing require minimum pre-processing. Karpathy et al. (2014*a*) proposed CNN models for video classification with large databases. Moreover, the feature extraction and classification tasks can be solved simultaneously by the network. These methods have provided promising results in the field of computer vision, machine learning and pattern recognition (Sharif Razavian et al. (2014)). Various implementations of CNN networks have been proposed for action recognition (Taylor et al. (2010), Tas & Koniusz (2018)). The 3D CNN feature based action recognition was proposed in Tran et al. (2014), Ji et al. (2013). A two-stream, spatiotemporal stream based approach has been proposed for action recognition in Simonyan & Zisserman (2014*a*). An Attention-based Temporal Weighted CNN (ATW) combines a visual attention model with a temporal weighted

[1] This section is published as M. A. Jalal, W. Aftab, R. K. Moore and L. Mihaylova, ”Dual Stream Spatio-Temporal Motion Fusion With Self-Attention For Action Recognition,” 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2019, pp. 1-7.

multi-stream CNN (Zang et al. 2018). The CNN based hand pose recognition approach was first proposed in Barros et al. (2014). A 77.5% accuracy of hand gesture recognition has been shown in Molchanov et al. (2015) on the VIVA dataset (Ohn-Bar & Trivedi (2014)).

Recurrent neural networks (RNN) have been achieved good results in temporal modelling of sequential data (Hochreiter & Schmidhuber (1997)). A visual action is a sequence of consecutive events that happens in a period of time. Modelling temporal context and modelling the relation between the visual sequences can give a rich representation. Visual sequence modelling has been carried out by several in the literature by Karpathy & Li (2014), Donahue et al. (2014). Veeriah et al. (2015) show that the salient motion feature between the consecutive frames (derivative of states between frames) can be used successfully with long-short-term-memory networks LSTMs (Hochreiter & Schmidhuber (1997)) to model time-series action sequence modelling.

Vaswani et al. (2017) propose an attention mechanism based architecture with feed-forward neural networks to show that dependencies in between the elements in a sequence can be learned by attention mechanism. Attention networks have become vastly popular for modelling long term dependencies (Xu, Ba, Kiros, Cho, Courville, Salakhudinov, Zemel & Bengio (2015), Jaderberg et al. (2015), Vaswani et al. (2017), Ba et al. (2015), Woo et al. (2018)). Wang et al. (2018) propose non-local networks to measure positional dependency within the same sequence.

4.1.2 Approach

In this research, the motivation is to investigate the dependency between spatial and temporal data for action recognition. The dual-stream architecture (Simonyan & Zisserman (2014*a*)) has been adopted for the base of this attention model. This architecture is a computational model of the two-stream hypothesis (Goodale & Milner (1992),

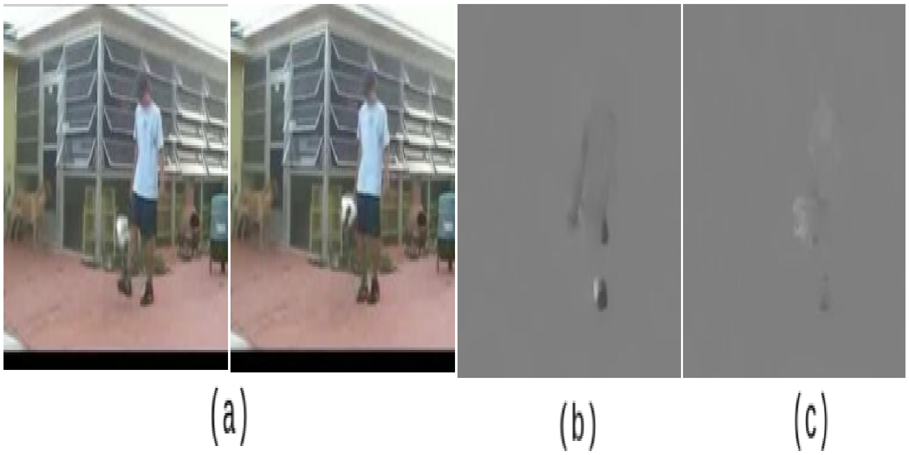


Figure 4.1: *(a) consecutive pair of images (b) and (c) horizontal and vertical component of optical flow*

Sedda & Scarpina (2012)), which states that the human visual cortex system consists of dual channels (dorsal and ventral) to process spatial and temporal information. Static RGB frames will be used for the spatial modelling and dense dynamic optical flow for temporal modelling.

Optical Flow

Optical Flow (OF) is a visual object tracking method that approximates the relative motion of an object and the observer (sensor). The OF algorithm assumes constant pixel intensity across consecutive frames and relatively small object motion (displacement). Based on these assumptions, the OF algorithm calculates a vector displacement field around each pixel for 2D tracking and each voxel for 3D tracking. Various techniques have been proposed to determine the OF such as Horn & Schunck (1981), Lucas et al. (1981) and Brox et al. (2004). In this research, the Brox et al. (2004) method is used to extract OF because the method can be applied with GPU parallel processing that saves significant time for optical flow processing of the video samples. The OF is a displacement vector between consecutive frames t and $t + 1$. The vector $d_t(u, v)$ is the displacement of point (u, v) from frame t to $t+1$. The horizontal u_t^x, u_{t+1}^x and vertical v_t^y, v_{t+1}^y are used as input channels in the CNN for temporal modelling. A sample OF frame is shown in Figure 4.1.

Convolutional Neural Network

Convolutional Neural Networks (CNN), like other neural networks, are multi-layer neural networks. A CNN consists of the convolution and other layers (such as sub-sampling, pooling, ReLU, fully connected, loss) working in a deep learning framework. The initial layers detect low-level features, and the last layers work on the high-level feature space. The characteristics of these networks are that each feature of the layer is connected to a local area, also called *local receptive field*, of the previous layer. These areas are overlapping and, when combined, provide an overall result for a given task. The motivation behind CNNs is reviewed in Section 2.2.2. A brief description of the CNN setup for this experiment is stated here. The feature maps are convoluted with kernels $f[x_b, y_b]$. The kernels are the local receptive field $f[x_a, y_a]$. While the kernels slide through the image, they extract visual features (edges, corners or more abstract features) and combine the set of outputs to form feature maps. If the kernels of size $[h \times w \times N]$ ([height × width × depth], and $n = 1, 2, \cdots, N$) are used, the n^{th} convolutional feature map would be:

$$y_n = f\left(\sum_j g_n * x_j\right), \tag{4.1}$$

where g_n is the n^{th} kernel and x_j $(j = 1, 2, \cdots, J)$ is the j^{th} input feature map of size $[A \times B]$ and $f(\cdot)$ is a nonlinear activation function. The kernel size is significant for preserving locality in the whole network as well as controlling representations (LeCun et al. 2015).

In this research, a smaller kernel size (2-5) is used to preserve locality by considering a small neighbourhood at a time. Smaller kernel sizes can increase non-linearity in the network and enable feature fusion as suggested by Lin et al. (2013), HasanPour et al. (2018). According to Wolpert (1992), stacking more than one CNN layer results in increased nonlinearity and richer representations. Generally, the CNN layer is followed by a pooling layer. The pooling layer improves the discriminability power of the network and robustness to shift and distortions (LeCun et al. (1999)). This means pooling brings invariance to the network. However, it is crucial to control the kernel size in pooling to keep it from losing information. The network learns faster when decorrelated network parameters have zero means and unit variances (LeCun et al. (1999)). Ioffe & Szegedy (2015) proposed batch normalisation for approaching this issue with reducing internal covariance shift in a batch. These methods are adapted to this work.

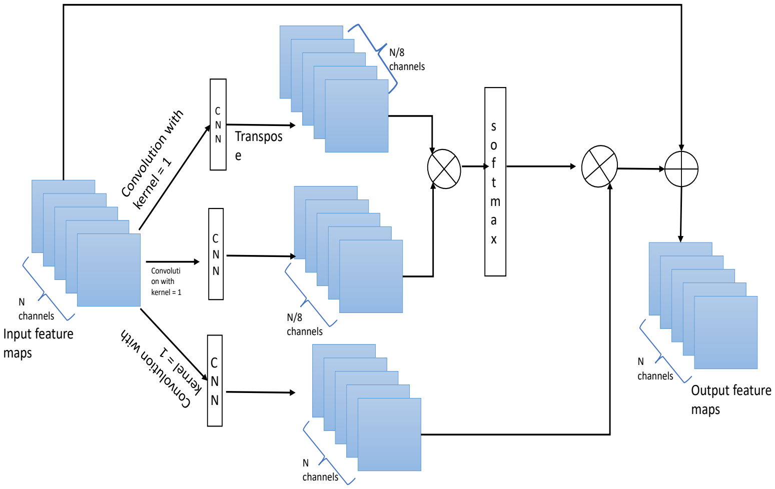


Figure 4.2: *Convolutional self-attention mechanism.*

Self Attention Networks

Self-attention networks flexibly models long-term inter-sequence dependencies. In this research, self-attention is used as non-local networks (Wang et al. (2018), Zhang et al. (2018)) to model the relationships between the regions in the feature maps from previous layers. The mechanism is shown in Figure 4.2. The CNN layer features $\mathbf{y}$ are transformed into feature spaces $\mathbf{j}$, $\mathbf{k}$ and $\boldsymbol{l}$, where

$$j\left(y\right) = W_j y \qquad k\left(y\right) = W_k y \qquad l\left(y\right) = W_l y. \tag{4.2}$$

Here W_j, W_k and W_l are network weights learned through back-propagation. The number of channels in W_j, W_k is less than the number of channels in the features. However, W_k has the same number of channels as input feature y. Dot product is used between j and k. Then normalization is done using the $softmax$ function.

$$e_{ij} = softmax\left(j\left(y_i\right)^T k\left(y_j\right)\right) \tag{4.3}$$

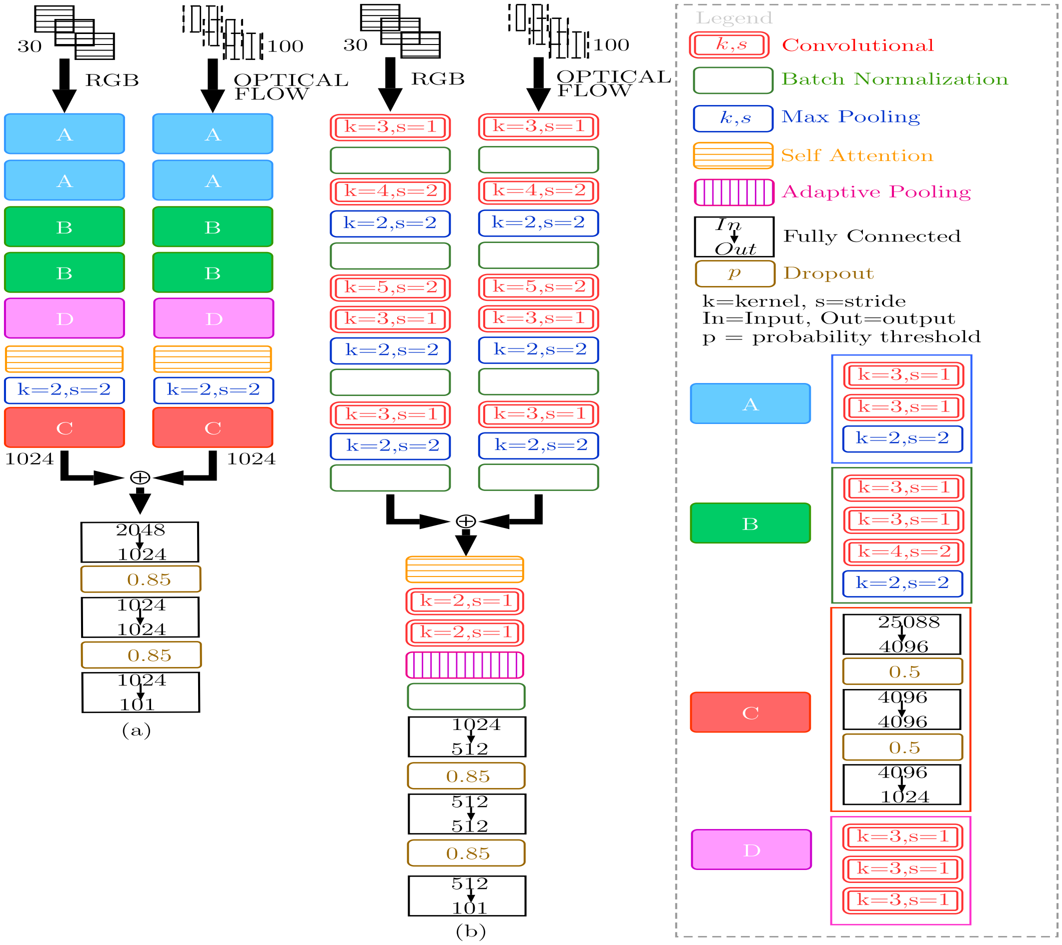
The attention map is calculated by doing matrix multiplication between e and $l(y)$. Scaling factor γ is multiplied with the attention map, which is added with the input feature map.

$$attention\_output = \gamma(el(y)) + y \tag{4.4}$$

In this work, γ is randomly initialised, which learns the non-local dependencies as well as the local neighbourhood dependent representation.

4.1.3 The Proposed Framework

The contribution is two-fold for this proposed architecture. Two models based on two different fusion techniques are proposed. The first model shows late fusion, and the



(a)

(b)

Figure 4.3: *Early and late fusion with attention network*

second model shows early fusion. The frameworks are shown in Figure 4.3. Each video V is divided into N frames $\{f_1, f_2, ..., f_N\}$. Consecutive frames are highly redundant, frames are chosen sequentially but having a small time distance with each other.

Late Fusion

This model is inspired from VGGNET (Simonyan & Zisserman (2014*b*)) for the late fusion model. The spatial stream operates on a sequence of RGB video frames. The frames are stacked and fused by interpreting each of the frames as an individual channel. The model consists of 5 different CNN layer blocks. Block A has two convolution layers with kernel size 3, followed by a maxpooling layer with kernel size 2. Block B has three convolution layers with kernel size $3, 3, 4$ respectively, followed by a maxpooling layer with kernel size 2. Block D has three convolution layers with kernel size 3. This followed by a convolutional attention layer. The attention layer (Section 4.1.2) fuses the spatial feature maps from block D of channels sized 128. The convolution layers in the attention layer have a kernel size of 1. Block C has three fully connected layers. The temporal stream has a similar architecture as shown in Figure 4.3. The output o_r of spatial o_s and temporal streams o_t are fused using concatenation

$$o_r = o_s \oplus o_t, \tag{4.5}$$

(where $\oplus$ denotes concatenation). Then o_r is fed to fully connected layers, and *dropout* is used for regularisation.

Early Fusion

The proposed early fusion model has a simplistic approach. Both the spatial and temporal streams have a smaller feature extraction layer compared to the late fusion model. The input channels are fed to a convolution layer with kernel size 3 followed by a batch normalisation layer. The number of output channels is 128. The feature maps are slowly down-sampled and then up-sampled in the following layers. Maxpooling has been used to introduce sparsity in the network parameters. Batch normalisation is used after each convolution layer to reduce correlation among the parameters at the same layer within the network. Both spatial and temporal stream produce 128 channels of feature maps. These feature maps are fused using a self-attention layer. The network weights W_j, W_k, W_l, mentioned in Section 4.1.2, are three convolutional layers with kernel size 1. To reduce computation time, the number of output channels for W_j and

W_k is one-eighth of the input channels (256) in the self-attention layer. The number of output channels for W_l is the same as the input channels in the self-attention layer. The scaling factor γ is randomly initialised. The kernel size in the attention layer is set to 1 to perform feature level fusion. The output feature maps from the attention layer are fed into an adaptive pooling layer to produce fixed-sized output feature maps. These feature maps are given as input to three fully connected layers. Similar to the previous model, *dropout* has been used as a regulariser.

4.1.4 Performance Evaluation

Dataset

In this work, the UCF-101 dataset (Soomro et al. (2012*a*)), described in Section 2.3.7, is used for evaluating the performance of the proposed method.

Experiment

The experiments have been conducted using the PyTorch (Paszke et al. (2017)) deep learning framework. 10 motion frames are taken with interval for the RGB stream training. In the training phase, the frame sequences are changed in an interval of 100 epoch. For example, if we take $f_5, f_{15}, f_{25}, f_{35}, ..., f_{95}$ frames in the the first 100 epochs, we change the sequence to $f_8, f_{18}, f_{28}, ..., f_{98}$ frames for 100 to 200th epoch. Also, the frames are augmented using adaptive scaling and random cropping at every epoch. The whole model training is done using augmented data with different sequences.

Learning

The Adam optimiser by Kingma & Ba (2014) is applied to a mini-batch of 25 videos with categorical cross-entropy loss. The momentum and weight decay are set to 0.9 and

0, respectively. Throughout the network, the learning rate is set to 0.0001. Dropout layers are used with the fully connected layers to prevent the network from overfitting. In the late fusion model, Figure 4.3, the dropout rate in block C is set to 0.5. The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

Data Augmentation

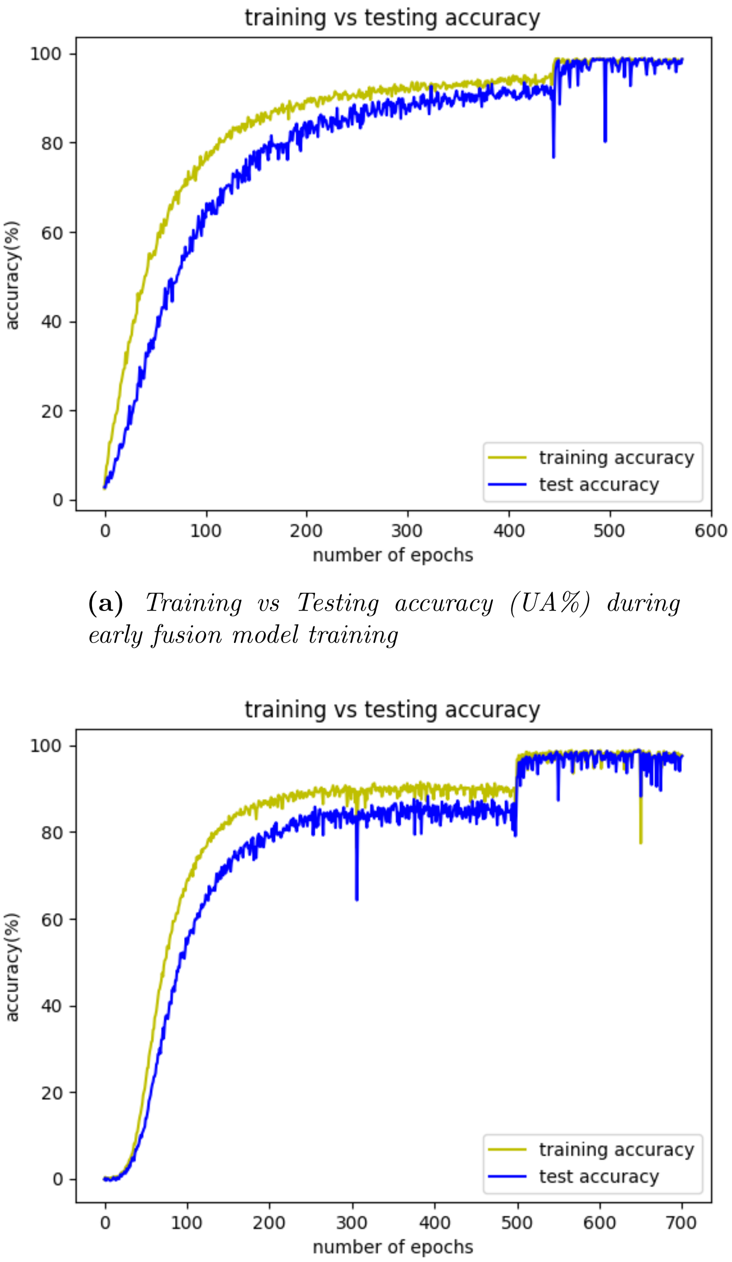
Random horizontal flipping and random cropping are applied to the frames for data augmentation to increase the diversity of the training samples. The frames are normalised and re-sized to [224 × 224] images.

Table 4.1: *Comparison of the proposed methods with other state-of-the-art methods in terms of unweighted accuracy (UA%)*

Method	**Accuracy (UA%)**
Two Stream Simonyan & Zisserman (2014*a*)	88.0
C3D (3 nets) Tran et al. (2014)	85.2
Two stream + LSTM Ng et al. (2015)	88.6
Two stream VGGNet-16	90.9
Long Term Temporal Convolution Varol et al. (2017)	91.7
KVMF Zhu et al. (2016)	93.1
TSN 3 Modaliites Wang et al. (2016)	94.2
Proposed Network I(late fusion with augmented data)	**98.8**
Proposed Network II(early fusion with augmented data)	**99.1**

Results

The proposed frameworks have been evaluated on UCF-101 dataset with the split *I*. The unweighted accuracy comparisons with the state-of-the-art systems are shown in Table 4.1. It can be clearly seen that both the proposed frameworks achieved a state-of-the-art results in UCF-101 test dataset. Two types of models, i.e. late fusion attention model and early fusion attention model, are demonstrated. With the late fusion attention model, very deep neural architecture is explored. The training vs testing accuracy is shown for both of the models in Figure 4.4.



(a) *Training vs Testing accuracy (UA%) during early fusion model training*

(b) *Training vs Testing accuracy (UA%) during late fusion model training*

Figure 4.4: *Training vs Test in different attention fusion hierarchy with UCF-101*

Two points can be clearly drawn from Figure 4.4. Firstly, the representation learning on training data with augmentation generalises for the test data. Secondly, from the initial training stage, both the networks do not over-fit. Furthermore, in this research, the training videos are processed in such a way that most of the frames are utilised. As mentioned in section 4.1.3, every 100th epoch the frame sequences have been changed, but the frame order remains intact. This allows using the maximum training data as well as provide good data augmentation.

The early fusion attention model converges faster than the late fusion attention model. Also, the early-fusion-attention model has fewer parameters than the popular deep neural networks, such as VGGNet, AlexNet, ResNet.

4.1.5 Discussion

The dual-stream learning is to model the dorsal and ventral stream learning hypotheses for human cognition. Two stages of fusion in between those streams are investigated. Also, a reduction of the computational cost with the early fusion attention model is seen, which has a smaller network size without compromising the performance. Finally, this work has tried to bring together good practices for designing CNN and deep networks.

4.2 Visual Emotion

Facial expression is a non-verbal signal for conveying emotions. Emotion carries paralinguistic cues and information about individual intention. Every research paper on this topic has different perspectives on the facial expression recognition (FER) problem. In this section, a deep neural network model is proposed, to understand the positional dependencies between the facial regions across the channels in CNN. The goal of this network is to learn those dependencies to produce the categorical distribution of emotions from static facial expressions [2].

The self-attention attention mechanism based on convolution blocks has been used here. This section is organised as follows, Section 4.2.1 is about the related works, Section 4.2.2 explains the approach towards the architecture and relevant models, Section 4.2.3 describes the proposed architecture, Section 4.2.4 discusses the experimental settings, Section 4.2.5 elaborates the implications of the results and Section 4.2.6 makes further discussion and propose possible future enhancements.

4.2.1 Related Works

FER systems can be either static or dynamic (Li & Deng (2018)). The static-based models (Mollahosseini et al. (2015), Liu, Han, Meng & Tong (2014)) encode the spatial features from a single image and the dynamics-based models encodes the spatiotemporal features over a time span (Zhao et al. (2016), Zhao & Pietikainen (2007)). The FER systems were built with hand-crafted feature descriptors (e.g histogram oriented gradient (HOG) (Orrite et al. (2009), Dahmane & Meunier (2011)), local binary pattern (LBP), local ternary pattern (LTP) (Gritti et al. (2008)) and Gabor Filter (Zhang et al. (1998)). These feature descriptors have their advantages and disadvantages (Carcagnì et al. (2015)). These features are used with generative or discriminative classifiers (eg. SVM, GMM Carcagnì et al. (2015), Tariq et al. (2013)). Deep learning has been incorporated in FER systems for a long time (Liu et al. (2015), Ijjina & Mohan (2014)). The high level of abstraction and non-linear representations have resulted in state-of-the-art results in FER systems. Deep networks can be trained for flexible tasks (Khalajzadeh et al. (2014)). Hybrid models with handcrafted features and neural net-

[2] This section is published as M. A. Jalal, L. Mihaylova and R. K. Moore, "An End-to-End Deep Neural Network for Facial Emotion Classification," 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2019, pp. 1-7.

works have recently become popular (Connie et al. (2017), Georgescu et al. (2018)). Due to the nature of these features, the facial emotion recognition was biased towards its subject, environment and colour. The generalisation of correlation between the facial parts is crucial for subject independent emotion classification. . Traditional CNN such as AlexNet performed very well in these problems (Mollahosseini et al. (2015), Mollahosseini et al. (2016)). Learning task based contextual representation was missing from these early FER models. These task based context generally involves the saliency regions and interdependencies among the regions for FER tasks. Sometimes these dependencies over time build emotion context. Some research proposed temporal context modelling for learning temporal dependencies by using recurrent neural networks (RNNs) (Chen & Jin (2015), Ebrahimi Kahou et al. (2015)). A dual channel network was also used for learning temporal features (Fan et al. (2016)). Fan et al. (2016) used VggNet (Simonyan & Zisserman (2014*b*)) and RNN in one channel and 3D CNN on the other channel. These hybrid models often result in high performance for FER tasks. These architectures combine the spatial local feature extraction capability of CNNs with RNNs temporal modelling (Kahou et al. (2013), Liu, Wang, Li, Shan, Huang & Chen (2014)). They proposed an average-based aggregation strategy for the features contrary to Jain et al. (2018), who employed feature level fusion. However, despite doing partial generalisation using the deep neural structures, these models did not have a selection mechanism (visual attention or saliency points) for representational learning.

Vaswani et al. (2017) proposed an attention mechanism based architecture with feedforward neural networks where they showed that the sole attention mechanism could learn the global dependencies between the input and output. Today attention networks have become vastly popular for modelling long term dependencies (Xu, Ba, Kiros, Cho, Courville, Salakhudinov, Zemel & Bengio (2015), Jaderberg et al. (2015), Vaswani et al. (2017), Ba et al. (2015), Woo et al. (2018)). Wang et al. (2018) proposed a method to measure positional dependency within the same sequence. Parikh et al. (2016) also had a similar approach, i.e. to decompose a sequence into a subsequence and compare it with the other subsequences in the same sequence.

4.2.2 Approach

Classifying emotions from facial expressions is based on the relative differences in facial muscles at a given instance. Traditionally, FER systems consider facial muscles (action units) and visually salient facial landmarks for classifying emotion. But, here, let us consider the problem a little differently. Rather than looking for particular regions in the face, I consider looking for the relational dependencies between each position with the others within the same image by adapting self-attention (Wang et al. (2018), Parikh et al. (2016), Zhang et al. (2018)). In order to build this approach end-to-end, a stack of CNN with the different kernel sizes is employed on top of the network to learn the spatial features.

Convolutional Neural Network

The detailed mechanism and motivation of CNNs are reviewed in Section 2.2.2.

For this experiment, I opt to use a smaller kernel size to preserve locality by considering a small neighbourhood at a time. Also, for doing feature level fusion, smaller (3-5) sized kernels are used (Lin et al. (2013), HasanPour et al. (2018)). Stacking more than one CNN layer results in increased non-linearity and richer representations (Krizhevsky et al. (2012*a*), Simonyan & Zisserman (2014*b*)). Sometimes, the CNN layer is followed by a pooling layer that focuses on improving the discriminability power of the network and robustness to shift and distortions (LeCun et al. (1999)). However, it is crucial to control the kernel size in pooling to keep it from losing information. The network

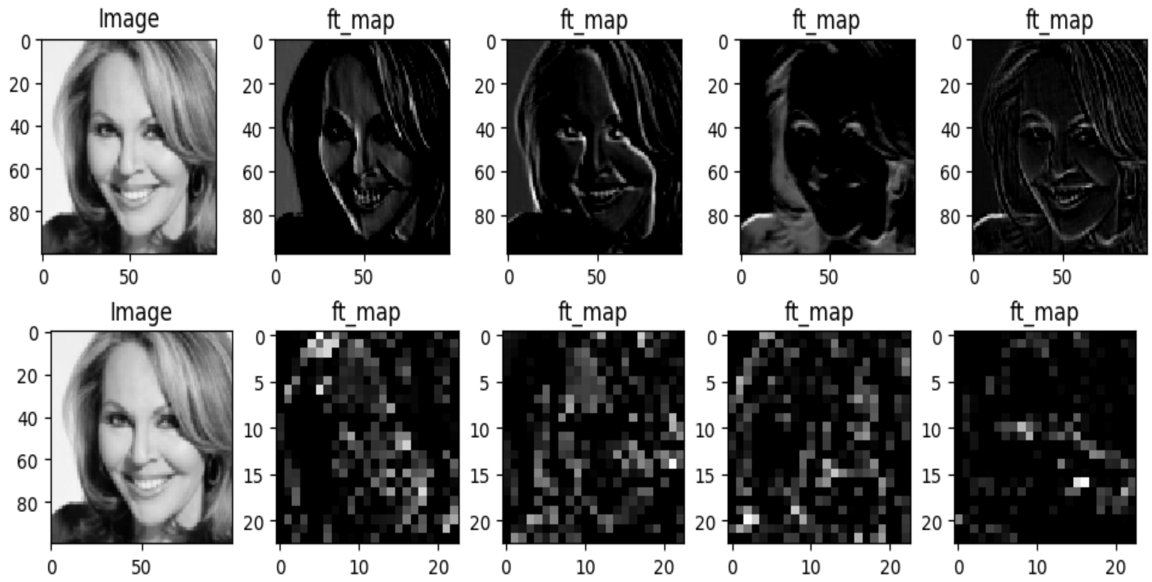


Figure 4.5: *Original Image and sample feature maps after c1 (top row) and after c2 (bottom row)*

learns faster if the network parameters are decorrelated, and they are linearly transformed to have zero means and unit variances (LeCun et al. (1999)). Ioffe & Szegedy (2015) proposed batch normalisation for approaching this issue with reducing internal covariance shift in a batch.

Self Attention Networks

Self-attention does model long-term inter-sequence dependencies. Self-attention as non-local networks (Wang et al. (2018), Zhang et al. (2018)) are used to model the relationships between the regions in the feature maps from previous layers. The self-attention mechanism used here is described in Section 4.1.2.

The self-attention learns non-local dependencies and biases among the local neighbourhood of the segment.

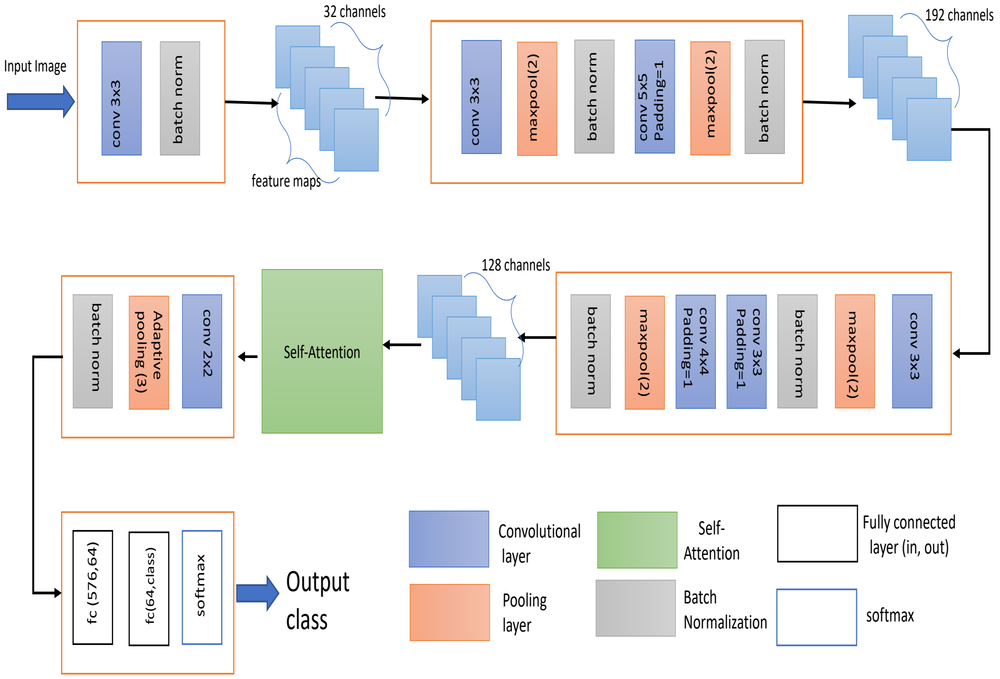


Figure 4.6: *The proposed self-attention network architecture for emotion classification*

4.2.3 Proposed DNN Architecture

The proposed framework is shown in Figure 4.6. The first CNN block (c1) has one convolutional layer of output channel size 32 and kernel size 3. This is followed by batch normalisation and rectified linear unit (ReLU) nonlinearity. The second CNN block (c2) has two convolution layers of kernel size 3 and 5. Maxpooling is used to introduce sparsity in the network. Two maxpooling layers are used with kernel size 2. Throughout the network, the same kernel size (2) has been used for maxpooling. The input feature channels and output feature channels in c2 are 32 and 192 respectively. The third CNN block (c3) has three convolutional layers. A max-pooling layer and batch normalisation follow the first convolutional layer. The remaining two convolutional layers are stacked together. The input feature channels and output feature channels in c3 are 192 and 128 respectively. Both upsampling and downsampling is performed in c3. Sample outputs feature maps have been visualized in Figure 4.5.

Three CNN blocks are followed by a self-attention layer (a1). The network weights W_j, W_k, W_l, mentioned in Section 4.1.2, are three convolutional layers with kernel size 1. The number of output channels for W_j and W_k is one-eighth of the input channels in the self-attention layer. The number of output channels is decreased to reduce the computation time. The number of output channels for W_l is the same as the input channels in the self-attention layer. The scaling factor γ is randomly initialised. The kernel size in the attention layer is set to 1 to perform feature level fusion.

The output feature maps from a1 are fed into the fourth CNN block (c4). This block has a convolution layer with an input channel size 128 and output channel size 64 with kernel size 2. This convolution is followed by an adaptive average pooling layer to produce a fixed length of 3×3 sized feature maps. Throughout the network, after each pair of convolution and max-pooling, batch normalisation is performed to reduce the correlation between the parameters in the network.

The final block (l1) is a dense layer. It (d1) has two parts. The first part contains one fully-connected layer, followed by a dropout (0.85) layer. ReLU activation function has been used with this. The second part has one fully connected layer, followed by a softmax layer. Dropout (0.85) is used for regularisation.

4.2.4 Experiment Setup

Dataset

AffectedNet corpus has been used for this research. AffectNet is described in Section 2.3.1.

Data Preparation

The number of samples in each category clearly shows that the classes are heavily imbalanced. The standard training split of the database is also imbalanced. A heavy up-sampling and down-sampling strategy have been adopted to cope with this problem. For training, I randomly chose 13000 images for each category from the official training split. The under-represented classes are up-sampled by replicating and augmentation. So, the number of samples for each class is neutral (13000), happy (13000), sad (13000), surprise (13000), fear (12756), disgust (11409), anger (13000), contempt (11250). However, it can be seen that the data is imbalanced contrary to the approach taken by Mollahosseini et al. (2017*b*) where they took 15000 samples and up-sampled heavily to make the data balanced thus making this experimental scenario more challenging.

At the time of doing this particular experiment, no official test set was released. The official standard validation split is used for testing the performance of the proposed framework as the validation set is published for testing purposes. Each of the classes

Table 4.2: *Test Accuracy (UA%) on AffectNet validation set*

Method	**Accuracy (UA%)**
Baseline (AlexNet)	58.0
Proposed Network	**93.8**

has 500 sample images for testing.

Experiment

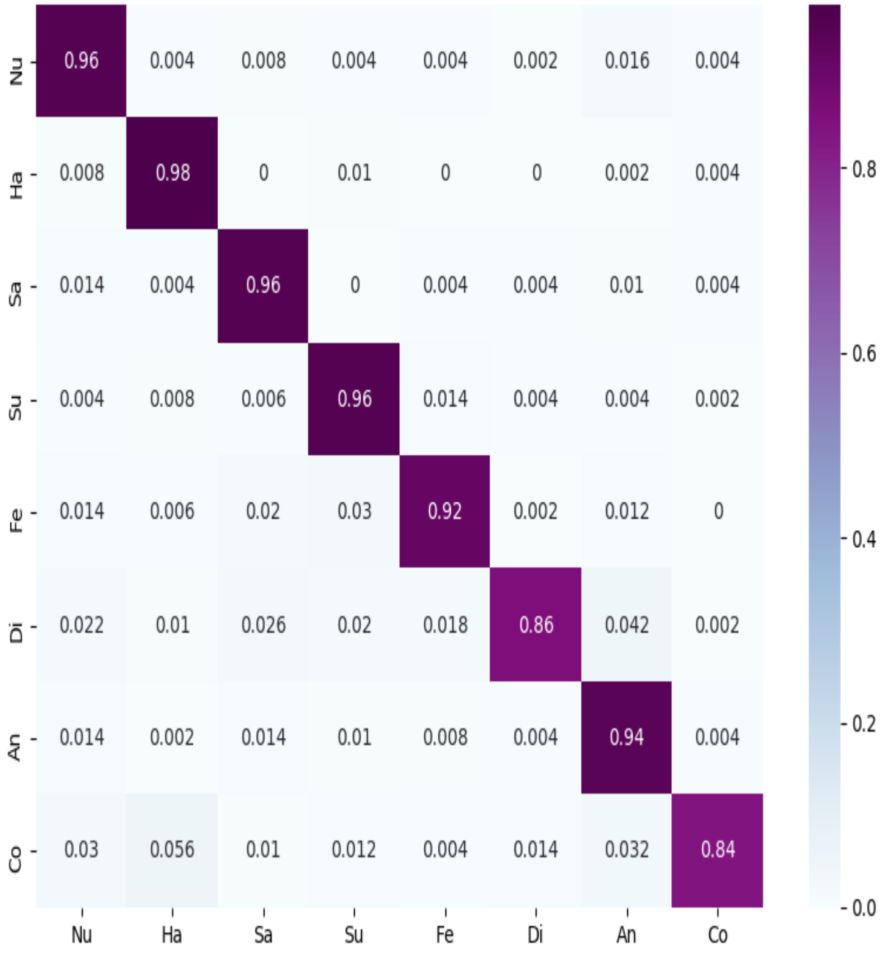
The experiments have been conducted using the PyTorch Paszke et al. (2017) deep learning framework. An Nvidia GTX 1080ti GPU has been used for executing the experiments. The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

Learning

The Adam optimiser Kingma & Ba (2014) has been applied to mini-batch of 500 images with categorical cross-entropy loss. The momentum and weight decay are set to 0.9 and 0 respectively. Throughout the network, the learning rate is set to 0.0001. To prevent the network from over-fitting dropout layers have been used with the fully connected layers. In the proposed model, at Figure 4.6, the dropout rate is set to 0.85 in the classifier section.

Data Augmentation

Random horizontal flipping and random cropping have been applied to the frames for data augmentation in order to increase the diversity of the training samples. The frames are normalised and re-sized to $[224 \times 224]$ images.



	Nu	Ha	Sa	Su	Fe	Di	An	Co
Nu	0.96	0.004	0.008	0.004	0.004	0.002	0.016	0.004
Ha	0.008	0.98	0	0.01	0	0	0.002	0.004
Sa	0.014	0.004	0.96	0	0.004	0.004	0.01	0.004
Su	0.004	0.008	0.006	0.96	0.014	0.004	0.004	0.002
Fe	0.014	0.006	0.02	0.03	0.92	0.002	0.012	0
Di	0.022	0.01	0.026	0.02	0.018	0.86	0.042	0.002
An	0.014	0.002	0.014	0.01	0.008	0.004	0.94	0.004
Co	0.03	0.056	0.01	0.012	0.004	0.014	0.032	0.84

Figure 4.7: *Confusion Matrix for the eight emotion category labels Nu (neutral), Ha (happy), Sa (sad), Su (surprise), Fe (fear), Di (disgust), An (anger), Co (contempt)*

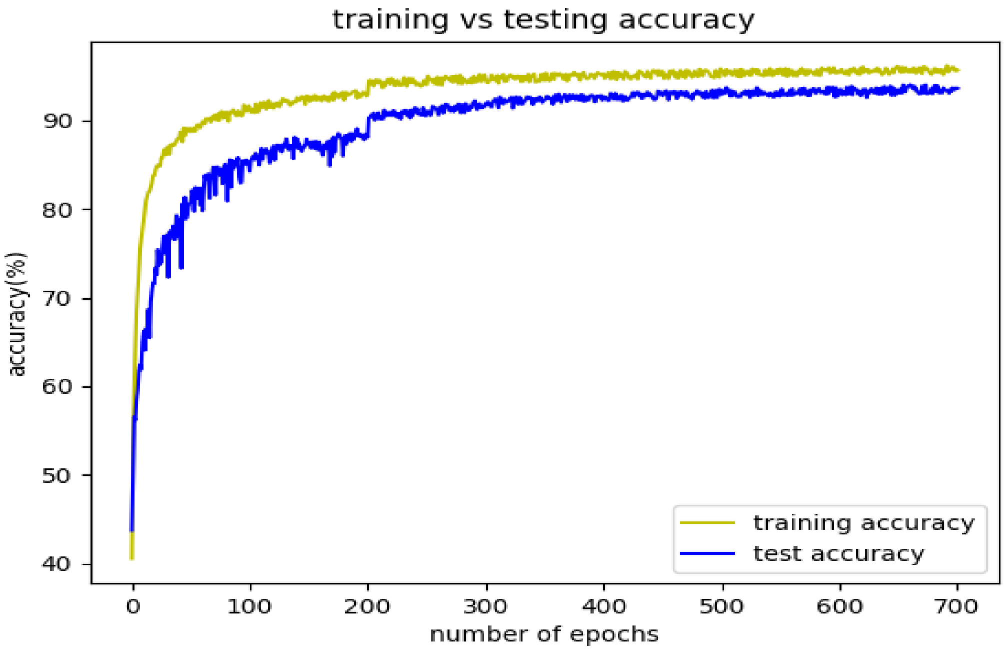


Figure 4.8: *Training and Test accuracy (UA%) in 700 epochs*

4.2.5 Results

The proposed framework has been evaluated with the 'AffectNet' (Mollahosseini et al. (2017*b*)) corpus. The creators of the 'AffectNet' database have not released any official test set yet. After contacting the authors of Mollahosseini et al. (2017*b*), they advised using the official validation set as the test set. According to Mollahosseini et al. (2017*b*), a baseline system has been implemented on AlexNet Krizhevsky et al. (2012*a*). The results on the validation set are reported officially in Mollahosseini et al. (2017*a*). Also, the architectural reference mentioned in the paper Mollahosseini et al. (2017*b*) is followed. AlexNet is built upon sixty-two million parameters, making it a very deep neural network architecture. The proposed framework has fewer parameters compared to the very deep neural network architectures. However, it is clear in Table 4.2 that the performance gain with the proposed framework is more than 30%. It showed that the performance of the proposed framework is not biased against any particular emotion category. In Figure 4.7 the confidence scores are given. The experiment is run five times and took the mean of the unweighted accuracy (UA) scores. The accuracy on AffectNet validation set is reported as 93.8%.

The network in the proposed framework has considerably fewer parameters in AlexNet and also most of the state-of-the-art deep neural networks. Hence, the computational cost is reduced without compromising the performance of the framework. Figure 4.7 shows the confidence values among the classes. The images with 'happy' labels show the highest confidence, followed by surprise, sad, neutral and anger. There has been a conflict among the disgust and contempt emotion labelled images as human annotators agree upon only 60.7% of the image labels (Mollahosseini et al. (2017*b*)). Figure 4.8 clearly shows that the network learns generalized deep representations on training data. Since the initial training phase, the network successfully managed to avoid overfitting.

4.2.6 Discussion

In this research, a self-attention based facial emotion recognition model has been demonstrated. The non-local feature dependencies are extracted in the image segments and prioritise image segments. Also, small neighbourhoods are considered with smaller convolutional kernels while creating feature maps. Massive data augmentation is also performed. These approaches have led to a balance in the number of samples in each category and richer representations in the network. At the time of this research, there was no official test set available. This section can be seen as a pilot study.

4.3 Summary

According to the dual learning cognitive theories discussed in Chapter 2, learning happens in dual neural pathways, i.e. dorsal and ventral pathways. A dual-stream spatiotemporal fusion architecture for human action classification is proposed in Section 4.1. The spatial and temporal data are fused using an attention mechanism. Two fusion techniques have been investigated, and the performance has been discussed with a comparison of different fusion techniques. Section 4.2 presents a self-attention based convolutional neural network with data augmentation for facial emotion recognition from images. An important takeaway from this chapter is about model adaptation in neural network and transfer learning. It is evident that all the tasks do not need very deep architectures, and the structure of the architecture is more important than the depth.

Chapter 5

Learning Spatio-Temporal Context for Speech Emotion

The conceptualisation of emotion involves the attentional cueing from spatial and temporal representations. The exogenous and endogenous mechanisms of selective attention relate the perceptual information with the learned experience or goal/task-oriented motivation to achieve relevant contextual cues. These salient context cues are essential to understand emotion. In this chapter, speech emotion is modelled into discrete sensory events (happy, sad, neutral, etc.) for obtaining emotional intelligence, which affects the understanding of context and meaning of speech. The meaning and intent of speech are dependent on emotion context about 'what' is said and 'how' / 'where' it is said. The above premises invokes some research questions about the spatial and temporal nature of emotion and its relation to the computational models. These are:

- The capsule networks in the previous chapter claim to learn task-oriented structural relations. Do speech emotions have such temporal representations, and what is the nature of those representations?
- How spatiotemporal context modelling affects emotion categorisations?

- Is speech emotion a static or dynamic phenomenon?
- Is there any universal representation of emotion among different variations of accents in a language?

There are mainly two different approaches for representing emotions, i.e. categorical and dimensional. In categorical representation, the emotions exist as discrete labels such as happy, angry, sad etc., whereas the dimensional approach emphasises understanding emotions in terms of valence and arousal. In this thesis, it has been assumed that emotion is a categorical perception representing discrete sensory events.

Section 2.3 describes the corpora used in this chapter. The rest of the chapter is dedicated to investigate and answer these questions and structured as follows: Section 5.1 presents a brief background about the computational models of emotions and their relevance to this work; Section 5.2 proposes a *rnn-cnn-capsule* based hybrid topology framework to learn temporal dynamics and understand capsule representation of internal states. However, capsule networks are massive and computationally expensive. Section 5.3 proposes a convolutional attention based spatiotemporal modelling technique *CSA* that is significantly smaller and faster than all the *state-of-the-art* models proposed with similar performance; Section 5.4 briefly discuss the impact of cross-corpus training and the implication of natural, elicited and acted emotions in emotion recognition. The convolution and self-attention blocks are discussed briefly with experiment specific motivations as the previous chapter for convenience. Finally, Section 5.5 brings the whole chapter together and summarise the chapter. Each section presents the motivation behind it, formulates the problem, describes the proposed framework and relevant literature in that section, analyses the results, and discusses the implications.

5.1 Computational Models of Speech Emotion

From a pattern recognition viewpoint, speech emotion recognition (SER) requires a front-end that extracts a set of features that ideally bear maximum correlation with emotion attribute while having the least sensitivity to other speech aspects. However, such signal parameterisation through feature engineering is challenging. In practice, the most popular features are eGeMaps (Eyben et al. (2016)), MFCCs (Nwe et al. (2003)) and filterbanks (Tin Lay Nwe et al. (2001)). These features are used with different classifiers such as hidden Markov models (HMMs) (Schuller et al. (2003)), support vector machines (SVMs) (Cao et al. (2015)), GMM (Loweimi et al. (2015)), deep belief networks (DBNs) (Le & Provost (2013)) and deep neural networks (DNNs), and treated as a standard categorical classification task.

Deep neural networks (DNNs) can solve the data representation problem by learning a series of task-specific transformations. The network layers extract abstract representations and also filter out the irrelevant information, which leads to a more accurate classification (Mirsamadi et al. (2017), Zhang et al. (2016)) and better generalisation (Kawaguchi et al. (2017), Huang et al. (2014)). Temporal models were also proposed for modelling sequential data with mid to long-term dependencies (Lim et al. (2016), Kim et al. (2017), Wu et al. (2019), Jalal, Loweimi, Moore & Hain (2019)

5.2 Temporal Clusters for SER

Emotion recognition from speech plays a significant role in adding emotional intelligence to machines and making human-machine interaction more natural. In this section, a novel temporal modelling framework is proposed for robust emotion classification using bidirectional long short-term memory network (BLSTM), CNN and Capsule networks. The BLSTM deals with the temporal dynamics of the speech signal by effectively representing forward/backward contextual information, the CNN along with the dynamic routing of the Capsule Net learn temporal clusters, which altogether provide a state-of-the-art technique for classifying the BLSTM representation. The goal is to build a deep temporal model of utterances through leveraging the information encoded in the speech dynamics and sequential nature. The experimental results prove the effectiveness of the proposed hybrid topology, leading to the state-of-the-art performance (at the time of publish of this research Jalal, Loweimi, Moore & Hain (2019)) on FAU-Aibo (Schuller et al. (2009), Steidl (2009), Batliner et al. (2008)) and RAVDESS databases (Livingstone & Russo (2018)) in both binary and 8-class emotion classification tasks. The proposed approach was compared with a wide range of architectures on the FAU-Aibo, and RAVDESS corpora and remarkable gain over state-of-the-art systems were obtained. For FAO-Aibo and RAVDESS 77.6% and 56.2% accuracy was achieved, which is 3% and 14% (absolute) [1].

The rest of this section is organised as follows. In Section 5.2.1, representation learning through RNNs, 1D-CNN and capsule routing networks are reviewed and discussed. Section 5.2.2 explains the proposed architecture and its advantages. Section 5.2.3 presents the experiment setup, datasets, features etc. In Section 5.2.4 the experimental results are presented along with discussion and Section 5.2.5 makes further discussion and concludes this section.

5.2.1 Approaches to Representation Learning

Speech is sequential data with high temporal dynamics (Lee & Tashev (2015)). The speaker-related properties like emotion are distributed in the utterance and vary at a slower pace than the lingual content. To adequately capture such attributes, the em-

[1] This section is published as Jalal, M. A., Loweimi, E., Moore, R. K., & Hain, T. (2019, September). "Learning temporal clusters using capsule routing for speech emotion recognition". In Proc. Interspeech (Vol. 2019, pp. 1701-1705).

ployed algorithm should be capable of handling sequence properties and go beyond mere short-term processing techniques. There are two main approaches in neural networks (NN) to deal with such sequential dynamics: augmenting the input by stacking the previous/next frames or using a network with some memory, representing the temporal evolution of the system's internal state. Feedforward NN (Neural Network) and RNN (Recurrent Neural Network) are examples of the first and second cases, respectively.

Recurrent Neural Networks

In a regular feedforward NN, which is a memoryless system, the temporal information is provided through the input by stacking neighbouring contextual frames. Setting the context length is done empirically and is a task and data-dependent practice (Graves & Schmidhuber (2005)). On the other hand, Recurrent Neural Networks (RNN) by utilising the internal state, keep track of what has happened in the past and consider such temporal evolution while making a decision at each time step. RNN is discussed with the architecture along with the motivations in Section 2.2.3. In the speech samples, for each time step t, the activations of the unfolded network are concatenated over the last $T-1$ time steps, providing a matrix of $T \times N$, where N is a number of nodes of the BLSTM layer, and T is the context length (including current frame t). The output activation will be $o_1, o_2, ..., o_{T-1}, o_T$. In this work these BLSTM activations are the first temporal representation in the hierarchy.

$$g_t = f(W_{xg}x_t + W_{gg}g_{t+1} + b_g). \tag{5.1}$$

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \tag{5.2}$$

where $W_{xg}, W_{gg}, W_{xh}, W_{hh}$ are weight matrices between the corresponding layers.

1D-Convolutional Capsules

The activation of each time-step is fed into CNNs that have been successfully applied to various speech-related tasks such as automatic speech recognition (Trigeorgis et al. (2016), Palaz et al. (2015)) and emotion classification (Huang et al. (2014), Mao et al. (2014)). CNNs can model patterns with high robustness to variations and distortions (LeCun & Bengio (1998))(see Section 2.2.2). A key component of this work is that it has applied 1D convolution to each of the temporal activation states of BLSTM for learning different abstract temporal representations. The output matrices of these convolution units are concatenated to form a capsule, and a squashing function is applied to get vector representation for each capsule. A capsule is a group of neurons. The output of the capsule j, $\mathbf{o}_j^t$, is

$$\mathbf{o}_j^t = \mathbf{g}(\mathbf{s}_j), \tag{5.3}$$

where $\mathbf{s}_j$ is the input (concatenated 1D-Conv output) to the capsule j and $\mathbf{g}(.)$ is a squashing function. The intuition behind squashing is to shrink short vectors (less likely ones) to zero and long vectors (more likely ones) to nearly (below) 1. In Sabour et al. (2017), $\mathbf{g}(\cdot)$ is defined as follows

$$\mathbf{g}(\mathbf{x}) = \frac{\|\mathbf{x}\|^\mathbf{2}}{\mathbf{1} + \|\mathbf{x}\|^\mathbf{2}} \frac{\mathbf{x}}{\|\mathbf{x}\|}. \tag{5.4}$$

where $\mathbf{x}$ is the input vector. Xi et al. Xi et al. (2017) introduced an alternative non-linear activation function $g(\cdot)$ as below

$$g(\mathbf{x}) = (1 - \frac{1}{e^{|x|}})\frac{\mathbf{x}}{|\mathbf{x}|}, \tag{5.5}$$

They hypothesised that such function is more sensitive to small changes in x.

In this research, each 1D Conv-Capsule consists of a group of neurons that collectively learn specific temporal entities presented by the BLSTM layer (shown in Fig 5.2). Contrary to the normal units which return a scalar, a capsule outputs a vector whose length is proportional with the likelihood of the entity presence, and direction represents the instantiation parameters.

Capsule Routing Network

After max-pooling the feature maps (outputs of the filters) in a CNN, an approximate translation invariant representation is achieved at the expense of losing orientational and relative spatial information about the parts or entities in an image (Sabour et al. (2017)). For classification tasks where the input should be mapped into a label, this information loss may not pose a serious issue. However, when some segmentation is required, the (approximately) translation invariance rendered by max-pool becomes problematic due to the information loss it brings about. Sabour et al. (2017) proposed a novel technique called *Routing by agreement* which yields a translation *equivariance* instead of the translation *invariance* in the CNNs. It deals with the problem mentioned above and better preserves the hierarchical relationships between lower and higher-level features.

In the routing layer, the previous layer capsules try to predict the output in the next layer. The capsules in the lower layer predict the output of the capsules n in the next layer. The input of the capsule $\mathbf{n}$, $\mathbf{s_n}$, is a weighted sum of such *predictions*

$$\mathbf{s_n} = \sum_{\mathbf{m}} \mathbf{c_{mn}} \hat{\mathbf{p}}_{\mathbf{n|m}}, \tag{5.6}$$

where $\hat{p}_{n|m}$ is the prediction of capsule m (in the lower level) about the output of capsule n and coefficients $c_{m|n}$ are weights. The weights are computed by a softmax function operating on $b_{m|n}$ coefficients which are learned through the *routing-by-agreement* algorithm (Sabour et al. (2017)):

$$c_{mn} = \frac{exp(b_{mn})}{\sum_k exp(b_{mk})}, \tag{5.7}$$

The *agreement* of the predicted output and actual output indicates the correctness of the prediction of the capsule m in the lower level about the capsule n's output in the higher level in the hierarchy, namely $\mathbf{a_{mn}}$. It determines how the lower and higher level features should be linked together, which is in contrast to the conventional networks where the higher level feature are merely a weighted sum of the lower level features. The prediction about capsule n is computed as a product of the transformation matrix W_{mn} and the output of the preceding layer p_m. Then, the agreement, a_{mn}, is computed using an inner product.

$$\mathbf{\hat{p}_{n|m}} = \mathbf{W_{mn}p_m}, \qquad \mathbf{a_{mn}} = \mathbf{p_n} \cdot \mathbf{\hat{p}_{n|m}}, \tag{5.8}$$

where $\mathbf{o_j}$ is the output of capsule $\mathbf{j}$.The inner product is added to b_{mn} (prior probablities initialised by zero) such as $b_{mn} = b_{mn} + a_{mn}$. The cost function is computed using the marginal loss as described in Sabour et al. (2017) and the transformation matrices, W_{mn}, are learned by backpropagation.

If capsule m (in the lower level) contains an instantiation of an entity represented by capsule n (in the higher level), the routing process makes the link between m and n capsules stronger and vice verse. Hence, the impact of the features from the m^{th} capsule on the n^{th} capsule is dynamically adjusted. Max-pooling is a *static* form of routing where only the most active unit in the pool is routed to the higher level, without considering the *dynamic* of the agreement between the low and high-level features in the hierarchy.

The transformation matrices, $\mathbf{W_{ij}}$, is learned by backpropagation.The cost function is computed using marginal loss as described in Sabour et al. (2017). The loss L_s for the class $s = 1, 2, \cdots, S$ equals

$$\begin{aligned} L_s = T_s\ max(0, m^+ - ||\mathbf{v}_s||)^2 \\ + \lambda(1 - T_s)\ max(0, ||\mathbf{v}_s\| - m^-)^2, \end{aligned} \tag{5.9}$$

where $T_s = 1$ if and only if an entity (emotion, here) of class s exists and $\|\mathbf{v}_s\|$ represents the length of vector $\mathbf{v}_s$. For the correct class s, loss is calculated when the length of the $\mathbf{v}_s$ is below m^+. Meanwhile, for the incorrect class s, loss is computed when the length of the $\mathbf{v}_s$ is beyond m^-. The λ is a control parameter and the total loss equals $L_{total} = \sum_d L_s$, which simply sums the losses across the final layer capsules.

Supervector Extraction Using Generative Models

In general, the generative models are more flexible than discriminative models in handling the variable-length patterns. If the whole utterance is represented through a fixed-length pattern or supervector, the classification via discriminative models such as NNs or SVMs would be more straightforward. Generative models can be useful in extracting a fixed-length representation and stacking the means of the maximum a posteriori (MAP) adapted Universal Background Model - Gaussian Mixture Model (UBMGMM) (Reynolds et al. (2000)) and iVector (Dehak et al. (2011)) are two examples of such methods used for supervector extraction.

For supervector extraction in building baseline systems, the method proposed in (Loweimi et al. (2015)) has been used. In this technique, first a GMM with M components is estimated for each class. Then, the posterior probabilities of all the components of GMMs are computed for each frame, averaged over all the utterance frames and finally stacked into a supervector.

The length of the supervector is $M \times C$ where C is the number of classes. For UBMGMM (Reynolds et al. (2000)) and iVector (Dehak et al. (2011)) the supervector length is $M \times D$ where D indicates the length of the raw feature vectors. For tasks where the number of classes is inherently small such as emotion recognition, C is notably less than D. As a result; this approach leads to a more compact representation which facilitates faster and more efficient learning. For more details about the advantages of this approach, please refer to Loweimi et al. (2015). eGeMAPS by Eyben et al. (2016) have also been used to extract supervector for comparison purposes.

5.2.2 Proposed DNN Architecture

The general architecture of the proposed framework is illustrated in Figure 5.1. In this architecture, BLSTM, Conv-Capsule and the capsule routing layer are playing complementary roles. BLSTM is used to deal with the sequential nature of the speech and its temporal dynamic. The Conv-Capsule model learns more abstract and richer representations of those temporal features. Finally, the capsule routing layer further distils the extracted patterns and maps them to a categorical distribution.

The Capsule net, in comparison with BLSTM, has a lower capability in handling and processing the forward/backward contextual information encoded in a sequential data like speech. On the other hand, BLSTM is not as powerful as the Capsule net in dealing with static patterns. For example, it is not translation invariant. The order, i.e. using 1D convolution capsules on top of the BLSTM is justifiable as follows: first, temporal features enriched by contextual information is extracted through BLSTM and then more abstract information distillation is carried out through the capsules and routing process. The process preserves all the temporal cell state sequences using 1D

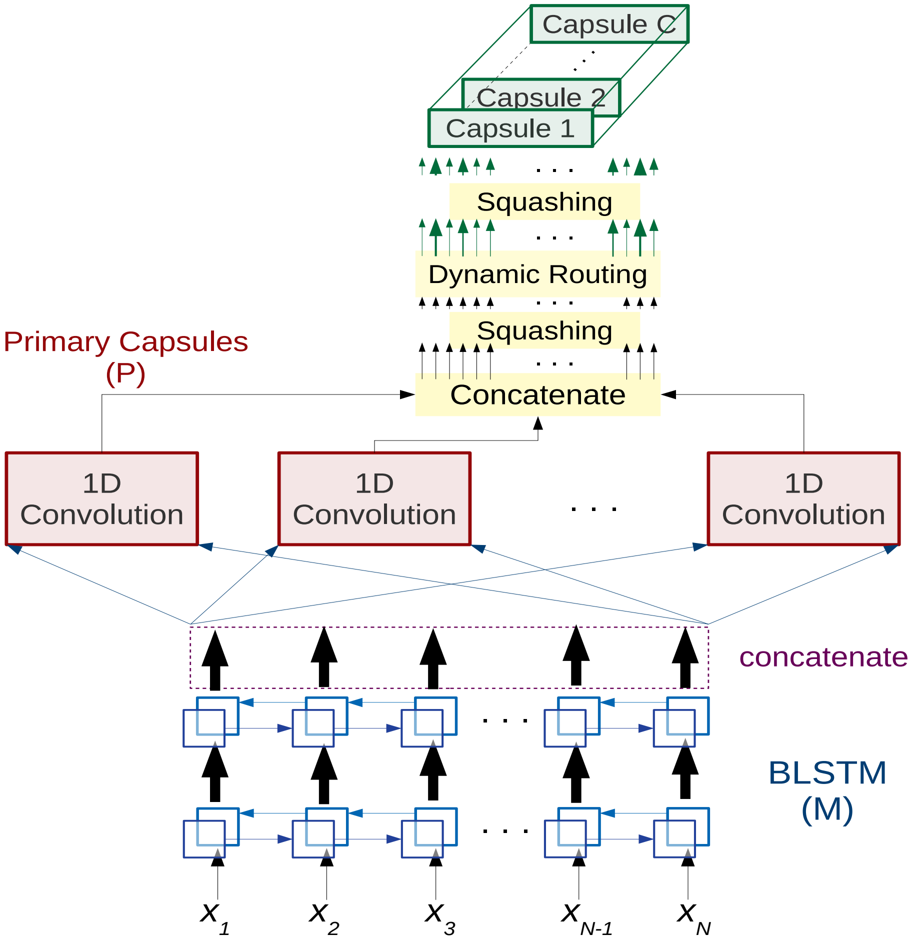


Figure 5.1: *The proposed architecture consisting bi-directional long-short term memory, 1D convolution and routing network.*

convolution. It was noticed that using 2D convolution distorts the temporal alignment. As such, each part is used in a task which it best fits, and the other component compensates for its shortcoming. This makes the structure *super-additive* and improves the overall performance as verified by the experiments.

Getting into more details, the overall network is comprised of two BLSTM layers, 1D Conv-Capsule layer consisting of capsules and a capsule routing layer. Input layer consisted of 70 nodes (length of the feature vector), and each BLSTM hidden layer (M in Figure 5.1) contained 256 units.

The next layer consists of four 1D-CNNs for two class categorisation task and

10 1D-CNNs for eight class categorisation task. Each of these CNNs has 90 filters and operates on the same receptive field of BLSTM temporal activation. The output feature maps are concatenated to form a capsule. The output of a capsule is squashed (equation 5.3, 5.4) for getting the vector representation of the feature learned by that capsule.

The routing capsule layer is connected to the previous layer through the transformation matrices which is similar to a fully-connected layer in the conventional NNs, except for replacing the scalar-to-scalar with a vector-to-vector transform. The number of capsules in this layer equals the number of classes, and each output layer capsule is connected to all the capsules in the previous layer. The previous layer capsules compute the prediction of the output of capsules in the next layer (eq. 5.8). The agreement coefficient is achieved by measuring the distance between the predicted output and the actual output (eq. 5.8). Finally, the output of these capsules is computed using eq. 5.6. The output layer capsules compute the posterior probabilities. The transformation matrices are learned by backpropagation.

5.2.3 Experimental Setup

Dataset

FAU-AiBo (Section 2.3.2) and RAVDESS (Section 2.3.3) have been used for training and testing the proposed framework.

Features

The eGeMAPS by Eyben et al. (2016) and supervector by Loweimi et al. (2015), was used in the baseline systems. The default parameters for feature extraction reported in their respective publications were applied. The log-spectrogram feature with 128 filterbanks (*FB*128) is used. The feature vector consists of F0, 23-dimensional MFCC and energy augmented by delta and delta-delta, denoted by *Feature**. To further enrich the input of the DNNs with contextual information, each frame's feature vector was appended with the feature vectors of the preceding/following 45 frames. This paves the way for better capturing the mid to long-term properties of the speech through processing a context of about 900 ms. Networks were trained by PyTorch (Paszke et al. (2017)) and optimisation was done by Adam (Kingma & Ba (2014)). length 23 was appended by delta and delta-delta coefficients as well as the pitch frequency and the log-energy resulting in a feature vector with 71 elements. and the performance evaluation of the networks is conducted in Python using PyTorch framework.

Setup

FAU-Aibo (Schuller et al. (2009), Steidl (2009), Batliner et al. (2008)) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) databases (Livingstone & Russo (2018)) have been used (Section 2.3.2 and 2.3.3). The RAVDESS is an audio-visual database, and only its speech part is utilised here, which covers eight acted emotional expressions: neutral, calm, happy, sad, angry, fearful, surprise and disgust, while FAU consists of 5 emotional classes: anger, emphatic, neutral, positive and rest (other categories). FAU consists of children's speech recordings who were communicating with Sony's pet robot Aibo, so the emotions are natural and spontaneous. FAU consists of two sets, namely Ohm and Mont, which cover 55% and 45% of the whole data, respectively, with totally disjoint speakers. More information about the corpora can be found at Section 2.3, 2.3.2 and 2.3.3.

For training the system, 75% of the data (randomly chosen) was employed, and the remaining 25% was used for testing. The RAVDESS speaker-independent scenario is performed by using 19 speakers for training and four different speakers for testing. For

Table 5.1: *UA(%) on FAU speaker independent scenario (Mont/Ohm as train/test) and vice versa*

Method	Train	Test	2-class UA(%)	8-class UA(%)
Supervector+SVM	Mont Ohm	Ohm Mont	62.8 56.5	29.8 36.3
Supervector+CNN	Mont Ohm	Ohm Mont	68.0 70.8	53.6 58.7
eGeMAPS+CNN	Mont Ohm	Ohm Mont	61.7 68.2	42.3 55.7
(*Feature**) + Capsule	Mont Ohm	Ohm Mont	70.5 71.3	53.3 59.0
(*Feature**)+BLSTM	Mont Ohm	Ohm Mont	71.7 72.2	53.4 58.7
(*Feature**) + Proposed Framework	**Mont** **Ohm**	**Ohm** **Mont**	**74.5** **75.3**	**55.3** **61.8**
(*Feature**)+ Capsule + BLSTM	Mont Ohm	Ohm Mont	70.4 71.2	53.8 58.6
(*Feature**)+ BLSTM + CNN	Mont Ohm	Ohm Mont	72.1 72.9	54.4 58.8

FAU corpus speaker disjoint training/testing, another approach was followed: since it consists of two subsets, i.e. Ohm and Mont, one was used for train and the other one to test. These two sets are disjoint in terms of speakers, which makes the test condition more challenging than 75/25% case and provides a better platform for evaluating the robustness of the system. The downside, however, is that a lower amount of data becomes available for training. In the second approach, namely choosing 75/25% for the train/test sets, experiments were run ten times, and the mean was calculated. The baseline systems trained with different hyperparameters, and the results have been reported for comparison purposes. No transfer learning mechanism used and the classifiers trained from scratch. The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

Table 5.2: *UA(%) on RAVDESS speaker independent scenario*

Method	2-class UA(%)	8-class UA(%)
***FB*128 + Proposed Framework**	**66.3**	**50.1**
***Feature** + Proposed Framework**	**70.4**	**56.2**
COVAREP + LSTM Beard et al. (2018)		41.2

Table 5.3: *Performance on RAVDESS (75%/25% for train/test)*

Method	2-class UA(%)	8-class UA(%)
Supervector+SVM	65.8	36.3
Supervector+CNN	65.9	34.6
eGeMAPS+CNN	71.4	33.0
(*Feature**) + Capsule	60.3	25.5
(*Feature**)+BLSTM	74.2	63.9
(*Feature**) + Proposed Framework	**79.5**	**69.4**
(*FB*128) + Proposed Framework	**73.9**	**68.1**
(*Feature**)+ Capsule + BLSTM	66.8	35.4
($F0 + MFCC^\star$)+ BLSTM + CNN	74.8	51.3

Table 5.4: *FAU (75%/25% for train/test) in binary emotion classification task.*

Method	**Proposed Framework**	DBN Latif et al. (2018)	Sparse AE+ SVM Latif et al. (2018)
UA(%)	**77.6** ±0.265%	74.11	71.73

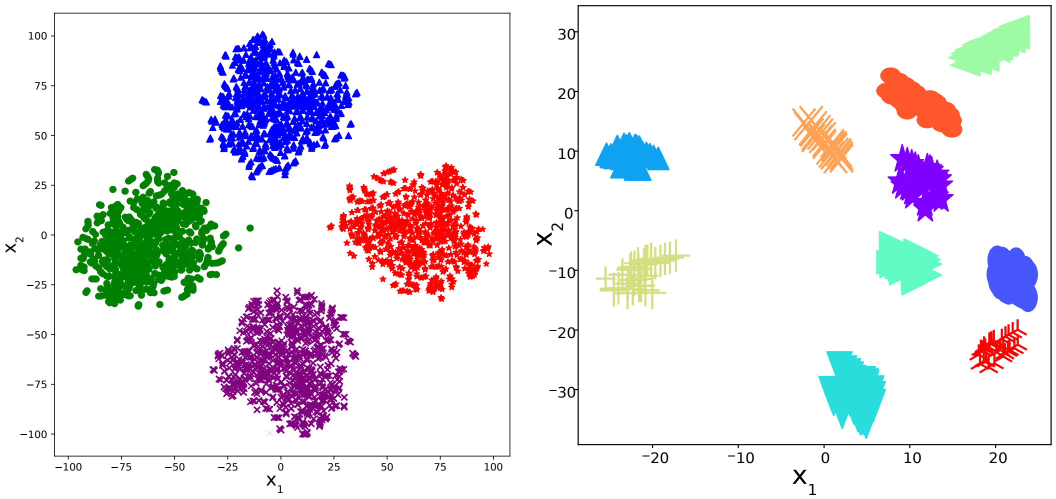


Figure 5.2: *Scatter plot of the four Conv-Capsule (left) and ten Conv-Capsule (right) outputs for RAVDESS after dimensionality reduction via t-SNE. Each color represents one cluster.*

5.2.4 Results

It is hypothesised at the beginning of this section that each 1D Conv-Capsule will learn different temporal properties for the same BLSTM temporal activation. The output of those capsules are extracted, and after dimensionality reduction, they are plotted in Fig 5.2. Each of these capsules learns totally different and unique temporal features. This is for both cases of four and ten capsules.

Tables 5.1, 5.2, 5.3, and 5.4 show the unweighted accuracy (UA) for binary (positive vs negative) as well as 5-class (FAU) and 8-class (RAVDESS) emotion classification tasks. The proposed approach in comparison with different systems and baselines leads to a notably better performance. To visualise how well the proposed network separates classes, dimensionality reduction using t-SNE (Maaten & Hinton (2008)) is performed on the output layer for RAVDESS test and train data. Fig. 5.2 illustrates the network successfully clusters the representations in the output layer.

The combination of the supervector (as input) with SVM and CNN is shown for FAU corpus in Table 5.1. SVM (with RBF basis) is outperformed by most of the DNN-

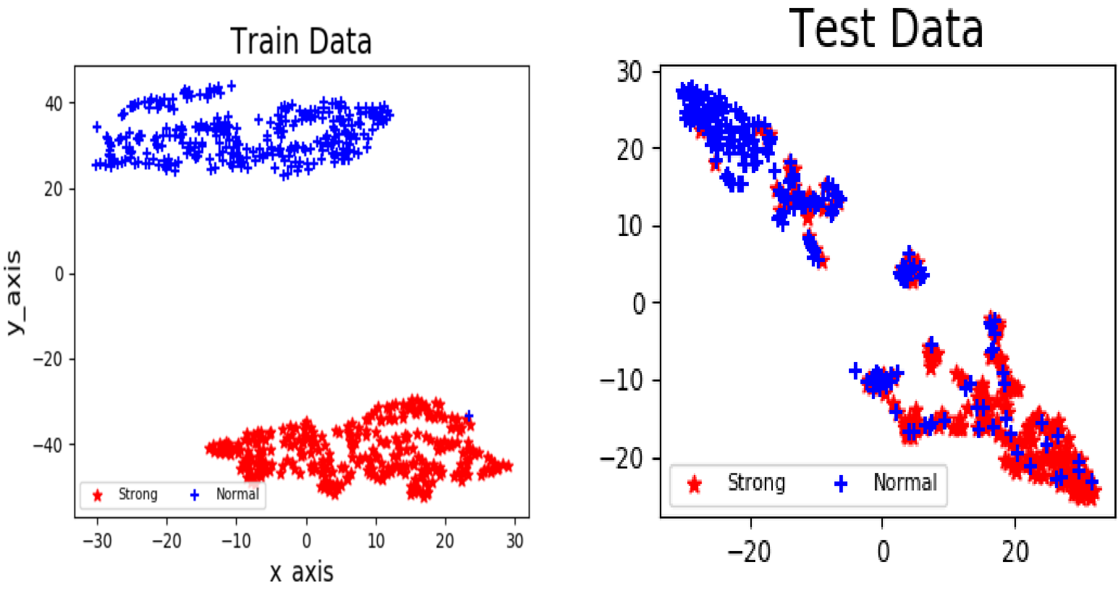


Figure 5.3: *Scatter plot of the proposed system's output for RAVDESS in 2-class task (strong/normal emotion) after dimensionality reduction via t-SNE.*

based back-ends. Comparing the Capsule net with CNN in combination with BLSTM shows the superiority of the Capsule network, which can be explained considering the advantages of the routing-by-agreement process over the max-pooling, as explained in Section 2.3. In Sabour et al. (2017), it is also claimed that the Capsule nets require less data than CNN, and this can be another reason for the superiority of the Capsule net in this task. It should be mentioned that the training time for the capsules on this proposed system was noticeably higher than CNN.

As seen in Table 5.1, the order of the BLSTM and the Capsule nets is also important and using the BLSTM on top of the Capsule net obviously degrades the performance and leads to a sub-additive combination. This can be explained based on the argument put forward in Section 5.2.1. Table 5.3 shows the results for RAVDESS database in which similar trends can be observed in terms of the ranking of the different systems. In all these experiments, $F0 + MFCC^{\star}$ features outperform filterbank features on average 3% accuracy.

At the time of publishing, the state-of-the-art accuracy for FAU (2-class) was 74.11% (Latif et al. (2018)). Latif et al. (2018) used deep belief network (DBN) and randomly selected 75% of the data for training purpose and 25% of the data for testing purpose. To do a fair comparison, similar to their approach, the data was divided into splits (75/25%) and run ten times, and the final result is the average of ten runs. As seen in Table 5.4, the proposed hybrid architecture leads to 77.6% ±0.265% accuracy (standard deviation ±0.265%) for FAU, which is 3.5% (absolute) higher than the state-of-the-art performance.

The state-of-the-art accuracy for 8-class RAVDESS audio speech emotion classification is 41.25% (Beard et al. (2018)), while the proposed system leads to remarkably higher accuracy of 56.2%. It has also been noticed that by increasing the number of frame embeddings for RAVDESS data, the accuracy for 2-class classification increases dramatically for BLSTM+Capsule and BLSTM+CNN model 76.9% and 77.1% respectively.

5.2.5 Discussion

In this section, a hybrid architecture consisting of a BLSTM, 1D Conv-Capsule and capsule routing layers was proposed for speech emotion recognition. The BLSTM is tasked with handling the temporal dynamics of the speech as sequential data and extracting contextually rich representations through forward/backward processing of the short-term features. The Capsule layers provide a state-of-the-art system for further distilling and processing the patterns extracted by the BLSTM. These techniques resulted in hierarchical temporal modelling that facilitates the better representation of clustering and categorisation. The proposed architecture was compared with a wide range of alternative networks, and a state-of-the-art performance was achieved. Applying this architecture to language and speaker recognition tasks are recommended for future research. Furthermore, the *routing-by-agreement* is a hierarchical attention mechanism in the capsules, which is decided by the agreement between capsules in two different hierarchies. This leads to the cluster forming in each layer of hierarchy. Each of the capsules generally represents a unique or unique group of attributes. It is not clear what kind of attributes they represent. The capsule layers are connected only to their neighbouring layers, i.e. there is no dense interconnection among different hierarchies. However, capsules are so computationally expensive, and it is not easy to use deeper and dense capsule networks. Nevertheless, the proposed model provides a powerful technique to model temporal dynamics for SER tasks.

5.3 Spatiotemporal context modelling for SER

Speech emotion recognition (SER) is a requisite for emotional intelligence that affects the understanding of speech. Learning emotional contextual feature representation independent of speaker and environment is essential. In this section, a novel spatiotemporal context modelling framework for robust SER is proposed to learn feature representation by using acoustic context expansion with high dimensional feature projection. The framework uses a deep convolutional neural network (CNN) and self-attention network. [2].

The rest of this section is organised as follows. The previous works related to this research are discussed in Section 5.3.1. In Section 5.3.2, the components of the proposed approach, i.e. CNN and self-attention network are reviewed, and the motivations behind using them are discussed. Section 5.3.3 explains the proposed architecture and its advantages over the current state-of-the-art. In Section 5.3.4, the experimental

[2] This section is published as M. A. Jalal, R. K. Moore and T. Hain, "Spatio-Temporal Context Modelling for Speech Emotion Classification," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 853-859, doi: 10.1109/ASRU46091.2019.9004037.

setup is presented. In Section 5.3.5, the results are presented along with discussion. Finally, Section 5.3.6 draws future research path and concludes this section.

5.3.1 Related Works

Emotion is a para-linguistic perception that could be presented either in a categorical or dimensional annotation scheme. However, one thing is clear; emotion is a long-term perception built upon context and familiarity over time (Wu et al. (2019), Jalal, Loweimi, Moore & Hain (2019)). Wu et al. (2019) proposed a model where they extracted the spatial features from speech using convolutional neural network (CNN) and used these features in a dual-stream capsule and bi-directional GRU network to get categorical distributions for the given features. Jalal, Loweimi, Moore & Hain (2019) proposed a temporal modelling approach by extracting the temporal dynamic features with BLSTM hidden layers and used capsule networks for doing hierarchical clustering of these features. Both Wu et al. (2019), Jalal, Loweimi, Moore & Hain (2019) try to learn the spatiotemporal context representation over time for SER tasks, and they were the best reported results for 4-class speech emotion recognition task on IEMOCAP (Busso et al. (2008)) and FAU AiBo (Schuller et al. (2009)). Beard et al. (2018) used LSTM and globally contextualised attention mechanism to learn features for SER tasks.

5.3.2 Approach

Different context modelling techniques were applied in previous research (Wu et al. (2019)). However, the results on SER tasks did not improve much (Table 5.5). This might indicate that rather than learning better feature representations, the models were getting better at the mapping between the source feature distribution and the target emotion annotation. In this section, the goal is to learn better feature representation by using acoustic context expansion with high dimensional feature projection. It has similarity with fMPE proposed by Povey et al. (2005). Spatial features y are extracted using convolutional neural network (CNN). The task-specific high dimensional feature expansion is done using a self-attention network, and it is projected (A) to the original feature dimension. The degree of projection is controlled using a parameter γ, which is also learned through backpropagation same as the rest of the network. After projection, the new feature $\hat{y}$ will be

$$\hat{y} = y + \gamma(A). \tag{5.10}$$

Convolutional Neural Network

Convolutional Neural Networks (CNN), like other neural networks, are multi-layer neural networks where the initial layers detect low-level features, and the final layers process more abstract high-level feature space. The special characteristic of these networks is that they share weights. So, each neuron of a layer is connected to a local area, also called *local receptive field or kernels*, of the previous layer. The *local receptive field* is replicated, and its weight matrix is multiplied all over the input space for detecting specific patterns at each layer. Section 2.2.2 describes CNNs in more detail.

In this study, a smaller kernel size (2-6) is used for preserving the locality by considering a small neighbourhood at a time. Smaller kernel sizes may increase non linearity in the network while enabling feature fusion (Lin et al. (2013), HasanPour et al. (2018)). Wolpert (1992) suggested stacking more than one CNN layer results in increased non-linearity and richer representations. The CNN layer is followed by applying subsampling with pooling layer for improving the discriminability power of the network, robustness to shift and distortions (LeCun et al. (1999)). So, pooling brings invariance to the network. Though, it is crucial to control the kernel size in pooling to keep it from losing information. The network learns faster with decorrelated network parameters that have zero means and unit variances (LeCun et al. (1999)). Ioffe & Szegedy (2015) proposed batch normalisation for approaching this issue with reducing internal covariance shift in a batch. These methods are adapted in this proposed *CSA* framework.

Self Attention Networks

Self-attention networks have the flexibility of modelling long-term inter-sequence dependencies. In this research, self-attention as non-local networks (Wang et al. (2018), Zhang et al. (2018)) is used to model the relationships between the regions in the feature maps. The mechanism is shown in Figure 5.4 block A_1. The features maps from previous CNN layers $\mathbf{y}$ are transformed into three feature spaces $\mathbf{j}$, $\mathbf{k}$ and $\boldsymbol{l}$, where

$$j\left(y\right) = W_j y \qquad\qquad k\left(y\right) = W_k y \qquad\qquad l\left(y\right) = W_l y. \tag{5.11}$$

Here $\boldsymbol{W_j}$, $\boldsymbol{W_k}$ and $\boldsymbol{W_l}$ are network weights learned through back-propagation. Each of these three projections perform downsampling of the input feature maps. The number of channels in $\boldsymbol{W_j}$, $\boldsymbol{W_k}$ is less than the number of channels in the input features. However, W_k has the same number of channels as input feature y. Dot product is used to calculate the positional relationship between the elements j and k. Therefore, the attention energy is calculated by using the $softmax$ function .

$$e_{ij} = softmax\left(j\left(y_i\right)^T k\left(y_j\right)\right). \tag{5.12}$$

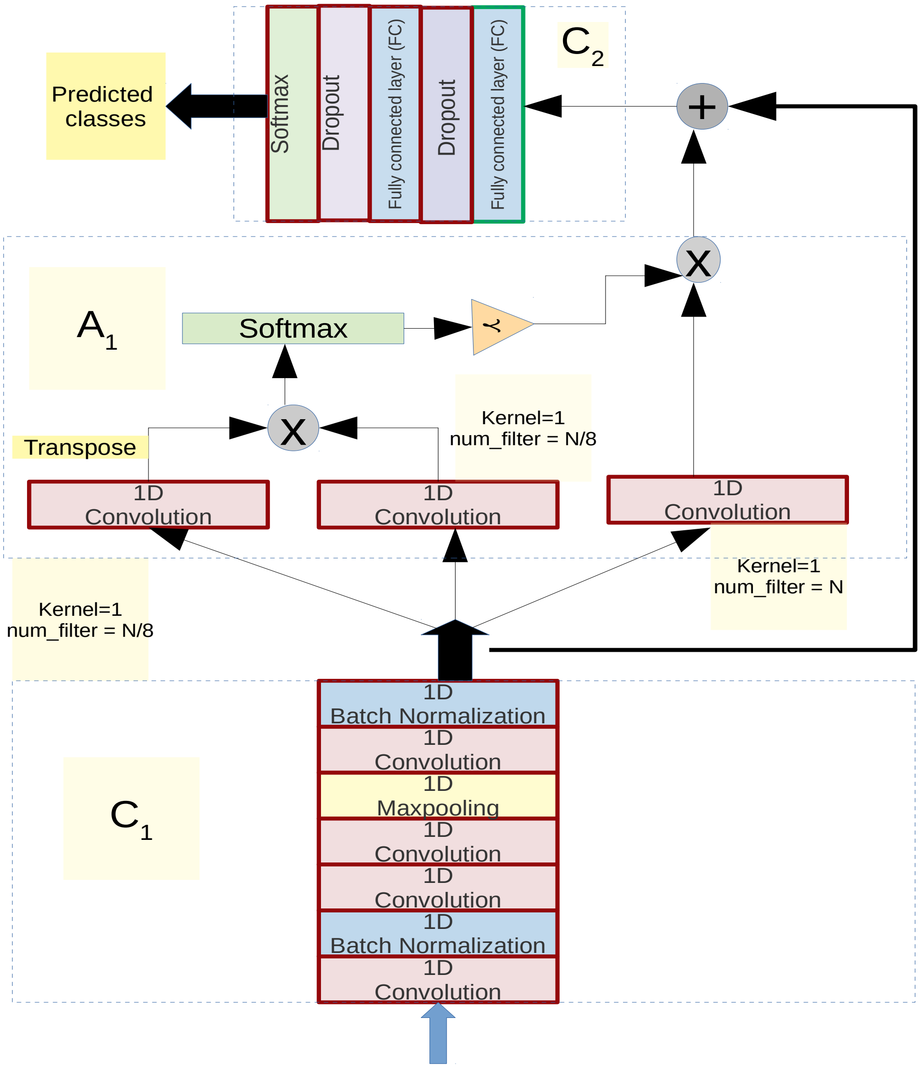


Figure 5.4: *Schematic diagram of the proposed CSA framework*

The attention map A is calculated by doing matrix multiplication between e and $l(y)$. The attention is projected into the same dimension as the original feature y.

$$A = e(l(y)) \tag{5.13}$$

The scaling factor γ is multiplied with the attention map to control the degree of attention. Attention A is then added with the input feature to increase the context information.

$$\hat{y} = \gamma(A) + y. \tag{5.14}$$

In this work, γ is randomly initialized.

5.3.3 Proposed DNN Architecture

The general architecture of the proposed convolutional self-attention framework (CSA) is illustrated in Fig 5.4. In this architecture, the C_1 and A_1 block is playing complementary roles. C_1 is a hierarchical structure combining several convolutional neural layers that learn deep feature representation, whereas A_1 learns task-specific high dimensional features from C_1's output feature vector and projects down to the dimension of C_1's output features (Equation 5.13). Thus, performing task specific context modelling.

The proposed model consists of a CNN block C_1 comprising CNN layer, batch normalisation layer and maxpooling layer. Rectified linear unit (ReLU) activation is applied for adding non-linearity. In C_1, the input features are fed to a convolution layer with kernel size 5 followed by a batch normalisation layer. The number of filters is 128. Two 1D convolution layers have been stacked with kernel size 3 and padding 1. Maxpooling with kernel size 2 is used to introduce sparsity in the network parameters (section 5.3.2). After each series of convolution process, batch normalisation is used extensively to reduce correlation among the parameters at the same layer within the network (section 5.3.2).

C_1 produces 128 channels of feature maps and they are fused using a self-attention layer A_1 (Section 5.3.2). The network weights W_j, W_k, W_l, mentioned in section 5.3.2, are three convolutional layers with kernel size 1. The number of filters for convolutional layers W_j and W_k is one-eighth of the input feature maps (128) in the self-attention layer. The number of filters is decreased to reduce the computation time. However, the filters for W_l remain the same as the input feature maps in the self-attention layer. The energy is calculated by applying softmax on the product of transformed feature matrices $\boldsymbol{W_j}$ and $\boldsymbol{W_k}$ (Equation 5.12). The attention map $el(\dot{y})$ is calculated with the energy(e)(Equation 5.13). The scaling factor γ is initialised as zero. The kernel size is 1 in the attention layer, which helps feature level fusion.

The contextually enhanced output features from the attention layer are given as input to the classifier block C_2. C_2 has three fully connected layers. ReLU activation function has been used to increase non-linearity in the feedforward neural network. *dropout* is used as a regulariser. Finally, softmax scoring is used for acquiring the predicted categorical distribution of the target class, which is used to calculate the cross-entropy loss. Network parameters are learned through backpropagation.

5.3.4 Experimental Setup

Baseline

Recurrent capsule net, Long Short-Term Memory (LSTM) network, bi-directional LSTM (BLSTM), and CNN based baselines are used for comparison. Since the same train/test set has used mentioned in the baselines and in Section 2.3, some of the results are directly compared with the results from those baselines Wu et al. (2019), Li et al. (2018*a*). The networks are trained with PyTorch framework Paszke et al. (2017) and optimisation is done using Adam Kingma & Ba (2014) with 0.0001 learning rate.

Table 5.5: *Performance UA(%) on IEMOCAP Busso et al. (2008) speaker independent scenario*

Method	4-class (UA%)
CNN_ LSTM Satt et al. (2017)	59.4
CNN_ GRU Wu et al. (2019)	51.84
CNN_ SeqCap Wu et al. (2019)	56.71
CNN_ RecCap Wu et al. (2019)	58.17
CNN_ GRU-SeqCap Wu et al. (2019)	59.71
Attention Pool Li et al. (2018*a*)	71.8
Proposed *CSA* Approach	**76.3**

Features

23-dimensional log mel filterbanks with energy features are used as they provide superior results shown in Milner et al. (2019), Jalal, Loweimi, Moore & Hain (2019) as well as Section 5.2.4.

To further enrich the input of the DNNs with contextual information, each frame's feature vector is appended with the feature vectors of the preceding/following frames. It paves the way for better capturing the mid to long-term properties of the speech through processing acoustic context. This framework gives the flexibility of understanding the role of context length and acoustic cues. Further experiments with the context length using the *CSA* model can be found in next sections (Section 7 and Table 7.1).

Dataset

IEMOCAP (IEM) corpus (Busso et al. 2008) is used for validating the proposed framework in this section. The train/test splits are discussed in Section 2.3.4.

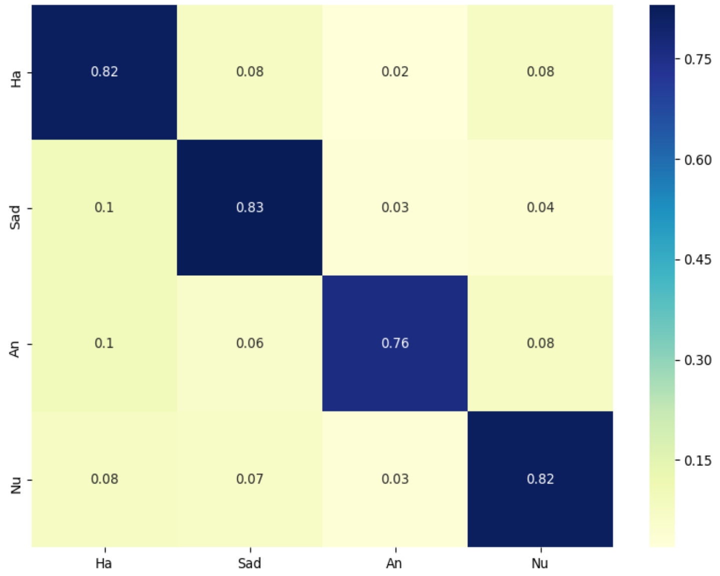


Figure 5.5: *Confusion matrix for IEMOCAP Busso et al. (2008) four class classification (Ha:Happy, Sad:Sad, An:Angry, Nu:Neutral emotion)*

5.3.5 Results

The results for IEMOCAP are compared with some of the best reported results in the same task. It was hypothesised at the beginning of the section that the proposed model would be able to generalise well due to learning better feature representations. The UA (%) results are reported in Table 5.5. Table 5.5 report the experimental results where the training and test set do not have any overlapping speakers. The proposed spatiotemporal modelling approach is compared to different DNN models with a different combination of architectures and features. Three observations can be made from the table. The confusion matrix for the best result experiment is shown in Figure 5.5.

Firstly, the proposed framework performed significantly better than the other DNN systems reported in the table. The proposed model has a performance improvement over 4.5% for IEMOCAP 4-class task.

Secondly, as it was hypothesised, the model learned speaker-independent emotional context representation, which is evident in the speaker-independent scenario in Table 5.5.

In section 5.3.3 the C_1 block learns deep feature representations and pass it to A_1. Then A_1 transforms it to higher dimensional feature space and project it to the same dimension of the original feature vector (feature vector from C_1) by adding $el(\dot{y})$ (Equation 5.11, 5.12, 5.13, 5.14). This approach has similarity with fMPE where Povey et al. (2005) uses gaussian posteriors for transforming into high dimensional feature space. However, in the proposed framework, the attention network can learn task-specific features which are eventually added with the other features for context expansion.

5.3.6 Discussion

In this section, a spatiotemporal context modelling technique is proposed for a speech emotion classification task that learns rich feature representation and improves the classification. Overall, the performance of the proposed network is significantly better than the previous approaches. Speech emotion context modelling is explored further in Chapter 6.

5.4 Crosscorpus SER

The speech emotion corpora that have been discussed in this chapter and Section 2.3, are from three categories such as: natural (FAU-AiBo, MOSEI), elicited (IEMOCAP) and acted (RAVDESS, eNTERFACE). Each of the corpora has participants of different gender, ethnicity and age. Although there was some research that investigated the crosscorpus impact, those research focused on the cross-lingual aspect. However, it is very crucial to know if the acted, elicited, and natural corpora within the same language can benefit each other. In other words, if there is a common latent representation and understanding about emotion across different corpora in a language. The crosscorpus study Milner et al. (2019) presents whether the emotion representation learning from one corpus can be used for speech emotion recognition in other corpora [3].

5.4.1 DNN Model

A BLSTM-Attention model has been used for this research. The attention mechanism has the flexibility of computing long-term inter-sequence dependencies. In this work, the attention value on the overall temporal feature space from the BLSTM is computed. The attention mechanism focuses the network onto specific parts of itself by computing the global mean, which captures global information. The global mean is multiplied over the whole temporal vector to compute the positional dependency of each element with $tanh$ non-linearity. The resulting vector is used to compute the attention weights using $softmax$ scoring. The attention mechanism is similar to Equation 6.5. The soft attention mechanism is adopted for this work, and the multiplicative method is applied as in Beard et al. (2018), similar to Nam et al. (2017), as the authors found their results similar to the standard concatenated approach of Bahdanau et al. (2014). The main challenge for crosscorpus SER is to minimise the long-term spatiotemporal variation between different corpora for a given categorical distribution, and one solution is to model global spatiotemporal dependencies between corpora. The proposed framework with attention modelling is an efficient way to approach the problem.

[3] The crosscorpus SER work is published as R. Milner, M. A. Jalal, R. W. M. Ng and T. Hain, “A Cross-Corpus Study on Speech Emotion Recognition,” 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 304-311. The thesis author contributed to the implementation of the attention model experiment used in crosscorpus experiment, and the presentation of the work.

The BLSTM contains two hidden layers of 512 nodes each. The output layer of size 1024 feeds into the attention mechanism computing a context vector of size 128, which is projected to 1024 nodes. This is then passed to the predictor stage that linearly projects to the number of classes. The cross-entropy loss function is applied, which is preceded by a $softmax$ layer in the PyTorch implementation.

5.4.2 Crosscorpus training

Baseline

Baseline results from the literature are difficult to find for this work since this work requires the 'big-six' emotions, and many datasets contain more or less. Hence, evaluation results are not equivalent. Also, not all research states both unweighted accuracy and weighted accuracy. For ENT, a convolutional recurrent neural network with an attention mechanism proposed by Huang & Narayanan (2017) achieves a UA of 91.7%. For 6-class classification, RAV and IEM are not generally evaluated this way, IEM results focus on the four classes (IEM4) mentioned in Section 2.3.4. The best UA (other than the research work in this chapter) found for the IEM4 test set is 71.8% Li et al. (2018*b*) presenting an attention pooling based representation learning method. A crosscorpus UA of 64.0% is shown in Luo et al. (2018) where a two-channel system approach is adopted combining high-level statistic features with a convolutional recurrent neural network. For WA, 56.1% is shown in Desplanques & Demuynck (2018) which also presents a crosscorpus WA of 48.4% applying factor analysis to find emotion factors to classify. For MOS, a WA of 52.5% is presented in Beard et al. (2018).

Dataset

Acted corpus eNTERFACE (Section 2.3.5 denoted as ENT), acted corpus RAVDESS (Section 2.3.3 denoted as RAV), elicited emotion corpus IEMOCAP (Section 2.3.4 denoted as IEM) and natural corpus MOSEI (Section 2.3.6 denoted as MOS). They are trained on 'big-six' emotions Ekman (1992), which are *happy*, *sad*, *anger*, *surprise*, *disgust* and *fear*. IEM has six classes and IEM4 has four emotion classes (Section 2.3.4).

Features

Log mel filterbank features with 23 dimensions are used in the experiments (Milner et al. 2019).

Experimental Setup

The model is trained on one corpus and tested on the remaining corpora. The standard train/test splits are used throughout the experiments. The Adam optimiser by Kingma & Ba (2014) is used with the initial learning rate of 0.0001. PyTorch approach of ReduceLROnPlateau setting is found to be four epochs with a multiplicative factor of 0.8. The models were trained to 200 epochs, and the best model chosen by averaging the results across the datasets as adding a stopping criterion could influence the effect on the different datasets. The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

5.4.3 Results and Discussion

The results from the crosscorpus experiments are shown in Table 5.6 and Table 5.7. Table 5.6 shows the accuracy results over each corpus. For example, the model is trained

Table 5.6: *Crosscorpus (CC) results across the five testsets, where each row refers to a single model. The asterisks refer to the results which can be compared to baseline systems and bold refers to the matched condition.*

Training	Unweighted Accuracy, UA%					Weighted Accuracy, WA%				
Data	ENT	RAV	IEM	IEM4	MOS	ENT	RAV	IEM	IEM4	MOS
ENT	**95.6***	74.7	77.9	66.2	54.8	**92.0**	54.5	56.2	58.1	48.5
RAV	74.7	**86.1**	82.2	71.8*	56.6	54.4	**75.0**	56.3	60.1*	49.4
IEM	72.9	75.6	**90.4**	**72.0***	66.3	51.2	56.2	**64.1**	**67.7***	49.4
MOS	78.4	73.3	82.0	70.2	**74.5**	61.2	52.0	54.4	58.5	**52.8***

Table 5.7: *Performances across the emotions for the crosscorpus (CC) models, where each row refers to a single model. Results in bold refer to the best performances for each emotion across the CC models.*

Training	Unweighted Accuracy, UA%							Weighted Accuracy, WA%						
Data	hap.	sad	ang.	sur.	dis.	fear	neu.	hap.	sad	ang.	sur.	dis.	fear	neu.
ENT	**74.1**	67.4	70.8	74.9	78.3	86.9	61.3	58.4	65.8	**69.0**	**61.4**	**59.3**	**57.9**	**59.1**
RAV	61.5	77.3	71.3	87.2	71.1	83.0	**69.1**	58.7	65.8	61.3	57.4	55.2	55.4	50.0
IEM	66.4	**77.4**	56.8	87.8	**85.6**	**88.4**	**69.1**	53.3	**68.1**	67.8	50.0	50.0	50.0	50.0
MOS	60.7	67.3	**75.8**	**88.1**	84.7	**88.4**	66.3	**58.9**	66.6	53.6	49.9	51.6	50.0	55.4

with ENT and tested with ENT, RAV, IEM, IEM4, MOS. Here ENT is the matched corpus, and the rest of them are mismatched. The results in the bold diagonal refer to the matched condition (trained and tested on the same dataset), and the non-bold performances are the mismatched condition (trained and tested on different datasets). Asterisks are displayed to compare results from the baseline results mentioned in Section 5.4.2. The crosscorpus and matched results achieve better results than the baseline results, which shows the SER model and is set up well for recognising emotions from speech and adequate to use for this study. However, it is expected that the best results are seen in the matched condition. However, the model trained on IEM produces the best-mismatched results for MOS and RAV test sets but not for ENT.

It can be inferred that IEM, being an elicited dataset, is close to both the acted and more natural datasets. While comparing the models trained on the acted datasets, it can be seen that the RAV model performs better for IEM, IEM4 and MOS than the ENT model. This shows that despite both being acted datasets, there seems to be more useful information to learn from a model trained on RAV than on ENT. The number of samples and recording hours are higher in RAV than ENT, which reinforces the challenges of datasets with different annotations and sizes.

Performance reduces when moving to the natural dataset MOS, showing the difficulty of classifying more naturally produced emotions. It is also possible that the emotions in MOS are not correctly annotated. The perception of annotators is very crucial here. Because the annotators can differ significantly within the same emotion segment (as seen in the IEMOCAP annotations). Among the classwise performance in Table 5.7, the bold numbers refer to the best performance for the emotion category. For the UA results, the models trained on IEM and MOS give better performances for all the emotions except for happy. This is a common problem in IEMOCAP; the models train on IEMOCAP with six emotions often confuses between happy and surprise. However, in the WA results, the model shows the opposite, that training on ENT is better across five of the seven emotions. It suggests that acted datasets are good at separating emotions, even across datasets with different emotion production types.

Crosscorpus experiments show the matched results outperform mismatched and that the model trained on the elicited dataset achieves best-mismatched performance in most cases. Overall, while looking across the emotion class results, ENT data shows the best WA results for most of the emotions. It implies that acted datasets may have their advantages with more natural datasets. These experiments also indicate a common contextual representation for the emotions among different styles of emotion corpora. Until now, the publicly available corpora for speech emotion recognition are small in terms of the number of recording hours, and there is an imbalance in the data categories. Therefore, it may be useful to use different (acted, elicited, natural) corpora together for training to learn common contextual representation for emotion, which is beneficial from a machine learning standpoint.

5.5 Summary

The chapter elaborates on the research questions presented in Section 5. All the experiments are conducted with a single modality, and no multimodal data is used. Therefore, the results are not comparable with multimodal systems, though most of the results produced in this chapter are significantly higher than the multimodal systems on the same datasets. Previously, in Chapter 2, Section 2.3 describes the corpora and the data splits used in this chapter. Section 5.1 presents a brief background about the computational models of emotions and their relevance to this work. Section 5.2 and 5.3 present two different types of hybrid topologies for context learning. Section 5.2 proposes a *rnn-cnn-capsule* based hybrid topology framework to learn temporal dynamics and understand capsule representation of internal states. The representation based internal temporal clusters is shown and discussed better to understand the capsule mechanism as well as context representation. Section 5.3 presents a different deep neural network model with self-attention mechanism. It learns from fixed-length segments. These models help to understand the implication of acoustic cue and context for time. Section 5.3 proposes a convolutional attention based spatiotemporal modelling technique *'CSA'* that is significantly smaller and faster than all the *state-of-the-art* models proposed with similar comparable performance. Furthermore, there is a common ground cross-culturally and cross-linguistically about the universal understanding of emotion. So, different corpora with different annotations and/or different accents may benefit from each other. Section 5.4 ponders upon the impact of crosscorpus training and shows the implication of natural, elicited and acted emotions in emotion recognition.

Chapter 6

Hierarchical Attention for Speech Emotion

Emotion in speech is a fundamental trait of human communication that reflects the meaning and intent and the distinction about 'what is said' and 'how it is said', which is not precisely defined for SER tasks. Typically, emotion is represented in either a categorical or a dimensional annotation scheme. Although the duration and the position of an emotion is not well defined in a sentence, it is clear that emotion is built upon on either short-term or long-term context (Wu et al. (2019), Jalal, Loweimi, Moore & Hain (2019), Beard et al. (2018)). As mentioned earlier in Section 2.2.4, learning sub-symbolic representation is often hierarchical and dynamic, which depend upon communication between different hierarchies. The subsymbolic representations are abstract and hard to interpret. In this chapter, a novel interpretable hierarchical mixture of multi-view attention is proposed for speech emotion classification (SER).

6.1 Removing Bias with Residual Mixture of Multi-View Attention for SER

The proposed model for speech emotion recognition performs a deep level feature transformation. It learns different task-specific feature representations from utterances and performs feature transformation in a high dimensional space. This is followed by projection to the original feature vector[1].

The feature projection aims to remove task-specific bias in the feature space. The experimental results show the effectiveness of the proposed computational model, leading to state-of-the-art results on the IEMOCAP (Busso et al. (2008)) corpus in a 4-class classification task.

The rest of this section is organised as follows. The previous work related to this section is discussed in Section 6.1.1. In Section 6.1.2, the components of this framework, i.e. long short term neural networks (LSTMs) and the proposed multi-projection self-attention network, mixture of multi-view attention (MOMA), are described. Section 6.1.3 also presents and explains the proposed architecture in terms of representation learning and the motivations behind it. In Section 6.1.4, the experimental setup is explained, and the results are presented along with a discussion in Section 6.1.5. Finally, Section 6.1.6 concludes the section and Section 6.2 summarises the chapter.

6.1.1 Related Work

In this work, acoustic context expansion has been carried out with high dimensional multi-instance feature projection. It has similarity with the context expansion technique in feature-space minimum phone error (fMPE) by Povey et al. (2005). Sequen-

[1] This section is accepted for publication in Interspeech 2020 as Jalal, M. A., Milner, R., Hain, T, & Moore, R. K. (2020, May). “Removing Bias with Residual Mixture of Multi-View Attention for Speech Emotion Recognition”. In Interspeech 2020.

tial and hybrid-hierarchical models were proposed to learn deep feature representations (Wu et al. (2019), Beard et al. (2018)), and task-specific feature clusters (Jalal, Loweimi, Moore & Hain (2019)). Variants of attention-based mechanisms have been proposed which performed significantly better than the previous models (Lian et al. (2019), Tarantino et al. (2019), Jalal, Moore & Hain (2019)). One of the possible reasons why attention models outperform others is that the models learn the biases for a specific task, or group of tasks, leading to improved generalisation. Recently, a sequence and attention-based domain adversarial system was presented in Milner et al. (2019) which investigated whether the information in acted datasets can be learnt to benefit emotion prediction for natural datasets and achieved state-of-the-art results.

6.1.2 Representation Learning

The features over time are extracted using a bi-directional long short-term memory network (BLSTM). Then, multiple instances of attention vectors are computed which are projected on to a representation space derived from the same features. The final 'smoothed' projection is applied to attain bias in the original feature space expansion.

BLSTM Encoder

Long short-term memory (LSTM) networks use the left to the right temporal order of the sequence, whereas studies show that future or forward contexts are useful for context-sensitive sequence modelling (Graves & Schmidhuber (2005), Schuster & Paliwal (1997)). BLSTMs model the input sequence forward and backwards in two separate recurrent neural networks (RNNs) as a way to exploit the contextual information from the past and the future (Graves & Schmidhuber (2005)). Applying these networks, a temporal feature distribution over the sequence can be obtained in the encoder layer which is stacked. This can be expressed by

$$y^{fwd}[t,h] = [LSTM\left(y^{t}[h], y^{fwd}[t,h-1]\right)] \tag{6.1}$$

$$y^{bck}[t,h] = [LSTM\left(y^{t}[h], y^{bck}[t,h+1]\right)] \tag{6.2}$$

$$y[t,h] = [y^{fwd}[t,h], y^{bck}[t,h]], \tag{6.3}$$

where t is the timesteps, h is hidden dimensions, The output y is stacked over time to form a matrix $\mathbb{Y} \in \mathbb{R}^{(T \times h)}$. More details can be found at Section 2.2.3.

Mixture of Multi-View Attention (MOMA)

Self-attention networks can flexibly learn representations for long-term inter-sequence dependencies (Vaswani et al. (2017)). In this work, the basic attention block is similar to Beard et al. (2018), Bahdanau et al. (2014). First, a global contextualised attention mean M is calculated by computing the global mean across time. The mean is then repeated as the same temporal domain length as $\mathbf{Y}$ to form a matrix which has same size as $\mathbf{Y}$. Both $\mathbf{Y}$ and $\mathbf{M}$ are projected on to fully-connected layers, namely $\mathbf{W}_h$ and $\mathbf{W}_m$. These fully-connected layers are multiplied to find non-local positional dependencies and the result is projected to another fully-connected layer, $\mathbf{W}_e$, to produce the attention vector over time frames.

$$E = tanh\left(\boldsymbol{W}_h \boldsymbol{Y}\right) * tanh\left(\boldsymbol{W}_m \boldsymbol{M}\right) \tag{6.4}$$

$$a_{att1} = Softmax\left(\boldsymbol{W}_e * \boldsymbol{E}\right), \tag{6.5}$$

where E is positional dependency or self-attention between $\mathbf{W}_h$ and $\mathbf{W}_m$, and a_{att1} is the attention. This attention is projected onto $\mathbf{Y}$ as $\boldsymbol{Y'}$ and added as a skip connection with $\mathbf{Y}$. The skip connection reduces the degradation problem and helps the network attain iterative non-local feature learning (Wang et al. (2017), Law et al. (2017)). The schematic diagram is shown in Figure 6.1.

$$\boldsymbol{Y} = \boldsymbol{Y'} + \boldsymbol{Y}. \tag{6.6}$$

Next, multiple attention blocks can be applied, and each of these blocks have different initialisation. These are projected to a common space through a control parameter, which acts like an attention mixture model and is referred to as *MOMA*. All the spaces are derived from the same source $\boldsymbol{Y}$. However, they learn different representations.

$$E_n = tanh\,(\boldsymbol{W}_{h_n}\boldsymbol{Y}) * tanh\,(\boldsymbol{W}_{m_n}\boldsymbol{M})\forall n = 1, 2, 3 \tag{6.7}$$

$$a_n = Softmax\,(\boldsymbol{W}_{e_n} * \boldsymbol{E_n})\forall n = 1, 2, 3. \tag{6.8}$$

where n is the number of attention units and a_n is attention at the nth attention block. Each of $\boldsymbol{W}_{h_n}$ and $\boldsymbol{W}_{m_n}$ are initialized differently but they share a common representation space, $\boldsymbol{Y}$. This means different instances of E_1, E_2, E_3 are obtained from a common representation space. The projection is controlled using γ_1, γ_2, γ_3 as seen in Equation 6.9. Here $\boldsymbol{W}_{h_n}$, $\boldsymbol{W}_{m_n}$ and $\boldsymbol{W}_{e_n}$ are fully-connected layers and the network weights are trained through back-propagation.

$$a_{att2} = \sum_{n=1}^{3} (\gamma_n \cdot a_n)\,, \tag{6.9}$$

where a_{att2} is the attention output from the *MOMA* attention blocks and n is the number of attention blocks in *MOMA* layer. Each of these attention vectors are time aligned with the input segment in the network.

Here it has been hypothesised that by projecting the mixture of attention scores into the common feature space, the model is learning loosely correlated task-specific attention representations and by adding them the model performs smoothing to improve robustness. To investigate this hypothesis, attention vectors are extracted and analysed with the input segments to investigate attention in the intermediate hierarchies (Figure 6.1). In this work, the γ_1, γ_2, γ_3 are initialised randomly. This layer obtains the non-local dependencies.

6.1.3 Proposed DNN Architecture

The overall architecture of the proposed framework is shown in Figure 6.1. The model is a hierarchical attention structure with LSTMs. The LSTM processes long term temporal sequential dependencies and produces an abstract sequential feature representation. The attention layers attain positional dependencies to capture dynamic acoustic cues.

The *BLSTM Encoder* contains two hidden layers of 512 nodes each. It outputs a stacked matrix of size *[number of frames]* × 1024. This output of size 1024 is fed into the first attention layer *Attention Layer 1*. The attention mechanism is computing a context vector of size 128. The attention projection is of size *[number of frames]*×1024).

The output from the encoder and the attention projection are added as residual skip connections and passed to the *MOMA* layer with three attention blocks i.e. *Attention_1_Layer2, Attention_2_Layer2, Attention_3_Layer2*. Each block in *Attention Layer 2* process it individually and projects it with a control parameter γ. Finally, these attention heads are added, and the result is projected to 1024 nodes. The *Attention Layer 2* obtains task-specific high dimensional features from *Attention Layer 1*'s output feature space and performs smoothing on task-specific multi-view attention. The components are explained in Section 6.1.2. The $\boldsymbol{W}_y$'s, $\boldsymbol{W}_m$'s and $\boldsymbol{W}_e$'s are fully-connected neural layers, and along with the γ's they are trained through backpropagation. This is then passed to the emotion classifier, which linearly projects to

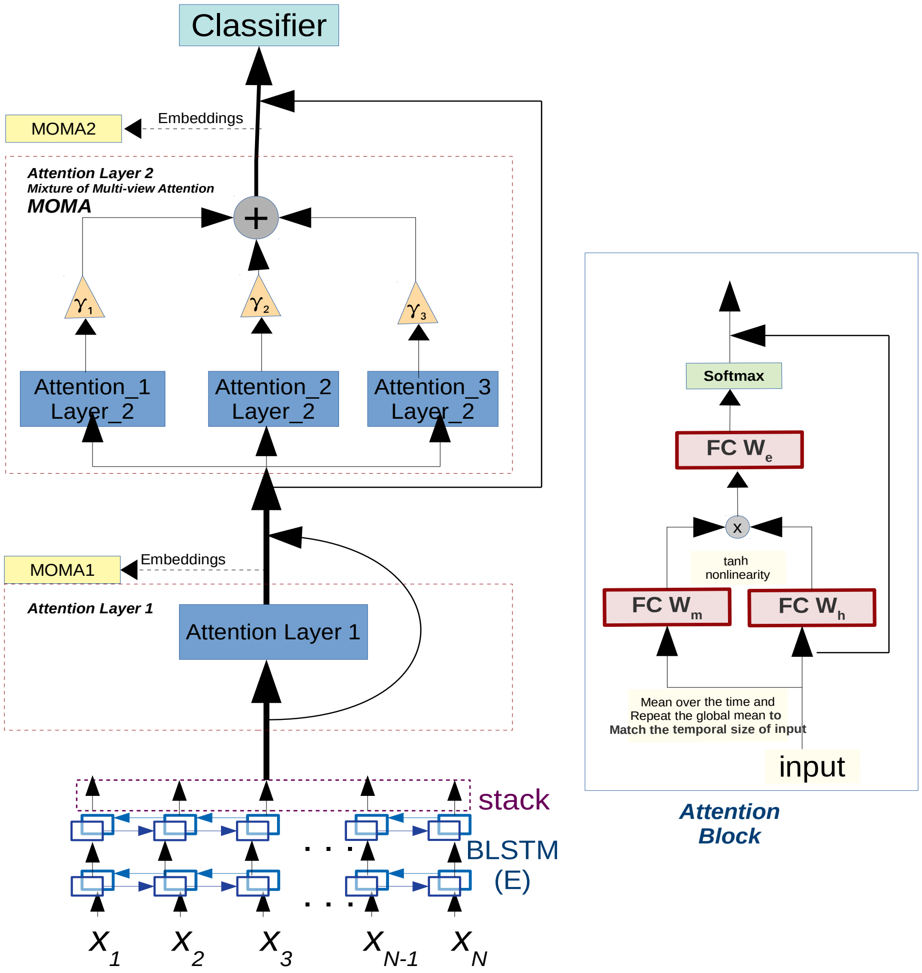


Figure 6.1: *Schematic diagram of the MOMA architecture.*

the number of classes. The cross-entropy loss function is applied, which is preceded by a $softmax$ layer.

6.1.4 Experimental Setup

Dataset

The IEMOCAP corpus (IEM4) Busso et al. (2008) is used for validating the proposed framework. In the literature it is common for IEMOCAP to be evaluated with four classes (IEM4): *happy*, *sad*, *anger* and *neutral* (where *happy* is combined with *excitement*) Li et al. (2018*b*). This is referred to as IEM4 in this section and in Milner et al. (2019). The details can be found in Section 2.3.4.

Features

Experiments in Jalal, Loweimi, Moore & Hain (2019), Milner et al. (2019) showed that the sequence model-based systems performed best with 23-dimensional log-Mel filterbank features which is therefore applied to the *MOMA* system as well.

Implementation

The Adam optimiser by Kingma & Ba (2014) is applied to the proposed model with the initial learning rate of 0.0001. As Adam adaptively optimises the learning rate, the PyTorch approach of ReduceLROnPlateau has been investigated. The optimum patience setting was found to be four epochs with a multiplicative factor of 0.1. Transfer learning mechanisms are not used. The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

Evaluation

Unweighted accuracy (UA) and the weighted accuracy (WA) described in Section 2.4 are used to evaluate the results.

Baseline

The results are compared directly with speech emotion recognition systems that use the IEM4 dataset. Four of these baselines process audio data only. For comparing UA, the results from a CNN-LSTM (Satt et al. (2017)) model, a deep capsule network with GRU (Wu et al. (2019)), and a deep attention pooling (Li et al. (2018*a*)) based

Table 6.1: *Performance of the MOMA model compared to baselines evaluated on the IEM4 dataset in terms of UA(%) and WA(%).*

Method	UA%	WA%
Factor analysis Desplanques & Demuynck (2018)	-	56.1
CNN_ LSTM Satt et al. (2017)	59.4	-
CNN_ RecCap Wu et al. (2019)	58.2	-
CNN_ GRU-SeqCap Wu et al. (2019)	59.7	-
Attention Pool Li et al. (2018*a*)	71.8	-
Convolutional self-attention Jalal, Moore & Hain (2019)	76.3	68.8
MULTIMODAL: Attention Lian et al. (2019)	*78.0*	-
MOMA	**80.5**	**74.8**

model are presented. The result from Desplanques & Demuynck (2018) has been cited to show WA baseline. A multimodal system Lian et al. (2019) carrying out SER on textual as well as audio data is also included to show how much the MOMA model could reach.

6.1.5 Result and Discussion

The baseline systems and performance of the proposed model are shown in Table 6.1. It is clear that the proposed *MOMA* model outperforms the baseline systems, including the multimodal approach, which uses lexical and audio data, as opposed to the *MOMA* only using audio data. The proposed system has achieved 80.5% UA and 74.8% WA on segment-level training.

It can be said that the model learns speaker-independent emotional context information. In Equations 7, 8 and 9, the different projections on the same derived feature space $\mathbf{Y}$ learn different variations of the same feature space and the γ's make it more flexible. As a result, the network becomes more robust to speaker and distortion variations (see Section 6.1.5).

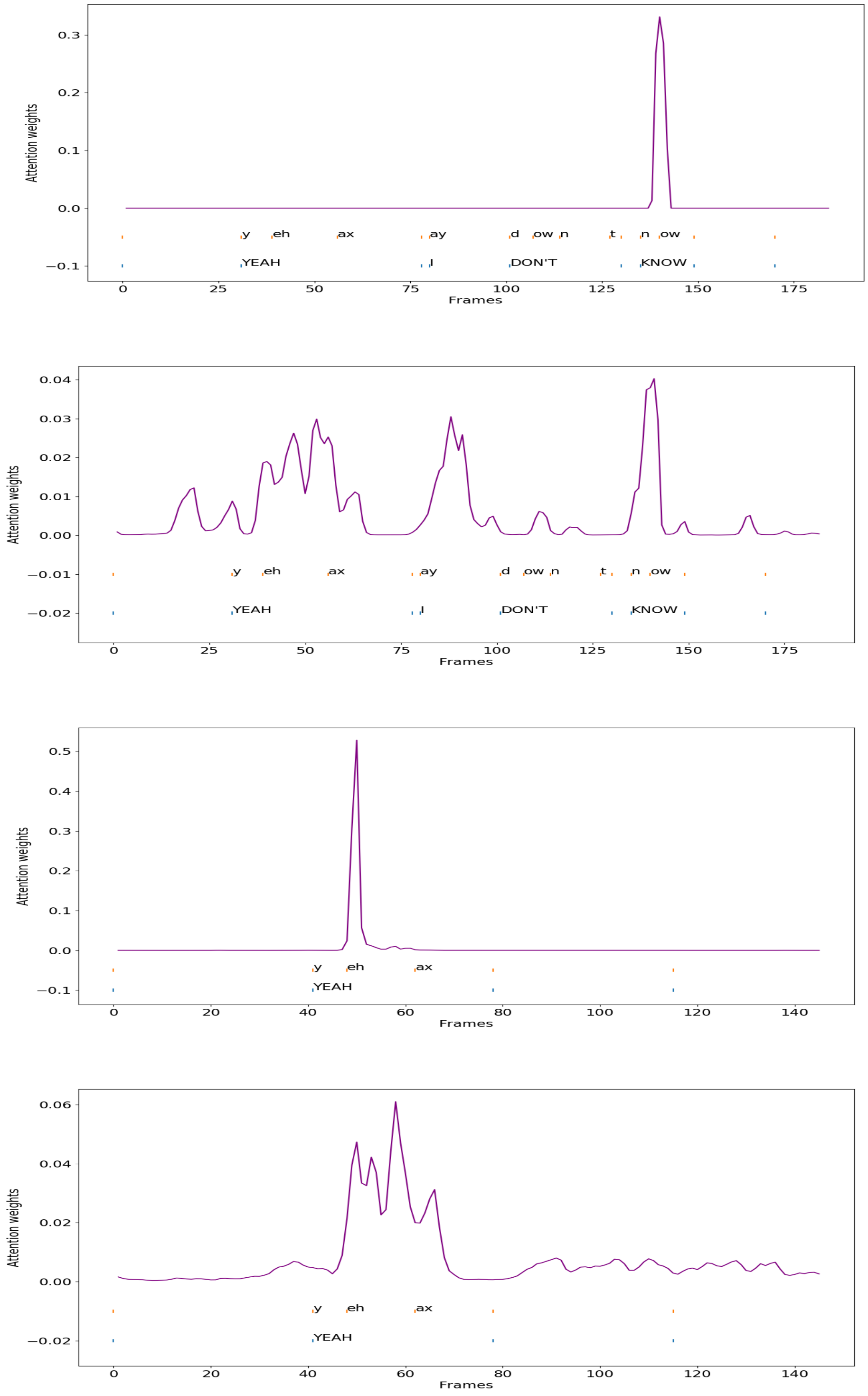


Figure 6.2: *Acoustic attention weights on four different segments from top to bottom (a) MOMA1:"Yeah, I don't know", (b)MOMA2:"Yeah, I don't know" (c) MOMA1:"Yeah" and (d) MOMA2:"Yeah"*

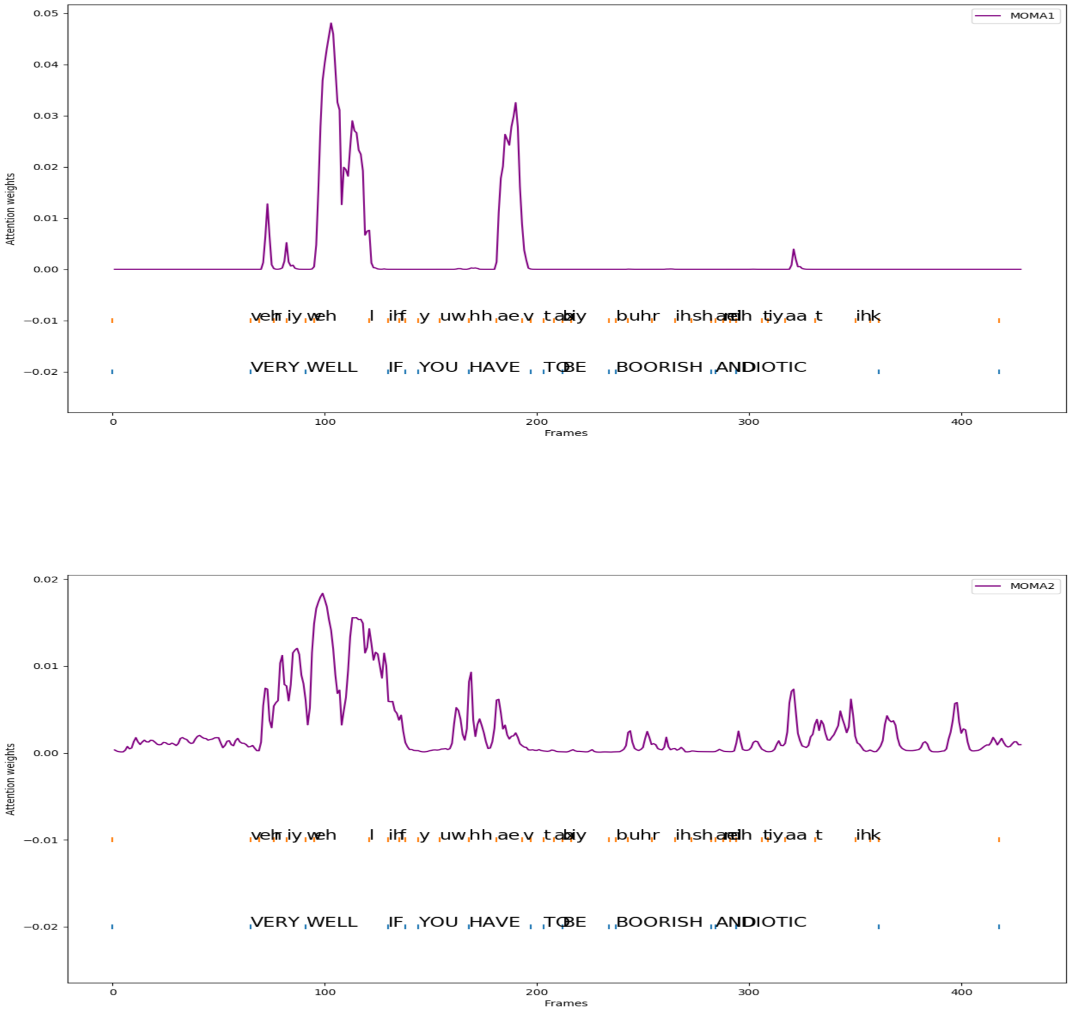


Figure 6.3: *Acoustic attention weights on four different segments from top to bottom (a) MOMA1:"Very well if you have to be boorish and idiotic", (b)MOMA2:"Very well if you have to be boorish and idiotic"*

Hierarchical Attention Weights

To further show how the learned attention representations improve the performance, Figures 6.2 and 6.3 compare the attention at different hierarchies over the same utterance. The audio segments are mapped to the attention to show the relative positions of the attention weights compared to the phones and words.

The projections over two utterances are shown for comparison. The embeddings are extracted from two stages of the network i.e *MOMA1* and *MOMA2*. *MOMA1* (Equation 6.5) is the extracted attention vector embedding at *Attention Layer 1* and *MOMA2* (Equation 6.9) is the attention embedding at *Attention Layer 2* from Figure 6.1. *Attention Layer 2* is the mixture of multi-view attention network.

According to Section 6.1.2, the mixture of attention network is learning loosely correlated task-specific attention representations because the attention blocks are projected onto a common feature space which is added in the end. Thus, it performs smoothing and improves overall robustness which is evident in Figures 6.2 and 6.3. From Figures 6.2a, 6.2c, and 6.3a, it is observed that the attention weights from *Attention Layer 1* embeddings, i.e. *MOMA1* are sensitive to particular regions and phones. However, Figures 6.2b, 6.2d, and 6.3b show that the attention weights from *Attention Layer 2* embeddings, i.e. *MOMA2* are well distributed over the phone boundaries. Thus, it is evident that there are different representations over the different stages of hierarchy in the network. Also, it can be observed that the attention weights of *MOMA2* are well distributed over the phone boundaries compared to *MOMA1*. Whereas the attention in *MOMA1* is sensitive to some regions, but *MOMA2* is smoothed over the overall boundary. These results strongly indicate that *MOMA2* is more robust than *MOMA1*.

Number of Attention Blocks

Although the mixture of multi-view attention shows a significant improvement of the attention weighting over the segments, the optimal number of such attention blocks is unclear. In this work, three blocks have been applied with three control parameters. A higher number of attention blocks may increase the performance of the model, but it can also overfit the model due to the higher number of parameters. Furthermore, it can cause the *degradation* problem in the model.

6.1.6 Discussion

In this section, a residual mixture of multi-view attention emotional context modelling technique, MOMA, using acoustic feature space expansion has been proposed. The model attains task-specific bias in the feature representation resulting in an improved classifier and *state-of-the-art* performance for this SER task. The model also features hierarchical attention. The interpretability of intermediate states of this particular type of attention mechanism has been explored in order to investigate the hypothesis that by projecting the mixture of attention scores into the common feature space, the model is learning loosely correlated task-specific attention spaces and by adding them, the model performs smoothing to achieve more robustness. This has inspired an empirical way to interpret speech-based emotion perception in computational models by plotting the attention weights with respect to the words and the phones. Exploring this network to adapt to different speech-related tasks would be interesting further work.

6.2 Summary

The hierarchical nature of cognition and learning is discussed in Chapter 2. Section 6.1 presents a deep neural network model that performs hierarchical learning with two different attention mechanism. The *MOMA* model produces the best performance in both weighted and unweighted accuracy. Section 6.1 also elaborates on the hierarchical learning by empirically interpreting the intermediate states of the model which opens a whole new way of understanding and building models for emotion recognition and paves the path for analysing the model perception (This is discussed at Chapter 7).

The psychology and cognitive theories are essential to know how to implement the protocols for emotion recognition. If it is possible to understand how the computational models interpret speech emotion, it will be easier to compare it with human perception and directly apply the cognitive theories in speech emotion recognition protocols and models.

Chapter 7

Interpretation of Attention and Context for Speech Emotion

The aim of speech emotion recognition (SER) is to automatically detect human emotions from spoken audio (Dellaert et al. (1996), Picard (1997)). Research in vocal expression recognition of emotion is interdisciplinary. There have been several reviews in the field (Koolagudi & Rao (2012), El Ayadi et al. (2011), Anagnostopoulos et al. (2015), Ververidis & Kotropoulos (2006)) and previous research in psychology has attempted to represent emotions using different models, such as Plutchik's wheel of emotion (Plutchik (1997)) or the hourglass of emotions (Susanto et al. (2020)). However, emotions are complex as they cannot be clearly defined, which makes it difficult to automatically detect them accurately.

Speech emotion recognition is essential for obtaining emotional intelligence which affects the understanding of context and meaning of speech. Other research also claimed that emotion information could exist in small overlapping acoustic cues. The most fundamental distinction that can be made in speech sounds is between vowels and consonants (Ladefoged & Broadbent (1957), Ladefoged (2005)). Harmonically structured vowel and consonant sounds add indexical and linguistic cues in spoken information. Previous research argued whether vowel sound cues were more important in carrying the emotional context from a psychological and linguistic point of view. These sounds also carry socio-linguistic information, and previous research on phonetics and psychology argue whether vowel or consonant sounds are more dominant in determining the underlying emotion. Some research suggests that consonants play a more vital role in delivering socio-linguistics information. However, some of the recent research shows that vowels play a vital role (Cole et al. (1996), Fogerty & Kewley-Port (2009)). According to Ladefoged & Broadbent (1957), vowels have higher variation in formant structures, allowing them to be richer as acoustic context. Some of the researchers hypothesised that vowel cues have better auditory memory representation than consonants for recognising speech emotions.

These claims have not been corroborated in computational speech emotion recognition systems, and it is not typical for current deep neural models to be interpreted with speech emotion data in this way. In this research, a convolution-based model and a long-short-term memory-based model, both using attention, are applied to investigate these theories of speech emotion on computational models.

The above premises invokes some additional research questions about the spatial and temporal nature of emotion and its relation to the computational models. These are

- What is the optimal length of the context to classify emotions?
- How is emotion propagated throughout the communication? Is it a static or dynamic phenomenon?
- How the computational models interpret emotions? If there is any similarity between human perception and machine perception of emotion?

The rest of the chapter is dedicated to investigate and answer these questions.

Most recently, Milner et al. (2019) presented a domain adversarial system for investigating whether the information in acted datasets can be learnt to benefit emotion prediction for natural datasets. The work aimed to be consistent by only considering datasets with adult English speakers with the big-six emotions: happiness, sadness, anger, surprise, disgust and fear. The method applies a bi-directional long short-term memory (BLSTM) with an attention layer and trains in a domain adversarial fashion. It uses sequence modelling, which is arguably more appropriate for use with emotions that change over time. Alternatively, Jalal, Moore & Hain (2019) presented a convolution-based self-attention (CSA) model for speech emotion recognition with fixed-length context sizes. Both the models achieved very high accuracy compared to the current state-of-the-art models.

In this work, two computational SER models based on the previous work (Milner et al. 2019, Jalal, Moore & Hain 2019) have been interpreted in light of the proposed phonetic, linguistic and psychological claims about the acoustics cues for speech emotion recognition in humans.

The role of acoustic context and word importance is demonstrated for the task of speech emotion recognition. The IEMOCAP corpus is evaluated by the proposed models, and 80.1% unweighted accuracy is achieved on pure acoustic data which is higher than current reported state-of-the-art models on this task. The phones and words are mapped to the attention vectors, and it is seen that the vowel sounds are more important for defining emotion acoustic cues than the consonants, and the model can assign word importance based on acoustic context. It is shown that the attention weights in the proposed networks are hugely inclined to the vowel sounds, and it imposes word importance based on preceding/following acoustic context and prosody. It is also shown that smaller acoustic contexts are vital in carrying emotions as previously hypothesised [1].

The rest of this chapter is organised as follows. The relevant cognitive and linguistic research work related to this work and emotion perception is discussed in Section 7.0.1. In Section 7.0.2, the models of this study, i.e. *BLSTMATT* and *CSA*, are described. In Section 7.0.3, the experimental setup is explained, and the results are presented along with a discussion in Section 7.0.5. Finally, Section 7.0.6 makes further discussion and concludes this section. Section 7.1 summarises the chapter.

7.0.1 Consonant-Vowel Boundaries and Perception

Traditionally, it was observed that consonant sounds carry more important speech information until recent studies questioned this claim (Cole et al. (1996), Fogerty & Kewley-Port (2009), Kewley-Port et al. (2007)). Cole et al. (1996) and Fogerty & Kewley-Port (2009) have found that among humans, vowel-only segments have higher intelligibility at sentence level stimuli than consonant-only segments. Owren & Cardillo (2006) performed a similar experiment on word-level stimuli which mostly agrees with Cole et al. (1996), that meaning is more comprehensible to listeners with vowels. Furthermore, Owren & Cardillo (2006) adds that vowel phones at the beginning of a word constitute understanding for listeners, but consonant cues add more acoustic context for meaning. In sentences, vowels occur 10% less than consonants (Ramus et al. (1999)), but the vowel proportions yield the maximum intelligibility in a sentence.

It is a fact that the acoustic cues for sentence intelligibility are distributed across consonant-vowel boundaries (Fogerty & Kewley-Port (2009)). The perceptual cues associated with the acoustics in terms of vowels and consonants interact with each other to build an acoustic-phonetic context for speech perception (Cooper et al. (1952),

[1] This section is accepted for publication in Interspeech 2020 as Jalal, M. A., Milner, R., & Hain, T, (2020, May). "Empirical Interpretation of Speech Emotion Perception with Attention Based Model for Speech Emotion Recognition". In Interspeech 2020.

Miller (1994)). So, these small overlapping cues also hold some portion of perceptual and socio-linguistic information though it is not clear whether these types of acoustic cues can be useful for computational SER tasks.

The role of vowels in perceiving socio-linguistic information and emotion can also be explained with the different harmonic structures of vowels (Ladefoged & Broadbent (1957)). These different harmonic variations in long periods of time constitute prosody which plays a vital role in delivering emotion. Waaramaa et al. (2010) used short vowel samples (150 ms) to remove the prosody effect of vowels and showed that it is possible to perceive emotion from shorter utterances. This implies SER systems could be trained on shorter segments or parts of longer utterances.

7.0.2 Approaches

In order to investigate the attention on the phones and the importance of short segments in regards to SER, two audio-only state-of-the-art approaches proposed in SER are considered. The first is based on sequence modelling and is trained using a bi-directional long short-term memory network (BLSTM) (Milner et al. (2019)), and the second is trained using convolutional neural networks (CNNs) (Jalal, Moore & Hain (2019)) and is not a sequence model. The BLSTM model has been used to investigate the attention put on phones with the sentence-level sample. The CNN based model is used to investigate the importance of the short overlapping acoustic cues for SER.

BLSTM with attention (BLSTMATT)

This approach applies a BLSTM followed by an attention layer and has been described in detail in Milner et al. (2019). LSTM networks ignore the future context and rely on the temporal order of the sequence, whereas BLSTMs in Graves et al. (2013) introduce a second layer of hidden connections which flows in the opposite temporal direction as a way to exploit the contextual information from the past and the future (Schuster & Paliwal (1997)). By applying these networks, a temporal feature distribution over the sequence can be obtained, which is useful for SER tasks.

Attention has the flexibility of computing long-term inter-sequence dependencies. By computing the global mean, the attention mechanism focuses the network onto specific parts of itself which in turn captures global information. The non-linearity $tanh$ is used to multiply the global mean over the whole temporal vector, which computes the positional dependency of each element. The resulting vector is used to compute the attention weights using $softmax$. The soft attention mechanism is also adopted for this work, and the multiplicative method is applied as in Beard et al. (2018).

Finally, the classifier stage of the network contains a fully connected linear layer which projects the attention output down to the number of emotions present. It passes through a $softmax$ layer before computing the loss.

Convolutional Self-Attention (CSA)

The approach in Jalal, Moore & Hain (2019) extracts a spatial feature y using a CNN and performs task-specific high dimensional feature expansion using a self-attention network, which is projected to the original feature dimension. The new feature $\hat{y}$ will be

$$\hat{y} = y + \gamma(A), \tag{7.1}$$

where the learnable parameter γ controls the degree of projection and A is the attention map. The CSA model is described in Section 5.3.2.

7.0.3 Experimental Setup

Data

The IEMOCAP by Busso et al. (2008) dataset has been used for evaluation. The utterances are split into a train set of 4290 (Sessions 1-4) and a test set of 1241 (Session 5) and referred to as IEM4 in this thesis and in Section 2.3.4.

Features

Experiments in Milner et al. (2019) showed how the *BLSTMATT* system performed best in terms of unweighted and weighted accuracy with 23-dimensional log-Mel filter-bank features which are applied to the *CSA* system as well (see Section 5.4).

7.0.4 Implementation

The two systems are implemented in PyTorch Paszke et al. (2017). The *BLSTMATT* performs segment-level classification and the *CSA* performs frame-level classification. The Adam optimiser Kingma & Ba (2014) is applied to the two models with the initial learning rate of 0.0001. As Adam adaptively optimises the learning rate but does not change it, the PyTorch approach of ReduceLROnPlateau has been investigated. The optimum patience setting is found to be four epochs with a multiplicative factor of 0.8. System combination is also explored to investigate whether these models have a complementary relationship for SER. The system combination is done with weighted combinations of the posteriors from the systems. The hyperparameters in the experiments and the models are set up empirically after extensive experimental work.

Segment Level

The *BLSTMATT* contains two hidden layers of 512 nodes each. The output layer of size 1024 is fed into the attention mechanism computing a context vector of size 128, which is projected to 1024 nodes. This is then passed to the emotion classifier which linearly projects to the 4 classes. The cross-entropy loss function is applied, which is preceded by a $softmax$ layer in the PyTorch implementation. The *BLSTMATT* produces a variable-length attention vector based on the input segment length, as mentioned in section 7.0.2. To interpret the acoustic attention, the attention vectors have been extracted and mapped with the phones in the input segments.

Frame Level

The *CSA* consists of three CNN blocks, each block has batch normalisation and rectified linear unit (ReLU) activation. These layers produce 128 channel feature maps which are fused in a convolutional self-attention layer where the number of channels are downsampled. The contextually enhanced output features from the attention layer are given as input to the classifier which linearly projects to the 4 classes.

To investigate the acoustic context length, the utterances are split into chunks with an overlap of 10 frames. The utterances which are less than the context length are not included in the training or test sets. The size of the chunks is varied from 20 to 120 frames.

Evaluation

Unweighted accuracy (UA) and the weighted accuracy (WA) are used to evaluate the results. The calculation method is described in Section 2.4.

Table 7.1: *Results for both model architectures and system combination compared to baseline results on IEM4 data.*

System	Context	UA%	WA%
Factor analysis Desplanques & Demuynck (2018)	-	-	56.1
CNN_LSTM Satt et al. (2017)	-	59.4	-
CNN_RecCap Wu et al. (2019)	-	58.1	-
CNN_GRU-SeqCap Wu et al. (2019)	-	59.7	-
Attention Pool Li et al. (2018*a*)	-	71.8	-
MULTIMODAL: Attention Lian et al. (2019)	-	*78.0*	-
BLSTMATT	Variable	**80.1**	**73.5**
CSA	20	75.8	**69.4**
	30	**76.3**	68.8
	40	75.1	68.0
	50	73.9	67.8
	60	75.1	67.0
	70	74.1	64.7
	80	73.2	67.4
	90	74.8	66.9
	100	74.6	65.9
	110	73.8	67.5
	120	72.2	64.2
SYSCOMB: *BLSTMATT* with *CSA*	V./30	**80.5**	**74.0**

Baseline

The results are directly compared with other SER systems which also use the IEM4 dataset and process only audio. For WA, Desplanques & Demuynck (2018) applies factor analysis in a cross-lingual approach. For UA, in Satt et al. (2017) a CNN-LSTM model is trained, Wu et al. (2019) applies a deep capsule network with gated recurrent units (GRU) for sequence modelling, Li et al. (2018*a*) used deep attention pooling for SER tasks and Li et al. (2018*a*) applies attention pooling. Finally, a multimodal system which is also attention-based and processes both audio and textual data Lian et al. (2019) is included to show the performance the presented audio-only systems could achieve.

7.0.5 Results and Discussion

The experimental results are shown in Table 7.1. The *BLSTMATT* system is trained and tested with whole segments from the corpus. Naturally, the context length for *BLSTMATT* is variable because the segment lengths are not fixed in IEMOCAP. On the contrary, the *CSA* system is trained with fixed-length samples.

The *BLSTMATT* system outperforms the baselines in terms of UA and WA on IEM4. It even outperforms the multimodal system which makes use of textual information as well as audio, which the two presented models do not use. The *BLSTMATT* system outperforms the *CSA* model by 2.7% absolute difference. One of the possible reasons is that the *BLSTMATT* is trained with the whole segment, taking in all the information possible. Typically, in emotion recognition corpora a whole segment is labelled as one emotion category. However, all the smaller acoustic cues from the segment do not necessarily belong to the same emotion category because emotions are dynamic entities and can change momentarily. This segment issue will be briefly discussed later in Chapter 8.

The *CSA* system outperforms the best baselines in terms of UA and WA. However, unlike the *BLSTMATT*, it does not outperform the multimodal Lian et al. (2019) baseline. When comparing the context lengths, *CSA* shows better performance with smaller context lengths, and the best result of UA 76.3% comes with context length of 30. This result does not mean that acoustic length 30 is the optimal acoustic cue length because this particular result is based on the model architecture. However, it can be clearly said that the smaller acoustic cues hold socio-linguistic emotion information as previously claimed by the cognitive studies.

The two best system outputs (*BLSTMATT* and *CSA* with context length 30) can be combined to investigate whether a gain can be achieved from the different training methods. With the *CSA* output posterior probabilities scaled by a factor of 0.4 and

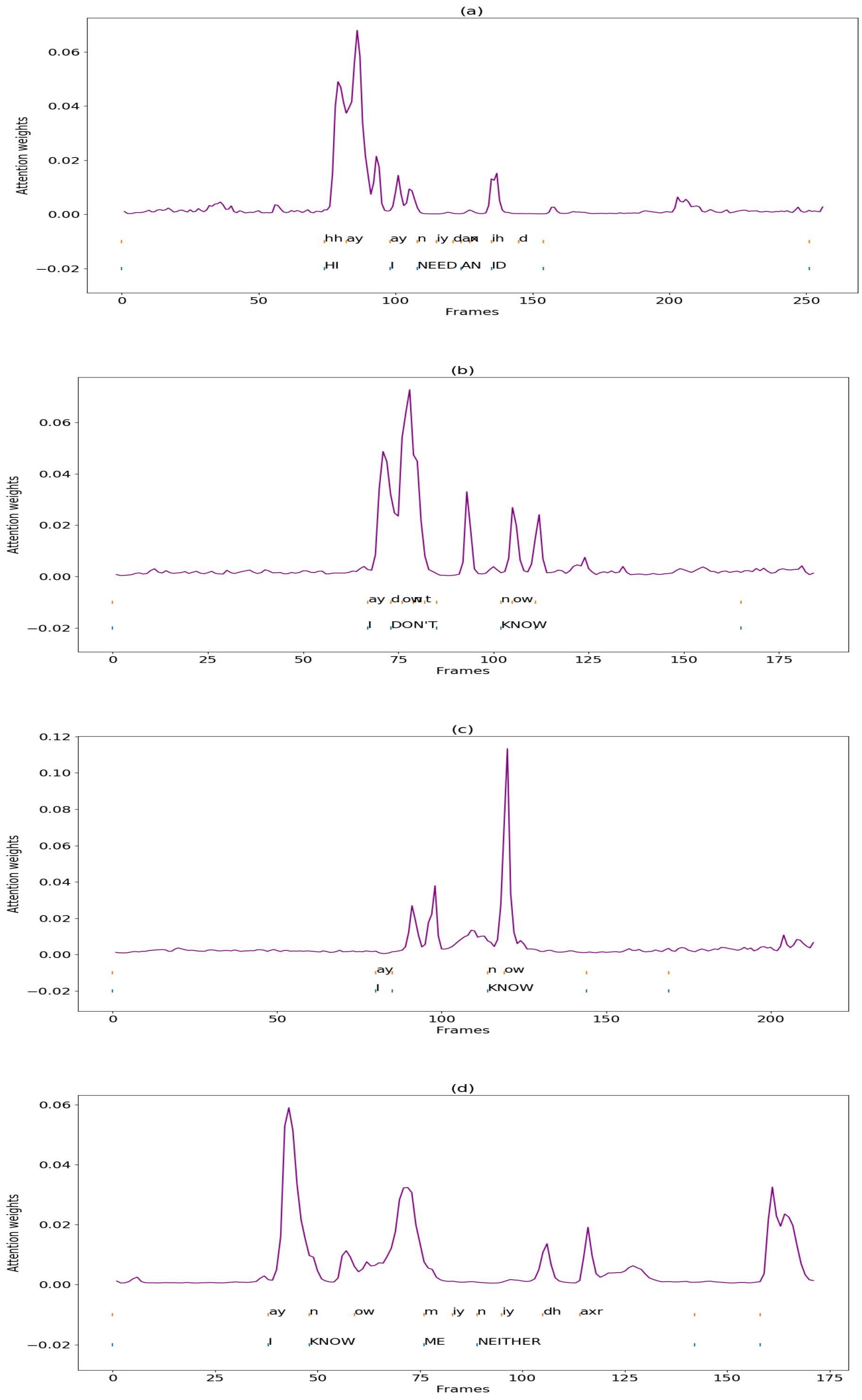


Figure 7.1: *Acoustic attention weights on four different segments from top to bottom (a) Neutral: "Hi, I need an ID", (b) Sad: "I don't know", (c) Sad: "I know" and (d) Happy: "I know, me neither"*

then multiplied by the posterior probabilities from the *BLSTMATT* gives a gain of 0.4% UA and 0.5% WA. This shows both systems learn the emotion classes in different ways, leading to overall improved performance when combining system outputs.

Further remarks in regard to emotion classification for the SER task, the trained neural network classifiers map the input sample to a categorical distribution. Therefore, the output of the supervised DNN SER classifiers is based on the ground truth provided by the annotators. It is likely that some emotions are redundant and challenging to infer based on the different voiced emotion portrayals across cultures. Also, perceptual differences can clearly be seen among the manual annotators of ground truth in IEMOCAP. Speech emotion is a continuous and dynamic process, and it is not logical to consider an emotion state over a long segment. So, using small overlapping acoustic cues for determining emotion state would be a pragmatic future path.

Attention to acoustic cues

For the *BLSTMATT* system, the attention weights for each test segment can be extracted and plotted against the aligned segment. These plots are shown in Figure 7.1 and 7.2. The phones are mapped to the attention vectors to show the relative positions of the attention weights compared to the phones and words. The segments displayed have some common words, but each segment falls under a different emotion category. Common words have been investigated from different emotion categories to demonstrate the word importance weights given by the attention vectors.

Firstly, it can be seen that the attention weights are higher and prominent near the vowel phones, which implies the vowels are incredibly significant for speech emotions. There is a strong correlation seen between vowels and high attention weights. The attention weights on the consonant phones are not high, but they are not negligible. The attention vector projections on consonants are dependent on the vowel cues, and

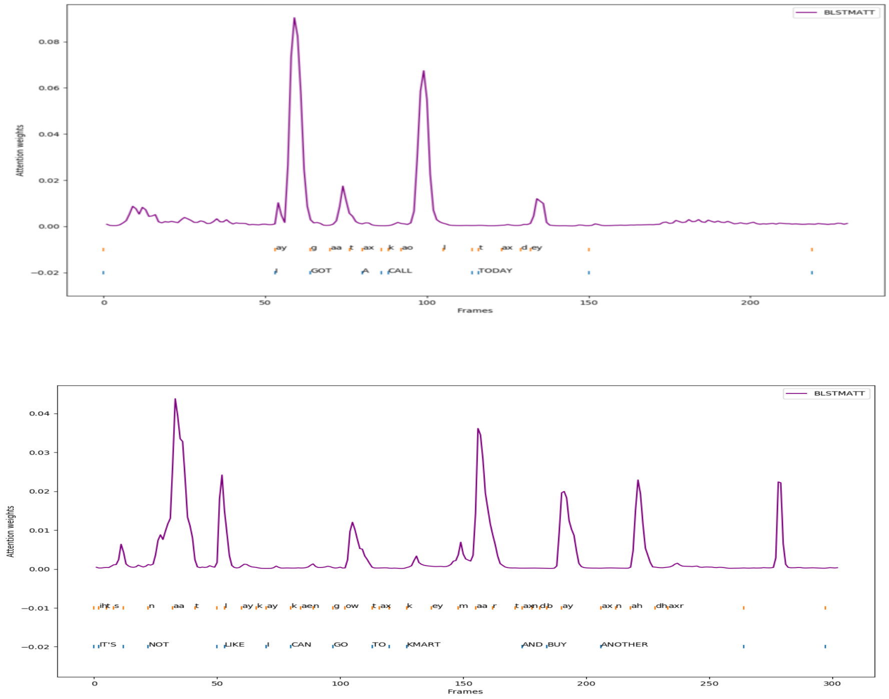


Figure 7.2: *Acoustic attention weights on four different segments from top to bottom (a) Sad: "I got a call today", (b) Anger: "It's not like I can go to KMART and buy you another"*

they constitute consonant-vowel boundaries in the context of emotion. These figures show similarity with the hypotheses and the claims about vowels and emotions from phonetics, psychology and linguistic studies mentioned in Section 7.0.1.

Secondly, the model gives an idea about the word importance in determining an emotion class. Here *word* is used from an acoustic point of view as the *BLSTMATT* model has not been trained with any language model. For example, the word “know” has three different representations over three different emotion categories. One possible reason can be that the preceding/following acoustic cue provides context information for a given region. The other reason lies with the role of prosody. The “I know” segment from both *sad*, Fig. 1c, and *happy*, Fig. 1d, categories have different representations, suggesting a strong relationship between the word importance and prosody for deciding emotion category. The attention on the phone “ay” across different segments and emotions shows the prosodic variation of the phone can constitute to different emotions.

7.0.6 Discussion

Two contrasting systems are presented and evaluated on the commonly used IEM4 dataset, which contains elicited emotions. In this research, the contribution is two-fold. Primarily, in this work, a novel empirical bridge between the cognitive, phonetic theories and the computational models have been demonstrated by interpreting the deep neural models over acoustic speech emotion data. The attention vectors are interpreted by mapping them in the plots with the phones and sentences. Secondly, the relevance of acoustic context information is investigated, and it has been shown that even smaller acoustic cues hold emotion information. The research also argues about the way speech emotion segments are labelled across long segments.

7.1 Summary

The understanding of emotion is subjective to the listeners with their individual perception. When the models learn from the data and their annotation, it also learns their biases. Human annotated data suffers from inconsistencies from different annotators which is a major problem with the available corpora. In most cases, we try to average the scores from all the annotators. Nevertheless, this is a real-world problem as well because humans differ in perception and understanding. So, rather than building a universal model for the whole utterance, the model should look in different instances with different duration within the utterance and calculate relative emotion over time. This approach will make the system more dynamic and less prone to error. Section 7 elaborates an analytical and interpretable framework. The role of vowel-consonant boundaries in an SER model has been explored. Additionally, it is shown that how prosody is related to the vowels and its representation in the intermediate attention representation. The results strongly indicate that vowel utterances play a very important role in building context for speech emotion.

Chapter 8

Implication for Human Robot Interaction

Initially, most of the previous research in human-robot interaction (HRI) was conducted to find human's perspective towards robots. Different studies assessed the way humans communicate with an artificial agent or robot to express their feelings (Turkle (2005), Turkle et al. (2006)). Researchers have found that interaction with social robots decreases stress levels and develops cognitive and emotional skills as well as social interaction (Turkle et al. (2006)). Turkle et al. (2006) mentions robots as *'relational artifacts'*. However, Kanda et al. (2004) conducted an experiment with child robot interaction, and they found that the children were enthusiastic in the beginning, but over one week, they lost interest in their robot companions. Kanda et al. (2004) suggested that this loss of interest is occurring due to the limited communication ability in robots, and they cannot keep up with a child's enthusiasm. Therefore some form of pseudo learning is needed in robots to learn the human counterpart's behaviour and respond according to it. Communication depends on the ability to create relationships between subjects. According to Turkle (2005), we connect to the entity we can nurture and this kind of relational communication increase engagement and mutual connection. The research strongly suggests the benefit of feedback-driven human-robot interaction.

Emotion is an internal factor in human communication. Social robots are created to make autonomous companion agents. Adaptive behaviour in social robots is a self-modification mechanism that maintains the impulse-response system to adapt or respond to an environment or a stimulus. Therefore, adaptive behaviour in social robots makes human-robot interaction more natural. One way to achieve that is by understanding emotion from the human's point of view and then adapting a robot's behaviour according to that understanding. Emotion could be inferred from different signals, i.e. visual, audio, touch etc. Previous research has proposed findings on emotional intelligence in human-robot interaction. However, very little research has focused on robot behaviour modelling based on the user's speech-related emotions.

This chapter suggests a research path for emotion-driven human-robot interaction. Here, the term *'human-robot interaction'* is used for both *virtual agents* and *zoomorphic robots.* The zoomorphic robots and the virtual agents are discussed as a social companion in Section 8.1. Section 8.2 presents the motivation behind the learning protocol in HRI. Section 8.3 presents how the proposed visual image and gesture recognition frameworks in this research may help to contribute to HRI research. Section 8.4 explains the challenges of understanding dynamic changes in emotional states for HRI and how to address them. Section 8.5 discusses possible research paths for emotion-based personalised behaviour models. Section 8.6 summarises the chapter.

8.1 Companion Robots and Virual Agents

Recently, with the advent of virtual agents, talking to a virtual presence in a mobile device is quite common (Jones & Schmidlin (2011), Saerbeck et al. (2010), Robins et al.

(2009), Mitchinson & Prescott (2016)). These entities have no physical presence, and most of the communication is held by speech. Virtual agents like Siri, Google agent, Bixby, Amazon echo, Cortana have become social companions. Unlike the past decade, now it is an established fact that there is no absolute necessity of physical embodiment for artificial agents to communicate with humans. The gestures, speech or text mostly operate these virtual agents. Zoomorphic robots such as Aibo, Paro, Miro are designed to be used as social companions and in the health and education sector (Mitchinson & Prescott (2016)). These virtual agents and zoomorphic robots do not follow the traditional learning approach by mimicking the learning process in children. A part of the learning process employs a modular supervised approach, and the rest use the personalised environment (human domain) to learn behaviour adaptively. For example, Miro has some predefined set of actions, and it also changes its behaviour using external stimuli. The primary goals in these agents are the completion of tasks and bond with humans.

Unlike virtual agents such as Siri, the zoomorphic robots have limited computational power and communication bandwidth. The communication protocols in zoomorphic robots are minimalistic and use limited visual cues, speech information and haptic feedback. The virtual agents also use visual cues (gestures) and speech information for cooperative task completion.

8.2 Approach for Human Robot Interaction

Human-robot interaction has some similarity with the early interaction with newborns. So to understand the nature of human-robot interaction and decision protocols of human-robot interaction, we need to understand how infants learn a skill, language and emotion and how they build the world model.

The goals of HRI are diverse because HRI has diverse application fields and problem-specific design protocols. In this chapter, the motivation is thinking towards companion robots or social robots that include devices and virtual agents to provide social support such as entertainment, teaching, comfort and assistance to children, adult and the elderly. That is the reason the approach is slightly different from the developmental approach, which is

> *"..to the autonomous design of behavioural and cognitive capabilities in artificial agents that takes direct inspiration from the developmental principles and mechanisms observed in the natural cognitive systems of children."* – Cangelosi & Schlesinger (2015)

The motivation is not to build a social companion that acts as a human or reasons as a human but to build a social companion that mimics the essential traits of human communication to please human caregiver. So, the approach is changed from *'autonomous design of behavioural and cognitive capabilities'* to *'supervised the goal-oriented adaptive design of behavioural and cognitive capabilities.'* The goal here is to please the human caregivers. Therefore the term *virtual social caretaker* is more applicable than *robot companion.*

From the arguments and pieces of evidence from Section 2.1.1, 2.1.2 and 2.1.3 it is clear that there is not a single independent theory of learning that explain the whole cognitive development process in babies. So, it is not practical to put only one theory in perspective. Overall, some ideas may be utilised from these theories, such as:

- All the theories agree that there is some form of innate learning structure and a priori knowledge.
- The cognition process can be overlapping and parallel, i.e. multi-sensory learning and processing as an inter-dependent neural process or independent neural

process (Section 2.1.5).

- Bornstein et al. (2012), Biringen & Robinson (1991), Sorce et al. (1985), Sorce & Emde (1981) suggest that learning in early child development is a reciprocal process that is motivated by emotional intelligence.

In HRI systems, each of the modalities may be trained separately as a form of a priori knowledge. The separated modalities are trained using different deep neural networks to map the source input with the goal-oriented specific categorical distribution. Each of these modules, such as vision, hearing, speech, haptics, can be trained separately following particular scenarios (culture, environment, language, human-caregiver such as children, adult, disabled people). These modules can be combined as learnt sensory modalities in a behaviour model. Emotion feedback may be used to tune the behaviour model as an adaptive reciprocal process that is motivated by emotional intelligence.

8.3 Visual Cues

Visual and gestural cues are essential for HRI tasks because these cues enable artificial agents to communicate cooperatively. Humans use different parts of the body for interaction. Therefore one crucial issue would be to control the gaze of the artificial agents (robots or virtual agents). Mohammad et al. (2010) compared different frameworks for gaze control in HRI and found that gaze control is grounded in a dynamic hierarchy of human-robot interaction in different timescales. Their results suggest that gaze control are learned by engaging in basic interactions and evaluating the hierarchy. Also, they found that *naturalness* of behaviour strongly correlates with the human-likeness of action and comfort of the human caregiver. Another approach to gaze control is tracking and selective attention. The tracking algorithm could be nativist or empiricist,i.e. the regions of tracking (salient attention regions) can be predefined, or the regions can be learned adaptively. Nickel & Stiefelhagen (2007), Bertsch & Hafner (2009) used fixed regions and detected skin colours to track the region of interest and employed interaction strategy to learn and detect gestures in an unsupervised manner. These approaches used additional prior knowledge such as head positions and other features such as hand estimation, skin colour etc. These experiments were performed in controlled environments, and in real-world instances, there will be much more challenges such as lighting variations, noises, camera quality, hand & face segmentation and different orientations of the image. Tracking with trajectory learning via optical flow and depth sensors are established techniques in gesture recognition and HRI (Bertsch & Hafner (2009), Nguyen-Duc-Thanh et al. (2012)). It is inevitable to use some form of gaze control to track the region of interest, but we need a robust gesture and posture recognition system that is robust to these environments and orientation changes. Thus, after a region of interest detection, the system should be able to recognise the visual cue irrespective of the environment.

In Section 3.1 a capsule-based hand sign language recognition system has been presented. It applies a representation based hierarchical learning mechanism, which helps to learn the relations between different aspects of the image. The system *routing-by-agreement* that helps to preserve the positional relations (Sabour et al. (2017)). American Sign Language dataset is used to evaluate the system, and 99.7% UA accuracy was achieved. This accuracy is higher than the baseline capsule neural network, convolutional neural network and deep capsule networks. The dataset is distorted with more than fifty variations such as rotation, random pixelation and lighting conditions. However, the results show that the proposed system is robust to these environmental changes. Similar behaviour is observed in Section 3.2 where extreme distortion is introduced to the test images, but the proposed model proved to be robust compared to the other systems. Although capsule networks are computationally complex, a small number of capsules are used in the models so that the proposed systems can perform active posture recognition in real-time.

These mentioned systems learn static postures in an efficient way. The same methods may be used to learn visual object categories and entities. Nevertheless, the artificial agents need to recognise postures too. Section 2.1.5 emphasises the dual neural path for learning in humans. The dual-learning hypothesis relies on the interaction between the ventral and dorsal neural pathways that can be generalised as the interaction between spatial and temporal representation. A similar approach is found in HRI. Section 4.1 presents the fusion techniques between the spatial and temporal stream with an attention mechanism. It shows that with a smaller number of parameters, the neural network can learn similar representations compare to the deeper networks. That implies that some parameters in the layers remain unaffected with the learning process in deeper neural networks, which is also supported empirically in Broughton et al. (2020). Section 4.1 uses a natural action recognition (UCF-101) dataset to show its usability (99% UA). Both dynamic and static visual cue recognition systems proposed in this research are robust in real-life scenarios and challenging conditions, and they have huge potential to be used in HRI.

8.4 Emotional Context

When humans and machines start to share tasks in cooperative interaction, emotions play a significant role as an intrinsic motivator. We know from Section 2.1.3 and 8.2 that emotional intelligence (EI) helps HRI in either side. Emotional gestures, emotion faces and other nuances of artificial agents give humans pleasure and a sense of *naturalness.* Emotion feedback from humans is used in robots as rewards and feedback.

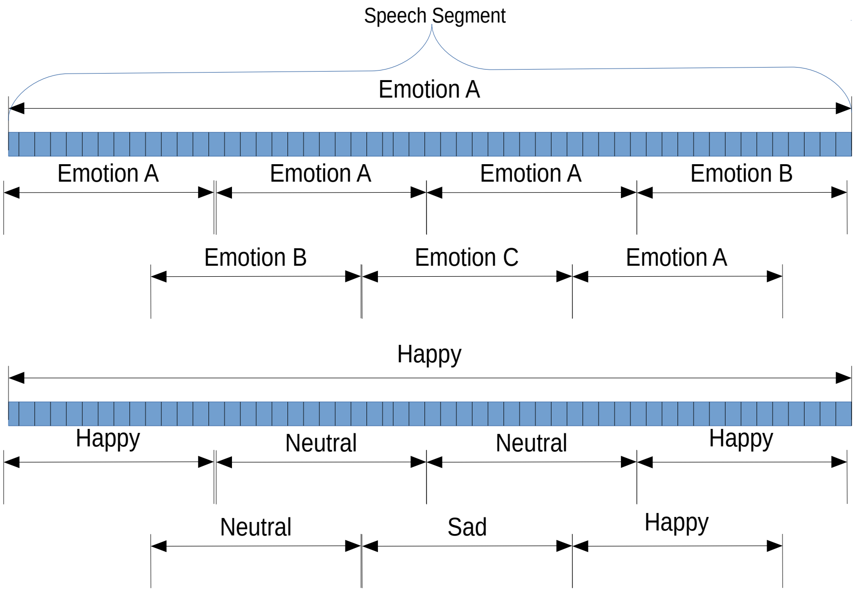


Figure 8.1: *Dynamic nature of emotion cues in speech segments*

However, in those studies, the length of those emotion episodes are not considered in artificial agents.

8.4.1 Overlapping Segments

Section 5.3 presents the *CSA* model to explore the relevance of socio-temporal context in speech emotion that provides real-time recognition of speech emotion. Section 7 uses the *CSA* framework to find the emotional boundaries and context boundaries within the segments. The results show that emotions can be detected even in small acoustic cues or windows (i.e. 200ms durations Table 7.1). More importantly, in the computational models used in the thesis, it has been observed in the experiments that there are different overlapping emotions with overlapping segments. The emotion states change quickly based on the immediate context and length of the acoustic cues or acoustic windows of focus.

The example is shown in Figure 8.1. Suppose the main segment is labelled as happy and the model is trained on the samples from that segment labelled as happy. However, when overlapping segments from that main segment is taken, various other emotions are detected sometimes. One reason may be that the model is sensitive and overfit the training data. However, similar behaviour is spotted with the *BLSTMATT* model and the *MOMA* model. Internal attention representation of the models are plotted against the corresponding speech segments, and they are analysed. It is found that the models are prioritising the vowels and vowel-consonant regions (Section 7.0.5), which is similar to human perception of emotions (Section 7.0.1). Therefore it may be said that different vowel-consonant region triggers the different perception of emotion within the same segment at a given duration.

Figure 8.1 describes that the overlapping acoustic boundaries may represent different 'sub-emotion' within a 'main-emotion' at a given time. This phenomenon explains the dynamic nature of emotions. The phenomenon is very likely to be true in facial emotions as well because, at a given time interval, the facial expressions do not stay the same.

This dynamic nature of emotion may be very helpful in HRI tasks. The tasks in HRI are goal-oriented, and they need constant feedback as an intrinsic reinforcement (Oudeyer & Kaplan (2009), Oudeyer et al. (2007)). These feedbacks use either external stimuli or internal comparison with the ground truth or both. However, dynamic emotion feedbacks will be useful to model human behaviour over the task completion of the artificial agent. As we already know from Field et al. (1986), Bornstein et al. (2012), Biringen & Robinson (1991), Sorce et al. (1985), Sorce & Emde (1981) that children use parental emotional feedback as motivation to model their world exploration and behaviour modelling.

8.4.2 Mixture of experts

Two different models are presented in Chapter 7, the first one is *CSA* and the second one is *BLSTATT* model. *BLSTATT* uses variable-length segments to recognise speech emotions and *CSA* uses small fixed-length segments. Table 7.1 shows that combining these two approaches gives a *state-of-the-art* technique for speech emotion classification. Therefore, as mentioned above, the models can work on a larger acoustic segment and its smaller sub-segments simultaneously. The models can combine their results over a larger time duration, and the result shows improvement with this approach. Thus, the mixture of experts approach is useful with the proposed models because it gives efficient short-term and long-term emotion prediction.

8.4.3 Adaptability

In the previous chapters, we have found out that context modelling helps preserve the universality of emotion across different languages' accents and dialects. Section 5.4 investigated the effects of context modelling over different corpora comprising different accents. Experimental results prove that attention context modelling is effective to learn a universal representation of emotion in language. Also, Section 6.1 presents a residual hierarchical mixture of multi-view attention network (*MOMA*) for context modelling that achieves *state-of-the-art* results on speech emotion recognition.

Virtual agents (Siri, Bixby) and social robots are commercially available, and they communicate with people from various backgrounds, accents and cultures. The context-oriented models will be beneficial for adaptive emotion learning in those circumstances.

8.4.4 Interpretability

The computational models are generally seen as black boxes. The representation of emotional intelligence in the machine models were also treated the same way. In this work, a framework in Section 7 has been proposed for interpreting the SER models with the input speech segments. The framework explained how the SER context model puts attention on the input speech segments to learn speech emotion categories. This open a whole new way of explaining and interpreting human-robot interaction because now the human and machine perception can be compared empirically.

8.5 Behavior Modelling

Oudeyer (2007) and Oudeyer et al. (2007) propose intrinsic motivation based developmental learning framework where the robots learn sensorimotor actions and behaviour. This research suggests learning mimicking children cognitive development. In Cangelosi & Schlesinger (2015), intrinsic motivation (IM) is further explained as 'artificial curiosity' that enables the artificial agent to learn to learn. IM enables the artificial agents for self-directed learning, promoting hierarchical learning and using learned schemas to learn new skills. Learning to distinguish between objects-attributes and to discover novelty is part of behaviour and knowledge development. The computational models use reinforcement learning to model intrinsic motivation, which also examines the motivations of extrinsic motivation on behaviour and learning. Emotions from different modalities may formulate a positive or negative reward. This section briefly discusses how discrete emotions from events may be used as a reward. Barto et al. (2004), Oudeyer (2007), Oudeyer et al. (2007), Cangelosi & Schlesinger (2015) describe reinforced learning model with a standard set of events, policies, environment and a virtual agent. Let use define e^n as the nth event of all possible event set E. The function

$SC(t)$ represents the sensory-perceptual information at time t and $SC(\rightarrow t)$ contains the present context but also the sensory context from the past. The predicted event with function P at time t is $e^{n'}$ and the ground truth of the prediction at time t to $t+1$ is e^{n}.

$$P\left(SC\left(\rightarrow t\right)\right)) = e^{n'}\left(t+1\right). \tag{8.1}$$

The loss would be

$$loss = \left\| e^{n'}\left(t+1\right) - e^{n}\left(t+1\right)\right\|^{2}. \tag{8.2}$$

We may use emotion feedback after the event at time t as a reward. We can calculate the emotion at the transition either by entropy (Equation 8.3) or a weighted sum (Equation 8.4) over the time on different modalities. The emotion reward will be the difference between the feedback at time t and $t+1$.

$$H_{Em}\left(Em, t\right) = -E\left\{\ln\left(Em, t\right)\right\} = -\sum_{em \in Em} P\left(em, t\right) ln\left(em, t\right). \tag{8.3}$$

$$E\left(Em, t\right) = \sum_{t_n \in t} \gamma_n \cdot E\left(Em, t_n\right). \tag{8.4}$$

The reward fuction r is achieved from Equaton 8.2 and 8.3.

$$r\left(SC\left(\rightarrow t\right)\right) = C \cdot e^{\left(\beta_1 \left\| e^{n'}(t+1) - e^{n}(t+1)\right\|^{2} + \beta_2 \left\| H_{Em}(Em,t) - H_{Em}(Em,t+1)\right\|^{2}\right)}, \tag{8.5}$$

where C, β_1, β_2 are constant scaling parameters. This reward formulation may be useful for learning a new skill in a goal-directed fashion where the world model is already learnt.

Another way to look at the behaviour in artificial agents with a limited number of event states (zoomorphic social robots) is to represent the events/states are set of vertices in a connected directed graph. The transition from one vertex to another may be determined by emotion feedback and multimodal sensory information. $\{v_1, v_2, ..., v_n\}$ are the set of actions and $\{e_1, e_2, ..., e_k\}$ are set of events initially unrelated to the action. After each action completion, the artificial agent will learn to execute events autonomously until the next action. It will use the feedback from humans as motivation for the transition. If the feedback function at time t is $F\left(t\right)$ and series of previous states are $E\left(t\right)$. The transition function $S\left(t+1\right) = \left(E \circ F\right)\left(t\right)$. Each state is a tuple of an event e_x and a emotion state $F\left(t\right)$. Here, the tuple is limited with only emotion feedback. However, it may contain multiple sensory information. We can compare different behaviour graphs and adapt to new transitions.

The *Behavioral graph* is a directed graph where the vertices represent actions, and an arc (u, v) denotes that the action v might occur after the occurrence of the action u in a predetermined time frame.

Let G and H be two behavioral graphs. We say G is *strongly similar* to H if $G \rightarrow H$ if there is a function $f\colon V(G) \rightarrow V(H)$ such that for each arc (u, v) of G there is an arc from $f(u)$ to $f(v)$.

We say G is *weakly similar* to H if there is a function $f\colon V(G) \rightarrow V(H)$ such that for each arc (u, v) of G there is a directed path from $f(u)$ to $f(v)$.

Let G be a behavioral graph and A be a subset of arcs of G. We say G is *semi similar* to H if there is a function $f\colon V(G) \rightarrow V(H)$ such that for each arc $(u, v) \in A$ there is an arc from $f(u)$ to $f(v)$, and for any other arc (u', v') of G there is a directed path from $f(u')$ to $f(v')$.

The behaviour graphs can be customised and modulated by the human user's emotional feedback. These graphs can be used to analyse and cluster behaviours among the human participants to personalise the communication between humans and robots/virtual agents.

Hypothesis 1 *If a behavioural graph G is strongly similar to a behavioural graph H, there is a strong correspondence between the actual behaviours of the corresponding entities.*

Hypothesis 2 *If a behavioural graph G is weakly similar to a behavioural graph H, there is a weak correspondence between the actual behaviours of the corresponding entities.*

8.6 Summary

This chapter discusses the possible ways of using emotional intelligence and context cues for HRI and its motivation. Section 8.2 presents a layout for the learning approach in HRI and uses arguments from Section 2.1.1, 2.1.2 and 2.1.3 to establish it. Section 8.3 and 8.4 use the findings from vision and emotion research to propose possible ways to use them in HRI. Section 8.4 justifies using overlapping cues and mixture models of multiple experts to address the dynamic and redundant nature of emotion in HRI. Section 8.5 discuss the possible use of emotion feedback as a reward function for learning action transitions and behaviour and behaviour graphs.

Chapter 9

Conclusions

The main goal of this research is to study the role of context and attention learning using vision, emotion information and novel deep neural network models and to be able to understand the sub-symbolic abstract learning in these learned representations. Speech emotion information has been used predominantly for spatiotemporal context modelling. As part of the experiments, different aspects of speech emotion in computational models have been explored. Furthermore, static and dynamic visual information (posture/gesture) is also used for context modelling. The research questions, introduced in Section 1.1, are addressed throughout the thesis, and some key findings and contributions from the research are given below.

- A novel capsule dynamic routing based American sign language recognition model and a vehicle logo recognition model, which learn hierarchical task-specific representations with dynamic routing and robust to rotation, noise and distortion compared to the CNN models (Presented in Chapter 3). While addressing the research question 1 about hierarchical structures and context cues in Section 1.1, I investigated the implications of hierarchical attention learning and attention fusion over different network hierarchies. Also, dual learning and dual fusion techniques using attention have been investigated. The results imply cross-modal cue learning is important for context modelling. The hierarchical structure plays a role too. However, it has been seen that data-dependent modelling plays a vital role in learning generalised representations (Presented in Chapter 4). However, I have performed experiments with augmented data. It will be a future work to use the un-augmented data and investigate the difference.

- The research question 2 and 3 about the emotion context in Section 1.1 inspired Chapter 5 and 6 with three different attention models. Firstly, a novel hybrid temporal capsule network for speech emotion recognition (SER) is proposed for children speech emotion and acted speech emotion. The intermediate capsule clusters show that the model learns distinct temporal context representations in separate capsules (Presented in Chapter 5, Section 5.3). Secondly, a novel spatiotemporal context modelling framework with convolutional self-attention (CSA) is proposed to model bias and context in the samples for SER (Presented in Chapter 5, Section 5.3). This spatiotemporal context modelling technique gives superior classification accuracy (76.3% UA) with fewer model parameters compared to the state-of-the-art DNN models. This framework has been used to show that emotion context cues can be present in a speech in shorter durations. Also, it is shown that emotional cues are overlapping and dynamic (Presented in Chapter 7). Finally, a novel mixture of multi-view attention (MOMA) mechanism is proposed for attention smoothing and context modelling by learning task-specific bias in data. (Presented in Chapter 6, Section 6.2). The intermediate embeddings are analysed with an attention alignment framework to visualise the findings. This MOMA model produces superior classification accuracy (80.5% UA, 74.8% WA) compared to the previous spatiotemporal context model and other state-of-the-art models trained on only acoustic speech data. However, the number of parameters in MOMA model is significantly higher than the convolutional self-attention model in Section 5.3. One thing is evident that the acoustic cue boundaries are dynamic with clear phonemic correlations. It is also noticed that the attention cues are gradually learned and refined over network hierarchies.

- It is shown with a simple BLSTM-Attention model (BLSTMATT) that there is a common emotion representation among all different types (acted, elicited, natural) of speech emotion corpora. This finding implies that it is possible to use joint corpora training to learn common context representations among different types of speech emotion data. Currently, speech emotion corpora have insufficient recording hours, an imbalanced number in sample distribution per class and conflict in the manual annotation. The results indicate that with the joint corpora training regime, these problems can be alleviated for SER model training (Section 5.5).

- The research question 3 in Section 1.1 enquires about the phone boundaries and context cue length for SER. In this thesis, the effects of consonant-vowel (CV) boundaries are shown in the attention representation using BLSTMATT model and attention alignment. The results show that the vowel phones have a significant impact on modelling context in speech emotion. These results imply the role of acoustic context and prosody for speech emotion context (Presented in Chapter 7). Briefly, I have explored joint model performance in Chapter 7, which shows a little performance gain. The CSA model and the BLSTMATT model train differently for context representation. The CSA model is trained on fixed-length frames, but the BLSTMATT model is trained on variable-length segments. It is shown that a mixture of experts learning may improve the overall performance (0.4% UA and 0.5% WA gain in this case).

- As an added discussion from research question 4 in Section 1.1, how these findings of vision and emotion may impact HRI are discussed. With strong cognitive, psychological and computational evidence, it is argued that emotional feedback may be useful in goal-oriented task and behaviour modelling in HRI. I have also discussed how the use of emotion and visual context in overlapping segments may be beneficial for HRI and behaviour modelling.

However, there are some limitations of this thesis that can be addressed as future work.

- The capsule networks are computationally expensive, and it has not been used in action recognition experiments due to computational complexity. The capsule based architectures are only used for smaller corpora.
- The effect of augmented data vs un-augmented data has not investigated in Chapter 4 fusion networks. It will be interesting to analyse the difference with un-augmented data.
- Also, it will be interesting to see the difference in training between temporally ordered frames and temporally unordered frames for gesture recognition, which has not been addressed in this thesis.
- It is not clear if the networks are generalising the data and memorising the data up to what extent. Canonical correlation analysis over these networks may be an interesting way forward.
- This thesis shows the cv boundaries and alignment for SER tasks. However, the attention alignment is for a particular attention mechanism. It is not clear from the work that every attention mechanism produce similar alignment or not.

- The research also argues about the way speech emotion segments is labelled across long segments. The precise temporal limitations of this phenomenon are not clear.

- Hyperparameter tuning for the proposed models can be performed to increase model performance further.

Additionally, findings of acoustic cue length and CV boundaries corroborate the cognitive theories about acoustic context boundaries and speech emotion perception; more research is necessary to confirm these theories in computational models promptly. The internal representation in the computational models depends on the structure of the networks, the training data, ground truth and objective function. Thus, many variabilities should be addressed before confirming the universality of these theories in neural network models.

Furthermore, another significant improvement would be applying domain and model adaptation in these models for learning new context representations in the existing structures. Adaptation would further distil the anomalies while learning the task-specific context representations. The structural effect of deep neural networks is a big area of research in machine learning. It is unclear how each layer and hierarchy in the neural networks contribute to the overall model learning. The results from Broughton et al. (2020) strongly suggest that in deep neural network structures, some layers in specific hierarchies do not learn significantly, and they remain nearly to the initialised state [1]. It will be great future work to examine the amount of learning in these attention layers and the dependency on specific structures and training regime. Sailor et al. (2019) shows that rapid embedding compression and expansion with the

[1] This research is submitted in Interspeech 2021 as Broughton, S. J., Jalal, M. A., & Moore, R. K. “Investigating Deep Neural Structures and their Interpretability in the Domain of Voice Conversion”. Interspeech 2021 (on review)

squeeze and excitation network produce a richer representation that helps better model adaptation [2]. Adding these structures in the existing SER models may result in better spatiotemporal context modelling and would be a promising research path.

Finally, I have discussed the implications of the emotion models in HRI. However, these are not experimentally proven. So, using these SER models in HRI research would also be essential work for the future.

[2] This research is published in ASRU 2019 as H. B. Sailor, S. Deena, M. A. Jalal, R. Lileikyte and T. Hain, "Unsupervised Adaptation of Acoustic Models for ASR Using Utterance-Level Embeddings from Squeeze and Excitation Networks," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 980-987

Bibliography

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, A. P., Toderici, G., Varadarajan, B. & Vijayanarasimhan, S. (2016), Youtube-8m: A large-scale video classification benchmark, *in* 'arXiv:1609.08675'.

Admasu, Y. F. & Raimond, K. (2010), Ethiopian sign language recognition using Artificial Neural Network, *in* 'Intelligent Systems Design and Applications ISDA 2010 10th International Conference on', pp. 995–1000.

Ameen, S. & Vadera, S. (2017), 'A convolutional neural network to classify american sign language fingerspelling from depth and colour images', *Expert Systems* **34**(3), e12197.

Anagnostopoulos, C.-N., Iliou, T. & Giannoukos, I. (2015), 'Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011', *Artificial Intelligence Review* pp. 155–177.

Aryanie, D. & Heryadi, Y. (2015), American sign language-based finger-spelling recognition using k-Nearest Neighbors classifier, *in* '2015 3rd International Conference on Information and Communication Technology, ICoICT 2015', pp. 533–536.

Association, B. D. (2018), 'Bsl statistics. in: Sign language week 2018.'.

Ba, J., Mnih, V. & Kavukcuoglu, K. (2015), 'Multiple object recognition with visual attention', *CoRR* **abs/1412.7755**.

Bahdanau, D., Cho, K. & Bengio, Y. (2014), Neural machine translation by jointly learning to align and translate, *in* 'Proc. ICLR'.

Baldwin, D., Andersson, A., Saffran, J. & Meyer, M. (2008), 'Segmenting dynamic human action via statistical structure', *Cognition* **106**(3), 1382–1407.

Barros, P., Magg, S., Weber, C. & Wermter, S. (2014), A multichannel convolutional neural network for hand posture recognition, *in* 'International Conference on Artificial Neural Networks', Springer, pp. 403–410.

Barto, A. G., Singh, S. & Chentanez, N. (2004), Intrinsically motivated learning of hierarchical collections of skills, *in* 'Proceedings of the 3rd International Conference on Development and Learning', Cambridge, MA, pp. 112–19.

Bates, E., Dale, P. S. & Thal, D. (1995), 'Individual differences and their implications for theories of language development', *Handb. Child Lang.* pp. 96–151.

Batliner, A., Steidl, S., Hacker, C. & Nöth, E. (2008), 'Private emotions versus social interaction: A data-driven approach towards analysing emotion in speech', *User Modeling and User-Adapted Interaction* **18**(1-2), 175–206.

Bay, H., Ess, A., Tuytelaars, T. & Van Gool, L. (2008), 'Speeded-up robust features (surf)', *Computer vision and image understanding* **110**(3), 346–359.

Beard, R., Das, R., Ng, R. W. M., Gopalakrishnan, P. G. K., Eerens, L., Swietojanski, P. & Miksik, O. (2018), Multi-modal sequence fusion via recursive attention for emotion recognition, *in* 'Proceedings of the 22nd Conference on Computational Natural

Language Learning', Association for Computational Linguistics, Brussels, Belgium, pp. 251–259.

Bengio, Y. (2009), 'Learning deep architectures for ai', *Found. Trends Mach. Learn.* **2**(1), 1–127.

Bengio, Y. & Lecun, Y. (2007), *Scaling learning algorithms towards AI*, MIT Press.

Bertsch, F. A. & Hafner, V. V. (2009), Real-time dynamic visual gesture recognition in human-robot interaction, *in* '2009 9th IEEE-RAS International Conference on Humanoid Robots', IEEE, pp. 447–453.

Bheda, V. & Radpour, D. (2017), 'Using Deep Convolutional Networks for Gesture Recognition in American Sign Language', *CoRR* **abs/1710.06836**.

Bichot, N. P., Heard, M. T., DeGennaro, E. M. & Desimone, R. (2015), 'A source for feature-based attention in the prefrontal cortex', *Neuron* **88**(4), 832–844.

Biringen, Z. & Robinson, J. (1991), 'Emotional availability in mother-child interactions: A reconceptualization for research', *American journal of Orthopsychiatry* **61**(2), 258–271.

Bornstein, M. H., Suwalsky, J. T. & Breakstone, D. A. (2012), 'Emotional relationships between mothers and infants: Knowns, unknowns, and unknown unknowns', *Development and psychopathology* **24**(1), 113–123.

Broughton, S. J., Jalal, M. A. & Moore, R. K. (2020), 'Investigating deep neural structures and their interpretability in the domain of voice conversion', *SLT 2020 (submitted)* .

Brox, T., Bruhn, A., Papenberg, N. & Weickert, J. (2004), High accuracy optical flow estimation based on a theory for warping, *in* 'ECCV'.

Bruner, J. (1987), 'Life as Narrative', *Soc. Res. An Int. Q. Reflections Self* **54**(1), 11–32.

Bruner, J. S. (1961), 'The act of discovery.', *Harvard educational review* .

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Provost, E. M., Kim, S., Chang, J. N., Lee, S. & Narayanan, S. (2008), 'Iemocap: interactive emotional dyadic motion capture database', *Language Resources and Evaluation* **42**, 335–359.

Campos, J. J. & Barrett, K. C. (1984), 'Toward a new understanding of emotions and their development', *Emotions, cognition, and behavior* pp. 229–263.

Cangelosi, A. & Schlesinger, M. (2015), *Developmental robotics: From babies to robots.*

Cao, H., Verma, R. & Nenkova, A. (2015), 'Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech', *Computer Speech and Language* .

Carcagnì, P., Coco, M. D., Leo, M. & Distante, C. (2015), Facial expression recognition and histograms of oriented gradients: a comprehensive study, *in* 'SpringerPlus'.

Chen, R., Hawes, M., Mihaylova, L., Xiao, J. & Liu, W. (2016), Vehicle logo recognition by spatial-sift combined with logistic regression, *in* '2016 19th International Conference on Information Fusion (FUSION)', IEEE, pp. 1228–1235.

Chen, R., Jalal, M. A., Mihaylova, L. & Moore, R. (2018), Learning capsules for vehicle logo recognition, *in* '2018 21st International Conference on Information Fusion (FUSION) (FUSION 2018)', Cambridge, United Kingdom (Great Britain).

Chen, S. & Jin, Q. (2015), Multi-modal dimensional emotion recognition using recurrent neural networks, *in* 'Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge', AVEC '15, ACM, New York, NY, USA, pp. 49–56.

Cho, K., Courville, A. & Bengio, Y. (2015), 'Describing multimedia content using attention-based encoder-decoder networks', *IEEE Transactions on Multimedia* **17**(11), 1875–1886.

Chomsky, N. (1959), 'A Review of B.F. Skinner's Verbal Behavior', *Readings Psychol. Lang. PrenticeHall* **35**, 142–143.

Chuan, C.-H., Regina, E. & Guardino, C. (2014), American Sign Language Recognition Using Leap Motion Sensor, *in* 'Proc. from the 2014 13th International Conference on Machine Learning and Applications', pp. 541–544.

Cloutman, L. L. (2013), 'Interaction between dorsal and ventral processing streams: Where, when and how?', *Brain and Language* **127**(2), 251–263.

Cole, R. A., Yan, Y., Mak, B., Fanty, M. & Bailey, T. (1996), The contribution of consonants versus vowels to word recognition in fluent speech, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)'.

Connie, T., Al-Shabi, M., Cheah, W. P. & Goh, M. (2017), Facial expression recognition using a hybrid cnn–sift aggregator, *in* S. Phon-Amnuaisuk, S.-P. Ang & S.-Y. Lee, eds, 'Multi-disciplinary Trends in Artificial Intelligence', Springer International Publishing, Cham, pp. 139–149.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. & Gerstman, L. J. (1952), 'Some experiments on the perception of synthetic speech sounds', *The Journal of the Acoustical Society of America* .

Dahmane, M. & Meunier, J. (2011), Emotion recognition using dynamic grid-based hog features, *in* 'Face and Gesture 2011', pp. 884–888.

Dalal, N. & Triggs, B. (2005), Histograms of oriented gradients for human detection, *in* '2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)', Vol. 1, IEEE, pp. 886–893.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. & Ouellet, P. (2011), 'Front-end factor analysis for speaker verification', *Trans. Audio, Speech and Lang. Proc.* **19**(4), 788–798.

Dellaert, F., Polzin, T. & Waibel, A. (1996), Recognizing emotion in speech, *in* 'Proc. ICSLP'.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* '2009 IEEE conference on computer vision and pattern recognition', Ieee, pp. 248–255.

Desplanques, B. & Demuynck, K. (2018), Cross-lingual speech emotion recognition through factor analysis, *in* 'Proc. Interspeech'.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. & Darrell, T. (2014), 'Long-term recurrent convolutional networks for visual recognition and description', *CoRR* **abs/1411.4389**.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R. & Pal, C. (2015), Recurrent neural networks for emotion recognition in video, *in* 'Proceedings of the 2015 ACM on International Conference on Multimodal Interaction', ICMI '15, ACM, New York, NY, USA, pp. 467–474.

Ekman, P. (1992), 'An argument for basic emotions', *Cognition & emotion* **6**(3-4), 169–200.

El Ayadi, M., Kamel, M. S. & Karray, F. (2011), 'Survey on speech emotion recognition: Features, classification schemes, and databases', *Pattern Recognition* **44**(3), 572–587.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S. & Truong, K. (2016), 'The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing', *IEEE Transactions on Affective Computing* .

Fahlman & Hinton (1987), 'Connectionist architectures for artificial intelligence', *Computer* **20**(1), 100–109.

Fan, Y., Lu, X., Li, D. & Liu, Y. (2016), Video-based emotion recognition using cnn-rnn and c3d hybrid networks, *in* 'Proceedings of the 18th ACM International Conference on Multimodal Interaction', ICMI '16, ACM, New York, NY, USA, pp. 445–450.

Fathi, A. & Mori, G. (2008), Action recognition by learning mid-level motion features, *in* '2008 IEEE Conference on Computer Vision and Pattern Recognition', pp. 1–8.

Feldman, J. & Ballard, D. (1982), 'Connectionist models and their properties', *Cognitive Science* **6**(3), 205–254.

Fernández-Isabel, A. & Fuentes-Fernández, R. (2015), 'Analysis of intelligent transportation systems using model-driven simulations', *Sensors* **15**(6), 14116–14141.

Field, T., Vega-Lahr, N., Scafidi, F. & Goldstein, S. (1986), 'Effects of maternal unavailability on mother-infant interactions', *Infant Behavior and Development* **9**(4), 473 – 478.

Fisher, C., Hall, D. G., Rakowitz, S. & Gleitman, L. (1994), 'When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth', *Lingua* **92**(C), 333–375.

Fogerty, D. & Kewley-Port, D. (2009), 'Perceptual contributions of the consonant-vowel boundary to sentence intelligibility', *The Journal of the Acoustical Society of America* **126**(2), 847–857.

Fong, T., Nourbakhsh, I. & Dautenhahn, K. (2003), 'A survey of socially interactive robots', *Robotics and Autonomous Systems* **42**, 143–166.

Fukushima, K. (1980), 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position', *Biological cybernetics* **36**(4), 193–202.

Garcia, B. & Viesca, S. a. (2016), 'Real-time American Sign Language Recognition with Convolutional Neural Networks', *Convolutional Neural Networks for Visual Recognition* .

Georgescu, M., Ionescu, R. T. & Popescu, M. (2018), 'Local learning with deep and handcrafted features for facial expression recognition', *CoRR* **abs/1804.10892**.

Gleitman, L. R., Gleitman, H., Landau, B. & Wanner, E. (1988), 'Where learning begins: Initial representations for language learning', pp. 150–193.

Gleitman, L. R. & Newport, E. L. (1995), The invention of language by children: Environmental and biological influences on the acquisition of language, *in* 'Development', number 1995, pp. 1–24.

Goldman-Rakic, P. S., Cools, A. R. & Srivastava, K. (1996), 'The prefrontal landscape: Implications of functional architecture for understanding human mentation and the central executive [and discussion]', *Philosophical Transactions: Biological Sciences* **351**(1346), 1445–1453.

Gomes, H., Molholm, S., Christodoulou, C., Ritter, W. & Cowan, N. (2000), 'The development of auditory attention in children', *Frontiers in Bioscience* **5**(1), D108–D120.

Goodale, M. A. & Milner, A. (1992), 'Separate visual pathways for perception and action', *Trends in Neurosciences* **15**(1), 20–25.

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep learning*, MIT press.

Graves, A., Jaitly, N. & rahman Mohamed, A. (2013), Hybrid speech recognition with deep bidirectional lstm, *in* 'Proc. ASRU'.

Graves, A. & Schmidhuber, J. (2005), Framewise phoneme classification with bidirectional LSTM networks, *in* 'Proceedings of the International Joint Conference on Neural Networks', Vol. 4, pp. 2047–2052.

Gritti, T., Shan, C., Jeanne, V. & Braspenning, R. (2008), Local features based facial expression recognition with face registration errors, *in* '2008 8th IEEE International Conference on Automatic Face Gesture Recognition', pp. 1–8.

Hanin, B. & Rolnick, D. (2018), How to start training: The effect of initialization and architecture, *in* S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi & R. Garnett, eds, 'Advances in Neural Information Processing Systems 31', Curran Associates, Inc., pp. 571–581.

HasanPour, S. H., Rouhani, M., Fayyaz, M., Sabokrou, M. & Adeli, E. (2018), 'Towards principled design of deep convolutional networks: Introducing simpnet', *CoRR* **abs/1802.06205**.

Hayden, B. Y. & Gallant, J. L. (2009), 'Combined effects of spatial and feature-based attention on responses of v4 neurons', *Vision Research* **49**(10), 1182 – 1187. Visual Attention: Psychophysics, electrophysiology and neuroimaging.

He, K., Zhang, X., Ren, S. & Sun, J. (2015), 'Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(9), 1904–1916.

Hebb, D. O. (1949), *The organization of behavior: A neuropsychological theory*, Psychology Press.

Heinke, D. & Humphreys, G. W. (2003*a*), 'Attention, spatial representation, and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (saim).', *Psychological review* **110**(1), 29.

Heinke, D. & Humphreys, G. W. (2003*b*), 'Attention, Spatial Representation, and Visual Neglect: Simulating Emergent Attention and Spatial Memory in the Selective Attention for Identification Model (SAIM)', *Psychological Review* **110**(1), 29–87.

Hinton, G. E. (1989), 'Connectionist learning procedures', *Artificial Intelligence* **40**(1), 185–234.

Hinton, G. E. & Anderson, J. A. (2014), Implementing semantic networks in parallel hardware, *in* 'Parallel models of associative memory', Psychology Press, pp. 201–232.

Hinton, G. E. & Salakhutdinov, R. R. (2006), 'Reducing the dimensionality of data with neural networks', *Science* **313**(5786), 504–507.

Hinton, G. F. (1981), A parallel computation that assigns canonical object-based frames of reference, *in* 'Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2', IJCAI'81, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 683685.

Hochreiter, S. (1998), 'The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **06**(02), 107–116.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Comput.* **9**(8), 1735–1780.

Hochreiter, S. & Urgen Schmidhuber, J. (1997), 'LONG SHORT-TERM MEMORY', *Neural Computation* **9**(8), 1735–1780.

Hoemann, K., Xu, F. & Barrett, L. F. (2019), 'Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis.', *Developmental psychology* **55**(9), 1830.

Horn, B. K. & Schunck, B. G. (1981), 'Determining optical flow', *Artificial intelligence* **17**(1-3), 185–203.

Huang, C., Liang, B., Li, W. & Han, S. (2017), 'A convolutional neural network architecture for vehicle logo recognition', *2017 IEEE International Conference on Unmanned Systems (ICUS)* pp. 282–287.

Huang, C. W. & Narayanan, S. S. (2017), Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition, *in* 'Proc. ICME'.

Huang, Y., Wu, R., Sun, Y., Wang, W. & Ding, X. (2015), 'Vehicle logo recognition system based on convolutional neural networks with a pretraining strategy', *IEEE Transactions on Intelligent Transportation Systems* **16**(4), 1951–1960.

Huang, Z., Dong, M., Mao, Q. & Zhan, Y. (2014), Speech emotion recognition using cnn, *in* 'Proceedings of the 22Nd ACM International Conference on Multimedia', MM '14, ACM, New York, NY, USA, pp. 801–804.

Hubel, D. H. & Wiesel, T. N. (1959), 'Receptive fields of single neurones in the cat's striate cortex', *The Journal of Physiology* **148**(3), 574–591.

Hubel, D. H. & Wiesel, T. N. (1962), 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex', *The Journal of Physiology* **160**(1), 106–154.

Hubel, D. H. & Wiesel, T. N. (1963), 'Receptive fields of cells in striate cortex of very young, visually inexperienced kittens', *Journal of Neurophysiology* **26**(6), 994–1002. PMID: 14084171.

Ijjina, E. P. & Mohan, C. K. (2014), Facial expression recognition using kinect depth sensor and convolutional neural networks, *in* '2014 13th International Conference on Machine Learning and Applications', pp. 392–396.

Ioffe, S. & Szegedy, C. (2015), 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', *CoRR* **abs/1502.03167**.

Izard, C. E. (1977), Differential emotions theory, *in* 'Human emotions', Springer, pp. 43–66.

Izard, C. E. (1994), 'Innate and universal facial expressions: evidence from developmental and cross-cultural research.'.

Jaderberg, M., Simonyan, K., Zisserman, A. & kavukcuoglu, k. (2015), Spatial transformer networks, *in* C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett, eds, 'Advances in Neural Information Processing Systems 28', Curran Associates, Inc., pp. 2017–2025.

Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P. & Zareapoor, M. (2018), 'Hybrid deep neural networks for face emotion recognition', *Pattern Recognition Letters* **115**, 101–106.

Jalal, M. A., Chen, R., Moore, R. K. & Mihaylova, L. (2018), American sign language posture understanding with deep neural networks, *in* '2018 21st International Conference on Information Fusion (FUSION)', pp. 573–579.

Jalal, M. A., Loweimi, E., Moore, R. K. & Hain, T. (2019), 'Learning temporal clusters using capsule routing for speech emotion recognition', *Proc. Interspeech 2019* pp. 1701–1705.

Jalal, M. A., Moore, R. K. & Hain, T. (2019), Spatio-temporal context modelling for speech emotion classification, *in* '2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)', pp. 853–859.

Jebali, M., Dalle, P. & Jemni, M. (2013), Hmm-based method to overcome spatiotemporal sign language recognition issues, *in* '2013 International Conference on Electrical Engineering and Software Applications, ICEESA 2013'.

Ji, S., Xu, W., Yang, M. & Yu, K. (2013), '3d convolutional neural networks for human action recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(1), 221–231.

Jones, K. S. & Schmidlin, E. A. (2011), 'Human-robot interaction: toward usable personal service robots', *Reviews of Human Factors and Ergonomics* **7**(1), 100–148.

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, c., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., Mirza, M., Jean, S., Carrier, P.-L., Dauphin, Y., Boulanger-Lewandowski, N., Aggarwal, A., Zumer, J., Lamblin, P., Raymond, J.-P., Desjardins, G., Pascanu, R., Warde-Farley, D., Torabi, A., Sharma, A., Bengio, E., Côté, M., Konda, K. R. & Wu, Z. (2013), Combining modality specific deep neural networks for emotion recognition in video, *in* 'Proceedings of the 15th ACM on International Conference on Multimodal Interaction', ICMI '13, ACM, New York, NY, USA, pp. 543–550.

Kalpagam Ganesan, R., Rathore, Y. K., Ross, H. M. & Ben Amor, H. (2018), 'Better teaming through visual cues: How projecting imagery in a workspace can improve human-robot collaboration', *IEEE Robotics Automation Magazine* **25**(2), 59–71.

Kanda, T., Hirano, T., Eaton, D. & Ishiguro, H. (2004), 'Interactive robots as social partners and peer tutors for children: A field trial', *Human–Computer Interaction* **19**(1-2), 61–84.

Kar, H. (1986), 'Convolution and matrix systems as con tent addressible distributed brain processes in perception and memory', *Journal of Neurolinguistics* **2**(2).

Karpathy, A. & Li, F. (2014), 'Deep visual-semantic alignments for generating image descriptions', *CoRR* **abs/1412.2306**.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. & Fei-Fei, L. (2014*a*), Large-scale video classification with convolutional neural networks, *in* 'Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition', CVPR '14, IEEE Computer Society, Washington, DC, USA, pp. 1725–1732.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. & Fei-Fei, L. (2014*b*), Large-scale video classification with convolutional neural networks, *in* 'CVPR'.

Kawaguchi, K., Pack Kaelbling, L. & Bengio, Y. (2017), 'Generalization in Deep Learning', *ArXiv e-prints* .

Kewley-Port, D., Burkle, T. Z. & Lee, J. H. (2007), 'Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners', *The Journal of the Acoustical Society of America* .

Khalajzadeh, H., Mansouri, M. & Teshnehlab, M. (2014), Face recognition using convolutional neural network and simple logistic classifier, *in* V. Snášel, P. Krömer, M. Köppen & G. Schaefer, eds, 'Soft Computing in Industrial Applications', Springer International Publishing, Cham, pp. 197–207.

Kim, J., Englebienne, G., Truong, K. P. & Evers, V. (2017), Deep temporal models using identity skip-connections for speech emotion recognition, *in* 'Proceedings of the 25th ACM International Conference on Multimedia', MM '17, ACM, New York, NY, USA, pp. 1006–1013.

Kim, R., Seitz, A., Feenstra, H. & Shams, L. (2009), 'Testing assumptions of statistical learning: Is it long-term and implicit?', *Neurosci. Lett.* **461**(2), 145–149.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *CoRR* **abs/1412.6980**.

Koolagudi, S. G. & Rao, K. S. (2012), 'Emotion recognition from speech: A review', *International Journal of Speech Technology* .

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012*a*), Imagenet classification with deep convolutional neural networks, *in* F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems 25', Curran Associates, Inc., pp. 1097–1105.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012*b*), 'ImageNet Classification with Deep Convolutional Neural Networks', *Advances In Neural Information Processing Systems* pp. 1–9.

Kuhl, P. K. (2007), ‘Is speech learning gatedby the social brain?’, *Developmental science* **10**(1), 110–120.

Ladefoged, P. (2005), *Vowels and consonants*, Blackwell Oxford, UK.

Ladefoged, P. & Broadbent, D. (1957), ‘Information conveyed by vowels’, *Journal of the Acoustical Society of America* .

Lahamy, H. & Lichti, D. D. (2012), ‘Towards real-time and rotation-invariant American sign language alphabet recognition using a range camera’, *Sensors (Switzerland)* **12**(11), 14416–14441.

Laptev, I. & Lindeberg, T. (2003), Space-time interest points, *in* ‘Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2’, ICCV ’03, IEEE Computer Society, Washington, DC, USA, pp. 432–.

Laptev, I., Marszalek, M., Schmid, C. & Rozenfeld, B. (2008), ‘Learning realistic human actions from movies’, *IEEE Conference on Computer Vision and Pattern Recognition* pp. 1–8.

Larochelle, H. & Hinton, G. E. (2010), Learning to combine foveal glimpses with a third-order boltzmann machine, *in* J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel & A. Culotta, eds, ‘Advances in Neural Information Processing Systems 23’, Curran Associates, Inc., pp. 1243–1251.

Latif, S., Rana, R., Younis, S., Qadir, J. & Epps, J. (2018), ‘Cross corpus speech emotion classification- an effective transfer learning technique’, *CoRR* **1801.06353**.

Law, S., Arpit, D., Ballas, N., Verma, V., Che, T. & Bengio, Y. (2017), ‘Residual connections encourage iterative inference’, *CoRR* **abs/1710.04773**.

Le, D. & Provost, E. M. (2013), Emotion recognition from spontaneous speech using hidden markov models with deep belief networks, *in* '2013 IEEE Workshop on Automatic Speech Recognition and Understanding', pp. 216–221.

LeCun, Y. & Bengio, Y. (1998), The handbook of brain theory and neural networks, MIT Press, Cambridge, MA, USA, chapter Convolutional Networks for Images, Speech, and Time Series, pp. 255–258.

LeCun, Y., Bengio, Y. & Hinton, G. E. (2015), 'Deep learning', *Nature* **521**(7553), 436–444.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**(11), 2278–2323.

LeCun, Y., Haffner, P., Bottou, L. & Bengio, Y. (1999), Object recognition with gradient-based learning, *in* 'Shape, Contour and Grouping in Computer Vision', Springer-Verlag, London, UK, UK, pp. 319–.

Lee, J. & Tashev, I. (2015), High-level feature representation using recurrent neural network for speech emotion recognition, ISCA - International Speech Communication Association.

Leslie, A. M. (1994), 'Pretending and believing: issues in the theory of ToMM', *Cognition* **50**(1-3), 211–238.

Li, P., Song, Y., McLoughlin, I. V., Guo, W. & Dai, L. (2018*a*), An attention pooling based representation learning method for speech emotion recognition, *in* 'Interspeech'.

Li, P., Song, Y., McLoughlin, I. V., Guo, W. & Dai, L.-R. (2018*b*), 'An attention pooling based representation learning method for speech emotion recognition'.

Li, S. & Deng, W. (2018), 'Deep facial expression recognition: A survey', *CoRR* **abs/1804.08348**.

Lian, Z., Tao, J., Liu, B. & Huang, J. (2019), Conversational emotion analysis via attention mechanisms, *in* 'INTERSPEECH 2019'.

Lim, W., Jang, D. & Lee, T. (2016), Speech emotion recognition using convolutional and recurrent neural networks, *in* '2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)', pp. 1–4.

Lin, M., Chen, Q. & Yan, S. (2013), 'Network in network', *CoRR* **abs/1312.4400**.

Lindsay, G. W. (2020), 'Attention in psychology, neuroscience, and machine learning', *Frontiers in Computational Neuroscience* **14**, 29.

Linsley, D., Shiebler, D., Eberhardt, S. & Serre, T. (2018), Learning what and where to attend, *in* 'International Conference on Learning Representations'.

Liu, M., Li, S., Shan, S., Wang, R. & Chen, X. (2015), Deeply learning deformable facial action parts model for dynamic expression analysis, *in* D. Cremers, I. Reid, H. Saito & M.-H. Yang, eds, 'Computer Vision – ACCV 2014', Springer International Publishing, Cham, pp. 143–157.

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z. & Chen, X. (2014), Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild, *in* 'Proceedings of the 16th International Conference on Multimodal Interaction', ICMI '14, ACM, New York, NY, USA, pp. 494–501.

Liu, P., Han, S., Meng, Z. & Tong, Y. (2014), Facial expression recognition via a boosted deep belief network, *in* 'Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition', CVPR '14, IEEE Computer Society, Washington, DC, USA, pp. 1805–1812.

Liu, Y., Zhang, Y. M., Zhang, X. Y. & Liu, C. L. (2016), 'Adaptive spatial pooling for image classification', *Pattern Recognition* **55**, 58–67.

Livingstone, S. & Russo, F. (2018), 'The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english', *PLOS ONE* **13**(5), 1–35.

Llorca, D. F., Arroyo, R. & Sotelo, M. A. (2013), Vehicle logo recognition in traffic images using hog features and svm, *in* '16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)', IEEE, pp. 2229–2234.

Lowe, D. G. (2004), 'Distinctive image features from scale-invariant keypoints', *International journal of computer vision* **60**(2), 91–110.

Loweimi, E., Doulaty, M., Barker, J. & Hain, T. (2015), Long-Term statistical feature extraction from speech signal and its application in emotion recognition, *in* 'Lecture Notes in Computer Science', Vol. 9449.

Lucas, B. D., Kanade, T. et al. (1981), 'An iterative image registration technique with an application to stereo vision'.

Luo, D., Zou, Y. & Huang, D. (2018), Investigation on joint representation learning for robust feature extraction in speech emotion recognition, *in* 'Proc. Interspeech'.

Luong, M.-T., Pham, H. & Manning, C. D. (2015), 'Effective approaches to attention-based neural machine translation', *arXiv preprint arXiv:1508.04025* .

Maaten, L. v. d. & Hinton, G. (2008), 'Visualizing data using t-sne', *Journal of machine learning research* **9**(Nov), 2579–2605.

Malinowski, P. (2013), 'Neural mechanisms of attentional control in mindfulness meditation', *Frontiers in Neuroscience* **7**, 8.

Mao, Q., Dong, M., Huang, Z. & Zhan, Y. (2014), 'Learning salient features for speech emotion recognition using convolutional neural networks', *IEEE Transactions on Multimedia* **16**(8), 2203–2213.

Martin, O., Kotsia, I., Macq, B. & Pitas, I. (2006), The enterface'05 audio-visual emotion database, *in* '22nd International Conference on Data Engineering Workshops (ICDEW'06)', IEEE, pp. 8–8.

Maye, J., Weiss, D. J. & Aslin, R. N. (2008), 'Statistical phonetic learning in infants: Facilitation and feature generalization', *Dev. Sci.* **11**(1), 122–134.

Maye, J., Werker, J. F. & Gerken, L. A. (2002), 'Infant sensitivity to distributional information can affect phonetic discrimination', *Cognition* **82**(3).

McCulloch, W. S. & Pitts, W. (1943), 'A logical calculus of the ideas immanent in nervous activity', *The bulletin of mathematical biophysics* **5**(4), 115–133.

Mehdi Y. N., S. A. K. (2002), 'Sign language recognition using sensor gloves', *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.* **5**, 2204–2206 vol.5.

Meltzoff, a. N. & Borton, R. W. (1979), 'Intermodal matching by human neonates.'.

Miller, J. L. (1994), 'On the internal structure of phonetic categories: A progress report', *Cognition* **50**(1-3), 271–285.

Milner, R., Jalal, M. A., Ng, R. W. M. & Hain, T. (2019), A cross-corpus study on speech emotion recognition, *in* '2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)', pp. 304–311.

Mirsamadi, S., Barsoum, E. & Zhang, C. (2017), Automatic speech emotion recognition using recurrent neural networks with local attention, *in* '2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)'.

Mishkin, M., Ungerleider, L. G. & Macko, K. A. (1983), 'Object vision and spatial vision: two cortical pathways', *Trends in Neurosciences* **6**, 414–417.

Mishra, N., Rohaninejad, M., Chen, X. & Abbeel, P. (2017), 'A simple neural attentive meta-learner', *arXiv preprint arXiv:1707.03141* .

Mitchinson, B. & Prescott, T. J. (2016), Miro: a robot mammal with a biomimetic brain-based control system, *in* 'Conference on Biomimetic and Biohybrid Systems', Springer, pp. 179–191.

Mnih, V., Heess, N., Graves, A. et al. (2014), Recurrent models of visual attention, *in* 'Advances in neural information processing systems', pp. 2204–2212.

Mohammad, Y., Okada, S. & Nishida, T. (2010), Autonomous development of gaze control for natural human-robot interaction, *in* 'Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction', EGIHMI 10, Association for Computing Machinery, New York, NY, USA, p. 6370.

Molchanov, P., Gupta, S., Kim, K. & Kautz, J. (2015), Hand gesture recognition with 3D convolutional neural networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition workshops', pp. 1–7.

Mollahosseini, A., Chan, D. & Mahoor, M. H. (2015), 'Going deeper in facial expression recognition using deep neural networks', *CoRR* **abs/1511.04110**.

Mollahosseini, A., Chan, D. & Mahoor, M. H. (2016), Going deeper in facial expression recognition using deep neural networks, *in* '2016 IEEE Winter Conference on Applications of Computer Vision (WACV)', pp. 1–10.

Mollahosseini, A., Hasani, B. & Mahoor, M. H. (2017*a*), 'Affectnet'.

Mollahosseini, A., Hasani, B. & Mahoor, M. H. (2017*b*), 'AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild'.

Moosavi-Dezfooli, S.-M., Fawzi, A. & Frossard, P. (2016), Deepfool: a simple and accurate method to fool deep neural networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2574–2582.

Nair, V. & Hinton, G. E. (2010), 'Rectified Linear Units Improve Restricted Boltzmann Machines', *Proceedings of the 27th International Conference on Machine Learning* (3), 807–814.

Nam, H., Ha, J.-W. & Kim, J. (2017), Dual attention networks for multimodal reasoning and matching, *in* 'Proc. IEEE CVPR'.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998), 'Language discrimination by newborns: toward an understanding of the role of rhythm.', *J. Exp. Psychol. Hum. Percept. Perform.* **24**(3), 756–766.

Nazzi, T., Jusczyk, P. W. & Johnson, E. K. (2000), 'Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity', *J. Mem. Lang.* **43**(1), 1–19.

Needham, A. & Baillargeon, R. (2000), 'Infants' use of featural and experiential information in segregating and individuating objects: A reply to Xu, Carey and Welch (2000)'.

Ng, J. Y., Hausknecht, M. J., Vijayanarasimhan, S., Vinyals, O., Monga, R. & Toderici, G. (2015), 'Beyond short snippets: Deep networks for video classification', *CoRR* **abs/1503.08909**.

Nguyen-Duc-Thanh, N., Lee, S. & Kim, D. (2012), 'Two-stage hidden markov model in gesture recognition for human robot interaction', *International Journal of Advanced Robotic Systems* **9**(2), 39.

Nickel, K. & Stiefelhagen, R. (2007), 'Visual recognition of pointing gestures for humanrobot interaction', *Image and Vision Computing* **25**(12), 1875 – 1884. The age of human computer interaction.

Noudoost, B., Chang, M. H., Steinmetz, N. A. & Moore, T. (2010), 'Top-down control of visual attention', *Current Opinion in Neurobiology* **20**(2), 183 – 190. Cognitive neuroscience.

Nwe, T. L., Foo, S. W. & De Silva, L. C. (2003), 'Speech emotion recognition using hidden markov models', *Speech Communication* **41**(4), 603–623.

Ohn-Bar, E. & Trivedi, M. M. (2014), 'Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations', *IEEE transactions on intelligent transportation systems* **15**(6), 2368–2377.

Olshausen, B. A., Anderson, C. H. & Van Essen, D. C. (1995), 'A multiscale dynamic routing circuit for forming size-and position-invariant object representations', *Journal of Computational Neuroscience* **2**(1), 45–62.

Orrite, C., Gañán, A. & Rogez, G. (2009), Hog-based decision tree for facial expression classification, *in* H. Araujo, A. M. Mendonça, A. J. Pinho & M. I. Torres, eds, 'Pattern Recognition and Image Analysis', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 176–183.

Oster, H., Hegley, D. & Nagel, L. (1992), 'Adult judgments and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas.', *Developmental Psychology* **28**(6), 1115.

Oudeyer, P., Kaplan, F. & Hafner, V. V. (2007), 'Intrinsic motivation systems for autonomous mental development', *IEEE Transactions on Evolutionary Computation* **11**(2), 265–286.

Oudeyer, P.-Y. (2007), 'What is intrinsic motivation? A ty pology of computational approaches', *Front. Neurorobot.* **1**(November), 1–14.

Oudeyer, P.-Y. & Kaplan, F. (2009), 'What is intrinsic motivation? a typology of computational approaches', *Frontiers in Neurorobotics* **1**, 6.

Owren, M. J. & Cardillo, G. C. (2006), 'The relative roles of vowels and consonants in discriminating talker identity versus word meaning', *The Journal of the Acoustical Society of America* .

Palaz, D., Magimai-Doss, M. & Collobert, R. (2015), Analysis of cnn-based speech recognition system using raw speech as input, *in* 'INTERSPEECH'.

Parikh, A. P., Täckström, O., Das, D. & Uszkoreit, J. (2016), 'A decomposable attention model for natural language inference', *arXiv preprint arXiv:1606.01933* .

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017), Automatic differentiation in pytorch, *in* 'NIPS-W'.

Peng, X., Wang, L., Wang, X. & Qiao, Y. (2016), 'Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice', *Computer Vision and Image Understanding* **150**, 109–125.

Piaget, J. (1952), 'When thinking begins', *The origins of intelligence in children* pp. 25–36.

Picard, R. W. (1997), *Affective Computing*, MIT Press, Cambridge, MA, USA.

Picard, R. W. (2000), *Affective computing*, MIT press.

Pigou, L., Dieleman, S., Kindermans, P.-J. & Schrauwen, B. (2015), Sign Language Recognition Using Convolutional Neural Networks, *in* 'Computer Vision - ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I', pp. 572–578.

Pinker, S. (2000), *The Language Instinct: How the Mind Creates Language*, Harper-Perennial modern classics, HarperCollins.

Pinker, S. (2009), *Language Learnability and Language Development, With New Commentary by the Author: With New Commentary by the Author*, Vol. 7, Harvard University Press.

Plutchik, R. (1997), 'Emotion: Theory, research, and experience: Vol. 1. theories of emotion', *New York: Academic* .

Posner, M. (2012), *Cognitive Neuroscience of Attention*, Guilford Publications.

Posner, M. I. (1980), 'Posner - 1980 - Orienting of attention', *Journal of Experimental Psychology* **32**(July 1979), 3–25.

Posner, M. I. & Boies, S. J. (1971), 'Components of attention', *Psychological Review* **78**(5), 391–408.

Posner, M. I. & Petersen, S. E. (1990*a*), 'The attention system of the human brain', *Annual review of neuroscience* **13**(1), 25–42.

Posner, M. I. & Petersen, S. E. (1990*b*), 'The attention system of the human brain'.

Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H. & Zweig, G. (2005), fmpe: Discriminatively trained features for speech recognition, *in* '2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005', pp. 961–964.

Psyllos, A. P., Anagnostopoulos, C.-N. E. & Kayafas, E. (2010), 'Vehicle logo recognition using a sift-based enhanced matching scheme', *IEEE transactions on intelligent transportation systems* **11**(2), 322–328.

Ramus, F., Nespor, M. & Mehler, J. (1999), 'Correlates of linguistic rhythm in the speech signal', *Cognition* **73**(3), 265–292.

Rauschecker, J. P. (1998), 'Cortical processing of complex sounds', *Current Opinion in Neurobiology* **8**(4), 516–521.

Rauschecker, J. P. & Scott, S. K. (2009), 'Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing', *Nature Neuroscience* **12**, 718–724.

Rauschecker, J. P. & Tian, B. (2000), 'Mechanisms and streams for processing of "what" and "where" in auditory cortex', *Proceedings of the National Academy of Sciences* **97**(22), 11800–11806.

Reed, C. L., Klatzky, R. L. & Halgren, E. (2005), 'What vs. where in touch: an fmri study', *NeuroImage* **25**(3), 718–726.

Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000), Speaker verification using adapted gaussian mixture models, *in* 'Digital Signal Processing', p. 2000.

Reynolds, J. H. & Heeger, D. J. (2009), 'The normalization model of attention', *Neuron* **61**(2), 168 – 185.

Robins, B., Dautenhahn, K. & Dickerson, P. (2009), From isolation to communication: A case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot, *in* '2009 Second International Conferences on Advances in Computer-Human Interactions', pp. 205–211.

Romanski, L. (2007), 'Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex', *Cerebral cortex (New York, N.Y. : 1991)* **17 Suppl 1**, i61–9.

Sabour, S., Frosst, N. & Hinton, G. E. (2017), Dynamic routing between capsules, *in* 'Advances in Neural Information Processing Systems', pp. 3859–3869.

Saerbeck, M., Schut, T., Bartneck, C. & Janse, M. D. (2010), Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor, *in* 'Proceedings of the SIGCHI conference on human factors in computing systems', pp. 1613–1622.

Sailor, H. B., Deena, S., Jalal, M. A., Lileikyte, R. & Hain, T. (2019), Unsupervised adaptation of acoustic models for asr using utterance-level embeddings from squeeze and excitation networks, *in* '2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)', pp. 980–987.

Samuels, R. (1998), 'Evolutionary Psychology and the Massive Modularity Hypothesis', *Br. J. Philos. Sci.* **49**(4), 575–602.

Sartre, J. (1962), *Sketch for a Theory of the Emotions*, University paperbacks, Methuen & Company, Limited.

Satt, A., Rozenberg, S. & Hoory, R. (2017), 'Efficient emotion recognition from speech using deep learning on spectrograms', *Proc. Interspeech 2017* pp. 1089–1093.

Schuldt, C., Laptev, I. & Caputo, B. (2004), Recognizing human actions: A local svm approach, *in* 'Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03', ICPR '04, IEEE Computer Society, Washington, DC, USA, pp. 32–36.

Schuller, B., Rigoll, G. & Lang, M. (2003), 'Hidden Markov model-based speech emotion recognition', *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).* **2**, 401–404.

Schuller, B., Steidl, S. & Batliner, A. (2009), The INTERSPEECH 2009 emotion challenge, *in* 'Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH', pp. 312–315.

Schuster, M. & Paliwal, K. K. (1997), 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing* **45**(11), 2673–2681.

Sedda, A. & Scarpina, F. (2012), 'Dorsal and ventral streams across sensory modalities', *Neuroscience Bulletin* **28**(3), 291–300.

Sharif Razavian, A., Azizpour, H., Sullivan, J. & Carlsson, S. (2014), Cnn features off-the-shelf: an astounding baseline for recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition workshops', pp. 806–813.

Sharma, R., Nemani, Y., Kumar, S., Kane, L. & Khanna, P. (2013), 'Recognition of Single Handed Sign Language Gestures using Contour Tracing Descriptor', *World Congress on Engineering* **2**, 754–758.

Shomstein, S. & Yantis, S. (2006), 'Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention', *Journal of Neuroscience* **26**(2), 435–439.

Sidney Fels, S. & Hinton, G. E. (1993), 'Glove-Talk: A Neural Network Interface Between a Data-Glove and a Speech Synthesizer', *IEEE Transactions on Neural Networks* **4**(1), 2–8.

Simonyan, K. & Zisserman, A. (2014*a*), Two-stream convolutional networks for action recognition in videos, *in* 'Advances in neural information processing systems', pp. 568–576.

Simonyan, K. & Zisserman, A. (2014*b*), 'Very deep convolutional networks for large-scale image recognition', *CoRR* **abs/1409.1556**.

Singer, W. (1995), 'Development and plasticity of cortical processing architectures', *Science* **270**(5237), 758–764.

Singh, A. K., John, B. P., Venkata Subramanian, S. R., Sathish Kumar, A. & Nair, B. B. (2017), A low-cost wearable Indian sign language interpretation system, *in* 'International Conference on Robotics and Automation for Humanitarian Applications, RAHA 2016 - Conference Proceedings'.

Singha, J. & Das, K. (2013), 'Indian Sign Language Recognition Using Eigen Value Weighted Euclidean Distance Based Classification Technique', *arXiv preprint arXiv:1303.0634* **4**(2), 188–195.

Smith, L. B. & Thelen, E. (2003), 'Development as a dynamic system', *Trends Cogn. Sci.* **7**(8), 343–348.

Soomro, K., Zamir, A. R. & Shah, M. (2012*a*), 'Ucf101: A dataset of 101 human actions classes from videos in the wild', *arXiv preprint arXiv:1212.0402* .

Soomro, K., Zamir, A. R. & Shah, M. (2012*b*), 'UCF101: A dataset of 101 human actions classes from videos in the wild', *CoRR* **abs/1212.0402**.

Sorce, J. F. & Emde, R. N. (1981), 'Mother's presence is not enough: Effect of emotional availability on infant exploration.', *Developmental Psychology* **17**(6), 737.

Sorce, J. F., Emde, R. N., Campos, J. J. & Klinnert, M. D. (1985), 'Maternal emotional signaling: its effect on the visual cliff behavior of 1-year-olds.', *Developmental psychology* **21**(1), 195.

Spence, C. & Driver, J. (2004), *Crossmodal Space and Crossmodal Attention*, Crossmodal Space and Crossmodal Attention, Oxford University Press.

Spence, C. & Santangelo, V. (2010), 'Auditory attention', *Oxford handbook of auditory science: Hearing* **3**, 249.

Starner, T. & Pentland, A. S. (1996), 'Real-Time American Sign Language Recognition Hidden Markov Models from Video Using', *AAAI Technical Report FS-96-05* pp. 109–116.

Steidl, S. (2009), *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Studien zur Mustererkennung, Isd.

Strang, G. (2019), *Linear Algebra and Learning from Data*, Wellesley-Cambridge Press.

Su, J., Vargas, D. V. & Sakurai, K. (2019), 'One pixel attack for fooling deep neural networks', *IEEE Transactions on Evolutionary Computation* **23**(5), 828–841.

Sundermeyer, M., Schl, R. & Ney, H. (2012), 'LSTM Neural Networks for Language Modeling', *Proc. Interspeech* pp. 194–197.

Susanto, Y., Livingstone, A. G., Ng, B. C. & Cambria, E. (2020), 'The hourglass model revisited', *IEEE Intelligent Systems* **35**.

Sze, V., Chen, Y. H., Yang, T. J. & Emer, J. S. (2017), Efficient Processing of Deep Neural Networks: A Tutorial and Survey, *in* 'Proceedings of the IEEE', Vol. 105, pp. 2295–2329.

Tarantino, L., Garner, P. N. & Lazaridis, A. (2019), Self-attention for speech emotion recognition, *in* 'INTERSPEECH 2019'.

Tariq, M. U., Yang, J. & Huang, T. S. (2013), 'Maximum margin gmm learning for facial expression recognition', *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* pp. 1–6.

Tas, Y. & Koniusz, P. (2018), 'Cnn-based action recognition and supervised domain adaptation on 3D body skeletons via kernel feature maps', *arXiv preprint arXiv:1806.09078* .

Taylor, G. W., Fergus, R., LeCun, Y. & Bregler, C. (2010), Convolutional learning of spatio-temporal features, *in* 'European conference on computer vision', Springer, pp. 140–153.

Thelen, E. & Smith, L. B. (1994), A dynamic systems approach to the development of cognition and action, *in* 'J. Cogn. Neuroscience.', Vol. 512, p. 376.

Tin Lay Nwe, Foo Say Wei & De Silva, L. C. (2001), Speech based emotion classification, *in* 'Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. TENCON 2001 (Cat. No.01CH37239)', Vol. 1, pp. 297–301 vol.1.

Tomasello, M. (2001), ‘First steps toward a usage-based theory of language acquisition’, *Cogn. Linguist.* **11**(1-2), 61–82.

Tomasello, M. & Rakoczy, H. (2003), ‘What Makes Human Cognition Unique? From Individual to Shared to Collective Intentionality’, *Mind Lang.* **18**(2), 121–147.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L. & Paluri, M. (2014), ‘C3D: generic features for video analysis’, *CoRR* **abs/1412.0767**.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B. & Zafeiriou, S. (2016), Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, *in* ‘2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 5200–5204.

Tryon, W. W. (1993), ‘Neural networks: I. theoretical unification through connectionism’, *Clinical Psychology Review* **13**(4), 341–352.

Tsai, Y. H., Hamsici, O. C. & Yang, M. H. (2015), Adaptive region pooling for object detection, *in* ‘Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition’, Vol. 07-12-June-2015, pp. 731–739.

Turkle, S. (2005), *The Second Self, Twentieth Anniversary Edition: Computers and the Human Spirit*, The MIT Press, MIT Press.

Turkle, S., Taggart, W., Kidd, C. D. & Dasté, O. (2006), Relational artifacts with children and elders: the complexities of cybercompanionship, Vol. 18, Taylor & Francis, pp. 347–361.

Varol, G., Laptev, I. & Schmid, C. (2017), ‘Long-term temporal convolutions for action recognition’, *IEEE transactions on pattern analysis and machine intelligence* **40**(6), 1510–1517.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), Attention is all you need, *in* 'Advances in neural information processing systems', pp. 5998–6008.

Veeriah, V., Zhuang, N. & Qi, G.-J. (2015), Differential recurrent neural networks for action recognition, *in* 'The IEEE International Conference on Computer Vision (ICCV)'.

Ververidis, D. & Kotropoulos, C. (2006), 'Emotional speech recognition: Resources, features, and methods', *Speech communication* **48**(9), 1162–1181.

Vrigkas, M., Karavasilis, V., Nikou, C. & Kakadiaris, I. A. (2014), 'Matching mixtures of curves for human action recognition', *Comput. Vis. Image Underst.* **119**, 27–40.

Vutinuntakasame, S., Jaijongrak, V. R. & Thiemjarus, S. (2011), An assistive body sensor network glove for speech- and hearing-impaired disabilities, *in* 'Proceedings - 2011 International Conference on Body Sensor Networks, BSN 2011', pp. 7–12.

Vygotsky, L. (1978), Interaction between learning and development, *in* 'Mind Soc.', pp. 79–91.

Waaramaa, T., Laukkanen, A.-M., Airas, M. & Alku, P. (2010), 'Perception of emotional valences and activity levels from vowel segments of continuous speech', *Journal of Voice* **24**(1), 30–38.

Waibel, a., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K. J. (1989), 'Phoneme recognition using time-delay neural networks', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**(3), 328–339.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. & Tang, X. (2017), 'Residual attention network for image classification', *CoRR* **abs/1704.06904**.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. & Van Gool, L. (2016), Temporal segment networks: Towards good practices for deep action recognition, *in* B. Leibe, J. Matas, N. Sebe & M. Welling, eds, 'Computer Vision – ECCV 2016', Springer International Publishing, Cham, pp. 20–36.

Wang, X., Girshick, R., Gupta, A. & He, K. (2018), Non-local neural networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 7794–7803.

Wilson, H. R. & Cowan, J. D. (1972), 'Excitatory and inhibitory interactions in localized populations of model neurons', *Biophysical Journal* **12**(1), 1–24.

Wolpert, D. H. (1992), 'Stacked generalization', *Neural Networks* **5**, 241–259.

Woo, S., Park, J., Lee, J.-Y. & So Kweon, I. (2018), Cbam: Convolutional block attention module, *in* 'Proceedings of the European conference on computer vision (ECCV)', pp. 3–19.

Worgan, S. & Moore, R. (2010), 'Speech as the perception of affordances', *Ecological Psychology* **22**(4), 327–343.

Wu, H. & Gu, X. (2015), Max-pooling dropout for regularization of convolutional neural networks, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Vol. 9489, pp. 46–54.

Wu, X., Liu, S., Cao, Y., Li, X., Yu, J., Dai, D., Ma, X., Hu, S., Wu, Z., Liu, X. & Meng, H. (2019), Speech emotion recognition using capsule networks, *in* 'ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 6695–6699.

Wynn, K. & Chiang, W.-c. (1998), 'Research Article LIMITS TO INFANTS ' KNOWLEDGE OF OBJECTS : The Case of Magical Appearance', **9**(6), 448–455.

Xi, E., Bing, S. & Jin, Y. (2017), 'Capsule Network Performance on Complex Data', *ArXiv e-prints* .

Xu, B., Wang, N., Chen, T. & Li, M. (2015), 'Empirical Evaluation of Rectified Activations in Convolution Network', *ICML Deep Learning Workshop* pp. 1–5.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. (2015), Show, attend and tell: Neural image caption generation with visual attention, *in* F. Bach & D. Blei, eds, 'Proceedings of the 32nd International Conference on Machine Learning', Vol. 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, pp. 2048–2057.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E. & Morency, L.-P. (2018), Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 2236–2246.

Zang, J., Wang, L., Liu, Z., Zhang, Q., Hua, G. & Zheng, N. (2018), Attention-based temporal weighted convolutional neural network for action recognition, *in* 'Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations', Springer, pp. 97–108.

Zeiler, M. D. & Fergus, R. (2014), Visualizing and understanding convolutional networks, *in* 'European conference on computer vision', Springer, pp. 818–833.

Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A. (2018), 'Self-Attention Generative Adversarial Networks', *arXiv e-prints* p. arXiv:1805.08318.

Zhang, S., Huang, T. & Gao, W. (2016), Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition, *in* 'Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16', pp. 281–284.

Zhang, Z., Lyons, M., Schuster, M. & Akamatsu, S. (1998), Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multilayer perceptron, *in* 'Proceedings of the 3rd. International Conference on Face & Gesture Recognition', FG '98, IEEE Computer Society, Washington, DC, USA, pp. 454–.

Zhao, G. & Pietikainen, M. (2007), 'Dynamic texture recognition using local binary patterns with an application to facial expressions', *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 915–928.

Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N. & Yan, S. (2016), Peak-piloted deep network for facial expression recognition, *in* 'Computer Vision – ECCV 2016', Springer International Publishing, Cham, pp. 425–442.

Zhu, W., Hu, J., Sun, G., Cao, X. & Qiao, Y. (2016), 'A key volume mining deep framework for action recognition', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 1991–1999.

Zhu, Y., Chen, W. & Guo, G. (2014), 'Evaluating spatiotemporal interest point features for depth-based action recognition', *Image Vision Comput.* **32**(8), 453–464.