



UNIVERSITY OF LEEDS

**Automatic, integrated and
structured reporting for radiology
image examinations**

Nurbanu Aksoy

Submitted in accordance with the requirements for the degree
of Doctor of Philosophy in Computer Science

The University of Leeds

Faculty of Engineering

School of Computing

July 2024

Intellectual Property

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

© 2024 The University of Leeds, Nurbanu Aksoy

Signed

A handwritten signature in black ink, appearing to be 'Nurbanu Aksoy', enclosed within a hand-drawn oval.

Abstract

The increasing availability of diverse data sources has expanded the potential for modality translation tasks in artificial intelligence, particularly in converting images into natural language descriptions. In the clinical domain, particularly in chest X-ray (CXR) analysis, advancing Computer-Aided Detection (CAD) and Diagnosis (CADx) technologies hold significant promise for improving patient outcomes and healthcare delivery. This research focuses on improving medical image representations by leveraging clinically relevant information and tasks to establish a more robust pipeline for the automated generation of radiology reports. By aligning with clinical pathways, we aim to generate accurate, contextually relevant reports that reflect real-world medical practices. This study addresses the limitations of current single-modality approaches, which often fail to capture complex relationships and complementary information across different data modalities. The primary objectives of this research are threefold: to develop efficient multi-input pre-processing mechanisms for diverse data types; to establish robust frameworks for modality fusion, combining visual, textual, and clinical data into unified embeddings; and to enhance representation learning capabilities through joint optimisation in multi-task learning. This thesis proposes novel multi-input multi-stream end-to-end networks demonstrating significant improvements in text generation accuracy and contextual relevance. It also includes comprehensive ablation studies, systematic analyses of different architectures, and the introduction of multi-task learning strategies to optimise feature learning and reduce hallucination in generated reports. The findings highlight the potential benefits of multi-modal and multi-task learning in medical applications, suggesting broader implications for other fields requiring integrated multi-modality. While the research faces challenges such as language variability, metric inadequacy, and computational demands, it presents a versatile framework with potential cross-domain applicability and a roadmap for future developments in multi-modal and cross-modal AI systems.

Acknowledgements

I would first like to express my sincere gratitude to my supervisors, Dr. Nishant Ravikumar and Professor Serge Sharoff, and my former supervisor, Professor Alejandro Frangi, for their unwavering support and invaluable guidance throughout my research journey.

I am thankful to my esteemed colleagues, Dr. Cynthia Maldonado Garcia and Dr. Rodrigo Bonazzola, for their thoughtful discussions and constructive critiques, which have not only enriched my research but also provided much-needed encouragement during challenging moments.

In addition, I would like to extend my heartfelt thanks to those who, *though not directly involved in this thesis*, have nonetheless made significant contributions to my personal and academic development. To my family—my father, Ömer Demir; my mother, Zekiye Demir; and my siblings, Fatma Nazli Demir and Ismail Demir—your constant support, belief in me, and inspiration have been a source of strength throughout my studies and life. I could not have achieved this without your love and encouragement.

I am also grateful to my dear friends, Sena Hilal Alev and Aybike Zeynep Genç, for their continued support and the joy they brought into my life, especially during the more challenging times. Your presence has been a true source of comfort.

My deepest gratitude goes to my beloved husband, Fatih Aksoy, and our precious daughter, Ece Aksoy. Their steadfast love, understanding, and support have been the foundation upon which I have built both my academic and personal life.

Finally, I would like to acknowledge the generous support of the Turkish Ministry of Education, whose funding made this research possible.

Contents

1	INTRODUCTION	1
1.1	Background	1
1.2	Rationale	3
1.3	Aims and objectives	4
1.4	Outline and Contributions	5
2	THEORETICAL BACKGROUND	7
2.1	Deep Learning Concepts	7
2.1.1	Fundamentals of Neural Networks and Learning Process	7
2.1.2	Advanced Neural Network Strategies	10
2.1.3	Principles of Deep Encoder-Decoder for Image-to-Text Generation	18
2.1.4	Concepts of Image Classification	19
2.1.5	Fundamentals of Deep Multi-Task Learning	21
2.2	Understanding X-ray reports and Medical Data Modalities	24
2.2.1	Chest X-ray Reporting	24
2.2.2	Medical Data Modalities	26
2.3	Datasets	28
2.3.1	Indiana University Chest X-ray Collection	28
2.3.2	Medical Information Mart for Intensive Care Database	29
2.4	Evaluation Metrics	30
2.4.1	Natural Language Generation Metrics	30
2.4.2	Human Analysis	33
2.4.3	Classification Metrics	33

3	LITERATURE REVIEW	36
3.1	Foundations of Natural Image Captioning	36
3.2	Advancements in Image Captioning: Attention and Semantic Features	37
3.3	Transformer Architecture	38
3.4	Automatic Radiology Reporting Baselines	40
3.5	Multi-modal and Multi-task Learning in Medical Imaging	41
4	MEDICAL ONTOLOGY-INFORMED RECURRENCE NETWORK	45
4.1	Introduction	45
4.2	Data Preliminary Processing	46
4.2.1	Data Formation	47
4.2.2	Data Pre-processing	49
4.3	Network Configuration	50
4.3.1	Encoder and Decoder Variants	50
4.3.2	Multi-view vs. Single View	51
4.4	Medical Ontology-Informed Multi-View Model	52
4.5	Experimental Setup	54
4.5.1	Training and Inference	54
4.5.2	Evaluation Details	55
4.6	Component-wise Performance Analysis	56
4.7	Qualitative and Quantitative Results	57
4.8	Discussion and Conclusion	59
5	MULTI-MODAL INTEGRATIVE ATTENTION NETWORK	61
5.1	Introduction	61
5.2	Dataset Formation	62
5.3	Feature Extraction and Pre-processing	63
5.3.1	Image Data	63
5.3.2	Clinical: Non-imaging Data	63
5.3.3	Non-clinical Data	64
5.4	Multi-modal Fusion Details	65
5.5	Experiments and Analysis	70
5.5.1	Experimental Setup	70

5.5.2	Evaluation Metrics	71
5.6	Quantitative Results	71
5.7	Qualitative Results	73
5.8	Radiologist Evaluation Results	76
5.9	Discussion and Conclusion	76
6	CROSS-TASK LEARNING FRAMEWORK	78
6.1	Introduction	78
6.2	Problem Formulation	79
6.3	Experiments and Analysis	81
6.3.1	Data and Feature Processing	81
6.3.2	Evaluation Metrics	83
6.3.3	Experimental Setup	83
6.4	Architectural Designs and Learning Strategies	84
6.4.1	Cross-modal MTL Approach	84
6.4.2	Results	86
6.4.3	Balanced Attentive MTL Approach	88
6.4.4	Results	90
6.5	Discussion and Conclusion	95
7	CONCLUSIONS AND FUTURE OUTLOOK	98
7.1	Overview	98
7.2	Significance and Implications	99
7.3	Nuances in Multi-Modal Learning	100
7.4	Limitations and Challenges	102
7.5	Future Directions	103
7.6	Final Remarks	107
	References	108
A	Integrative Attention Network: Natural Image Application	119
A.1	Introduction	119
A.1.1	Natural Image Datasets	119
A.1.2	Data Pre-processing and Dataset Formation	120

A.1.3	Network Design	121
A.1.4	Results and Discussion	124
B	Comparative Results of CM-MTL Model in Classification	127

List of Figures

2.1	Single processing unit	7
2.2	Backpropagation of errors through the network.	9
2.3	Convolutional Neural Network with one hidden layer	11
2.4	An unrolled Recurrent Neural Network with input units (x_0, x_1, \dots, x_t) , hidden units (h_0, h_1, \dots, h_t) , and a recurrent unit (A)	12
2.5	An illustration of LSTM and GRU units (Phi 2018), including the cell state, hidden state, input gate, forget gate, input gate, output gate, and associated operations.	12
2.6	Overview of the full Transformer architecture	14
2.7	The figure illustrates fusion strategies using deep learning. The model architecture varies for each strategy: early fusion (left), joint fusion (middle) and late fusion (right)(Huang et al. 2020)	17
2.8	A Deep Neural Encoder-Decoder Framework for Generating Text Descriptions from Image Features	18
2.9	An Overview Illustration of Multi-Task Framework	21
2.10	Examples of Chest X-Ray Images with Corresponding Radiologist-Generated Findings (Sirshar et al. 2022)	24
2.11	Random chest radiographs of 25 different patients from the IU CXR collection.	28
2.12	An example frontal and lateral chest X-ray imaging and corresponding report from the IU chest X-ray collection.	29
2.13	A sample X-ray report from the MIMIC-CXR database, comprising images captured from both the frontal and lateral perspectives of the patient.	30

4.1	Example of Raw XML Data Structure from the Indiana University Chest X-ray Collection	47
4.2	Example of the final processed text data(with start/end tokens) and input images for model input	50
4.3	The overall multi-view multi-input recurrence report generation framework . . .	52
4.4	Illustration of Ground Truths and Example Reports Generated by MeSH-Enriched Multi-view Model	58
5.1	The overall multi-modal data fusion with the cross-attention framework of the proposed CXR report generation model.	67
5.2	GT and GR report from the proposed FullFusion CXR report generation model.	75
6.1	The overall architecture of Cross-modal Multi-Task Learning (CM-MTL) Model .	85
6.2	Illustrative comparison of generated reports from different learning approaches with ground truth	88
6.3	The overall framework of our proposed Balanced Attentive multi-task learning network	90
6.4	Illustrative comparison of generated reports from different learning approaches with ground truth	94
A.1	An example image with different captions from Flickr8K dataset	120
A.2	An overall framework of using additional context	122
A.3	The captions generated from our semantically-enhanced captioning model	126

List of Tables

4.1	Most Frequently Occurring Mesh Terms in the Final Dataset	48
4.2	Top 5 Reports (Impression and Findings Compilation) with Highest Occurrence in the Final Dataset	49
4.3	The average BLEU scores obtained from the five-fold cross-validation results across different input views.	56
4.4	Comparative Performance Analysis of Baseline and Pretrained Models Using BLEU-n Scores, with 'B_n' Representing BLEU-n Scores	57
4.5	Qualitative Results of Baseline and Mesh-enriched Model Models	57
5.1	List of Variables Used in Multi-Modal Fusion Strategy	65
5.2	Quantitative Comparison of Fusion Methods: Performance Evaluation Across Multiple Metrics. B_n for BLEU-n, R_L for ROUGE-L, BS _{F1} for BERT Score F1Score and CBS _{F1} for Bio-ClinicalBERT Score F1Score.	72
5.3	Performance Comparison of Singular Data Models	73
5.4	Comparison between our Full Fusion Model and state-of-the-art methods on the MIMIC-CXR dataset, referencing results from their published literature.	73
5.5	Radiologist Evaluation Results on a 1-5 Scale	76
6.1	Performance comparison of a report generation using different training approaches. B_n for BLEU-n, R_L for ROUGE-L, BS _{F1} for BERT Score F1Score and CBS _{F1} for Bio-ClinicalBERT Score F1Score. STL denotes Single Task Learning, CM- MTL-TP represents Multi-Task Learning with Task Prioritisation for text gener- ation and CM-MTL-EQ indicates Multi-Task Learning with equal task weights for each task	86

6.2	P-values from pairwise t-test between STL and CM-MTL-TP approaches (rounded to 5 decimal places)	87
6.3	Performance comparison of a single task report generation (STL) and Balanced-Attentive Multi-task Learning (BA-MTL).	91
6.4	P-values from pairwise t-test between STL and BA-MTL approaches	91
6.5	Comparison with state-of-the-art radiology report generation methods on MIMIC-CXR. The best results are highlighted in bold.	91
6.6	Comparing performance of the ordinal classifier in Single-Task and Balanced Attentive Multi-Task Learning	92
6.7	Comparing performance of the multi-label classifier in Single-Task and Balanced Attentive Multi-Task Learning	93
A.1	A sample of pre-processed text data for LDA model	120
A.2	Performance Summary of EfficientNet, InceptionV3, and VGG-19, with 'B_n' for BLEU-n Scores and denoting BERTScore as BS with P for Precision, R for Recall, and F1 for F1 Score	125
A.3	Base Captioning Model Performance Across 5 Sets of Experiments	125
A.4	Semantically Enhanced Captioning Model Performance Across 5 Sets of Experiments	125
A.5	Statistical Significance (p-values) Comparing Evaluation Metrics for Captioning Models	125
A.6	Image Captioning Performance Analysis: BLEU and BERTscore on Flickr8K and MSCOCO datasets	126
B.1	Comparing performance of the ordinal classifier in Single-Task, Task-Prioritised Multi-Task and Equal-Weight Multi-Task Learning	127
B.2	Comparing performance of the multi-label classifier in Single-Task, Task-Prioritised Multi-Task and Equal-Weight Multi-Task Learning	127

List of Publications

Aksoy, N., Ravikumar, N. and Frangi, A.F., 2023, April. Radiology report generation using transformers conditioned with non-imaging data. In *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications* (Vol. 12469, pp. 146-154). SPIE.

Aksoy, N., Ravikumar, N. and Frangi, A.F., 2023, July. Poster presentation of Radiology report generation using transformers conditioned with non-imaging data. At *IndabaX Morocco 2023*, from 29-30 Julliet 2023, part of Deep Learning Indaba

Aksoy, N., Sharoff, S., Baser, S., Ravikumar, N. and Frangi, A.F., 2024. Beyond images: an integrative multi-modal approach to chest x-ray report generation. *Frontiers in Radiology*, 4, p.1339612.

Aksoy, N., Sharoff, S. and Ravikumar, N., 2024 , May. Enhancing Image-to-Text Generation in Radiology Reports through Cross-modal Multi-Task Learning. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*

Chapter 1

INTRODUCTION

1.1 Background

Advancements in technology have enabled the accumulation and storage of information in different formats, thereby expanding the foundational opportunities for the utilisation of the data (Gao et al. 2020; Safitra et al. 2024). With the increasing availability of diverse data sources, modality translation tasks in the field of artificial intelligence and deep learning have gained significant attention. These tasks aim to convert information from one representation format to another while bridging the gap between diverse data modalities, such as translating images into natural language descriptions or generating audio from textual inputs. Image-to-text modality translation task, in particular, integrates two important aspects of artificial intelligence, computer vision and natural language processing, therefore, it requires a deeper understanding of the image-text relationship in order to transform one form into another (Stefanini et al. 2022).

These tasks typically involve converting both modalities into compatible representations before framing the problem as a sequence-to-sequence task, treating modality translation as a single-modality process. However, this approach fails to capture the complex relationships and complementary information that exist across different data modalities. This oversight can lead to incomplete or inaccurate translations due to the under-utilisation of crucial contextual information. Modality translation necessitates the development of techniques that can effectively fuse heterogeneous data sources, leveraging their collaborative interactions to produce coherent and accurate outputs.

The significance of effectively managing this multi-modality is particularly evident within the clinical domain. Computer-Aided Detection (CAD) and Diagnosis (CADx) systems are increasingly recognised for their potential to enhance medical imaging analysis (Chan et al. 2020). The use of medical imaging is widespread across various branches of health sciences, serving multiple crucial purposes: diagnosing diseases, developing effective treatment plans, providing patient care, and predicting disease outcomes.

Radiologists, as key players in this field, are responsible for interpreting these medical images and creating comprehensive, full-text radiology reports based on their findings, integrating other relevant clinical data and information. This process, while essential, can be time-consuming and prone to human error, especially when dealing with large volumes of images.

Focusing on chest X-rays (CXR), in particular, are the most commonly used medical imaging techniques due to their accessibility, cost-effectiveness, and ability to provide valuable initial insights into a wide range of lung diseases (Rajpurkar et al. 2017). As such, CXRs are usually the first step in evaluating patients for various pulmonary conditions, making them a critical focus for improving diagnostic accuracy and efficiency. Therefore, advancing CAD and CADx technologies for CXR interpretation holds significant promise for improving patient outcomes and optimising healthcare delivery.

The reports generated from CXR examinations typically include the radiologists' observations and indicate normal and abnormal features in the images. By providing radiologists with a baseline analysis to validate and amend as needed, automation can reduce repetitive workflows. This would allow radiologists to focus their expertise on higher-level clinical thinking and improve efficiency in the clinical pathway. Moreover, these medical reports serve as crucial documentation in both computational and clinical domains. They provide comprehensive insights into patients' conditions, diagnostic findings, and treatment plans (Montagnon et al. 2020). The detailed information in these reports offers many opportunities for various tasks, connecting computational advances with real-world medical practice.

1.2 Rationale

The automated generation of accurate medical reports is often challenging due to the disparate nature of medical data (Jing et al. 2017). This task is cross-modal and multi-modal in nature, requiring the effective integration and processing of information from multiple sources. Therefore, generating a narrative text from radiology images presents unique challenges that highlight the need for a comprehensive multi-modal representation learning approach. (Moon et al. 2022).

Exclusively training models on medical images and their imaging variations pose significant limitations, as X-ray imaging inherently captures two-dimensional projections of three-dimensional anatomical structures. Consequently, preserving the entirety of relevant contextual information within these 2D representations becomes challenging, as vital spatial and contextual details may be lost during the feature extraction process. Furthermore, in certain instances, while similarities in medical imaging between male and female patients may be nearly identical in terms of visual patterns, differences in patient demographics can have a noteworthy clinical impact on the assessment and diagnosis (Sandstede et al. 2000).

Existing literature in automated radiology report generation often neglects to address this multi-modal complexity. Many approaches treat the task as a natural modality translation problem, focusing solely on translating medical images into textual reports or relying on limited supplementary data sources. However, medical report generation is a more complex task than natural modality translation tasks, like image captioning, as (1) most chest X-rays may appear similar at first glance, but subtle differences can lead to divergent generated reports and (2) generating coherent and structured paragraphs from medical images requires capturing higher-level semantics and context beyond short image captions. Therefore, such approaches cannot capture the contextual information that radiologists and clinicians leverage during the report generation process.

In light of these considerations, the following section outlines the aims and objectives of this research, focusing on developing techniques to overcome these limitations and improve the quality of automated radiology report generation.

1.3 Aims and objectives

The general purpose of this research is to improve image representations by leveraging clinically relevant information and tasks to establish a more robust pipeline for the automated generation of radiology reports. By aligning with clinical pathways, we aim to generate accurate, contextually relevant reports that reflect real-world medical practices. Specifically, the objectives of this research are threefold. First, we aim to develop efficient multi-input pre-processing mechanisms, focusing on creating techniques to effectively ingest and initially process multiple inputs. This involves designing pre-processing pipelines and input-specific encoders that prepare diverse data types for further processing in a downstream task. Second, we aim to establish a robust framework for modality fusion, emphasising the design of fusion methods that combine the pre-processed inputs from various modalities, including patient demographics and clinical information. The goal is to produce a unified embedding that optimally represents the integrated features from all data sources, enhancing the joint processing of diverse data for comprehensive report generation. Third, we intend to enhance representation learning capabilities by utilising joint optimisation techniques in multi-task learning to improve the model's ability to capture complex relationships between different modalities, leading to better performance in generating detailed and accurate radiology reports.

These objectives are guided by the two central research questions:

1. How can robust frameworks be established to enhance the utilisation of multiple modalities, and what methods can be employed to ensure seamless integration and joint processing of diverse data?
2. How can the representation learning capabilities of neural network models be improved through joint optimisation for relevant tasks, and how does this impact the generalisation and performance of radiology report generation?

In this context, this thesis examines the various methodological approaches I have undertaken along with the technological advancements in artificial intelligence and deep learning in the field of radiology report development during this period. Additionally, it explores how these developments have contributed to improving the accuracy, efficiency, and overall quality of radiology reports, thereby enhancing clinical decision-making processes.

1.4 Outline and Contributions

This section outlines the organisation of the chapters and summarises their key contributions. The theoretical background chapter aims to establish the theoretical and contextual groundwork for this research. It covers fundamental deep learning concepts like neural networks, encoder-decoder models, and multi-task learning strategies. It delves into the specifics of medical imaging and reporting, particularly chest X-ray reports and various medical data modalities. This chapter also introduces the datasets and the evaluation metrics utilised in the experiments. Its objective is to provide a comprehensive background with the necessary information to ensure that the research problem, proposed methodologies, and evaluation approaches can be effectively conveyed.

Chapter 3 presents key methodologies and advancements in natural image captioning, automatic radiology reporting, and the application of multi-modal and multi-task learning in medical imaging. It begins by discussing the standard encoder-decoder architecture used in image captioning, highlighting the integration of computer vision and natural language processing techniques, including attention mechanisms. It then explores the evolution of automatic radiology reporting, from CNN-RNN frameworks to transformer-based architectures, emphasising advancements in semantic feature integration. It critically analyses recent studies that leverage multi-modal data and multi-task learning techniques, highlighting their potential to enhance performance in radiology report generation tasks. Throughout this chapter, particular attention is given to how these existing approaches inform and contrast with the novel multi-modal report generation methods developed in this thesis.

Chapter 4 proposes attention-based vision encoding and recurrent decoding that use structured labels to improve the semantic alignment between visual and textual data while emphasising the need for specialised solutions in the clinical domain. Key contributions include:

1. The development of a novel multi-input, multi-view end-to-end network enriched with medical ontology terms to bridge the semantic gap between visual and textual representations.
2. An extensive ablation study uncovers the impact of various model configurations, offering insights for more accurate clinical text generation.

Chapter 5 presents a novel multi-modal, data-driven learning framework that integrates complementary information from diverse data modalities to generate comprehensive and input-specific radiology reports. It also addresses the limitations observed in the ontology-augmented approaches by introducing an integrative fusing approach. Key innovations include:

1. A multi-stream encoding pipeline that processes visual, textual, and scalar clinical data through tailored featurisation modules to derive modality-specific embeddings.
2. A cross-modal attention fusion mechanism that effectively combines the heterogeneous modality embeddings into a unified, capturing complementary perspectives. This general multi-modal architecture is extensible beyond radiology to jointly encode and generate from arbitrary combinations of data modalities.
3. Systematic ablation analysis quantifying the performance contributions of individual modalities and modality groups, demonstrating synergistic effects of multi-modal fusion.

Chapter 6 proposes a novel unified framework for report generation, ordinal and multi-label classification tasks by leveraging multi-modal data and multi-task learning (MTL) strategies. The proposed approach aims to improve representation learning capabilities through concurrent training of the relevant tasks while evaluating the impact of different training strategies. Key innovations include:

1. Two architectural designs, *Cross-Modal Multi-task Learning (CM-MTL)* and *Balanced-Attentive Multi-task Learning (BA-MTL)*, trained using single and multi-task learning strategies. CM-MTL prioritises the report generation task, while BA-MTL employs a balanced approach to weight each task to improve the feature learning for all tasks.
2. Demonstration of how multi-task learning enhances image representation for report generation by jointly optimising related tasks and assessing the impact of diverse training and weighting configurations on generated reports.

Lastly, Chapter 7 presents a comprehensive overview of the thesis, deriving a set of conclusions from the research findings. It also examines the inherent limitations of the study and discusses future directions.

Chapter 2

THEORETICAL BACKGROUND

This chapter aims to establish a solid foundation for the research by providing relevant theoretical concepts, contextual information, and necessary background details. It functions as an extensive introductory resource, providing the essential information needed to fully understand the research problem, related works, the suggested techniques and methods, and the criteria used for assessments - all of which will be explored in detail in the following chapters.

2.1 Deep Learning Concepts

2.1.1 Fundamentals of Neural Networks and Learning Process

Neural networks are a class of machine learning models that are inspired by the structure of the brain's neurons. They have been widely adopted and have achieved great success in various domains, including computer vision, natural language processing, and medical imaging analysis.

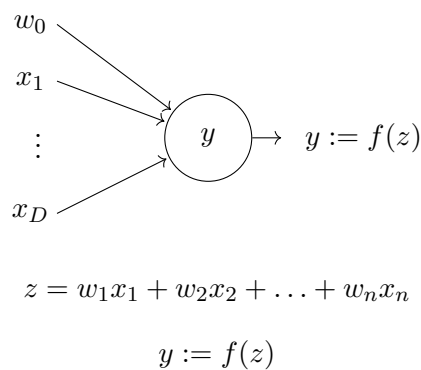


Figure 2.1: Single processing unit

In this model, the basic building block is called a node. A neural network is composed of layers, including an input layer, one or hidden layers and an output layer. These layers consist of interconnected nodes. Each node, in the network, receives inputs (x_1, x_2, \dots, x_n) , from the preceding layer and assigns them weights (w_1, w_2, \dots, w_n) . It then calculates the sum and applies an activation function to produce an output (y) .

The activation functions are used to introduce non-linearity into the network, allowing it to model complex relationships. Sigmoid Function, Softmax Function and Rectified Linear Unit are essential components in neural networks, each serving distinct purposes based on the nature of the tasks they are applied to.

Sigmoid Function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2.1)$$

The sigmoid activation function is particularly advantageous in classification tasks due to its ability to map input values to an output range between 0 and 1. This characteristic makes it suitable for tasks where probabilities or binary classifications are needed. By compressing the output to a bounded range, the sigmoid function ensures that each neuron's output is interpretable as a probability, which is essential in scenarios requiring clear decision boundaries.

Softmax Function:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad \text{for } i = 1, 2, \dots, K \quad (2.2)$$

The softmax activation function is used in tasks that require the prediction of multiple mutually exclusive classes, such as in attention mechanisms and sequence decoders. Softmax transforms the input values into a probability distribution over multiple classes, ensuring that the sum of the output probabilities equals one. This feature is particularly useful for generating sequences of words or selecting the most probable class in a classification task. By calculating probability scores, the softmax function allows the network to make decisions based on the highest probability, facilitating accurate and coherent predictions in complex tasks such as natural language processing.

Rectified Linear Unit:

$$\text{ReLU}(z) = \max(0, z) \quad (2.3)$$

The ReLU (Rectified Linear Unit) activation function is widely used due to its simplicity and effectiveness in mitigating the vanishing gradient problem. ReLU outputs zero for negative input values and the input value itself for positive inputs, which introduces sparsity and helps in efficient learning. By allowing gradients to flow more directly through the network, ReLU facilitates faster convergence during training. It is particularly useful in the hidden layers of neural networks, where it enables the network to model complex data patterns effectively. Applying ReLU after normalisation layers ensures that the activations are well-scaled, further enhancing the training process and overall network performance.

In summary, the sigmoid function is utilised for its probabilistic interpretation in binary and multi-label classifications, the softmax function for generating probability distributions over multiple classes, and the ReLU function for its simplicity and efficiency in promoting faster and more effective learning in neural networks.

The network is then trained on many samples to learn patterns. The learning objective is defined using a loss function that quantifies the difference between actual and expected outputs. By minimising this loss value across training iterations, the network parameters are updated to make progressively better predictions. The most common method used to minimise the loss is back-propagation combined with gradient-based optimisation techniques like gradient descent. This is used to adjust the weights between nodes by calculating the error between the predicted and expected outputs. This error value is then propagated backwards to determine gradients, adjust the weights and update parameters in the direction that reduces loss.

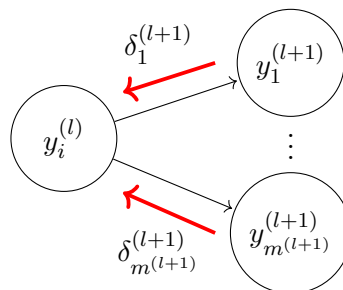


Figure 2.2: The flow of information and the error update process during the training of a neural network

In Figure 2.2, the notation $y_i^{(l)}$ represents the output of a neuron in layer l (the current layer), where the subscript i denotes the specific neuron. Similarly, $y_1^{(l+1)}, \dots, y_{m^{(l+1)}}^{(l+1)}$ denote the outputs of neurons in layer $l + 1$ (the next layer), with subscripts indicating individual neurons. The terms $\delta_1^{(l+1)}, \dots, \delta_{m^{(l+1)}}^{(l+1)}$ represent the error terms associated with neurons in layer $l + 1$ during backpropagation. The arrows illustrate connections between neurons, with red arrows indicating the backpropagation of errors from layer $l + 1$ to layer l .

In addition to loss functions and weight optimisation, neural networks have various hyperparameters that can be tuned to improve performance, such as the learning rate for gradient descent, number of hidden layers, number of nodes per layer, type of activation functions between nodes, and regularization parameters. Finding the right combination of hyperparameters requires experimenting with different values, tracking evaluation metrics, and tweaking the parameters accordingly.

2.1.2 Advanced Neural Network Strategies

Convolutional Feature Extraction

Convolutional Neural Network (ConvNet or CNN) is a type of artificial neural network that was introduced by Yann LeCun in 1998 for a handwritten document recognition task (LeCun et al. 1998). It was originally designed to process two-dimensional image data; however, its success has also been demonstrated for image encoding of three-dimensional data. CNNs have seen widespread application in machine vision tasks such as image classification, recognition, processing, and captioning.

A traditional CNN architecture consists of two major parts: feature extraction and connection. In a simple ConvNet, the feature extraction part has a convolutional layer and a pooling layer (Figure 2.3). The convolutional layer performs convolution operations between a set of independent kernels and an array of input by sliding kernels over the image. This operation ensures that the spatial relationship between pixels is maintained. The output of this layer is called the “Feature Map”, “Activation Map” or “Convolved Feature”. The pooling layer subsamples (or downsamples) the activation maps in order to reduce their dimensions while preserving important information. The pooling layer also helps to reduce computational complexity and control any overfitting of the network. Different types of pooling (Max, Average and Sum) can be selected based on model design. In the connection, a fully connected layer, which is basi-

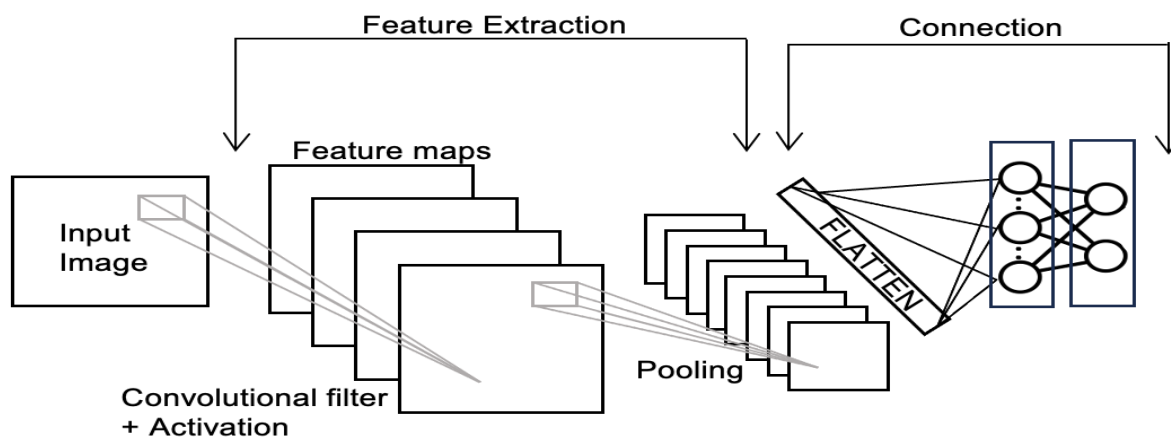


Figure 2.3: Convolutional Neural Network with one hidden layer

cally a Multi-Layer Perceptron with an activation function, uses extracted high-level features to conduct a given task. The output can be the desired result and/or an input for a subsequent network.

In general, most CNN architectures use several feature extraction parts followed by fully connected layers to generate output. The first layers of the network represent the low-level features of an image that can be easily generalised for similar tasks. On the other hand, the final layers include high-level features that are more particular to the application of the model. Parameters such as the number of convolutional and pooling layers, the number of neurons, kernel shape and number, and the type of activation function are fine-tuned by the researcher according to the problem definition and network design.

Recurrent Sequence Generation

To effectively process sequential and contextual data, it is important to keep previously learned information and consider it when generating the following output because the order in which the information is presented is as significant as the meaning it conveys. On the other hand, in a feed-forward neural network, data only flows in a single direction, from input to output, therefore it is only capable of processing current data and does not have a “memory”, as such, to store such information.

The Recurrent Neural Network (RNN) is an effective algorithm for addressing this problem since it has internal memory. The loop in RNN models ensures that the output of the preceding timeframe is fed into the subsequent one. In this context, existing information can be transferred and used in subsequent steps. Although internal memory allows the network to connect and

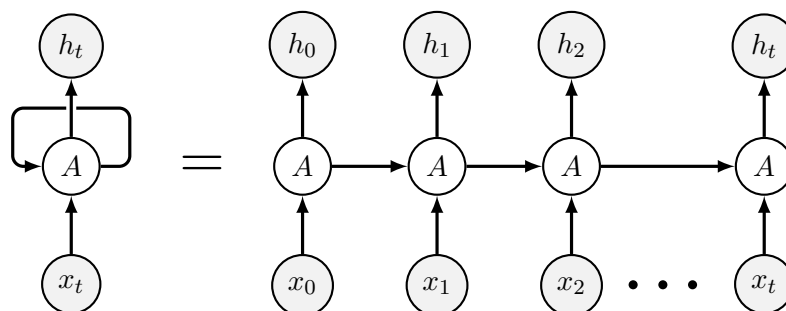


Figure 2.4: An unrolled Recurrent Neural Network with input units (x_0, x_1, \dots, x_t) , hidden units (h_0, h_1, \dots, h_t) , and a recurrent unit (A) .

utilise past information, vanilla RNN still suffers from the vanishing and exploding gradient problem and short-term memory dependencies. Two types of RNN, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have been developed to overcome these drawbacks. In contrast to traditional RNNs, LSTM units have a cell state (aka “second state vector” or

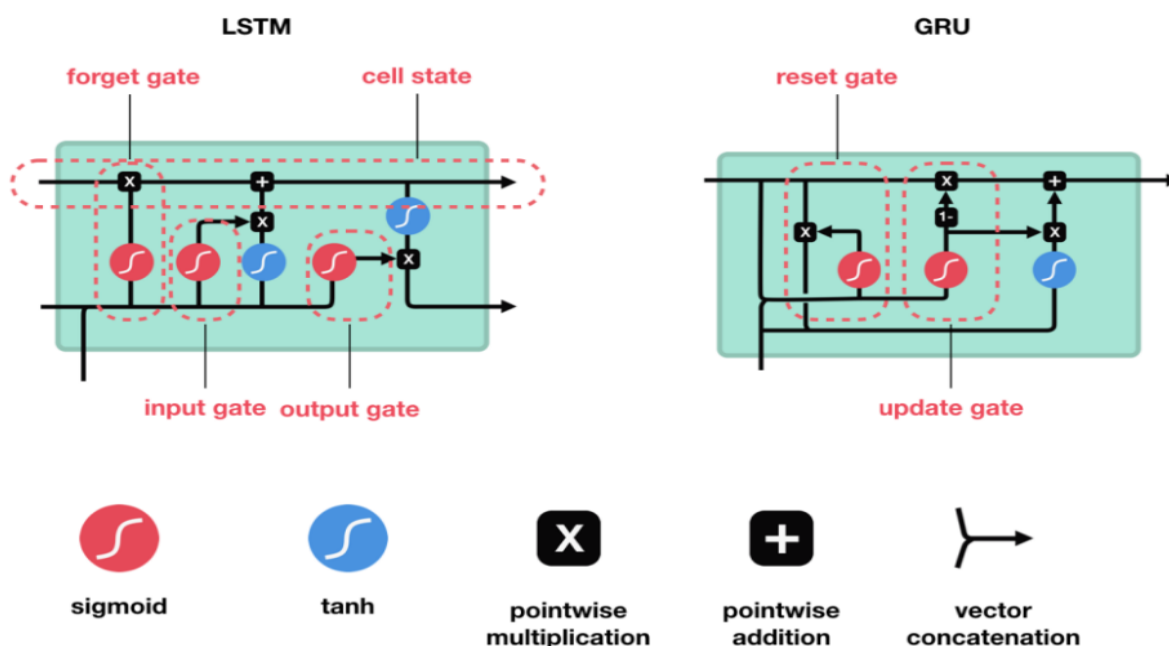


Figure 2.5: An illustration of LSTM and GRU units (Phi 2018), including the cell state, hidden state, input gate, forget gate, input gate, output gate, and associated operations.

“long-term memory—” or “memory cell”) that allows the network to retain information in memory for a long time. Furthermore, the cell state is regulated by Input, Output, and Forget gates, which determine what information should be passed through the network, updated, or forgotten. GRU has also a similar structure to LSTM but only has Reset and Update gates to control information flow (Figure 2.5).

Bidirectional Recurrent Neural Networks (RNNs) enhance traditional RNN architectures by processing data in both forward and backward directions. This bidirectional approach allows the network to capture dependencies and context from both past and future states simultaneously, which is particularly advantageous in tasks requiring a comprehensive understanding of sequential data. In contrast to traditional RNNs, which process data sequentially from one direction, bidirectional RNNs, such as bidirectional Gated Recurrent Units (biGRUs), mitigate the limitations of unidirectional models by leveraging information from both ends of the sequence.

In the domain of multi-modal learning, where accurate integration of data from multiple sources, biGRUs offer distinct advantages. By leveraging bidirectional processing, biGRUs can effectively align textual descriptions with visual features, ensuring that the generated reports are not only accurate but also contextually relevant. Moreover, biGRUs address computational efficiency concerns compared to their LSTM counterparts, making them well-suited for real-time applications in medical image analysis and report generation. Although the vanishing and exploding gradient problem has been addressed by this approach to some extent, long-term dependency continues to be a challenge.

Attention Mechanism

Attention is an effective and prominent strategy designed to improve the performance of the deep neural network. It was originally proposed for another sequence-to-sequence task, neural machine translation, however, it has been successfully fine-tuned in various image processing and natural language processing studies. The main principle behind the algorithm is to dynamically allocate different weights to different parts of the input sequence during each decoding step. In other words, it ensures that the prediction of each time step is based on the associated part rather than the entire input.

The attention mechanism introduces attention weights a_{ij} for each pair of input and output positions (i, j) . These weights are computed using a scoring function e_{ij} that measures the relevance of the input at position i to the output at position j . The attention weight a_{ij} is obtained by applying a softmax function to the scores:

$$a_{ij} = \frac{e^{e_{ij}}}{\sum_{k=1}^T e^{e_{ik}}}$$

The context vector c_j for each output position j is then computed as the weighted sum of the encoder hidden states:

$$c_j = \sum_{i=1}^T a_{ij} h_i$$

where h_i represents the hidden state of the encoder at position i .

The attention context vector c_j is used in conjunction with the decoder hidden state to generate the output y_j :

$$y_j = \text{Decoder}(c_j, s_j)$$

where s_j is the decoder hidden state at position j .

The attention mechanism allows the model to focus on relevant parts of the input sequence when generating each element of the output sequence, improving the model's ability to capture long-range dependencies and handle variable-length sequences effectively.

Transformer Architecture

The Transformer is a neural network architecture for sequence-to-sequence tasks, intended to address long-term dependencies while achieving fast training. It uses an encoder-decoder structure with stacked self-attention and feed-forward layers. Let the input sequence have n tokens, represented by vector embeddings $x_1, \dots, x_n \in \mathbb{R}^d$, where $d = d_{model}$ for the base model.

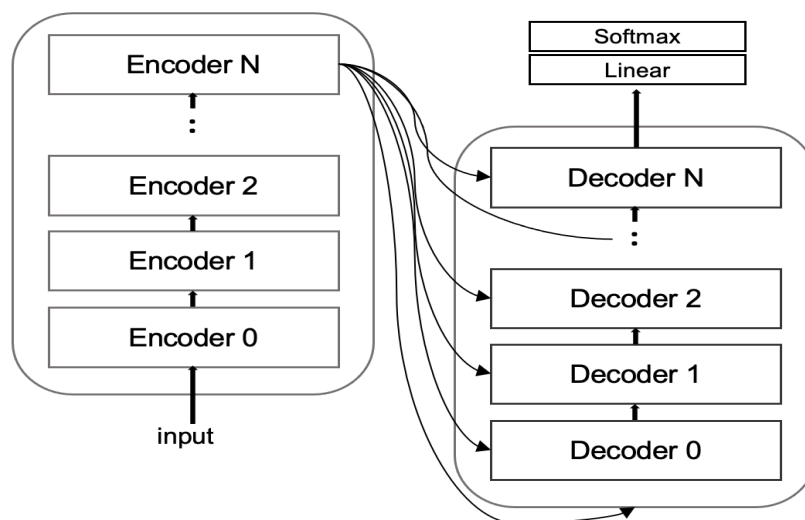


Figure 2.6: Overview of the full Transformer architecture

The encoder has N identical layers. Each layer has two sub-layers; Multi-Headed Self-Attention and Feed Forward Network. Input and Output Embeddings are learned embeddings that convert input and output tokens to vectors of dimension d_{model} . The same weight matrix is used for the input embedding layer and the pre-softmax linear transformation in the output layer.

Multi-Headed Self-Attention allows every token to attend to every other token in the sequence. This is done by first projecting the embeddings into a query (Q), key (K), and value (V) vectors using learned projections $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

Where $X = [x_1, \dots, x_n]$ and d_k is the key dimension. Self-attention is then calculated using a scaled dot product:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \in \mathbb{R}^{n \times d_v} \quad (2.4)$$

This is done for h parallel heads, whose outputs are concatenated to get the multi-headed self-attention output $Z \in \mathbb{R}^{n \times dh_v}$.

Feed Forward Network is a simple network with two linear transforms and ReLU activation that operates on each token separately:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad \text{where } W_1 \in \mathbb{R}^{d \times d_{ff}}, W_2 \in \mathbb{R}^{d_{ff} \times d}.$$

Each sub-layer in the encoder and decoder employs a residual connection followed by layer normalisation. The output of each sub-layer is:

$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

where $\text{Sublayer}(x)$ is the function implemented by the sub-layer itself.

The decoder also has N identical layers. In addition to the sub-layers in each encoder layer, the decoder inserts a multi-headed attention sub-layer between the self-attention and feedforward sub-layers to attend to the output of the encoder: Multi-Headed Attention over Encoder.

It takes the target token embeddings $Y \in \mathbb{R}^{m \times d}$ as queries, and encoder output $Z \in \mathbb{R}^{n \times d_v}$ as keys and values to focus on relevant parts of the input sequence:

$$\text{Attention}(Y, Z, Z) = \text{softmax}\left(\frac{YK^\top}{\sqrt{d_k}}\right)Z$$

Where $K = ZW^K$ and the dimensions are as defined previously.

Since self-attention layers do not have recurrence or convolution, positional information needs to be injected via encodings $PE \in \mathbb{R}^{n \times d}$. These can be constructed using sine and cosine functions:

$$PE(pos, i) = \begin{cases} \sin\left(\frac{pos}{10000^{2i/d}}\right) & \text{if } i \text{ is even,} \\ \cos\left(\frac{pos}{10000^{2i/d}}\right) & \text{if } i \text{ is odd.} \end{cases} \quad (2.5)$$

Where pos is the position and i is the dimension. The computed positional embeddings P are added to the token embeddings X before feeding to the encoder. By incorporating positional information, the decoder can use the order of sequence tokens. Stacking N decoder layers allows the modelling of complex relationships in output generation. By stacking such layers that jointly attend to the full sequence, the Transformer can effectively model long-range dependencies.

In the original paper (Vaswani et al. 2017), the model uses the Adam optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate varies over training according to:

$$\text{lrate} = d_{model}^{-0.5} \cdot \min(\text{step_num}^{-0.5}, \text{step_num} \cdot \text{warmup_steps}^{-1.5})$$

Dropout with a rate of 0.1 is applied to the output of each sub-layer before it is added to the sub-layer input and normalized. Label smoothing with $\epsilon_{ls} = 0.1$ is also employed. The base model has $d_{model} = 512$, $d_{ff} = 2048$, 6 encoder and decoder layers, 8 attention heads, and 65M parameters. A larger model with $d_{model} = 1024$, $d_{ff} = 4096$, 6 encoder and decoder layers, 16 attention heads, and 213M parameters.

Data Fusion Strategies

Data fusion refers to the integration of different data modalities that provide separate perspectives on a problem to be addressed, and using multiple modalities has the potential to decrease the number of errors compared to approaches that only use one type of data (Stahlschmidt et al. 2022). Deep learning fusion strategies can be broadly classified into three categories: early fusion, late/decision fusion, and hybrid/joint fusion.

In the process of early fusion, the original or transformed features are combined at the input level before being fed into a single model that can handle all the information. There are various methods of joining data, but early fusion commonly involves concatenation or pooling.

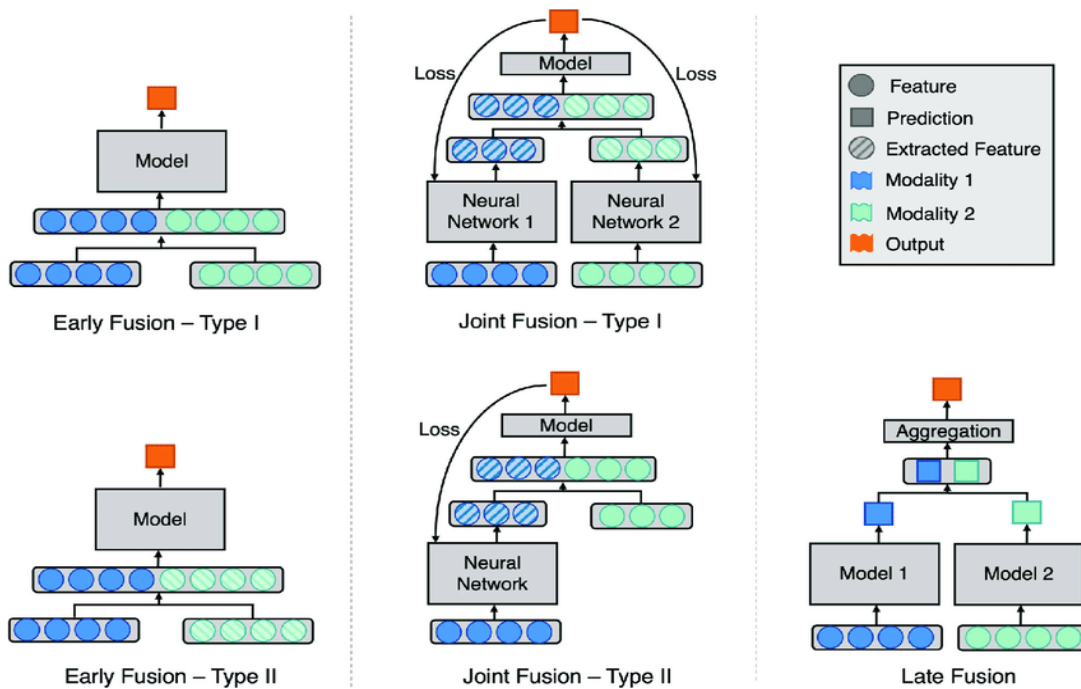


Figure 2.7: The figure illustrates fusion strategies using deep learning. The model architecture varies for each strategy: early fusion (left), joint fusion (middle) and late fusion (right)(Huang et al. 2020)

In late fusion, the input data is processed independently through separate networks. The outputs from these networks are then combined at a later stage to form a joint decision. Late fusion strategies learn modality-specific features separately and then integrate them downstream in the model (e.g. just before the prediction/output layer). Lastly, joint fusion involves combining the features extracted from different modalities at different stages of the network architecture.

2.1.3 Principles of Deep Encoder-Decoder for Image-to-Text Generation

Encoder-decoder architectures (as illustrated in Figure 2.8) are powerful for deep learning-based solutions to image-to-text generation owing to their ability to divide the problem into separate but interlinked subtasks. The encoder portion first extracts visual representations from input images. Encoders leverage deep neural networks tailored to computer vision feature extraction, whether convolutional or transformer-based.

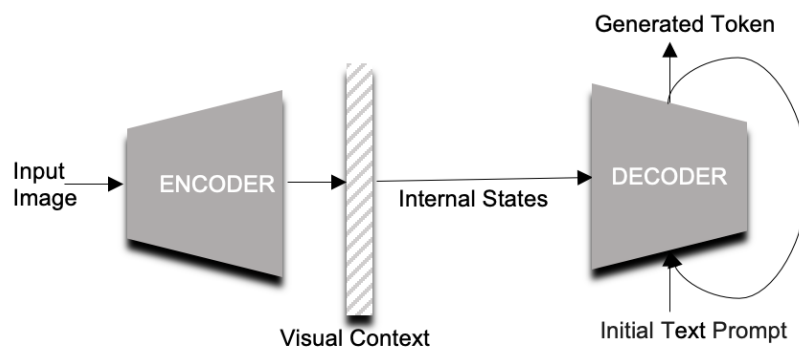


Figure 2.8: A Deep Neural Encoder-Decoder Framework for Generating Text Descriptions from Image Features

Meanwhile, the decoder generates textual sequences auto-regressively using its internal deep network focused on natural language processing, predicting output tokens conditioned on the encoder context. Attention mechanisms may optionally connect the two, allowing the decoder to emphasise encoder features most useful when generating each next word.

Functionally, the segmented encoder-decoder pipeline permits individual optimisation of image analysis and text synthesis stages. Encoders can integrate innovations in representation learning and computer vision without needing to balance competing text generation constraints. Decoders incorporate advances in sequence modelling, language coherence, and context handling from raw encoder output.

Together, specialised deep encoder-decoder networks coordinate to translate encoded image concepts into text descriptions accurately capturing visual scan characteristics. Critically, the modularity of this approach allows for improving the encoder, decoder, and attention mechanisms independently to strengthen overall system performance iteratively.

2.1.4 Concepts of Image Classification

Image classification is a fundamental problem in computer vision that an algorithm needs to determine the objects or scenes present in a given image and assign the image to its corresponding class or classes. There are different types of classification tasks based on their desired outputs or the complexity of the labels. The main types are binary classification (present/absent), multi-class (one label from many), multi-label (belonging to multiple labels), ordinal (order-based sorting) and hierarchical classification into nested categories. This section outlines the specifics of two classification tasks, multi-label and ordinal classification, that are undertaken in this study.

Multi-label Classification

Multi-label image classification is a problem where an algorithm needs to predict the labels that represent objects, characteristics, or other elements found within an image. The task is to assign the given image to multiple labels or categories at the same time. The primary objective of multi-label classification is to train the network that can accurately predict all the relevant labels for a given image. This task is more complex and challenging than single-label classification because the model needs to learn to recognise and identify multiple objects, concepts, or scenes within the same image.

In multi-label image classification, the goal is to learn a function f that maps an input image x to a set of relevant labels $Y \subseteq \{1, 2, \dots, L\}$, where L is the total number of possible labels.

The objective function to be optimised during training can be formulated as follows:

$$\min_f \frac{1}{N} \sum_{i=1}^N \mathcal{L}(f(x_i), Y_i)$$

Here, N is the number of training examples, and \mathcal{L} is a suitable loss function that measures the discrepancy between the predicted labels $f(x_i)$ and the true labels Y_i for each training example (x_i, Y_i) .

There are two common methods to solve multi-label classification; transforming the problem into binary classification or a multi-class classification problem. In the binary classification problem transformation method, each label is treated as a separate binary classification problem. This method is useful when there is no correlation between the labels.

On the other hand, the multi-class classification approach transforms the multi-label problem into a single multi-class classification task, where each combination of labels represents a unique class. If there are correlations between the labels or the order of the labels is important, this method is more suitable. It is important to consider the dataset and the relationships between the labels when choosing the most appropriate method.

A commonly used loss function for multi-label classification is the binary cross-entropy loss:

$$\mathcal{L}(f(x_i), Y_i) = - \sum_{j=1}^L y_{ij} \log \sigma(f_j(x_i)) + (1 - y_{ij}) \log(1 - \sigma(f_j(x_i)))$$

In this equation, y_{ij} is a binary indicator that takes the value 1 if the j -th label is relevant for the i -th example, and 0 otherwise. σ is the sigmoid function, and $f_j(x_i)$ is the output of the model for the j -th label and the i -th input example. The goal is to find the parameters of the function f that minimise the overall loss across the entire training dataset.

Ordinal Classification

Ordinal image classification is the task of assigning the image or objects identified within the image to predefined classes/labels that can be used to establish a superiority/preference relationship with each other. The primary goal of this task is to keep the order of the labels when training the network and making the prediction.

It can be formulated as $\mathcal{Y} = \{1, 2, \dots, K\}$ where the input space of images is \mathcal{X} , ordinal label space is \mathcal{Y} , and K is the number of ordinal classes or labels. The goal is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that maps an input image $x \in \mathcal{X}$ to an ordinal label $y \in \mathcal{Y}$, while preserving the order/rank of the labels.

The ordinal classification problem can be considered as a constrained optimisation problem, where the objective is to minimise a loss function $\mathcal{L}(f(x), y)$ that takes into account the order of the labels. The loss function needs to be chosen considering the problem formulation and dataset characterisation. The objective is to find the function f that minimises the overall loss while ensuring that the predicted ordinal scores respect the order of the true labels.

2.1.5 Fundamentals of Deep Multi-Task Learning

Multi-task learning is a method in deep learning that aims to improve the performance of several related tasks by jointly and simultaneously training a single model to perform multiple tasks. In MTL, multiple tasks are trained in parallel by sharing features between tasks, which allows more generalised representations. This contrasts with single-task learning, where separate models are designed to perform individual specific tasks. MTL also helps reduce overfitting by improving the regularisation and data efficiency.

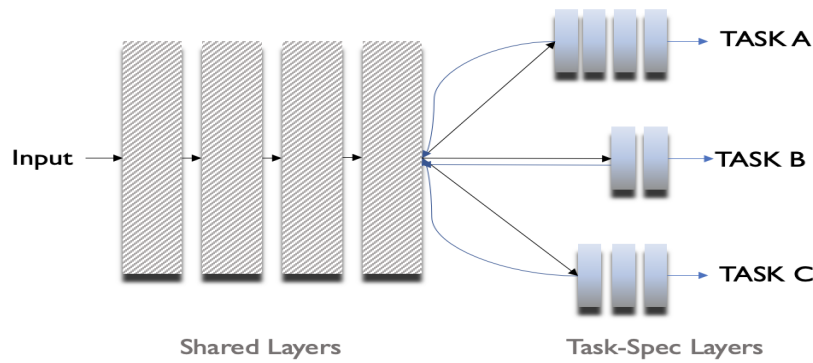


Figure 2.9: An Overview Illustration of Multi-Task Framework

Multi-task learning differs from transfer learning, where a model trained on one task is repurposed for a second task. In MTL, both the weights and knowledge are transferred and shared between tasks, whether to use a pre-trained model or not, to find the optimal set of parameters that works well for all tasks.

The general MTL equation can be defined as:

$$\min_{\theta} \sum_{t=1}^T \lambda_t \mathcal{L}_t(X^t, Y^t; \theta)$$

Where: T is the number of tasks X^t and Y^t are the input and output data for task t \mathcal{L}_t is the loss function for task t λ_t is the weight or importance of task t θ represents the shared model parameters across tasks. The goal is to find the optimal parameters θ that minimise the weighted sum of task-specific losses.

Broadly, there are three types of MTL architectures; Hard Parameter Sharing, Soft Parameter Sharing and Task-Specific Layers. In hard parameter sharing, all tasks share the same set of parameters θ , except for task-specific output layers W^t . The shared representation is denoted as $\phi(X; \theta)$, and the output for task t is computed as $f^t(X; \theta, W^t) = W^t \phi(X; \theta)$. The hard

parameter sharing function can be described as:

$$\min_{\theta, W^1, \dots, W^T} \sum_{t=1}^T \lambda_t \mathcal{L}_t(f^t(X^t; \theta, W^t), Y^t)$$

In soft parameter sharing, each task has its own set of parameters θ^t , but these parameters are regularised to be similar to each other through a distance metric $\Omega(\theta^t, \theta^s)$ between tasks t and s . The soft parameter sharing function becomes:

$$\min_{\theta^1, \dots, \theta^T} \sum_{t=1}^T \lambda_t \mathcal{L}_t(f^t(X^t; \theta^t), Y^t) + \gamma \sum_{t=1}^T \sum_{s=t+1}^T \Omega(\theta^t, \theta^s)$$

Where γ controls the strength of the parameter similarity regularisation.

Task-Specific Layers approach has shared layers $\phi(X; \theta_s)$ across tasks, as well as task-specific layers $g^t(h; \theta_t)$, where $h = \phi(X; \theta_s)$ is the shared representation. The objective function can be calculated as:

$$\min_{\theta_s, \theta_1, \dots, \theta_T} \sum_{t=1}^T \lambda_t \mathcal{L}_t(g^t(\phi(X^t; \theta_s); \theta_t), Y^t)$$

In summary, hard parameter sharing enforces a single set of shared parameters across tasks, soft parameter sharing encourages parameter similarity through regularisation, and task-specific layers allow for both shared and task-specific representations.

Furthermore, in order to successfully generalise the representation across all tasks, choosing appropriate loss functions is an important step in MTL networks. The overall network may have a **joint loss function** that combines the losses from all tasks into a single objective while the model learns to optimise all tasks simultaneously, balancing their contributions during training. The joint loss function can be calculated as:

$$\mathcal{L}_{joint} = \sum_{t=1}^T \lambda_t \mathcal{L}_t(f_t(X; \theta), Y_t)$$

Where: T is the number of tasks λ_t is the weight or importance of task t \mathcal{L}_t is the loss function for task t $f_t(X; \theta)$ is the model's output for task t given input X and shared parameters θ Y_t is the ground truth for task t . Another approach is to use **task-specific loss** where each task

can have its own loss function decided based on its characteristics. It can be formulated as

$$\mathcal{L}_{task_specific} = \sum_t \lambda_t \mathcal{L}_t(f_t(X; \theta), Y_t)$$

Where the notations are the same as in the joint loss function, but without task weights λ_t .

Another way is to use the **weighted loss** approach by assigning different weights to each task's loss. It allows for fine-grained control over the model's behaviour, prioritising certain tasks over others based on their significance. The weighted loss function can be expressed as:

$$\mathcal{L}_{weighted} = \sum_t \lambda_t w_t \mathcal{L}_t(f_t(X; \theta), Y_t)$$

Where: w_t is the weight assigned to task t . Other notations are the same as in the previous approaches.

Multi-task learning has been successfully applied to various vision and text-related tasks, demonstrating its effectiveness in improving model performance and generalisation. In terms of network design, the choice of architectural framework and loss function approaches depends on the characteristics of the problem definition. Overall, multi-task learning provides a powerful framework for training deep networks to simultaneously tackle multiple related tasks.

In the field of computer vision, multi-task learning has been applied to simultaneously tackle multiple image understanding tasks that can jointly handle low, mid, and high-level vision tasks. For instance, researchers have developed models that can perform boundary detection, normal estimation, saliency detection, semantic segmentation, and object detection within a single framework (Kokkinos 2017). Recent works have shown that models can jointly perform high-level vision tasks, including instance segmentation, panoptic segmentation, surface normal estimation, depth estimation, 2D keypoint detection and so on, leveraging the interrelated nature of these tasks to improve overall performance and computational efficiency (Vandenhende et al. 2021; Bhattacharjee et al. 2022).

The field of Natural Language Processing has seen significant advancements in multi-task learning, particularly with the rise of large language models. Models have been developed to simultaneously perform various language processing tasks such as part-of-speech tagging, chunking, named entity recognition, and semantic role labelling. By sharing a common representation

for related tasks, these multi-task models have shown improved performance compared to their single-task counterparts. Recent research has focused on developing models that can perform multiple NLP tasks such as sentiment analysis, named entity recognition, and question answering simultaneously, while also being able to generalise to new, unseen tasks with minimal fine-tuning (Wei et al. 2021).

Multi-task learning also made good progress at the intersection of vision and language tasks. Visual Question Answering models have been designed to jointly learn visual attention and question attention, similarly, in image captioning, multi-task approaches have been used to simultaneously optimise for multiple caption quality metrics, resulting in more accurate and diverse captions (J. Li, Selvaraju, et al. 2021; P. Wang et al. 2022). These recent advancements highlight the potential of multi-task learning in vision-text tasks.

2.2 Understanding X-ray reports and Medical Data Modalities

2.2.1 Chest X-ray Reporting

A chest X-ray is a diagnostic imaging procedure that uses a focused beam of radiation to generate detailed images of the chest's internal structures. It is commonly to detect and assess conditions such as pneumonia, emphysema, or Chronic Obstructive Pulmonary Disease(COPD). X-ray reports typically follow a standardised structure to communicate the findings of an X-ray test (Figure 2.10).

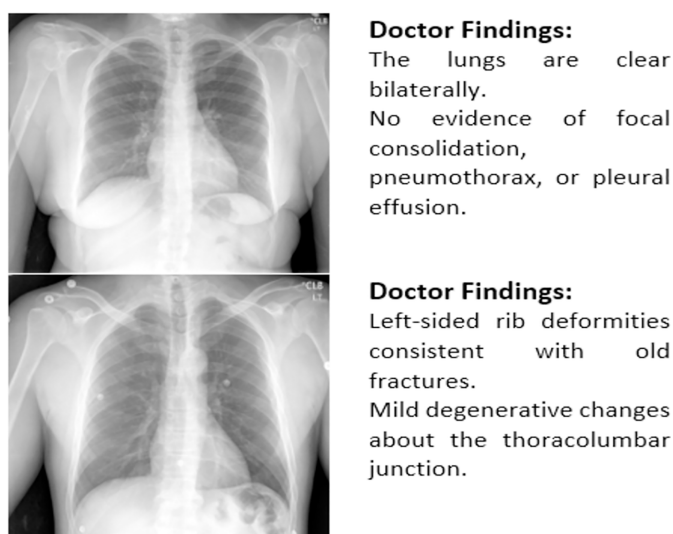


Figure 2.10: Examples of Chest X-Ray Images with Corresponding Radiologist-Generated Findings (Sirshar et al. 2022)

The examination results and notes of any irregularities (or other clinically relevant findings) are described in these reports. Using a consistent format helps ensure all essential clinical details are included, which can enhance the clarity, precision, and readability of the reports.

Given considerations of conciseness, this section exclusively focuses on the most common sections observed within standardised chest X-ray reporting. These report elements are, in alignment with those published in open-source and authorised datasets, making it easier to perform computational analyses that are aligned. However, it's worth mentioning that real-world clinical environments may necessitate radiologists to provide specific information based on their professional judgment to effectively communicate nuanced diagnostic details to referring healthcare providers.

Standardised X-ray reports are typically semi-structured and consist of multiple sections. Within many available databases, a single report is often associated with multiple X-ray projections. The "Examination" section details the specific X-ray procedure performed, outlining the imaging modality and protocols. The "Indication" section provides context by explaining the clinical reasons prompting the X-ray, such as suspected conditions. "Technique" describes the technical aspects of the imaging process, including patient positioning and equipment details. The "Comparison" section, when applicable, contrasts current findings with prior imaging studies for temporal context. The "Findings" section offers the radiologist's observations, detailing any abnormalities or important features. Lastly, the "Impression" section summarises the overall findings and provides a diagnostic conclusion.

This format of X-ray reports provides a structure for computational analysis. The labelled sections and clinical terminology are used to supply the lexical and ontological framework for training machine learning models. Computer-aided systems can extract radiological observations, diagnostic impressions, technical parameters, and patient indications as distinct data representations. Encoding these standardised semantic segments enables reliable information extraction as well as generating synthetic reports.

Having large datasets with consistent reporting formats helps in training machine learning models to automate radiology workflows. Therefore, explaining the common X-ray report structures provides an important basis for researchers to develop automated systems that can interpret medical images and generate reports. However, standardised structure alone does not fully capture the complexities of radiological language. There remains substantial linguistic diversity

in how observations and impressions are expressed between different practitioners.

2.2.2 Medical Data Modalities

In healthcare, deep learning methods often depend on analysing different types of data, each providing unique insights, to understand/extract information and guide clinical decision-making. These modalities, including textual data, imaging data, genomic data and so on, are in different formats and structures, requiring specialised encoding strategies for meaningful interpretation within deep learning models.

This section provides a concise overview of three categories of medical data used in this research: imaging data, structured health records, and unstructured clinical notes and reports.

Imaging Data

Modalities like X-rays, CT scans, and MRIs offer invaluable visual information, but their large size and inter-scanner variability necessitate specialised processing for meaningful analysis. Focusing on X-rays, are a form of electromagnetic radiation that can penetrate the body and produce images of its internal structures. When X-rays are directed at the body, different tissues absorb varying amounts of radiation. Bones, for instance, absorb more X-rays and appear white on the resulting image, while softer tissues absorb fewer X-rays and appear in shades of grey. This differential absorption creates a contrast that highlights the structures within the body, allowing radiologists to diagnose conditions such as fractures, infections, and tumours.

Chest X-rays (CXR) can be categorised into various types based on the views and techniques used. The most common types are the posteroanterior (PA) and lateral views. The PA view is taken with the patient standing facing the X-ray film, while the X-ray machine is positioned behind them. This view is considered the standard because it provides a clear image of the lungs, heart, and chest wall. The lateral view is taken from the side and is often used in conjunction with the PA view to provide a more comprehensive understanding of the chest's anatomy. Other specialised views include the anteroposterior (AP) view, typically used for bedridden patients, and the decubitus view, which can help in identifying pleural effusions by showing fluid movement when the patient is lying on their side.

Structured Health Records

Electronic Health Records (EHRs) are digital versions of patients' paper charts and are comprehensive records of a patient's medical history, maintained over time. These records include

a wide range of data types that are critical for patient care, research, and health management. They are continuously updated as patients interact with the healthcare system. The acquisition process begins with patient registration, where basic demographic information is recorded. During clinical encounters, healthcare providers document medical histories, physical examinations, diagnoses, treatment plans, and outcomes.

EHRs also include laboratory and imaging results, prescriptions, and any interventions performed, mostly in a structured format. As patients receive care from multiple providers and facilities, their records are aggregated into a unified record that reflects their complete history. Structured data within EHRs include heart rate, respiratory rate, oxygen saturation, temperature, level of acuity, gender, ethnicity, diastolic blood pressure (DBP), systolic blood pressure (SBP), and ICD titles, especially when coded as numeric or categorical values.

Unstructured Clinical Notes and Reports

Clinical notes and reports are essential components of patient records, providing detailed and contextual information about patient care. Unlike structured data in EHRs, clinical notes are typically unstructured, written in free text by healthcare providers. Clinical notes are generated during various points of patient care, starting from the initial patient consultation to follow-up visits and specialist referrals, including details from imaging reports (examination report) and patients' complaints (chief complaint). Physicians, nurses, and other healthcare providers document their observations, assessments, and plans in these notes.

The unstructured nature of clinical notes captures the richness and complexity of patient care but presents challenges for analysis due to variations in terminology, writing styles, and documentation practices. Advanced natural language processing techniques are often required to extract meaningful information from these texts for use in deep learning models, however, their unstructured nature, free text variations, and coding inconsistencies pose challenges for deep learning models.

These different modalities are recorded in various ways, each presenting unique encoding challenges for analysis. For example, imaging data requires handling of size and pixel values, whereas EHRs store information in a structured textual format suitable for manipulation with natural language processing techniques. Clinical notes contain unstructured free text that captures subtle details, however, lack consistency and structure. Appropriate encoding strategies need to be developed and applied based on the type, format and structure of each modality.

2.3 Datasets

2.3.1 Indiana University Chest X-ray Collection

The Indiana University (IU) Chest X-ray collection (Demner-Fushman et al. 2016) is a publicly available dataset that contains 7,470 chest X-ray images (.png format) taken from frontal and lateral views. It also includes 3,955 corresponding radiology reports (.XML format) that have been anonymised. Each report provides findings, impressions, comparisons, and indications for each patient case and multiple X-ray images can relate to a single report. Additionally, the dataset has Medical Subject Heading (MeSH) annotations, each MeSH term consists of a pair: a finding and its corresponding description.

To illustrate examples of the image data, Figure 2.11 shows a random sample of frontal and lateral chest X-rays from the collection. Figure 2.12 then provides a full report for one patient case with anonymised information marked as ‘XXXX’. More explanation on the content of the report sections is available in Section 2.2.

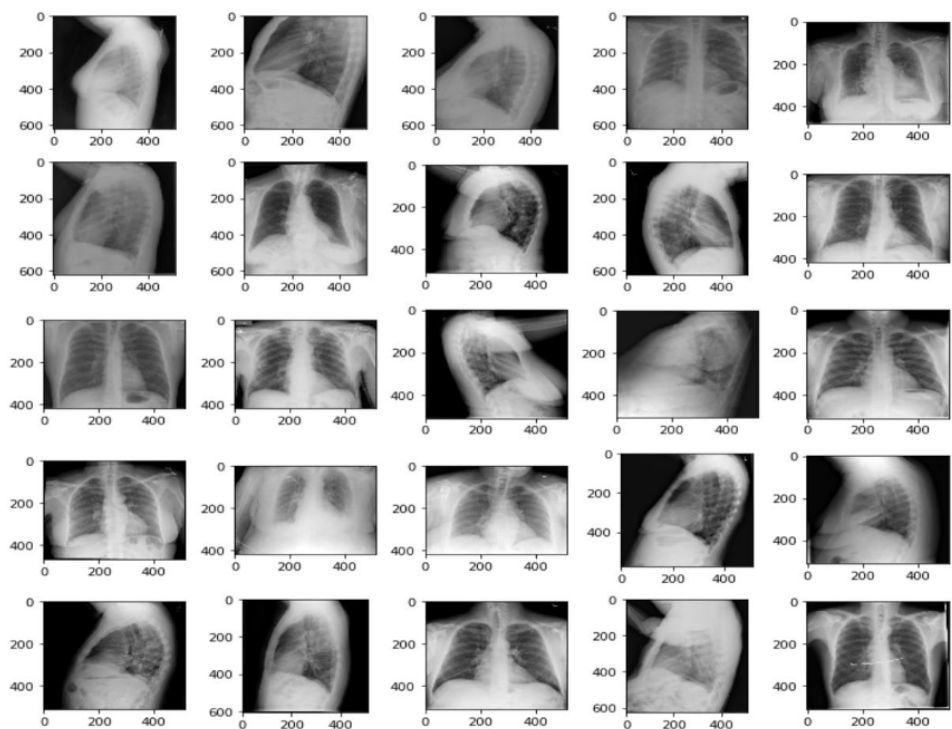


Figure 2.11: Random chest radiographs of 25 different patients from the IU CXR collection.

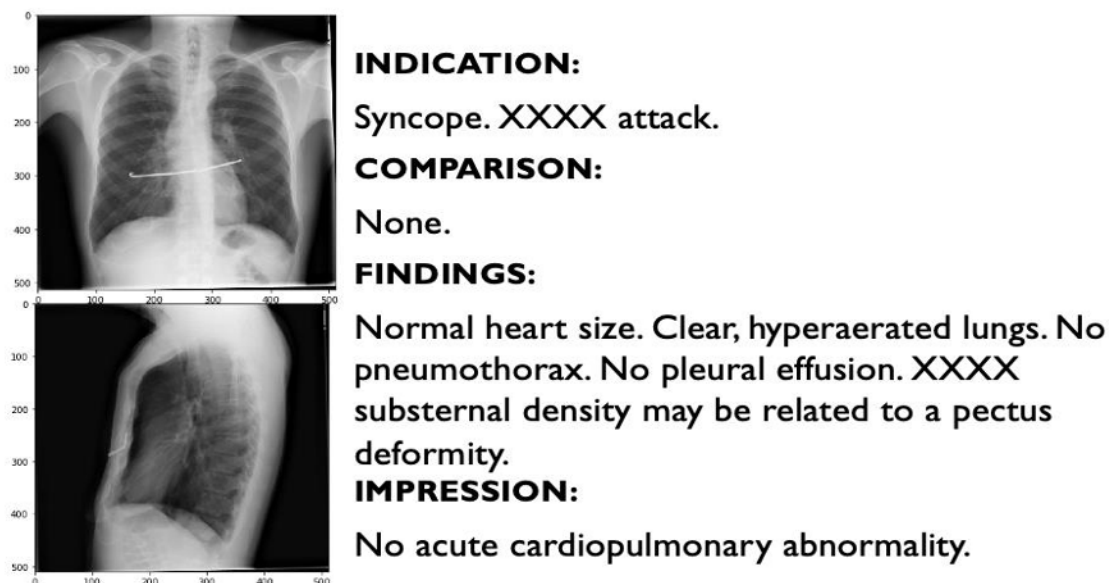
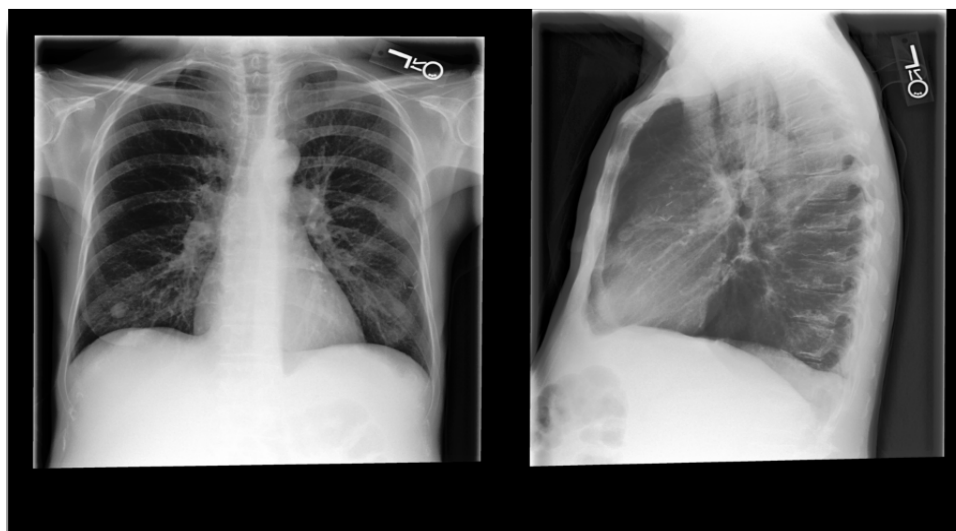


Figure 2.12: An example frontal and lateral chest X-ray imaging and corresponding report from the IU chest X-ray collection.

2.3.2 Medical Information Mart for Intensive Care Database

The MIMIC (Medical Information Mart for Intensive Care) database is a large, freely available database comprising de-identified health data associated with patients who stayed in care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. In the MIMIC database, there are several datasets designed for various applications. The dataset used in this study was created by using MIMIC-CXR (Johnson, Pollard, et al. 2019), MIMIC-IV (Johnson, Bulgarelli, et al. 2023), and MIMIC-IV Emergency Department (MIMIC-IV-ED) datasets.

MIMIC-CXR (version 2.0) encompasses a vast collection of 377,110 CXR images captured from multiple views, together with 227,835 de-identified radiology reports, of 63,473 patients. As shown in Figure 2.13, each report contains 'examination', 'indication', 'technique', 'comparison', 'findings', and 'impressions'. Meanwhile, MIMIC-IV (version 2.0) comprises de-identified patient data, including characteristics like age, gender, ethnicity, and marital status, extracted from individuals. Furthermore, MIMIC-IV-ED (version 2.2) is an extensive dataset of emergency department (ED) admissions at the BIDMC between 2011 and 2019, which contains detailed clinical information, including diagnosis, medication, triage, and vital signs.



FINAL REPORT

EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___F with new onset ascites // eval for infection

TECHNIQUE: Chest PA and lateral

COMPARISON: None.

FINDINGS:

There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomeastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.

IMPRESSION:

No acute cardiopulmonary process.

Figure 2.13: A sample X-ray report from the MIMIC-CXR database, comprising images captured from both the frontal and lateral perspectives of the patient.

2.4 Evaluation Metrics

2.4.1 Natural Language Generation Metrics

The quality of the automatically generated reports was evaluated using several metrics to compare the generated text to the reference reports. The first set of metrics utilised was the **BLEU** (**Bilingual Evaluation Understudy**) scores, which measure the N-gram precision between the candidate and reference texts Papineni et al. 2002. Specifically, BLEU-1 to BLEU-4 scores were calculated, which assess the precision for unigrams to 4-grams respectively. The BLEU score is computed as:

$$\text{BLEU} = \text{BP} \times \exp \left(\sum_{n=1}^N w_n \log(\text{precision}_n) \right)$$

where

$$\text{BP} = \begin{cases} 1 & \text{if length of generated text} > \text{length of reference text} \\ e^{(1 - \frac{\text{length of reference text}}{\text{length of generated text}})} & \text{otherwise} \end{cases}$$

and

$$\text{precision}_n = \frac{\text{Number of n-grams in the generated text that match a reference}}{\text{Total number of n-grams in the generated text}}$$

BP is a brevity penalty that penalises short candidates, precision_n is the modified n-gram precision which averages the precision for all order n-grams up to length N , w_n are positive weights summing to 1 that allow flexible weighting of the different order n-grams, and N is the maximum n-gram order. The BLEU scores measure the local word-level similarity between the generated and reference texts, with higher scores indicating greater similarity using n-grams.

The second metric was the **ROUGE_L (Recall-Oriented Understudy for Gisting Evaluation)** score (C.-Y. Lin 2004), which calculates the longest common subsequence (LCS) between the generated and reference summaries as:

$$\text{ROUGE}_L = \frac{\text{LCS}(\text{generated}, \text{reference})}{\text{length}(\text{reference})}$$

The ROUGE_L score measures the recall by the ratio of LCS to the length of the reference summary. It also implicitly measures precision since LCS can be viewed as a sequence of consecutive matches between the generated and reference. Thus, the ROUGE_L score assesses the quality of the generated summary by comparing the longest co-occurring in-sequence n-grams to the reference summary. Higher ROUGE-L scores indicate better-quality text generation.

In our final approach, we additionally compute the METEOR score to facilitate a more comprehensive comparison with the existing literature, as it is frequently used in conjunction with others. **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** score (Banerjee and Lavie 2005), which evaluates the quality of generated text by comparing it to reference texts based on the harmonic mean of unigram precision and recall, with an

additional alignment step. The METEOR score is calculated as:

$$METEOR = \frac{10 \cdot P \cdot R}{R + 9 \cdot P} \cdot \left(1 - \frac{c}{l}\right) \quad (2.6)$$

where P is precision, R is recall, c is the number of chunks, and l is the length of the longest common subsequence. The METEOR metric considers synonyms, stemming, and word order, providing a more nuanced assessment of text similarity compared to simple n-gram overlap metrics. Higher METEOR scores indicate better-quality text generation due to better alignment with the reference text.

We further also evaluated semantic similarity between the generated report and ground truth using **BERTScore** (T. Zhang et al. 2019) and **Bio-ClinicalBERT Score** (Equation 2.7).

$$BScore = \frac{1}{N} \sum_{i=1}^N (F1(y_i, \hat{y}_i) + Suff(y_i, \hat{y}_i) + Flu(y_i, \hat{y}_i)) \quad (2.7)$$

Where: N is the number of sentence pairs in the evaluated dataset y_i is the i^{th} reference sentence \hat{y}_i is the i^{th} generated sentence $F1(y_i, \hat{y}_i)$ is the F1 score between y_i and \hat{y}_i using both BERT and ClinicalBERT embeddings separately $Suff(y_i, \hat{y}_i)$ is the sufficiency score between y_i and \hat{y}_i $Flu(y_i, \hat{y}_i)$ is the fluency score between y_i and \hat{y}_i

These metrics use contextual embeddings from BERT and Bio-ClinicalBERT models to provide a more nuanced assessment of meaning compared to strict n-gram matching. The BERT-based metrics were able to capture whether the generated reports conveyed clinically coherent descriptions despite differing word usage compared to the reference. These automated evaluation metrics quantified linguistic similarity at word level, sentence level, and semantic meaning levels.

Overall, BLEU scores and ROUGE-L, while commonly used, have limitations as they rely on exact word matching and do not capture clinical meaning. A report using different but medically equivalent terms would return a lower score. METEOR performs slightly better as it considers synonyms and stemming, but still struggles with medical terminology. BERTScore shows improved correlation due to its contextual embeddings, but isn't specifically trained for medical language. BioClinical BERTScore, on the other hand, addresses these limitations by using models pre-trained on biomedical texts. It is more suitable for capturing domain-specific semantic similarities and clinical accuracy in radiology reports.

However, no single metric comprehensively captures all aspects of report quality - a combination of metrics alongside expert human evaluation remains ideal.

2.4.2 Human Analysis

In addition to the automated metrics, we conducted a human evaluation to analyse the generated text results. We employed two approaches for this analysis. The first approach involved comparing the generated text with the corresponding ground truth reports to identify differences and similarities in grammar, sentence structure, word choice, and the correctness of detecting medical abnormalities.

Furthermore, for a portion of the thesis, a board-certified radiologist was involved to provide insights. The radiologist assessed both the generated text and ground truth in three criteria: language fluency, content selection, and the accuracy of identifying abnormal findings.

2.4.3 Classification Metrics

For evaluating model performance on **multi-label classification**, several metrics were utilised including Precision, Recall, F1 Score, Hamming Loss, and Exact Match Ratio.

Precision measures the accuracy of positive label predictions out of all predicted positive labels:

$$Precision = \frac{TP}{(TP + FP)}$$

Whereas Recall calculates the percentage of true positive labels that were correctly predicted:

$$Recall = \frac{TP}{(TP + FN)}$$

Here, TP , FP , and FN are the numbers of true positives, false positives, and false negatives respectively. The F1 Score provides a balance between Precision and Recall through their harmonic mean:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{(Precision + Recall)}$$

Hamming Loss evaluates how many times on average, the model incorrectly predicts labels:

$$HammingLoss = \frac{1}{N} \sum_{i=1}^N \frac{|y_i \Delta \hat{y}_i|}{|L|}$$

Where y_i is the true set of labels, \hat{y}_i is the predicted set of labels, L is the set of all labels, and Δ is the symmetric difference between the true and predicted label sets.

Finally, Exact Match Ratio measures how often the model correctly predicts all labels for a given sample:

$$ExactMatchRatio = \frac{\# \text{ samples with all labels correctly predicted}}{N}$$

Ordinal Classification Accuracy, The Mean Absolute Error and Accuracy-Correlation Hybrid Metric were used to evaluate model performance on **ordinal classification**.

Ordinal Classification Accuracy measures the accuracy by computing the percentage of correct predictions:

$$OrdinalAccuracy = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(y_i = \hat{y}_i)$$

Where N is the total number of samples, y_i is the true ordinal label, \hat{y}_i is the predicted ordinal label for the i^{th} sample, and $\mathbf{1}(\cdot)$ is an indicator function that returns 1 if its argument is true (correct prediction).

The Mean Absolute Error (MAE) calculates the average magnitude of errors:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Similarly, the Mean Squared Error (MSE) computes the average squared differences between the true and predicted ordinal values:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Finally, the Accuracy-Correlation Hybrid Metric combines an accuracy measure with Spearman's rank correlation coefficient:

$$ACH = 2 \cdot \frac{\text{OrdinalAccuracy} \cdot \rho_s}{\text{OrdinalAccuracy} + \rho_s}$$

Where ρ_s is the Spearman rank correlation between the true ordinals y_i and predicted ordinals \hat{y}_i . This metric assesses both accurate prediction and preserving the relative ordering of ordinal categories.

Together these metrics evaluate model performance on ordinal classification from different aspects - accuracy of predictions, error magnitudes, and preserving ordinality relationships. The metrics provide a comprehensive quantification of the model's ability to predict ordered values.

Chapter 3

LITERATURE REVIEW

3.1 Foundations of Natural Image Captioning

Image Captioning, also known as Visual Captioning, is the task of automatically generating a natural language description of a given image. It requires the recognition and detection of objects located in the image, the identification of attributes, and the determination of their relationships and interactions. Subsequently, it must generate coherent sentences based on the features extracted. Since this task entails the incorporation of computer vision and natural language processing, it has attracted tremendous attention in the artificial intelligence community. In only a short period of time, lots of research has been proposed to fulfil this task (Vinyals et al. 2015; Donahue et al. 2015; Karpathy and Fei-Fei 2015; Mao et al. 2015).

The most common deep learning architecture used by researchers follows a standard encoder-decoder baseline that consists of two phases. Broadly, the encoder part receives an image and sends it to the image feature extractor model which is usually pre-trained on large datasets for classification and recognition tasks. The region-based visual features are extracted and these high-level image representations are used as input by the text generation decoder to produce a relevant caption. This framework allows to training of the entire network end-to-end, meaning that the system can learn all parameters during training. The main differences between the studies based on this framework are the type of neural network employed and different encoder-decoder configurations (X. Chen and Lawrence Zitnick 2015).

One of the prominent studies in image captioning (Vinyals et al. 2015) employs a similar maximised probabilistic framework with neural machine translation (NMT) but it makes use of CNN for encoding while state-of-art NMT models employ RNN-encoder. They proposed an end-to-end system that is composed of a vision CNN image embedder and an LSTM-based sentence generator. Donahue et al. 2015 introduced a task-specific Long-term Recurrent Convolutional Network (LRCN) model for the activity recognition and the generation of images and video descriptions. For the image captioning task, they have used an integrated system consisting of a single CNN model, based on the Caffe reference model (Jia et al. 2014) which has a very similar architecture to AlexNet, and LSTM-based language model. Moreover, Karpathy and Fei-Fei 2015 have used a CNN-based visual model and multimodal RNN to propose a visual-semantic alignment image captioning model that can compute the latent alignment between images and captions. Mao et al. 2015 have also used a base CNN and RNN image captioning model for a specific Novel Visual Concept learning from Sentences (NVCS) task. The model adopts a cumulative concept learning strategy; it can transfer previously learned concepts and expand the dictionary without start-to-end network training when a new concept is introduced to the system.

3.2 Advancements in Image Captioning: Attention and Semantic Features

Although RNN-based text generators have achieved noteworthy results, they cannot access all previous hidden states from the encoder as the sequence gets longer. This is still challenging for LSTM and GRU networks although they have a longer reference window and more capacity than simple RNNs. Therefore, even though this powerful approach has yielded promising results, it had the drawback of having a limited reference window and identifying only one part of the image while generating the next word. Inspired by the neural machine translation (Bahdanau et al. 2014), the attention mechanism has been employed by researchers to overcome this problem. In principle, it can consider the entire sequence, therefore, the decoder can focus on the relevant part of the image while generating the next word. Adoption of the attention mechanism has resulted in significant success in many sequence-to-sequence models and has been effectively included in visual captioning models (Xu et al. 2015; Z. Yang et al. 2016; Lu et al. 2017; Anderson et al. 2018).

Xu et al. 2015 presented an image caption generation model by the use of both stochastic and deterministic attention mechanisms along with convolutional feature extraction and LSTM-based sentence generation. Z. Yang et al. 2016 developed the Review Network that has an additional reviewer component to attentive encoder-decoder frameworks. In each review step, a thought vector is generated by using an attention mechanism and generated vectors are used by the RNN-decoder. Furthermore, Lu et al. 2017 also improved the neural encoder-decoder frameworks with adaptive attention by incorporating the sentinel gate and spatial attention into the system. Anderson et al. 2018 combined both bottom-up and top-down attention mechanism that allows more attention to the notable objects and regions rather than operating in the equally-sized image regions.

Considering that image captioning is a multi-modal activity, researchers have introduced semantic features to the network to better capture the relationship between these two modalities. You et al. 2016 proposed a new attention-based approach by introducing the system of semantic attention. They have used the GoogleNet CNN model (Szegedy, W. Liu, et al. 2015) for extracting objects' regions, etc. in addition to the visual representation of the images. The semantic attention model allows to incorporation of the extracted representations and visual concepts in the language model for producing the description of the given image. Gan et al. 2017 also considered using the semantic concepts and developed A Semantic Compositional Network (SCN) that extracts tags from the image and is used in an LSTM-based language model.

3.3 Transformer Architecture

The recurrent-based text generation models have a well-known vanishing and exploding gradients problem which means that recent input sequence causes a bias as there is limited access to the previous inputs and there is no direct access to all inputs. Leveraging recent advances of transformers (Vaswani et al. 2017), which are self-attention-based neural networks, in the natural language processing (NLP) area, state-of-the-art image captioning architectures have tended to substitute their model components with the transformers (W. Liu et al. 2021).

The main advantages of the transformers over other architectures are that it does not use recurrence and is entirely based on an attention mechanism. It takes and executes the input sequence as a whole, allows more parallelisation, and learns the relationship between words in the sequences by the use of multi-head attention mechanisms and positional embeddings. Since

more context is included in the network, the transformer-based architectures can learn faster and more effectively.

In the wake of this phenomenon, several variants of the transformers were developed for the image caption generators. The original transformer architecture also inherits an image-text embedding type framework which consists encoder and decoder. The most commonly adapted transformer-based encoder-decoder architecture for image captioning has three main components; visual feature extraction model, Transformer-based encoder and Transformer-based decoder. As in previous studies, pre-trained CNN models are also employed for high-level feature extraction. However, in this approach, the output of the visual model is used by the Transformer-based encoder to map the visual features and to generate a sequence of image representations. Then, the transformer-based decoder receives the encoder's results to generate a corresponding caption for the given image.

X. Zhu et al. 2018 presented the Captioning Transformer (CT) model that uses the ResNeXt CNN model (Xie et al. 2017) as an encoder and Transformer as a decoder. W. Zhang et al. 2019 also used the Transformer model as a decoder along with the ResNet CNN model, additionally, they have improved the network with a combination of spatial and adaptive attention. G. Li et al. 2019 enhanced the vanilla Transformer architecture with Entangled Attention (ETA) and Gated Bilateral Controller (GBC), their proposed model allows to processing of semantic and visual concepts concurrently.

Moreover, Cornia et al. 2020 introduced a fully-attentive model called Meshed-Memory Transformer that consists of a Memory-Augmented Encoder, which has enriched with learnable the keys and values with a priori information and used a learnable gating mechanism to perform a mesh connectivity, and Meshed Decoder that performs a meshed connection between all encoding layers. On the other hand, W. Liu et al. 2021 proposed a full Transformers network without having any convolutional operation in the encoder. Different from previous studies, their model, CaPtion TransformeR (CPTR), uses the raw image and adjusts it according to the accepted input form of the Transformer-encoder by dividing the original image into N patches. After reshaping the patches, the obtained patch embedding is incorporated with positional embedding.

Despite advancements that come with transformers and semantic features, directly applying natural image captioning methodologies to radiology report generation remains challenging due to the nature of the medical data. The following section reviews the literature on radiology reporting and how current approaches have adapted these methods for the medical domain while addressing the associated challenges.

3.4 Automatic Radiology Reporting Baselines

In recent years, many studies have had great success in fine-tuning deep neural networks to generate medical reports. Earlier existing radiology report generation studies adopt image captioning approaches for medical report generation and leverage the CNN-RNN framework (Jing et al. 2017; Xue et al. 2018; Yuan et al. 2019; Shaokang Yang et al. 2020; Singh et al. 2021).

Jing et al. 2017 employed a pre-trained VGG-19 model to learn visual features and use the extracted features to predict relevant tags for any given chest X-ray. The predicted tags are used as semantic features in the network and both semantic and visual features are fed into the Co-attention Network. Hierarchical LSTM has used the context vector provided by the Co-attention Network to generate the topic and description of the given X-ray. Although the model obtained promising results and achieved great success in its field, the repetitive sentences in reports and the generation of different results for the same patient undermined its credibility from both medical and computational perspectives.

Xue et al. 2018 improved the pre-trained Resnet-152 encoder using multi-view content (both lateral and frontal view) and incorporated them to ensure the consistency of the results. They also generated a report with a sentence decoder and additionally used the first predicted sentence as a joint input along with image encoding.

Another major study was proposed by Yuan et al. 2019, who pre-trained their multi-view encoder from scratch using the CheXpert dataset (Irvin et al. 2019) instead of using ImageNet pre-trained models. In order to enhance the decoder, they extracted and applied medical concepts from the reports. Applying medical concepts conveyed the semantics in the content of the report and they achieved noteworthy results. The idea of using medical concepts was also employed by Shaokang Yang et al. 2020. While they applied a similar approach in principle, they proposed a reward term to extract more precise concepts. Although the medical concepts

obtained were more accurate compared with other studies, they were still not very informative about the given X-ray.

On the other hand, Singh et al. 2021 argued that the format of the normal and abnormal reports differs, therefore, a single framework cannot handle both styles accurately. To overcome this limitation, they followed a slightly different approach and first, they classified the reports as normal and abnormal. Then, they generated the Findings section and summarised it to acquire the Impression section for both normal and abnormal reports. They also adopted a pre-trained CNN model, InceptionV3, for visual feature extraction and used attention-based LSTM for text generation.

Later studies have taken advantage of Transformer for medical report generation, after its success for text generation based on non-linguistic representation. Xiong et al. 2019 designed a hierarchical Transformer model which contains a novel encoder that can extract the regions of interest from the original image by using a region detector and uses these regions to obtain visual representations. Moreover, Z. Chen et al. 2020 introduced a medical report generator via a memory-driven Transformer. They have used a relational memory to keep the knowledge from the previous case, in this manner, the generator model can remember similar reports when generating the current report.

Nooralahzadeh et al. 2021 proposed a progressive Transformer-based report generation framework that produces high-level context from the given X-ray and converts them into radiology reports by employing the Transformer architecture. Their proposed model consists of pre-trained CNN as a visual backbone, mesh-memory Transformer (Cornia et al. 2020) as a visual language model and BART (Lewis et al. 2019) as a language model.

3.5 Multi-modal and Multi-task Learning in Medical Imaging

Within the medical imaging field, the utilisation of multi-modal data has the potential to enhance performance in addressing complex tasks that exceed the capabilities of a single imaging modality. Concentrating on chest X-ray modality, multiple tasks such as image classification, image retrieval, and modality translation have leveraged data fusion strategies.

X. Wang et al. 2018 introduced a CNN-RNN architecture called the text-image-embedding network (TieNet) to extract discriminative representations of both chest radiographs and their accompanying reports by combining visual and textual information through joint fusion. The experimental results indicate that TieNet’s multimodal approach outperforms its unimodal counterpart in multi-label disease classification. Chauhan et al. 2020 employed a semi-supervised approach to train the network on chest radiographs and associated radiology reports to evaluate the severity of pulmonary edema. This study demonstrated that joint learning of image-text representations enhances the performance of models designed to predict the severity of pulmonary edema, compared with supervised models that relied solely on image-derived features.

Alfarghaly et al. 2021 employed a transformer-based language model conditioned on both visual features from the input medical image and embeddings representing relevant clinical attributes or findings. The key idea is to guide the report generation process by explicitly providing condition embeddings that encode specific clinical conditions as additional input, along with the image features.

Hayat et al. 2022 discussed the challenge of integrating data from different sources in healthcare due to the asynchronous collection of modalities. They proposed an LSTM-based fusion module, called MedFuse, that accommodates uni-modal and multi-modal input for mortality prediction and phenotype classification tasks. In contrast with intricate multi-modal fusion techniques, MedFuse yields considerably better performance on the fully paired test set, furthermore, it demonstrates robustness when dealing with the partially paired test set, which includes instances of missing chest X-ray images.

With increasing interest in this application domain, studies have become more attentive to the distinctions between image captioning and report generation tasks. As a result, researchers have begun to develop more knowledge-informed networks tailored specifically to the task of image-guided radiology report generation. L. Wang et al. 2022 introduced a task-aware framework that is designed to be adaptable to different imaging types and medical scenarios. It prioritises understanding specific diagnostic tasks related to various medical conditions, ensuring accurate and contextually relevant report generation.

Shuxin Yang et al. 2022 highlighted the significance of both input-independent general medical knowledge and input-dependent specific contextual information in generating accurate chest radiology reports. They proposed a knowledge-enhanced method that leverages this information

along with visual features to improve the quality and accuracy of generated reports for chest X-rays.

F. Liu et al. 2022 utilised a different learning approach, they proposed a method called competence-based multimodal curriculum learning for improving the generation of medical reports. They introduced a curriculum learning strategy that incorporates multimodal data, text and images, to enhance the competence of medical report generation models. The method aims to sequentially expose models to increasingly complex tasks based on their current performance level, thereby improving their overall competency in generating accurate and comprehensive medical reports.

Z. Wang et al. 2022 proposed a pure transformer-based model called TransRAD that consists of a vision transformer encoder for extracting visual features from the input image, and a transformer decoder. The model adopted multi-task learning, optimising for 1) generating accurate reports, 2) ensuring consistency with expert annotations, and 3) matching radiologists' writing styles. It combines cross-entropy loss for report generation, focal loss for finding classification, and cycle-consistency loss for style preservation. Wu et al. 2023 introduced a technique called multi-modal contrastive learning, which aims to enhance the synergy between different modalities of data. By leveraging contrastive learning, the proposed method aligns and embeds visual and textual representations in a shared space, facilitating the generation of more informative and accurate radiology reports.

Zhao et al. 2023 introduced a method for generating radiology reports that integrate medical knowledge and utilise multilevel alignment between medical images and textual reports. The approach aims to improve the accuracy and relevance of generated reports by aligning specific regions of interest in images with corresponding descriptive sections in the reports. Experimental verification demonstrates the effectiveness of the method in enhancing the coherence and informativeness of generated radiology reports, thereby potentially improving clinical decision-making processes.

Tanida et al. 2023 proposed an interactive and explainable approach that employs a region-grounded vision-language model that generates reports grounded on image regions identified by a region proposal network. The proposed model allows radiologists to interactively refine the generated reports by providing feedback on the region-text alignments. An explainable module interprets the model's decision process by visualising attention maps over the image and text.

Additionally, a reinforcement learning-based revision module enables the model to iteratively revise its outputs based on the radiologist’s feedback, promoting more accurate and coherent report generation. Their approach combines multimodal learning from images and text with interactive refinement, aiming to improve report quality and provide insights into the model’s reasoning process.

More recently, Jin et al. 2024 presented a PromptMRG model that employs a pre-trained vision-language model and generates radiology reports by conditioning it on both the input medical image and a dynamically constructed prompt. The prompt is designed to guide the model towards generating reports that are tailored to the specific diagnosis present in the image. The prompts are automatically constructed using a diagnosis prediction model, which identifies the most relevant diagnoses from the input image. The generated prompts are then concatenated with the image embeddings and fed into the vision-language model for report generation.

Performance on downstream tasks can be improved by training text generation models on multiple objectives simultaneously (Y. Zhang et al. 2019; Su et al. 2021). While multi-task learning has demonstrated its potential to enhance text generation performance, effectively training a neural network to produce coherent narrative text from diverse multi-modal and cross-modal inputs remains a complex and challenging task. Cross-modal learning, which involves models understanding connections across modalities, presents inherent difficulties. Nonetheless, multi-task learning can mitigate some of these challenges in cross-modal text generation by enabling models to jointly learn representations across modalities while optimising multiple objectives.

More specifically, image-to-narrative language generation in the medical domain is a more challenging task due to the diversity of objects in medical images, and the requirement for additional contextual information to analyse and interpret medical images, unlike the relatively straightforward nature of natural images. While multi-modal approaches have proven valuable in tackling some of these challenges in various vision language tasks such as visual question answering or medical report generation, there is a notable gap in exploring diverse multi-modal data and multi-task learning techniques for radiology report generation.

Chapter 4

MEDICAL ONTOLOGY-INFORMED RECURRENCE NETWORK

4.1 Introduction

Medical image captioning has started to attract attention from both natural language processing (NLP) and computer vision communities, following the success of image captioning. Researchers have attempted to generate radiology reports autonomously, which have the potential to improve the way of analysing and interpreting medical images. Traditional methods of image captioning, such as report retrieval and template-based generation, are limited in their ability to produce flexible and comprehensive textual descriptions that can be applied to new images. However, researchers have explored the automatic generation of medical reports using image captioning methods, which involves the use of deep learning models that can automatically write the findings and impression parts of medical reports of chest X-rays (CXRs) (Pang et al. 2023; Akhter et al. 2023).

In the literature, most existing deep learning approaches proposed for radiology report generation for given CXR images, leverage networks comprising a convolutional encoder and recurrent decoder, which was originally introduced for the task of image captioning. Although existing studies have shown promising results, they often treat the task of report generation as a

captioning problem.

However, report generation differs in several significant respects. Highlighting the similarities and differences with natural image captioning can help to identify challenges arising in the medical domain. Furthermore, these insights serve to emphasise key problems and identify gaps in current research efforts. Presenting the nature of the radiology report generation task by establishing encoder-decoder models grounded in natural image captioning and language processing techniques can assist in revealing the challenges involved in generating accurate reports from medical imaging.

This chapter aims to address the limitations of existing radiology report generation models by incorporating structured medical knowledge into the image captioning paradigm. By leveraging the power of medical ontologies, we aim to enhance the semantic representation of medical images and improve the accuracy of generated radiology reports.

The key contributions of this chapter are:

1. The development of a multi-input, multi-view encoder-decoder architecture enhanced with medical ontology terms, aiming to bridge the semantic gap between visual and textual representations. To the best of our knowledge, this is the pioneering study that utilises Medical Subject Headings (MeSH) indexing and integrates them within a unified framework for radiology report generation.
2. An extensive ablation study to uncover the impact of various model configurations by evaluating different architectural configurations. This analysis provides valuable insights to guide the development of more accurate solutions for generating clinical text from multi-modal inputs.

4.2 Data Preliminary Processing

The IU chest X-ray collection (Demner-Fushman et al. 2016) was used to carry out the experiments and evaluate the proposed approaches. The dataset provides findings, impressions, comparisons, indications sections for given chest X-rays and a list of labels (MeSH terms) identified by the Medical Text Indexer (MTI). Both comparison and indication sections cannot be completed by a radiologist in the absence of previous records, therefore, report generation algorithms are also not able to generate these sections. In addition, there was a lot of missing

information in both sections due to the anonymisation of the data. Therefore, the Impression and Findings sections along with the MeSH terms were primarily used in this phase of the study.

4.2.1 Data Formation

Text Data

To begin, the raw XML data (Figure 4.1) underwent parsing, denoted as $P(X)$, followed by the extraction of nodes yields the set: $E(P(X)) = \{A, PI, M\}$, where A represents the Abstract node, PI represents the ParentImage node, and M represents the MeSH node.

```

      .
      .
      .
<Abstract>
    <AbstractText Label="COMPARISON">None.</AbstractText>
    <AbstractText Label="INDICATION">Positive TB test</AbstractText>
    <AbstractText Label="FINDINGS">The cardiac silhouette and mediastinum size
are within normal limits. There is no pulmonary edema. There is no focal consolidation.
There are no XXXX of a pleural effusion. There is no evidence of
pneumothorax.</AbstractText>
    <AbstractText Label="IMPRESSION">Normal chest x-XXXX.</AbstractText>
      .
      .
      .
</Abstract>
</MedlineCitation>
<MeSH>
    <major>normal</major>
  </MeSH>
  <parentImage id="CXR1_1_IM-0001-3001">
      .
      .
      .

```

Figure 4.1: Example of Raw XML Data Structure from the Indiana University Chest X-ray Collection

The Abstract node encompassed sections such as Comparison, Indication, Findings, and Impression. In contrast, the ParentImage node contained identification names for each report, and the MeSH node provided a list of mesh terms represented as tags.

Addressing the issues mentioned earlier, the Indication and Comparison sections were removed from the primary dataset and excluded from the experiment. Furthermore, 104 reports were eliminated due to the unavailability of relevant X-rays within the collection.

The data cleaning process started with an initial step, which involved identifying and removing data points with missing or NaN values within the Impression, Findings, or Mesh sections. Subsequently, $D = \{d_1, d_2, \dots, d_n\}$ be the set of n data points in the initial dataset, where each data point d_i consists of three components: Findings f_i , Impression i_i , and Mesh terms m_i .

After applying the series of operations using regular expressions, we have $f'_i = \text{regex}(f_i)$, $i'_i = \text{regex}(i_i)$, and $m'_i = \text{regex}(m_i)$ where $\text{regex}(\cdot)$ represents the sequence of operations performed on the input text, including:

1. Converting all characters to lowercase
2. Eliminating punctuation, numbers, and undesired tags (e.g., "years old," "xxxx," etc.)
3. Expanding contractions (e.g., changing "won't" to "will not")

The Report variable for further experiments concatenates the Findings and Impression sections. Let F' , I' , and M' be the sets of unique Findings, Impression, and Mesh terms, respectively, $F' = \{f'_i \mid i = 1, 2, \dots, n'\}$, $I' = \{i'_i \mid i = 1, 2, \dots, n'\}$, $M' = \{m'_i \mid i = 1, 2, \dots, n'\}$. "Report" is $r_i = f'_i \oplus i'_i$, where \oplus denotes the concatenation operation. The dataset formation resulted in the unique numbers of Report and Mesh terms $|R'| = 2314$, and $|M'| = 1552$, respectively.

As a case in point, the Report derived from the data illustrated in Figure 4.1 was transformed to "the cardiac silhouette and mediastinum size are within normal limits there is no pulmonary edema there is no focal consolidation there are no of a pleural effusion there is no evidence of pneumothorax normal chest".

Index	Mesh Term	Count
1	normal	1124
2	no indexing	68
3	lung hypoinflation	39
4	thoracic vertebrae degenerative mild degenerative change	21
5	spine degenerative mild degenerative change	17
6	thoracic vertebrae degenerative degenerative change	15
7	spine degenerative degenerative change	14
8	granulomatous disease granulomatous disease	12
9	cardiomegaly mild cardiomegaly	10
10	lung hypoinflation markings bronchovascular	8

Table 4.1: Most Frequently Occurring Mesh Terms in the Final Dataset

Index	Report	Count
1	no acute disease the heart is normal in size the mediastinum is unremarkable the lungs are clear	45
2	no active disease the heart and lungs have in the interval both lungs are clear and expanded heart and mediastinum normal	43
3	no acute cardiopulmonary abnormality the lungs are clear bilaterally specifically no evidence of focal consolidation pneumothorax or pleural effusion cardio mediastinal silhouette is unremarkable visualized osseous structures of the thorax are without acute abnormality	35
4	normal chest heart size normal lungs are clear are normal no pneumonia effusions edema pneumothorax adenopathy nodules or masses	35
5	no active disease both lungs are clear and expanded heart and mediastinum normal	29

Table 4.2: Top 5 Reports (Impression and Findings Compilation) with Highest Occurrence in the Final Dataset

Image Data

Throughout the dataset, the number of images associated with a single report ranged from a minimum of one to a maximum of five, with the most common count being two. These X-rays exhibited varying shapes and were captured in either lateral or frontal views (please refer to Figure 2.11). To address this variability, each frontal image was paired with a related lateral image, creating new data points. However, in cases where no lateral view was available, the data point(image-report pair) was excluded. As a result, 3532 samples with multi-view X-rays were generated.

4.2.2 Data Pre-processing

In the text data pre-processing phase, the first step involved tokenisation, breaking the texts into manageable units. Special "`<start>`" and "`<end>`" tokens were added to the beginning and end of each Report sequence. Let S be the input sequence, which consists of n characters. After tokenisation, S is divided into a sequence of tokens $T = \{\langle\text{start}\rangle, t_1, t_2, \dots, t_m, \langle\text{end}\rangle\}$, where each token t_i represents a meaningful unit in the context of the application.

$$S = c_1 c_2 \dots c_n \rightarrow T = \{\langle\text{start}\rangle, t_1, t_2, \dots, t_m, \langle\text{end}\rangle\}$$

where c_i denotes the i -th character in the input sequence S , and t_i represents the i -th token in the tokenised sequence T . These markers facilitate the decoding process. Additionally, any out-of-vocabulary words were promptly replaced with "`<unk>`" to ensure a consistent and comprehensible vocabulary. Finally, each Report sequence was padded to a fixed length to meet

the language model's requirement for consistent input size. The Mesh terms also underwent preprocessing steps similar to the Report sequences. They were tokenised, and the resulting sequences were padded to a fixed length.

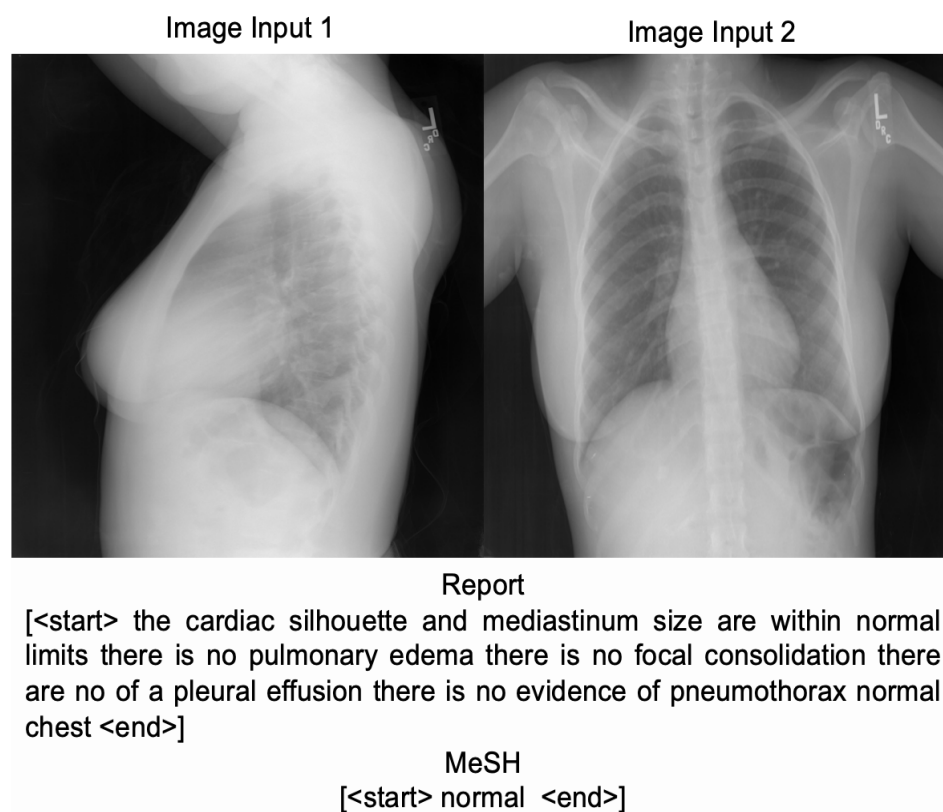


Figure 4.2: Example of the final processed text data(with start/end tokens) and input images for model input

For image processing, all images were resized using interpolation to match CNN's expected format of px by px, and each pixel value was normalised by dividing it by 255. Finally, a data loader object was created to fetch data from the dataset, and it was then fed into the model in batches.

4.3 Network Configuration

4.3.1 Encoder and Decoder Variants

Several CNN-based deep neural network models were developed and successfully fine-tuned for various medical image understanding tasks such as classification, segmentation, detection, localisation, and diagnostic captioning (Sarvamangala and Kulkarni 2022). ImageNet CNN models are commonly preferred for diagnostic captioning due to their good image encoding and

feature extraction performance (Pavlopoulos et al. 2022). Although these models are pre-trained on natural images, they have made significant contributions to medical image captioning tasks. Considering that, we implemented the two most commonly used ImageNet architectures, along with one domain-specific and one straightforward CNN architecture:

- A custom CNN with 3 convolutional layers and 2 dense layers, trained from scratch on our dataset.
- InceptionV3 (Szegedy, Vanhoucke, et al. 2016) - a pre-trained CNN known for efficiency and modest computational requirements.
- VGG19 (Simonyan and Zisserman 2014)- a pretrained CNN with greater depth and representational power.
- ChexNet (Rajpurkar et al. 2017) - an architecture that was pre-trained with the ChestX-ray14 dataset (Szegedy, Vanhoucke, et al. 2016), which consists of over 112,120 chest X-rays for the task of pneumonia detection.

For the decoder, we experimented with Bidirectional GRU by the use of visual attention and pre-trained word embeddings. Bidirectional RNNs allow for capturing context from both past and future directions. This bidirectional nature enables the model to better understand long-range dependencies between words in the caption, leading to more consistent descriptions of the image. Additionally, Bidirectional RNNs are better combined with visual attention mechanisms and pre-trained word embeddings which provides it with a good starting point for capturing syntactic relationships between words. Several studies have demonstrated that incorporating Bidirectional RNNs can lead to improved performance on standard evaluation metrics compared to using unidirectional RNNs (X. Chen and Lawrence Zitnick 2015; C. Wang et al. 2016).

4.3.2 Multi-view vs. Single View

Radiologists typically generate diagnostic reports by examining X-rays from various angles, as multiple X-rays can enhance the accuracy and reliability of their findings. In this context, we investigated three primary strategies: single view, duplication, and multi-view, aiming to determine the most effective approach.

- Single view: Each X-ray image and the corresponding report was treated independently, without linkage between different views of the same patient.

- Duplication: Both views for each patient were combined and presented to the model. In this approach, the model expects two input points. If the patient has only one X-ray, a replica of the existing image was created to obtain two data points.
- Multi-view: This approach neglects the single-view X-rays, and each of the frontal views is assigned as the first input and the lateral view is fed into the model as the final input.

4.4 Medical Ontology-Informed Multi-View Model

The network architecture was developed through experimentation with multiple approaches mentioned in Section 4.3. To keep it concise, only the final model components (illustrated in Figure 4.3) are presented with all implementation details.

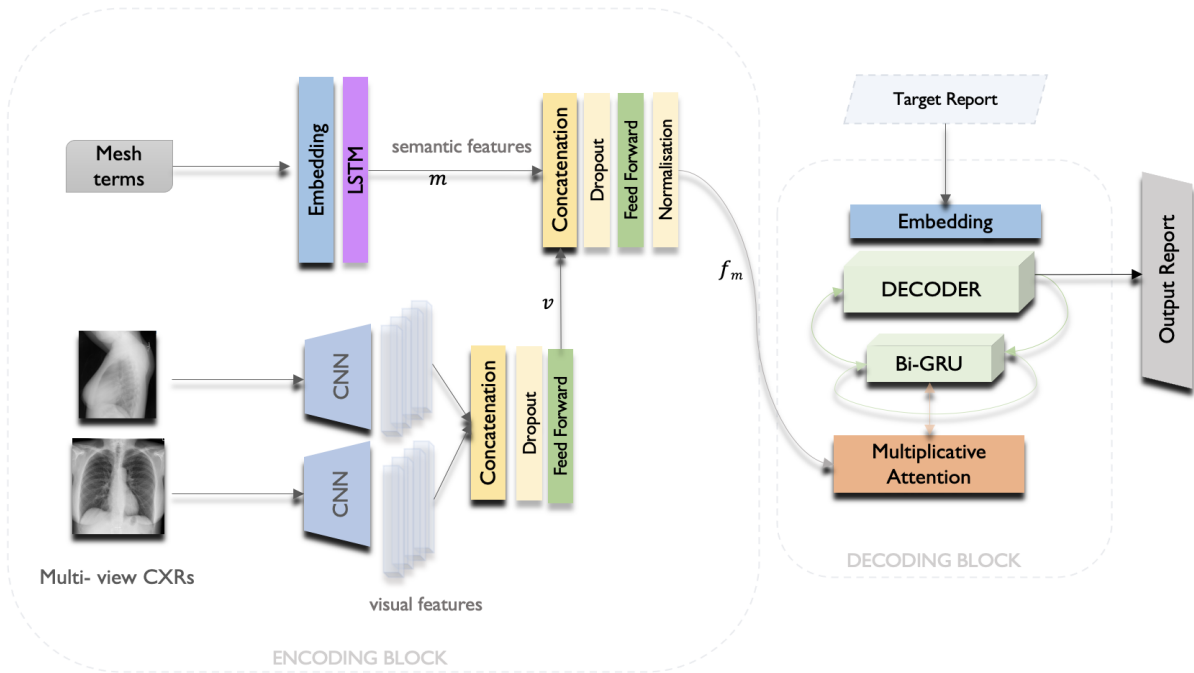


Figure 4.3: The overall multi-view multi-input recurrence report generation framework

The CNN model functions as the main encoder component for image representation extraction. Firstly, the pre-trained CheXNet model is loaded and reconstructed to output the activations of the layer just before the global average pooling layer. It is also configured to freeze all layers except for the last 10 layers. Both input views, frontal (I_f) and lateral (I_l), are processed through the CheXNet model, denoted as $f_{CheXNet}(\cdot)$, to obtain their respective feature representations.

These features from both views are then concatenated (denoted by \oplus) and passed through dense layers $f_{dense}(\cdot)$, resulting in the final encoded vector v representing the image features: $v = f_{dense}(f_{CheXNet}(I_f) \oplus f_{CheXNet}(I_l))$, where v is the final encoded vector representing the image features.

The resulting vector is then merged with Mesh features into a single vector representation and normalised using a normalisation function $\text{Norm}(\cdot)$ to obtain the mesh-enriched feature vector f_m , where M is associated mesh-terms and $m = \text{LSTM}(f_{\text{embedding}}(M))$, given by: $f_m = \text{Norm}(f_{dense}(m \oplus v))$.

The Target Report is first processed by an embedding layer, which converts the tokens into dense vector representations. These embedded vectors are then fed into the Decoder Block. A bidirectional GRU layer further processes the embedded input sequence within the Decoder Block. The output of this bidirectional GRU is concatenated with a context vector obtained from the attention mechanism.

The attention mechanism, specifically the Multiplicative Attention layer, computes this context vector. It does so by calculating a weighted sum of the encoder output (f_m), where the weights (attention scores) are determined by the similarity between the current decoder state and each encoder output.

This context vector provides additional information to the decoder, helping in the generation of the output sequence. Therefore, it is used as an additional input to the decoder, alongside the embedded input sequence. The Decoder Block preserves the ground truth sequences, forward and backwards hidden states of the bidirectional GRU, as well as the attention weights computed by the attention mechanism. It also applies dropout regularisation to the output of the bidirectional GRU, helping to mitigate overfitting.

Finally, the output logits, representing the predictions for the next token in the sequence, are obtained by passing the output through a dense layer.

In summary, this end-to-end approach combines; the processing of MeSH terms to provide semantics, extraction of image representations using CheXNet, and finally, a two-step decoding process to generate structured, detailed radiology reports from the medical images. The hybrid combination of components aims to provide both key medical semantics and visual feature understanding.

4.5 Experimental Setup

4.5.1 Training and Inference

The training of the proposed model was conducted to optimise its performance on the training and validation datasets. We followed a systematic approach for the selection of hyperparameters and implementation of the suitably optimised pipeline. The hyperparameters were tuned to balance between model complexity and generalisation.

Additionally, teacher forcing was employed during the training process in the text generation phase. It involves feeding the ground truth output from the previous time step as input to the model for the next time step in the decoder, rather than the model’s own predicted output. Let $\mathbf{x} = (x_1, x_2, \dots, x_T)$ be an input sequence of length T , and $\mathbf{y} = (y_1, y_2, \dots, y_T)$ be the corresponding target output sequence.

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^T P(y_t|y_1, \dots, y_{t-1}, \mathbf{x}) \quad (4.1)$$

Specifically, the conditional probability at time step t is computed as:

$$P(y_t|y_1, \dots, y_{t-1}, \mathbf{x}) = f_{\theta}(y_1, \dots, y_{t-1}, \mathbf{x}) \quad (4.2)$$

During inference, the model generates the output sequence one step at a time, using its own predicted output from the previous time step as input for the current time step: $\hat{y}_t = \operatorname{argmax}_{y_t} P(y_t|\hat{y}_1, \dots, \hat{y}_{t-1}, \mathbf{x})$, where \hat{y}_t is the predicted output at time step t . This provides additional guidance to the model and can help to stabilise training for sequence generation. For the loss function, we used SparseCategoricalCrossentropy loss, which allowed the model to effectively learn text patterns in the data. It is calculated as: $L(y_{true}, y_{pred}) = -\log(y_{pred}[y_{true}])$, where y_{true} is the true word index (label) at the current time step of the sequence, y_{pred} the predicted probability distribution over the entire vocabulary (output of the softmax function) at the current time step, and $y_{pred}[y_{true}]$ is the predicted probability for the correct word index at the current time step.

The key hyperparameters stated below are the final values selected after experimentation with different values to optimise performance:

- **Optimiser and Learning Rate:** Adam optimiser with a learning rate of $1e-3$ was chosen to ensure a stable convergence of the model during training.
- **Batch Size:** A batch size of 32 was employed to optimise computational efficiency.
- **Epochs:** The model was trained for 50 epochs with an early stop triggering value set to 10, ensuring sufficient exposure to the entire dataset while preventing overfitting.
- **Dropout Rate:** A dropout rate of 0.2 was applied to all dense layers, contributing to regularisation and preventing overfitting.
- **Maximum Sequence Length:** Sequences of a maximum length of 80 were considered during both the training and validation phases.
- **Embedding and Dense Dimension:** A dense and embedding dimension of 64 was chosen, influencing the internal representation of the model.

4.5.2 Evaluation Details

We conducted an ablation study to evaluate the impact of different architectural components mentioned in Section 4.3. Specifically, we kept the MeSH component and decoder constant across all experiments, while systematically varying the encoder-decoder variant and input-view representation.

Then, we assessed the performance with and without incorporating MeSH terms, using the optimal input and network configuration identified through our experiments. This analysis provided insights into the importance of architectural selection and the utility of leveraging domain-specific knowledge through MeSH term integration.

To quantitatively assess the model’s performance and facilitate comparative analysis across different configurations, we employed BLEU-1 to BLEU-4 scores as the evaluation metrics. This choice was motivated by the widespread adoption of BLEU scores as the standard evaluation metric in the domains of report generation and image captioning tasks.

To complement the quantitative evaluation and gain deeper insights into the model’s performance, we also conducted a qualitative human analysis on a sample of 50 generated outputs.

4.6 Component-wise Performance Analysis

As stated in 4.3.2, radiologists create a single report by using different views of X-rays, multiple X-rays may be required in some cases to obtain more accurate and reliable results. This approach was mimicked while determining image input points, and three main strategies, single view, duplication, and multi-view, were implemented to find the most effective result.

Data/Sample Size	Input Image Format	Average BLEU Score
7470	Single-view	0.267
3978	Multi-view (with duplication)	0.284
3532	Multi-view (without duplication)	0.271

Table 4.3: The average BLEU scores obtained from the five-fold cross-validation results across different input views.

In the single view approach, each X-ray (without distinction as frontal or lateral) and related report were treated as independent data. However, in this approach, inconsistent results were obtained for different X-rays of the same patient. Furthermore, the characteristic differences in views caused difficulties in training, leading to reduced performance. In order to negate this issue, both views for each patient were combined and presented to the model. In this case, the model expects two input points. If the patient has only one X-ray, a replica of the existing image is created, and two data points are thus obtained. With this duplication method, the problem of inconsistency between the results has been solved and the model performance increased by 6.5%.

However, since this technique is not accurate from a clinical perspective and does not provide additional information to the network, it was not used in the final model. The final method, the multi-view strategy, neglects the single X-rays, and each of the frontal views is assigned as the first input and the lateral view is fed into the model as the final input.

All hyperparameters were kept constant across all experiments, with only the input view and sample amount changing accordingly. Table 4.3 displays the average Bleu Score of the five-fold cross-validation results for each input view shape.

After exploring different strategies for handling input views, the next focus was on the impact of different CNN architectures on model performance. Table 4.4 reports the average Bleu n-gram scores for the baseline CNN model versus pre-trained models, evaluated using three-fold cross-validation, whereas the simple CNN model was only evaluated in one validation set. The simple CNN employs an approach of training from scratch, without incorporating any transfer learning or pre-trained models.

CNN Model	B_1	B_2	B_3	B_4
Simple CNN	0.25	0.11	0.09	0.04
InceptionV3	0.31	0.20	0.15	0.10
VGG-19	0.29	0.19	0.14	0.10
CheXNet	0.36	0.25	0.18	0.13

Table 4.4: Comparative Performance Analysis of Baseline and Pretrained Models Using BLEU-n Scores, with 'B_n' Representing BLEU-n Scores

Among visual feature extraction pre-trained models, CheXNet produced the most consistent and reliable results, followed by InceptionV3 and VGG-19.

4.7 Qualitative and Quantitative Results

After finalising the model architecture by selecting the input-view and encoder-decoder components, the next step explored was incorporating Medical Subject Headings (MeSH) terms into the training process, to help the model better understand key semantic relationships in the reports. It was hypothesised that utilising these MESH terms as additional context would enhance the model's relevance and accuracy.

	B_1	B_2	B_3	B_4
Baseline Model	0.32	0.22	0.17	0.12
Mesh-enriched Model	0.36	0.25	0.18	0.13

Table 4.5: Qualitative Results of Baseline and Mesh-enriched Model Models

Table 4.5 presents the quantitative evaluation results, contrasting the performance of the baseline model against the proposed MeSH-enriched model. The baseline configuration comprises the ChexNet vision feature extractor and a multi-view input representation without duplication. The MeSH-enriched model corresponds to the final proposed architecture (Figure 4.3), which incorporates the MeSH-generator component into the pipeline.

The results show that the Mesh-enriched model performed better in all metrics compared to the baseline model. Therefore, we assessed the statistical significance of the observed performance improvements by conducting pairwise t-tests for each evaluation metric. The results indicate that the improvement in BLEU-1 scores achieved by the MeSH-enriched model is statistically significant, while the differences in other metrics did not show statistical significance.

We further evaluated the Mesh-enriched model by comparing the generated text with the ground truth. As the model’s predictions were not meaningful for completely abnormal cases and did not provide useful insights, we excluded those examples from the visual results presented in Figure 4.4.

The figure demonstrates that some diagnoses were missing or inaccurate in the generated report or statements that were not in the original report were produced. In the first example displayed, “the mediastinum is unremarkable” is generated by the model, however, it was not mentioned in the original report. In the second sample, although the original report has the statement “The lungs are clear without infiltrate”, the generated report does not contain any information about the lungs.

	Ground Truth	Generated Report
	<p>Normal chest. Lungs are clear. The cardiomeastinal silhouette is normal. No pleural effusion is identified.</p>	<p>No acute cardiopulmonary abnormality. Heart size is normal. The mediastinum is unremarkable. Lungs are clear.</p>
	<p>The heart and mediastinum are unremarkable. The lungs are clear without infiltrate. There is no effusion or pneumothorax.</p>	<p>Heart size within normal limits. No focal airspace consolidation no pleural effusion or pneumothorax is seen. There is no evidence of pneumothorax.</p>
	<p>The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.</p>	<p>The heart size and mediastinal contours appear within normal limits. No focal consolidation suspicious pulmonary opacity pneumothorax or definite pleural effusion seen. No typical findings of pulmonary edema.</p>

Figure 4.4: Illustration of Ground Truths and Example Reports Generated by MeSH-Enriched Multi-view Model

Human Analysis

A 50-sample subset of the test data was evaluated through human analysis for each network configuration. This analysis compared the model-generated text to the ground truth radiology reports, examining differences and similarities in grammar, sentence structure, word choice, and the correctness of detecting medical abnormalities.

Notable improvements were not observed between multi-view and single-view approaches, despite evaluation metrics indicating enhancement. Among pre-trained models, a slight improvement in normal cases was observed with the ChexNet model incorporation. However, performance in detecting abnormal or unique cases remained largely similar across all pre-trained models used. Additionally, semantically identical sentences were formed in structurally different ways due to linguistic diversity in the training dataset. For instance, “within normal limit” and “normal” mostly convey equivalent meanings, however, they are represented differently which resulted in a lower score. Qualitative analysis on a 50-sample subset from the MeSH-enriched model’s output revealed that 36% of the generated texts matched the corresponding report exactly and 34% exhibited some missing information. The model hallucinated at 30%, resulting in producing extraneous statements not present in the original reports. Furthermore, as the target reports become longer, the generated reports exhibit limited context awareness. Among the 32 reports that show missing information or hallucinations, 24 contain long sentence structures.

Based on this analysis, the MeSH-enriched model demonstrated better performance compared to the image-alone baseline for only normal cases. However, utilising MeSH terms did not significantly contribute contextual information. This could be because the terms predominantly consisted of brief descriptive information that already existed in the radiology reports.

4.8 Discussion and Conclusion

This chapter demonstrates several key challenges and insights related to medical image captioning, especially, compared to natural image captioning. Generating coherent and structured paragraphs from medical images requires capturing higher-level semantics and context beyond short image captions. In this context, the choice of decoder architecture becomes crucial to handle the long sequences effectively. However, recurrent models with memory still suffer from vanishing and exploding gradients and have limitations on parallel computing. Also, the image representations are only utilised in the first phase of the decoder, and the further layers do not

use the visual information.

The distinctions between medical images are subtle and are not as straightforward to deal with as natural images. Furthermore, while the information needed to describe a natural image is often embedded within the image itself, additional contextual information is typically required to effectively analyse and interpret a medical image.

To leverage more visual insights and potentially improve the performance, we have explored the use of multi-view inputs compared to single-view inputs. Despite observing better quantitative evaluation metrics with the multi-view approach, the qualitative improvements in the generated captions were minimal. The integration of pre-trained models like CheXNet yielded only slight improvements for normal case reporting, while abnormal cases remained challenging.

While incorporating medical ontology information helped provide supplementary context, it did not offer significantly different knowledge than what was already present in the medical reports themselves. It is also important to note that most of the Mesh annotations indicated the normality. The network requires more explicit and descriptive information to fully understand and interpret the given image input. As evidenced by 30% of the generated reports containing extraneous statements, and 34% of them exhibiting some missing information, the model struggled with comprehensively understanding the key information needed from the images.

The human analysis demonstrates the importance of evaluating AI-generated medical reports, showing that qualitative assessment can reveal nuances that might be missed by automated metrics alone. It highlights that improvements in Natural Language Generation (NLG) metrics don't necessarily translate to clinically significant improvements. This is an important contribution as it emphasises the need for caution when interpreting automated evaluation metrics in medical AI applications.

To achieve more accurate and comprehensive medical image captioning, integrating more descriptive domain knowledge sources is required. This could involve leveraging structured medical knowledge bases, incorporating more detailed visual annotations, or developing techniques to better fuse multi-modal information from images and text reports. Additionally, advances in model architectures and training strategies that can effectively capture long-range dependencies and handle complex multi-modal inputs are necessary.

Chapter 5

MULTI-MODAL INTEGRATIVE ATTENTION NETWORK

5.1 Introduction

Deep learning techniques have been increasingly employed in clinical medicine over the years, due to their ability to process significant volumes of medical images, and large datasets, and guide/assist evaluate the effectiveness of diagnosis and treatment for many important diseases (Pei et al. 2023). By combining different types of medical data, deep learning models can efficiently extract and combine salient features from multi-modal sources. This data fusion approach improves real-world applicability in medical diagnosis and evaluation, enables quantitative analysis and informs treatment planning.

However, the majority of current deep learning approaches use methods for CXRs that solely consider the radiology image as input and disregard the non-imaging information that radiologists have access to during image interpretation. Only a limited number of studies integrate additional data into the network such as medical concepts, high-level contexts or categories of the images/reports. While these methods have shown some level of success, they mainly focus on enhancing the model with data derived from existing semantics rather than supplementing the training context with additional data. Furthermore, due to the nature of CXR images as 2D representations of 3D entities, there is a loss of important/relevant information in the data available for learning by networks or algorithms.

Therefore, we hypothesised that employing a data-driven learning-based approach, which integrates information from multiple modalities, would provide a more comprehensive representation of the patient’s clinical condition. By leveraging complementary perspectives from diverse data sources, this multi-modal approach could mitigate the limitations inherent to relying solely on visual features extracted from imaging data. Consequently, a holistic representation would allow the model to generate more informative and accurate reports compared to using only CXR images.

5.2 Dataset Formation

The MIMIC database was used to carry out the experiments and evaluate the proposed approaches. MIMIC-CXR, MIMIC-IV, and MIMIC-IV-ED datasets comprise distinct tables containing varying details related to a patient’s hospitalisation. An individual patient is assigned a unique identifier, referred to as the subject ID. However, since a single patient might have multiple hospitalisations, or a single stay may generate several records, linking these databases using subject ID proved unfeasible.

Moreover, as the aim is to generate an accurate report, it is imperative that non-imaging data be collected within the same time frame as the chest x-ray. Consequently, we resorted to record linkage between MIMIC-CXR and MIMIC-IV-ED databases and extracted data only if the patient was in the ED while the report was being generated and did not leave during that period.

After performing data cleaning procedures (following the same approaches as described in Section 4.2.1), filtering the images to only include anteroposterior and posteroanterior projections, keeping only one study if there is more than one record, removing replications, the resulting dataset contains 65813 entries and 11 features including acuity level, oxygen saturation, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, temperature, patient’s chief complaint, ICD title, gender and ethnicity.

One challenge with the dataset for this task is its biases, particularly, its skewed distribution towards normal cases and the presence of numerous identical reports for different patients. To minimise these issues, we selected a subset of 65813 entries, by identifying and cataloguing unique medical reports, ensuring that each distinct group was represented in the curated dataset to a similar extent. This subset consisted of 3000 total samples, which we further divided into a

training set comprising 2100 data points and a validation set comprising 900 data points. 3,000 samples were selected due to computational efficiency, as utilising the full dataset resulted in excessive training time without yielding significant improvements in the results. Subsequently, we evaluated the performance of our models on a holdout test set comprising 1173 unseen examples. As there is currently no comparable comprehensive dataset encompassing similar non-imaging data, we exclusively employed this specific dataset to train and assess the proposed approach.

5.3 Feature Extraction and Pre-processing

This section describes the pre-processing and encoding of different data modalities used in this study. The main objective is to align the data used in the study and the data typically encountered in medical practice while minimising potential biases that may arise.

5.3.1 Image Data

Each image went through resizing to 299 pixels x 299 pixels, followed by min-max normalisation to scale the intensity values to a range of 0 to 1. The process of obtaining the representation of each image can be described as a two-step procedure. In the first step, the EfficientNet model is utilised as the base model to extract the visual features of the image. In the second step, this feature vector is employed as the input for a transformer-based encoder which extracts higher-level features and fuses this information with clinical and non-clinical(demographic) data. A detailed explanation of the fusion process can be found in Section 5.4

5.3.2 Clinical: Non-imaging Data

This study exclusively employed clinical data that clinicians considered during patient evaluations, wherein, a chest X-ray examination was conducted if any disease/abnormalities were suspected. These data included heart rate, respiratory rate, oxygen saturation, temperature, level of acuity (severity), primary symptoms or complaints, as well as known or suspected diseases.

The acuity level of a patient is determined based on the triage assessment, and an integer value, *between 1 and 5*, is assigned to each case where 1 indicates the least severe and 5 is the most severe. The higher acuity levels are typically associated with the presence of abnormalities in the patient's case, therefore, utilising the acuity level may assist the network in determining

normal and abnormal cases while generating the report.

The integer-based variables including oxygen saturation, heart rate, respiratory rate, systolic blood pressure (sbp), and diastolic blood pressure (dbp) were initially treated to remove outliers. Subsequently, the values have been normalised within the range of 0 to 1, based on their respective minimum and maximum values. As for temperature data, a conversion to Fahrenheit scale was performed, and similar to the integer-based variables, the values were normalised between 0 and 1.

The text-based variables, namely the chief complaint and ICD title variables, are initially processed by converting characters to lowercase and removing unnecessary punctuation, such as commas, periods, and newline characters, utilising regular expressions. Consecutive periods are condensed into single spaces, and double periods are substituted with single spaces, contributing to a more consistent text format. The resultant text undergoes further standardisation by substituting shorthand phrases or abbreviations with their corresponding full-text counterparts. For example, 'cp' is replaced with 'chest pain', 'sob' or 'shortness of breath' is replaced with 'dyspnea' and so on. Standardisation also includes converting phrases like "chest pain, dyspnea" into "chest pain and dyspnea" as well as fixing typos and pluralisation issues such as changing "fevers" to "fever."

5.3.3 Non-clinical Data

In addition to clinical data, patient records often include non-clinical metadata that can provide valuable insights. This study concentrates on two commonly collected non-clinical variables: gender, and ethnicity. These variables have been demonstrated to have an impact on health outcomes (Aksoy et al. 2023) and are therefore of particular interest in this study.

As the gender data is already in binary format, the only necessary pre-processing step was to convert the data to a numerical representation by replacing 'Male' with 0 and 'Female' with 1.

The ethnicity data was initially categorised into 5 broad groups consisting of the most frequently occurring values and this initial categorisation slightly improved model performance. The data was then categorised in a more granular fashion into 9 groups: White, African American, Hispanic/Latino, Black, Asian, White/European, Russian, Other, and Unknown. We hypothesised that employing these more detailed ethnicity categories would enable more accurate report generation. Subsequently, the categorical ethnicity data was mapped to integer values and reshaped

into a 2D array to allow for input into the encoder.

5.4 Multi-modal Fusion Details

This section provides the details of the multi-modal data fusion strategy utilised in our proposed model. As shown in Figure 5.1, our approach employs a cross-attention mechanism to fuse the textual, visual, and scalar modalities. The key variables used in the fusion are defined in Table 5.1. Specifically, the scalar patient data, comprising attributes like heart rate, oxygen saturation, respiratory rate, blood pressure, temperature, acuity level and gender, is concatenated to form a continuous representation. This continuous data is then passed through a dense layer to produce a scalar output.

Variable	Description
$x_{\text{text_data}}$	2D input tensor of text data indices
$V_{\text{text_data}}$	Vocabulary size of text data
E	Embedding dimension
W_{emb}	Embedding weight matrix
f_{embed}	Embedding function
$X_{\text{chief_embed}}$	Embedded chief complaint data
$X_{\text{icd_embed}}$	Embedded ICD title data
X_{scalar}	Processed scalar patient data
X_{eth}	One-hot encoded ethnicity data
X_{patient}	Unified patient representation
Q	Query matrix for attention
K	Key matrix for attention
V	Value matrix for attention

Table 5.1: List of Variables Used in Multi-Modal Fusion Strategy

Each ethnicity group variable is transformed using the one-hot encoding (Equation 5.1), resulting in a matrix where each individual’s ethnicity is represented as a binary vector.

$$X_{\text{eth}} = [\delta(\text{eth}, 1), \delta(\text{eth}, 2), \dots, \delta(\text{eth}, 9)]. \quad (5.1)$$

where the function $\delta(i, j)$ is defined as

$$\delta(i, j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The chief complaint and ICD title data consist of text sequences with varying lengths and

vocabulary sizes. Therefore, these data are separately embedded using the following embedding technique before being further processed through dense layers.

Let,

$x_{\text{text_data}} \in \mathbb{Z}^{N \times M}$ — 2D input tensor of indices

$V_{\text{text_data}}$ — vocabulary size

E — embedding dimension

$W_{\text{emb}} \in \mathbb{R}^{V_{\text{text_data}} \times E}$ — embedding weight matrix

Then,

$f_{\text{embed}}(x_{\text{text_data}}) \in \mathbb{R}^{N \times M \times E}$

$$f_{\text{embed}}(x_{\text{text_data}})_{i,j,k} = W_{\text{emb}}[x_{\text{text_data}_{i,j}}, k] \quad (5.2)$$

where

$f_{\text{embed}}(x_{\text{text_data}})_{i,j,k}$ — embedding vector for token at position (i, j)

$x_{\text{text_data}_{i,j}}$ — integer index of token at position (i, j)

$W_{\text{emb}}[x_{\text{text_data}_{i,j}}, k]$ — k -th value from row of W_{emb} for index $x_{\text{text_data}_{i,j}}$

$$X_{\text{chief_embed}} = f_{\text{embed}}(x_{\text{chief_data}})$$

$$X_{\text{icd_embed}} = f_{\text{embed}}(x_{\text{icd_data}})$$

Where: $x_{\text{chief_data}}$ and $x_{\text{icd_data}}$ are the respective input indices tensors and f_{embed} is the embedding function defined in Equation 5.2.

After feature extraction and transformation of patient data inputs, the representations are concatenated into a unified patient representation vector. The processed scalar data output $X_{\text{scalar}} \in \mathbb{R}^{N \times M_{\text{scalar}}}$, one-hot encoded ethnicity output $X_{\text{eth}} \in \mathbb{R}^{N \times M_{\text{eth}}}$, and embedded chief complaint and ICD title outputs $X_{\text{chief_embed}}, X_{\text{icd_embed}} \in \mathbb{R}^{N \times M \times E}$ are concatenated for each

patient giving:

$$X_{\text{patient}} = \text{Concatenate}(X_{\text{scalar}}, X_{\text{eth}}, X_{\text{chief_embed}}, X_{\text{icd_embed}}; \text{axis} = 1) \quad (5.3)$$

Where the $\text{Concatenate}()$ operation joins the input tensors along the specified dimension, in this case, $\text{axis}=1$, yielding:

$$X_{\text{patient}} = \begin{bmatrix} X_{\text{scalar}} & X_{\text{eth}} & X_{\text{chief_embed}} & X_{\text{icd_embed}} \end{bmatrix} \in \mathbb{R}^{N \times (M_{\text{scalar}} + M_{\text{eth}} + 2M \times E)} \quad (5.4)$$

The resulting X_{patient} contains a unified representation of each patient's data for further use, combining structured scalar variables, categorical encodings, and semantically rich embedded features into a single vector. This concatenation enables the joint modelling of heterogeneous data types into an integrated patient representation.

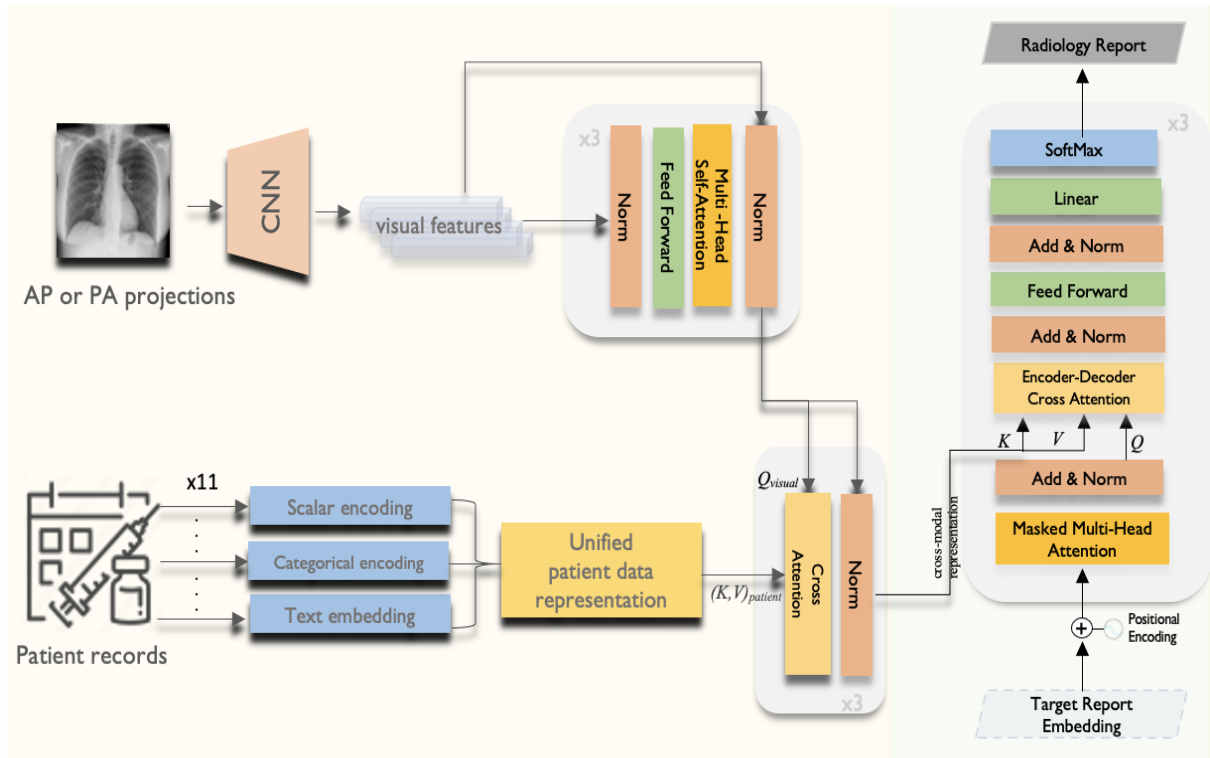


Figure 5.1: The overall multi-modal data fusion with the cross-attention framework of the proposed CXR report generation model.

Then, an EfficientNetB0 CNN backbone (Tan and Le 2019), pre-trained on ImageNet, extracts features from $299 \times 299 \times 3$ RGB input images. The choice of EfficientNet was motivated by its computational efficiency and scalable architecture, which are important considerations in

medical applications considering that computational resources might be limited or costly.

EfficientNet models maintain computational efficiency through a combination of neural architecture search and compound scaling techniques (Tan and Le 2019). This is achieved by concurrently scaling up the resolution, depth, and width of the network, ensuring a good balance between model size and accuracy. Furthermore, EfficientNet is used because it is capable of extracting meaningful hierarchical features from images due to its depth and width scaling factor, which is necessary for generating accurate and descriptive medical reports (Marques et al. 2020).

The CNN outputs $N \times D$ image embeddings, where N is batch size and D is the feature dimension. This 1280-length visual feature vector is transformed via layer normalisation and a dense layer to refine the image representation. Before starting to fusion operation, multi-headed self-attention (Equation 5.6) is then applied to enable the model to jointly focus on different positions in the image via parallel heads.

The self-attention outputs are then added to the original embedded image via residual connection, and normalised by a layer norm layer.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5.5)$$

where $\sqrt{d_k}$ is the dimension of the key vector k and query vector q

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (5.6)$$

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

The final output image embedding is further contextualised with information from the entire set of patient data via a cross-attention mechanism. In the cross-attention module, the convention is to take the image features as the query (Q) and the unified patient representation X_{patient} as the key (K) and value (V). This allows each part of the encoded image embedding to attend to relevant semantics from the full patient data:

$$Q = \text{Image Features} \in \mathbb{R}^{N \times D}$$

$$K = V = X_{\text{patient}} \in \mathbb{R}^{N \times D}$$

Where N is the batch size and D is the common embedding dimension across modalities.

Multi-headed scaled dot-product attention is again applied between Q and K to obtain attention weights representing the relevance of each part of the patient data to each part of the image. The weighted value matrices are concatenated and projected to obtain the cross-attention outputs allowing the model to condition each part of the image embedding on relevant unified patient representation. The cross-attention outputs are residually connected and normalised in a similar manner via element-wise addition with the output image embedding from the previous self-attention block and layer norm.

We adopt the canonical Transformer decoder architecture as it has proven effective in various sequence-to-sequence tasks, including image captioning (Vaswani et al. 2017). The decoder starts by embedding the input sequence using both target (token) embeddings and positional embeddings. Target embeddings provide the meaning of words, while positional embeddings provide information about the order of tokens in the sequence.

The initial layer employs self-attention, which is the key component that enables the Transformer decoder to effectively generate the output sequence. Self-attention allows the decoder to capture long-range dependencies within the text and ensures the generation of contextually relevant reports. This is achieved by having each output token attend to previously generated tokens in the sequence. This auto-regressive nature allows the model to condition on its past predictions. Importantly, the self-attention in the Transformer decoder is "masked," meaning that each output token can only attend to the tokens that come before it in the sequence. This prevents the model from attending to future tokens that it has not yet generated.

Next, the decoder performs attention over the encoded cross-modal representation obtained from the encoder. In this Encoder-Decoder Cross Attention module, the normalised output from the Masked MHA layer is used as the query (Q), while the encoded cross-modal representation is used as the key (K) and value (V). This allows each part of the decoder output to attend to the relevant semantic concepts and modalities from the image, facilitating more effective fusion and reasoning across the input modalities. The cross-attention mechanism enables the decoder to condition its generation on the rich visual information, leading to more coherent and contextually appropriate output sequences.

5.5 Experiments and Analysis

5.5.1 Experimental Setup

The model undergoes training through a custom loop that involves the following key steps: data retrieval, image embedding, encoding of clinical and non-clinical data, calculation of loss and accuracy, computation of gradients, weight updating, and tracker adjustment.

The sequence lengths of text data are standardised as follows: 43 tokens for reports, 2 tokens for chief complaints, and 6 tokens for ICD codes, calculated by averaging across all variables. The vocabularies contain over 6,000 unique tokens for radiology reports and over 3,000 tokens each for chief complaints and ICD codes. The vocabulary size and fixed sequence length were determined based on the complete dataset, not just the balanced subset of 3,000 samples used for training and validation. Both image features and text tokens are represented using 512-dimensional embeddings. The Transformer encoder and decoder layers include feed-forward networks with 512-dimensional units each, and Transformer layers utilise multi-headed attention with 3 attention heads. During training, a batch size of 64 is employed, and training proceeds for 100 epochs with early stopping triggered by validation loss stagnation over 5 epochs.

The model's training employed the Adam optimiser with a learning rate of $3e-4$ and linear warmup for the first 500 steps. After the warm-up phase, the learning rate remains constant, stabilising training and facilitating effective model fine-tuning. Loss is calculated using the Sparse Categorical Cross-Entropy loss function defined in Equation 5.7, and accuracy is assessed by matching predicted tokens with true tokens. Let: y_{true_i} be the ground truth for the radiology reports. y_{pred_i} be the output from our report generation model. The equation for cross-entropy loss for each report without reduction is given by:

$$\text{loss}_i = - \sum (y_{\text{true}_i} \cdot \log((y_{\text{pred}_i}))) \quad (5.7)$$

Where: i represents the index of the report. y_{true_i} is the ground truth for report i . y_{pred_i} is the generated report for index i . This loss calculation is performed for each report separately, without any reduction.

5.5.2 Evaluation Metrics

To evaluate the linguistic quality of the generated radiology reports, we computed several automatic evaluation metrics comparing the generated text to the reference reports. First, BLEU scores were calculated to assess n-gram precision for unigrams up to 4 grams. Second, the ROUGE-L score was used to measure the longest common subsequence, assessing the quality of the generated text in terms of recall and precision. Additionally, we evaluated semantic similarity using the BERT Score and Bio-ClinicalBERT Score. These metrics provide a more nuanced assessment of meaning compared to strict n-gram matching. The BERT-based metrics can capture whether the generated reports convey clinically coherent descriptions despite differing word usage compared to the reference. Collectively, these automated evaluation metrics quantify linguistic similarity at word level, sentence-level, and semantic meaning levels.

Finally, to ensure comprehensive evaluation, a board-certified radiologist examined a representative subset of the evaluation set, offering a human expert’s assessment to complement and validate the automated metrics’ analysis.

5.6 Quantitative Results

In this study, we leveraged 11 distinct clinical features along with chest X-rays to generate more accurate and informed radiology reports. The baseline model only employed chest X-ray images as input to generate corresponding reports, serving as our benchmark reference where the sole source of information was the visual data. In order to analyse the contribution of each distinct data feature to model performance, we conducted an ablation study by incrementally presenting different features alongside the chest X-ray images.

For a fair comparison, all data features were encoded in the same way across all experiments, and model hyperparameters, as well as dataset splits remained consistent.

We evaluated four main approaches:

1. The singular model incorporated a single additional feature to show individual performance. Oxygen saturation (O2Sat) is chosen for comparison as it demonstrated the highest performance among the singular models, as illustrated in Table 5.3.
2. The TextFusion model explored fusing textual features of reported primary symptoms

and ICD diagnostic codes with chest X-rays.

3. The ScalarFusion approach combined multiple predictive scalar features with the images, including O2Sat, diastolic blood pressure, temperature, patient acuity scores, and gender. Each of these scalars individually demonstrated performance improvements in singular models.
4. At the core of our study, the FullFusion model takes a holistic approach by fusing all available and relevant data points. This multi-modal fusion aims to effectively incorporate the diverse sources of information at hand, including chest x-ray images, structured clinical data, and unstructured text notes.

Table 5.2 presents a quantitative comparison based on the performance across multiple evaluation metrics. The metrics utilised for the assessment include BLEU-n (B_1 to B_4), ROUGE-L (R_L), BERT F1Score (BS_{F1}), and Bio-ClinicalBERT F1Score ($Bio-CBS_{F1}$). The highest-performing results are highlighted in bold in the table.

The SingularO2Sat method displayed notable improvements across multiple metrics compared to the baseline, while the TextFusion and ScalarFusion methods showcased marginal increases. The FullFusion method emerged as the top performer, showing substantial enhancements in various metrics, and highlighting the benefits of multi-modal fusion.

Method	B_1	B_2	B_3	B_4	R_L	BS _{F1}	Bio-CBS _{F1}
Baseline	0.326	0.205	0.138	0.084	0.301	0.192	0.787
SingularO2Sat	0.343	0.222	0.151	0.096	0.321	0.199	0.789
TextFusion	0.326	0.209	0.141	0.086	0.307	0.181	0.784
ScalarFusion	0.343	0.219	0.145	0.090	0.320	0.198	0.786
FullFusion	0.351	0.231	0.162	0.107	0.331	0.218	0.794

Table 5.2: Quantitative Comparison of Fusion Methods: Performance Evaluation Across Multiple Metrics. B_n for BLEU-n, R_L for ROUGE-L, BS_{F1} for BERT Score F1Score and CBS_{F1} for Bio-ClinicalBERT Score F1Score.

Among the test samples, approximately 33% of them have BLEU-1 scores between 0.1 and 0.3, around 54% have scores between 0.3 and 0.5, about 11% have scores between 0.5 and 0.7, and a mere 0.26% have scores between 0.7 and 1, indicating high similarity. Our BLEU-1 results exhibit strong concordance with existing report generation literature, which has established scoring norms averaging 0.3 to 0.4 for this metric.

Singular Model	B_1	B_2	BS _{F1}	Bio-CBS _{F1}	R.L
ETHNICITY	0.328	0.212	0.174	0.782	0.321
HEARTRATE	0.333	0.213	0.170	0.786	0.295
ICD	0.328	0.207	0.186	0.785	0.301
RESPRATE	0.329	0.213	0.185	0.788	0.309
SBP	0.336	0.219	0.197	0.786	0.319
DBP	0.338	0.220	0.188	0.784	0.322
O2SAT	0.343	0.222	0.199	0.789	0.321
ACUITY	0.342	0.222	0.200	0.789	0.309
TEMPERATURE	0.336	0.219	0.199	0.789	0.326
GENDER	0.341	0.224	0.197	0.787	0.331
CHIEF COMPLAINT	0.326	0.212	0.185	0.785	0.302

Table 5.3: Performance Comparison of Singular Data Models

We compared our model against relevant state-of-the-art models with the best results shown in bold (Table 5.4). In terms of the ROUGE-L score, which indicates the model’s effectiveness in achieving document-level linguistic coherence, our approach achieved a score of 0.331—the highest among all models evaluated. This result suggests that our model is particularly strong in capturing the extended linguistic context required for medical reports.

MODEL	B_1	B_2	B_3	B_4	R.L
Nooralahzadeh et al. 2021	0.378	0.232	0.154	0.107	0.272
L. Wang et al. 2022	0.395	0.253	0.170	0.121	0.284
Shuxin Yang et al. 2022	0.363	0.228	0.156	0.115	0.284
Z. Wang et al. 2022	0.351	0.223	0.157	0.118	0.287
Wu et al. 2023	0.340	0.212	0.145	0.103	0.270
Jin et al. 2024	0.398	x	x	0.112	0.268
Our FullFusion Model	0.351	0.231	0.162	0.107	0.331

Table 5.4: Comparison between our Full Fusion Model and state-of-the-art methods on the MIMIC-CXR dataset, referencing results from their published literature.

5.7 Qualitative Results

For a better interpretation of the results, we illustrated the samples in Figure 5.2 that showcase diversity such as accurate prediction, different expressions, missing and false arguments, and completely false prediction. We compared the ground truth with our FullFusion model, and the correctly predicted diagnoses highlighted in bold for emphasis.

The results show promise in producing reports that capture many of the key findings described in the ground truth reports. To begin with, in all cases, the order of findings aligns with the reports written by the radiologists and the generated reports are structurally correct. The results

also reveal a generally positive alignment in terms of language and grammar, however, some of the generated reports exhibit repeated words or phrases, which can affect the overall coherence. Additionally, the usage of "and" at the beginning of sentences and concluding paragraphs with "the" or "is" reveals grammatical inconsistencies. Furthermore, the FullFusion model accurately identifies normal cardiac, mediastinal, and hilar contours when present in the ground truth.

It also reliably notes the presence or absence of abnormalities like pulmonary edema, pleural effusion, focal consolidation, pneumonia, and pneumothorax which are crucial in radiology analysis. In some cases, the generated reports exhibit a reduced level of detail compared to the ground truth, omitting certain specific observations.

In the first sample, the model missed the right middle lobe atelectasis that was noted in the ground truth. In sample 2, the model hallucinated mediastinal clips not present in the ground truth or image. Sample 3 shows that the model did not fully capture the enlarged cardiac silhouette and vessels described in the ground truth. In sample 4, the model missed details about the interval removal of a central venous line and differences in positioning compared to a prior exam that provided important clinical context.

In sample 5, the model demonstrated enhanced detail compared to the ground truth by providing additional descriptive findings. Sample 6 shows that the model failed to identify the surgical clips in the right upper quadrant indicating a prior cholecystectomy that was noted in the ground truth. The model also incorrectly identified findings suggestive of chronic obstructive pulmonary edema in the upper quadrant.

Sample 7 had repetitive phrasing about no acute osseous abnormalities and failed to note the subsegmental atelectasis in the left lung base documented in the ground truth. Otherwise, the report accurately stated that the heart size was normal, the lungs were clear, and no effusions or pneumothorax were present. Lastly, in sample 8, the model did not fully capture the moderate cardiac enlargement and aortic tortuosity described in the ground truth, instead stating mild cardiac enlargement. The predicted report also repeats "stable mediastinal silhouettes are stable" incorrectly. However, it accurately notes the lack of pulmonary effusion, pneumothorax or consolidation similar to the ground truth.

Overall, these qualitative results demonstrate good progress for the radiology report generation model, with accurate high-level identification of key findings, but also room for improvement

in capturing more nuanced details and clinical context. While the baseline model was capable of providing results, it did not exhibit the same level of detail and accuracy as the enhanced model.

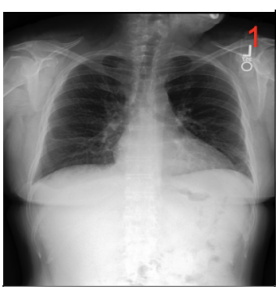
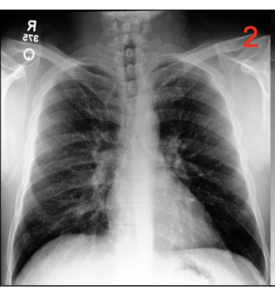
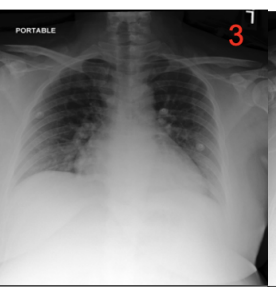
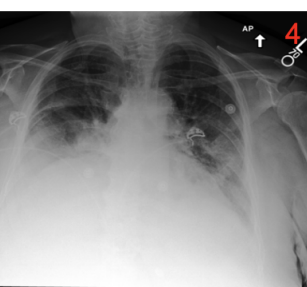
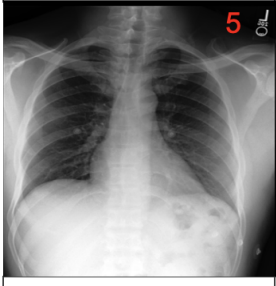
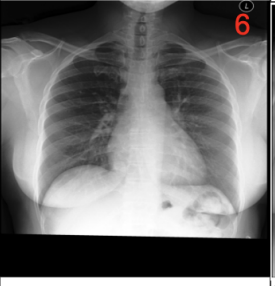
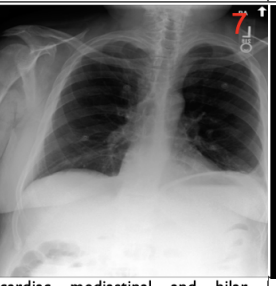
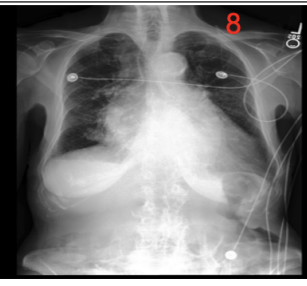
Input Image				
Ground Truth	cardiac mediastinal and hilar contours are normal the pulmonary vasculature is normal minimal atelectasis is seen in the right middle lobe no focal consolidation pleural effusion or pneumothorax is present there are no acute osseous abnormalities	pa and lateral views of the chest were obtained demonstrating clear lungs without focal consolidation effusion or pneumothorax cardiomeastinal silhouette appears normal the imaged osseous structures are intact no free air below the right hemidiaphragm	enlargement of the cardiac silhouette and increase caliber pulmonary vessels is in keeping with the patients pregnant status allowing for low lung volumes which results in crowding of bronchovascular structures there is no evidence of pulmonary edema or focal consolidation to suggest pneumonia no definite pleural effusion	interval removal of a right internal jugular central venous line lung volumes have diminished with crowding of the pulmonary vasculature and interval improvement in pulmonary vascular congestion bilateral layering effusions with associated airspace opacities likely reflecting partial lower lobe atelectasis although pneumonia cannot be excluded no pneumothorax overall cardiac and mediastinal contours are likely unchanged given differences in positioning
Generated Report	mediastinal and hilar contours are normal the pulmonary vasculature is normal streaky opacities in the lung bases likely reflect atelectasis no focal consolidation pleural effusion or pneumothorax is seen there are no acute osseous abnormalities the left upper abdomen is unremarkable there	and mediastinal clips are noted in the right upper lobe there is no focal consolidation effusion or pneumothorax the cardiomeastinal silhouette is normal imaged osseous structures are intact no free air below the right hemidiaphragm is seen in the right hemidiaphragm is	enlargement of the cardiac silhouette is mildly enlarged the mediastinal and hilar contours are unremarkable pulmonary vasculature is normal lungs are clear with no pleural effusion or pneumothorax is seen there are no acute osseous abnormalities are again noted in the right	lung volumes are low exaggerating heart size and bronchovascular markings interstitial markings distributed throughout both lungs which may be due to mild pulmonary edema which may be due to moderate pulmonary edema there is mild pulmonary edema with relative elevation of the
Input Image				
Ground Truth	the cardiomeastinal silhouette is normal there is no focal consolidation there is no pleural effusion or pneumothorax	the lungs are clear the cardiomeastinal silhouette is normal surgical clips in the right upper quadrant suggest prior cholecystectomy	cardiac mediastinal and hilar contours are unchanged with the heart size normal the pulmonary vasculature is not engorged re demonstrated within the left lung base is subsegmental atelectasis without focal consolidation the right lung is clear no pleural effusion or pneumothorax is present no acute osseous abnormality is detected	the heart is moderately enlarged the aortic arch is calcified the aorta is also moderately tortuous the cardiac mediastinal and hilar contours appear stable the chest is hyperinflated there is no pleural effusion or pneumothorax the lungs appear clear
Generated Report	there is no focal consolidation pleural effusion or pneumothorax the cardiomeastinal silhouette is normal there is no acute osseous abnormalities are seen there is no free air below the right hemidiaphragm is seen in the right	chest film demonstrates no focal consolidation pleural effusion or pneumothorax the cardiac and mediastinal silhouettes are unremarkable no acute osseous abnormalities are seen within the diaphragm suggesting chronic obstructive pulmonary edema is seen at the upper quadrant of uncertain chronicity pulmonary edema	tip of the heart size is normal cardiomeastinal contours are normal the pulmonary vasculature is normal lungs are clear without focal consolidation pleural effusion or pneumothorax is normal no acute osseous abnormalities identified no acute osseous abnormalities identified there is identified there	the cardiac silhouette is mildly enlarged the aorta is tortuous there is no focal consolidation pleural effusion or pneumothorax the cardiac and mediastinal silhouettes are stable and stable mediastinal silhouettes are stable the aorta is tortuous there is no pulmonary edema is

Figure 5.2: GT and GR report from the proposed FullFusion CXR report generation model.

5.8 Radiologist Evaluation Results

We evaluated the model using 158 randomly selected samples from the unseen test set, covering diverse medical conditions reflecting the full distribution. A board-certified radiologist assessed three criteria: language fluency, content selection, and correctness of abnormal findings (AF).

For language, the radiologist evaluated sentence structure, terminology, and overall clarity. For content, they compared the report’s level of detail, key findings, and image coverage to the true findings. They assessed the accuracy of abnormal findings by comparing them to the true conditions. The radiologist assigned 1-5 scores and noted preference between reports.

This methodology enabled quantitative and qualitative assessment of language generation, content selection, and diagnostics. The radiologist also noted that while performing well overall, some shortcomings were observed.

The model often missed surgical materials like catheters and clips and it fails to capture anomaly variations when the patient is inclined to the right or left. Sensitivity to bone lesions was lacking, overlooking non-urgent findings like scoliosis. However, it’s worth noting that these are not extensively covered in the ground truth as well. For normal X-rays, it occasionally included non-definitive elements.

While these additions may be accurate, there is a slight possibility that they may not be. This evaluation methodology provided valuable insights into model strengths and areas needing improvement.

Language Fluency	Content Selection	Correctness of AF
4.24	4.12	3.89

Table 5.5: Radiologist Evaluation Results on a 1-5 Scale

5.9 Discussion and Conclusion

This chapter presents a multi-modal data-driven integrative approach to enhance the precision and clinical relevance of radiology reports generated in conjunction with chest X-ray images. In the section pertaining to data selection, we ensured temporal alignment between the input data modalities and the target radiology reports to closely imitate real-world clinical workflows.

Additionally, we aimed for a balanced representation of each type of report in our sample selection to mitigate potential biases skewed towards normal cases. Although this approach resulted in a smaller dataset compared to existing literature, it was a crucial step for preventing biased results and validating the results in real clinical settings.

Recent literature highlights the potential of multi-modal learning techniques in advancing the quality of automated radiology report generation. However, a majority of medical report generation models primarily focus on target reports within specific information or incorporate image findings as supplementary inputs. Given that Chest X-rays present three-dimensional objects in a two-dimensional form, some valuable spatial and contextual information is lost, leading to semantic gaps in the data provided to the network. Furthermore, radiologists possess more data beyond images during report generation.

To address this limitation, we bridged the semantic gap between vision and language models by capturing uncodified information essential to the diagnosis process. We achieved this by introducing an ensemble of 11 supplementary features in conjunction with the chest X-ray data. These features were thoughtfully selected to enhance both accuracy and clinical insight in the generated reports. The results indicate that incorporating non-imaging clinical and non-clinical data positively impacts the quality of the generated reports, as measured by automatic evaluation metrics and human evaluation studies.

Our ablation study further demonstrates that providing all data simultaneously yields higher accuracy compared to using individual data components separately. This finding suggests that introducing data with no significant standalone impact on the model, when combined with other modalities through attention-based fusion, can lead to improved performance by capturing complementary information.

However, there are some limitations to our study. While the multi-modal deep neural network framework holds potential strength, its complexity and resource-intensive nature may pose challenges. This might hinder its real-time application in medical settings, especially those with limited resources and hardware accelerators like GPUs. Furthermore, our data solely originates from databases within a single institution, lacking a comparable comprehensive dataset that combines imaging and non-imaging data (both clinical and non-clinical) with linked radiology reports. Enhancing data diversity from various sources could enhance the overall robustness of the study.

Chapter 6

CROSS-TASK LEARNING FRAMEWORK

6.1 Introduction

With the development of deep learning techniques, the fields of computer vision and natural language processing have started to converge, as both images and texts can be represented via compatible embeddings. This convergence has led to success in the challenging cross-modal tasks and it has numerous real-world applications, such as image captioning (Stefanini et al. 2022), medical report generation (Ramirez-Alonso et al. 2022), and assisting the visually impaired (SS et al. 2023).

In such modality translation tasks, the objective is to learn complex non-linear mappings, typically between visual representations derived from an input image and complementary or target data in the form of text. This process requires the effective transfer of information across modalities, preserving as much relevant content as possible.

Within this context, the use of multi-modal data has provided an efficient way to improve the coherence and accuracy of the generated text conditioned on the given image input. Specifically, multi-modal deep learning architectures aim to capture and fuse the complementary information present across heterogeneous data modalities, such as images and structured data (e.g., medical records, demographic information, or clinical measurements).

Additionally, multi-task learning, where a single model is trained to perform multiple related tasks simultaneously, has played an important role in harnessing the full potential of multi-modal data in image-to-text generation (Bayouhd et al. 2021). It's important to note that this approach differs from traditional methods where separate models are trained for individual tasks. Unlike single-task learning, where models are tailored to a specific job, and transfer learning, where a pre-trained model is repurposed, multi-task learning takes a unified approach.

In Chapter 5, we took a step forward by combining patient information with medical images to improve the quality of automatically generated radiology reports. Although this approach showed promise in creating accurate and coherent reports, there's room for enhancement, particularly in capturing all relevant clinical findings.

Building on this groundwork, we explored the impact of multi-task learning on the accuracy and efficiency of radiology report generation. Chapter 6 introduces a new statistical system for cross-modal multi-task learning, that is for contemporary learning of various tasks (report generation, ordinal classification, multi-label classification) from multi-modal data. Our primary objectives encompass evaluating the model's performance in all tasks and investigating potential synergies between these parallel learning processes.

6.2 Problem Formulation

The proposed approach is designed to concurrently perform three tasks: Report Generation, Ordinal Classification, and Multi-Label Classification. These tasks involve processing various inputs and producing meaningful outputs while optimising for different loss functions. By integrating them within a single framework, the overarching objective is to leverage information and features from the other tasks, ultimately resulting in more precise and context-aware text generation capabilities.

Given an image I , unified additional features F , ground truth radiology report text sequences Y , ground truth ordinal acuity levels T (where T is an integer value between 1 to 5 indicating the severity of the patient's condition), and ground truth labels for findings Z (where Z are multi-label categorical values with 5 possible labels indicating the presence or absence of specific medical findings), the objective is to learn an encoder-decoder model to minimise the loss for the three tasks:

Report Generation (REPGEN), generates a radiology report \hat{Y} that maximises the probability of the ground truth text sequence Y given the image I and features F . The loss is defined as the cross-entropy loss between \hat{Y} and Y .

Ordinal Classification (OC) predicts ordinal acuity level \hat{T} given I and F . The loss is defined as the binary cross-entropy between \hat{T} and T .

Multi-Label Classification (ML) predicts multiple labels for findings \hat{Z} given I and F . The loss is the binary cross-entropy between \hat{Z} and Z .

Overall Loss Function, denoted as L , is composed of three task-specific loss components, α , β , and γ are hyperparameters that control the relative weighting of these losses:

$$L = \alpha \cdot L_{\text{REPGEN}}(\hat{Y}, Y) + \beta \cdot L_{\text{OC}}(\hat{T}, T) + \gamma \cdot L_{\text{ML}}(\hat{Z}, Z)$$

Where:

$L_{\text{REPGEN}}(\hat{Y}, Y)$ is the loss function for REPGEN.

$L_{\text{OC}}(\hat{T}, T)$ is the loss function for OC.

$L_{\text{ML}}(\hat{Z}, Z)$ is the loss function for ML.

We also applied gradient clipping during the update of task weights to prevent the exploding gradient problem. θ denoted as the threshold value for clipping, after computing the gradients ($\nabla \mathcal{L}$), the clipped gradients represented as:

$$\nabla_{\text{clipped}} = \text{clip}(\nabla \mathcal{L}, -\theta, \theta)$$

The updating of the weights of each task including gradient clipping calculated as:

$$\alpha = \alpha - \text{lr} \cdot \nabla_{\text{clipped}}[0]$$

$$\beta = \beta - \text{lr} \cdot \nabla_{\text{clipped}}[1]$$

$$\gamma = \gamma - \text{lr} \cdot \nabla_{\text{clipped}}[2]$$

Where lr represents the learning rate of the optimiser. During the training process, the model’s encoder-decoder parameters θ are optimised by minimising the overall loss L over the training dataset using the following objective:

$$\begin{aligned} \min_{\theta} L(\theta) = \min_{\theta} & \left(\alpha \cdot L_{\text{REPGEN}}(\hat{Y}, Y; \theta) \right. \\ & + \beta \cdot L_{\text{OC}}(\hat{T}, T; \theta) \\ & \left. + \gamma \cdot L_{\text{ML}}(\hat{Z}, Z; \theta) \right) \end{aligned}$$

6.3 Experiments and Analysis

6.3.1 Data and Feature Processing

The MIMIC-CXR database also contains structured labels for chest X-ray images generated by CheXpert (Irvin et al. 2019), an open-source rule-based tool. CheXpert operates in three main steps: Firstly, it identifies all mentions of a label, and then each mention is classified as positive, uncertain, or negative based on the local context. Finally, it resolves multiple mentions of the same label by giving priority to positive mentions, followed by uncertain mentions, and lastly, negative mentions.

In total, there are 14 structured labels generated by the CheXpert classifier: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomedastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pleural Other, Support Devices, and No Finding. For each condition, the classifier assigns one of four possible values:

- 1.0: Positive finding (condition is present)
- -1.0: Uncertain finding
- 0.0: Negative finding (condition is absent)
- NAN: No mention (missing data)

To address the class imbalance, we consolidated the original 14 CheXpert conditions into 5 broader classes. A class is marked as present (1) if any associated condition has a label (whether positive, negative, or uncertain), and absent (0) if completely unlabeled. Specifically:

- No Finding: Retained as is from the original labels
- Support Devices: Retained as is from the original labels
- Fracture: Retained as is from the original labels
- Lung Opacity: Generated by combining any presence of lung-related conditions (Atelectasis, Consolidation, Edema, Pneumonia, and Lung Lesion)
- Pleural: Generated by combining any presence of pleural-related conditions (Pleural Effusion, Pleural Other, and Pneumothorax)

After cleaning the missing data, the resulting dataset consisted of 60,933 data points. From this dataset, we selected a balanced subset of 10,000 samples, with 7,000 for training, 2,000 for validation, and 1,000 for the test set. The curated dataset consists of image data paired with findings, as well as clinical and non-clinical data (further detailed in 5.3), structured labels, and patient acuity level (also referred to as severity level).

Every image was resized and then normalised to ensure uniform intensity levels. Subsequently, a pre-trained model extracted visual features from the images. The resulting visual feature vector was then input into a transformer-based encoder to gain a deeper understanding of the recognised elements in the image.

Numerical variables were initially preprocessed to remove potential outliers based on domain knowledge and clinical perspectives. The remaining data for each numerical feature was then standardised to a 0 to 1 range by rescaling based on the minimum and maximum observed values. Binary variables were converted to numerical representations. Integer values were assigned to each group in categorical data, followed by one-hot encoding. The encoder expects input data in matrix form (samples \times features), therefore, the processed data was reshaped into a 2D array where each row represents a sample and each column represents a feature.

Text-based variables were pre-processed by converting to lowercase, removing unnecessary punctuation using regular expressions, and condensing consecutive periods into single spaces. Double periods were replaced with single spaces to maintain consistent text formatting. Additional standardisation involved replacing shorthand phrases or abbreviations with their full-text equivalents, correcting errors, and addressing inconsistencies in pluralisation. This pre-processing pipeline standardised various data types for input into the model.

Lastly, acuity levels were encoded using cumulative one-hot representation where ordinal levels are mapped to binary vectors. Each vector has a length equal to the number of ordinal levels minus one. The presence of a '1' in a specific position within the binary vector indicates the corresponding ordinal level. This encoding preserves the ordinal relationships between levels.

6.3.2 Evaluation Metrics

Several metrics are used to evaluate model performance on the report generation, multi-label and ordinal classification tasks. The linguistic and contextual quality of the generated radiology reports have been evaluated using BLEU-1 to 4, ROUGE_L, METEOR, BERT Score and BioClinicalBERT Score.

For multi-label classification, the metrics we employed included Precision, Recall, F1 Score, Hamming Loss, and Exact Match Ratio. Precision and Recall evaluated the accuracy of predicting positive labels and capturing all true positives respectively. F1 Score provided the balance between Precision and Recall. Hamming Loss quantified label prediction errors, and Exact Match Ratio measured how often the model correctly predicted all labels for a given instance.

For ordinal classification, our metrics consisted of Ordinal Classification Accuracy, Mean Absolute Error, Mean Squared Error, and the Accuracy-Correlation Hybrid Metric. Ordinal Classification Accuracy measured the accuracy by computing the total number of correct predictions divided by the total number of predictions. Mean Absolute Error and Mean Squared Error quantified the average magnitude of errors in predicted ordinal values. The Accuracy-Correlation Hybrid Metric combined aspects of accuracy and correlation to evaluate the preservation of the ordinal relationship.

6.3.3 Experimental Setup

All models were implemented in TensorFlow 2.3.0 and Keras. Transformer layers implemented with 3 attention heads, and 256 dimensional feedforward layers.

Each model was trained on 7,000 data using an Adam optimiser with a learning rate warmup over 10% of steps up to $3e-5$ and a batch size of 32. The validation, and test sets consist of 2,000, and 1,000 data, respectively. Training continued for 100 epochs with early stopping monitoring the validation loss with `patience=10`.

For parity in model optimisation, we maintained consistency in the choice of hyperparameters (e.g. learning rate, batch size, etc.) when training each of the assessed models. All models were trained on NVIDIA Tesla A100 GPUs with 40GB memory.

6.4 Architectural Designs and Learning Strategies

6.4.1 Cross-modal MTL Approach

Multi-task learning is a promising approach to enhance representation learning; by sharing information across tasks, it can improve the generalisation of the learned representations. However, one of the challenges in multi-task learning is determining the appropriate complexity of the network architecture, especially the decoders for each individual task. Complex decoders may actually degrade the quality of the learned representations by introducing noise and confusion. In this context, we initially employed a foundational multi-label classifier and an ordinal classifier.

In our experiments to determine the optimal architecture and loss hyperparameter values, we explored several approaches. Firstly, we prioritised increasing the accuracy of the main task; report generation (REPGEN). The other tasks, ordinal classification (OC) and multi-label classification (MLC) were secondary objectives. The Cross-modal MTL (CM-MTL) model leveraged multi-task learning to enhance performance primarily for the main task.

As illustrated in 6.1, the ordinal classifier starts with applying a dropout regularisation by randomly setting a fraction of input units to zero during training to prevent overfitting. Then, batch normalisation is performed to improve the stability and performance of the model during training. Finally, it passes the output through a fully connected layer with a sigmoid activation function. The sigmoid activation ensures that the output values are suitable for ordinal classification tasks with the correct order. The last layer also applies L2 regularisation to prevent overfitting.

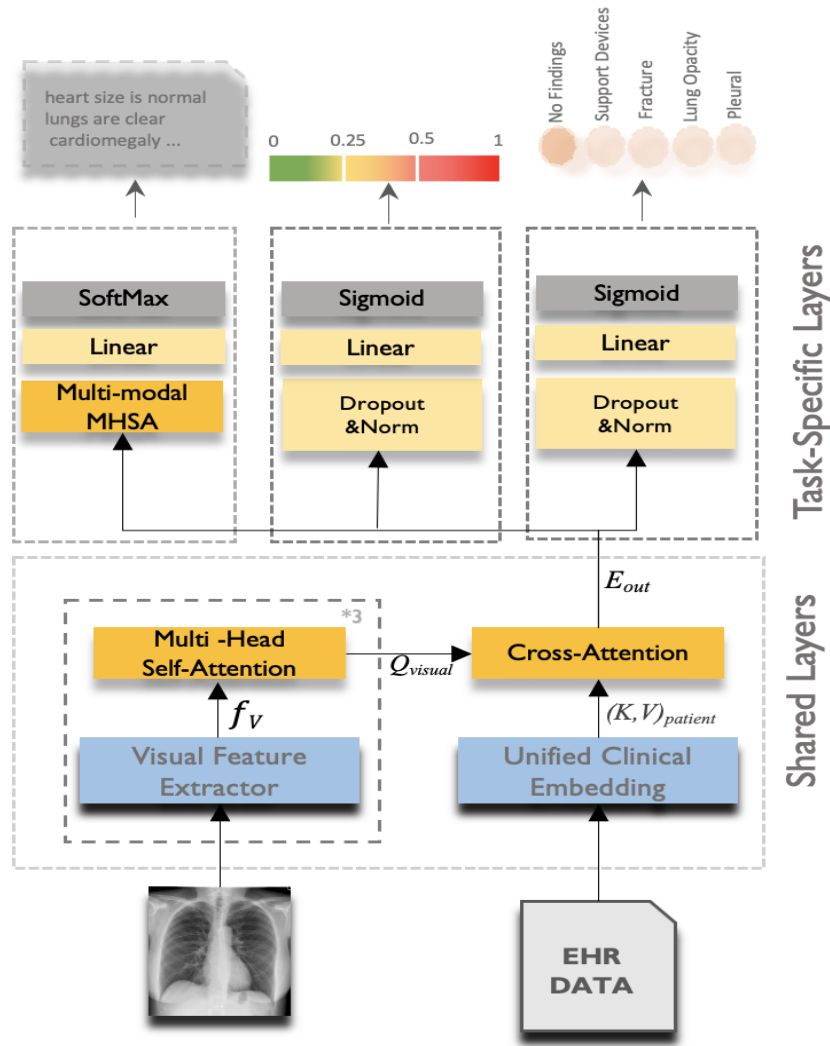


Figure 6.1: The overall architecture of Cross-modal Multi-Task Learning (CM-MTL) Model

For multi-label classification, the model begins with a dropout layer set at 0.5, which introduces regularisation by randomly deactivating input units during training to prevent overfitting. Following this, a batch normalisation layer is utilised to normalise activations within each batch. The model then employs a fully connected layer with five output units, activated using the sigmoid function. Similar to the ordinal classification, this layer is regularised using L2 regularisation to penalise large weight values, thus preventing overfitting.

The text generator decoder employs causal masking, which is combined with padding masks. It comprises two consecutive multi-head self-attention layers. The first layer attends exclusively to the target sequence, while the second layer attends to the encoder outputs using the decoder inputs as queries. The attended representations are processed through a two-layer position-wise feed-forward network. Finally, a linear layer produces predictions, which are used as inputs for

the next time step.

The approach employed two different strategies for handling multiple tasks. The first strategy treats all tasks equally by assigning them the same weight, while the second strategy prioritises the loss contribution (α) of the main task, report generation. Despite achieving good results in report generation, the ordinal classification (OC) and multi-label classification (MLC) models encountered convergence issues after a few training epochs.

6.4.2 Results

Quantitative Results

We evaluated the performance of the model on three tasks: report generation, ordinal classification, and multi-label classification. The CM-MTL model was further trained in two configurations for better assessment: equal weighting across tasks (CM-MTL-EQ) and task prioritisation for the text generation task (CM-MTL-TP). Then, we compared these CM-MTL models to single-task learning baselines (STL).

Table 6.1 displays the results for text generation tasks, measured in terms of BLEU scores (B.1 to B.4), BERT Score F1Score (BS_{F1}), Bio-ClinicalBERT Score F1Score ($Bio-CBS_{F1}$), and ROUGE-L (R.L). When employing the CM-MTL-EQ approach, the model exhibits slightly improved performance across most metrics compared to STL.

Method	B.1	B.2	B.3	B.4	BS_{F1}	$Bio-CBS_{F1}$	R.L
STL	0.3326	0.2159	0.1488	0.0950	0.2056	0.7857	0.3096
CM-MTL-EQ	0.3352	0.2229	0.1570	0.0983	0.1958	0.7883	0.3235
CM-MTL-TP	0.3424	0.2295	0.1616	0.1035	0.2065	0.7898	0.3366

Table 6.1: Performance comparison of a report generation using different training approaches. B_n for BLEU-n, R.L for ROUGE-L, BS_{F1} for BERT Score F1Score and CBS_{F1} for Bio-ClinicalBERT Score F1Score. STL denotes Single Task Learning, CM-MTL-TP represents Multi-Task Learning with Task Prioritisation for text generation and CM-MTL-EQ indicates Multi-Task Learning with equal task weights for each task

Notably, the CM-MTL-TP model achieved the best performance, outperforming STL and CM-MTL-EQ on all metrics. This demonstrates the benefits of multi-task learning with proper task weighting for improving text generation quality and it also validates the capability to leverage representations learned across related tasks.

To assess the statistical significance of the improvements achieved by the CM-MTL-TP approach

	B_1	B_2	B_3	B_4	BS_{F1}	Bio-CBS_{F1}	R.L
p-value	0.00117	0.00006	0.00035	0.02569	0.83824	0.00611	0.00000

Table 6.2: P-values from pairwise t-test between STL and CM-MTL-TP approaches (rounded to 5 decimal places)

over the Single-Task Learning (STL) baseline, we conducted a pairwise t-test for each metric. Table 6.2 presents the p-values from these significance tests, rounded to 5 decimal places.

The small p-values obtained for most metrics, particularly BLEU-2, BLEU-3, and ROUGE-L, indicate that the improvements in text generation quality are statistically significant. This analysis focuses on the text generation task, as it was the primary task where substantial improvements were observed with the CM-MTL-TP approach.

Multi-label classification and Ordinal classification models suffered from overfitting after only a few epochs. Despite attempts to optimise hyperparameters, STL classification models continued to overfit within a few epochs. In contrast, multi-task learning approaches helped prevent overfitting for 5-6 additional epochs compared to STL on this task. So even though computed multi-label classification metrics show better STL performance (Please refer to Table B.1 and Table B.2), this overfitting was inevitable with single-task training in this approach.

Qualitative Results

For a better comparison of the results, we have colour-coded the illustrated samples to match their respective ground truth labels, see Figure 6.2. Ambiguous or repeated expressions are denoted in italics, while incorrect predictions or expressions not present in the original report are underlined. If the statement is for both, such as when the expression is repeated and not in the original report, we used both italics and underlines.

The results indicate promise in generating reports that capture many of the main findings mentioned in the ground truth. All approaches show a generally positive alignment in terms of grammar, however, some of the generated reports exhibit repeated words or phrases, which can affect the overall coherence.

Particularly, The STL approach struggles with unnatural wording like "[UNK] for comparison" and hallucinates findings that are not present in either the image or the ground truth. The CM-MTL-EQ output has disjointed phrasing and repetition indicating a lack of narrative coherence.

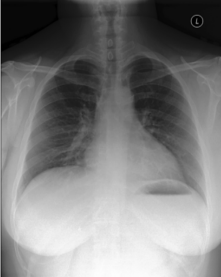

Input Image	Ground Truth	STL	MTL_EQ	MTL_TP
	the cardiac silhouette size is top normal mediastinal and hilar contours are normal lungs are clear and the pulmonary vascularity is normal no pleural effusion or pneumothorax is present no acute osseous abnormalities are detected	[UNK] for comparison there is no focal consolidation pleural effusion or pneumothorax the heart is normal in size the mediastinal and hilar contours are normal the pulmonary vascularity is normal there are no acute osseous abnormalities surgical clip is seen in	trace pleural effusion is seen the cardiac and mediastinal silhouettes are unremarkable there is no pneumothorax or pneumothorax the cardiac and hilar contours are normal there is no pneumothorax is no free air below the right hemidiaphragm is seen no focal	the lungs are clear the cardiac silhouette is normal in size the mediastinal and hilar contours are normal the pulmonary vasculature is normal no pleural effusion or pneumothorax is seen no acute osseous abnormalities identified there is seen the lungs are
	minimal basilar atelectasis is seen there is no focal consolidation no pleural effusion or pneumothorax is seen the cardiac and mediastinal silhouettes are unremarkable	change in the left lower lobe no focal consolidation pleural effusion or pneumothorax is seen the cardiac and mediastinal silhouettes are unremarkable no displaced rib fracture is identified the visualized upper abdomen is unremarkable no displaced fracture is seen beneath the	or size is normal the mediastinal and hilar contours are normal there is no pleural effusion or pneumothorax no free air below the right hemidiaphragm is seen no acute osseous abnormality is seen the right hemidiaphragm is seen no free air	aortic arch is again seen with known lower lobe atelectasis no focal consolidation pleural effusion or pneumothorax cardiac and mediastinal silhouettes are unremarkable pulmonary edema is detected no acute osseous structures are intact no acute osseous abnormalities identified no pulmonary edema

Figure 6.2: Illustrative comparison of generated reports from different learning approaches with ground truth

In contrast, the CM-MTL-TP generates smooth, logical statements more similar to the ground truth, with some minor repetition. In the second example, the CM-MTL-TP text exhibits a clearer structure, with sentences covering distinct findings. It includes details like "pulmonary edema", "aortic arch is again seen" and "no acute osseous abnormalities identified" not in the original text but present in the image.

Overall, results demonstrate good progress for the radiology report generation model, with accurate identification of key findings but also room for improvement. The STL model sometimes seems to include extraneous or inaccurate details where the equal-weighted MTL shows improvements in content quality over STL, but suffers from repetitiveness and disorganised narratives.

6.4.3 Balanced Attentive MTL Approach

In multi-task learning, there is often a trade-off between the different tasks, and one task may converge faster than others, leading to imbalanced learning, as evidenced by the results of the CM-MTL model. Improving the performance across all tasks leads to better representation learning of each modality.

To address this convergence issue and improve task-specific representation learning capabilities, we enhanced both classification architectures by integrating separate attention mechanisms.

These task-specific attention layers allow each classifier (ordinal and multi-label) to dynamically focus on the most relevant input features for its particular task during training and inference. For instance, the ordinal classification attention might emphasise features related to severity, while the multi-label classification attention might focus on features indicative of multiple conditions.

As illustrated in Figure 6.3, the final model comprises two primary blocks: shared layers that are common to all tasks and task-specific layers tailored to each individual task. The details of Visual Feature Extractor and Unified Clinical Embedding are presented in 5.4.

The cross-attention block focuses on the Unified Clinical Embedding, utilising the image features as queries. This allows the network to contextually focus on relevant aspects of the unified data conditioned on the image content.

The ordinal classification employs an attention module that computes attention weights a_w , which are then used to obtain a weighted representation of the input sequence x . This can be expressed as $z = \sum_i (a_{w_i} \cdot x_i)$, where z is the attended output. This attended output is passed through a dense layer W followed by a sigmoid activation to produce probabilities for each ordinal threshold, which are then averaged across the sequence dimension: $y_{oc} = \text{avg}(\sigma(Wz))$, where σ is the sigmoid function.

Meanwhile, the multi-label classification uses an attention module to obtain an attended encoding of the input sequence, $A(x)$. This attended encoding undergoes further processing through 5 parallel fully connected layers with sigmoid activations, each corresponding to a distinct label in the multi-label task.

The outputs of these sigmoid layers, representing the presence or absence of each label, are then concatenated to form the overall multi-label probability distribution. This can be represented as $y_{ml} = [\sigma(FC_1(A(x))), \sigma(FC_2(A(x))), \dots, \sigma(FC_5(A(x)))]$, where FC_i is the i -th fully-connected layer, σ is the sigmoid function.

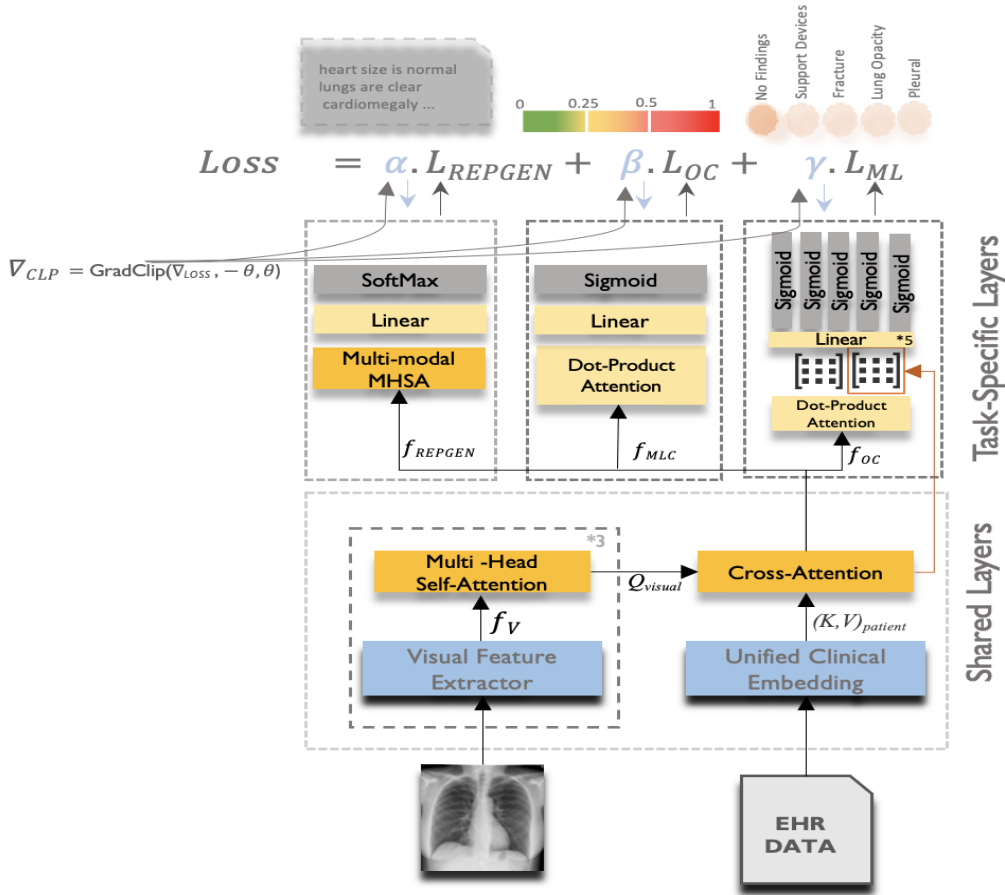


Figure 6.3: The overall framework of our proposed Balanced Attentive multi-task learning network

In summary, in this approach, the ordinal classifier uses attention to produce a single distribution over ordered classes, while the multi-label classifier leverages attention and parallel sigmoid layers to predict the presence of multiple labels simultaneously. The decoding process of the text generator follows the same approach as the CM-MTL model.

6.4.4 Results

Quantitative Results

We trained and evaluated the proposed approach by comparing it against single-task learning results. Initially, each task's decoders were trained independently, and then they were trained within a unified framework to assess the improvement brought by multi-task learning for each task.

Table 6.3 presents the performance comparison between single-task learning and Balanced-Attentive Multi-task Learning across several NLG metrics. The results demonstrate that multi-

task learning using the Balanced-Attentive approach significantly enhances the performance of the report generation model and consistently outperforms the single-task approach across all metrics.

Method	B_1	B_2	B_3	B_4	BS _{F1}	Bio-CBS _{F1}	R_L	METEOR
STL	0.330	0.215	0.148	0.095	0.201	0.787	0.304	0.157
BA-MTL	0.351	0.235	0.167	0.108	0.216	0.791	0.333	0.168

Table 6.3: Performance comparison of a single task report generation (STL) and Balanced-Attentive Multi-task Learning (BA-MTL).

The table 6.4 displays p-values resulting from pairwise t-tests comparing results between STL and BA-MTL approaches across BLEU1-4, BertScore and ROUGE scores. The p-values, ranging from 3.64×10^{-13} to 1.36×10^{-12} , indicate highly significant differences between the two approaches for each metric. In all metrics, except BS_{F1}, the p-values are exceedingly small, suggesting robust evidence against the null hypothesis. Despite BS_{F1} having relatively larger p-values compared to other variables, statistical significance remains evident.

	B_1	B_2	B_3	B_4	BS _{F1}	R_L
p-value	3.64×10^{-13}	3.33×10^{-10}	3.66×10^{-8}	7.05×10^{-4}	3.47×10^{-3}	1.36×10^{-12}

Table 6.4: P-values from pairwise t-test between STL and BA-MTL approaches

We further assessed our model’s performance alongside state-of-the-art methods in radiology report generation (detailed methodologies of these models are available in Chapter 3), using evaluation metrics commonly reported in these studies. Original scores from the respective papers were used for comparison.

MODEL	B_1	B_2	B_3	B_4	R_L	METEOR
Z. Chen et al. 2020	0.353	0.218	0.145	0.103	0.277	0.142
Nooralahzadeh et al. 2021	0.378	0.232	0.154	0.107	0.272	0.145
F. Liu et al. 2022	0.344	0.217	0.140	0.097	0.281	0.133
L. Wang et al. 2022	0.395	0.253	0.170	0.121	0.284	0.147
Shuxin Yang et al. 2022	0.363	0.228	0.156	0.115	0.284	x
Z. Wang et al. 2022	0.351	0.223	0.157	0.118	0.287	x
Wu et al. 2023	0.340	0.212	0.145	0.103	0.270	0.139
Tanida et al. 2023	0.373	0.249	0.175	0.126	0.264	0.168
Zhao et al. 2023	0.399	0.242	0.158	0.109	0.275	0.152
Jin et al. 2024	0.398	x	x	0.112	0.268	0.157
FullFusion Model ¹	0.351	0.231	0.162	0.107	0.331	0.157
BA_MTL Model	0.351	0.235	0.167	0.108	0.333	0.168

Table 6.5: Comparison with state-of-the-art radiology report generation methods on MIMIC-CXR. The best results are highlighted in bold.

Table 6.5 displays the natural language metrics of various models evaluated on the MIMIC-CXR dataset. The metrics used are BLEU scores (B_1, B_2, B_3, B_4), METEOR and ROUGE-L (R_L), which are standard metrics for evaluating natural language generation tasks. While models such as Zhao et al. 2023 and L. Wang et al. 2022 excel in specific BLEU scores, indicating strong performance in exact n-gram matching, our proposed BA_MTL Model demonstrates competitive performance across all metrics and notably outperforms in terms of ROUGE-L and METEOR scores.

This suggests superior overall summary quality and content fidelity. The consistent performance of our models across various metrics indicates robustness, although there is room for improvement in exact phrase matching as reflected by the BLEU scores. Interestingly, different models show strengths in different areas, with Tanida et al. 2023 leading in B_3 and B_4 scores, suggesting better performance in longer n-gram matches.

This variability in performance across metrics underscores the complexity of the task and the different focuses of various approaches. Our BA_MTL Model’s strong ROUGE-L score (0.333) implies that it captures the overall content and structure of the reference reports more effectively than other models, making it particularly suitable for tasks prioritising comprehensive content summarisation over exact phrase reproduction.

For the ordinal classification task, using the BA-MTL approach shows a reduction of 2.52% in accuracy. In terms of Mean Absolute Error (MAE), STL achieved a lower value than achieved by BA-MTL, this suggests that STL is better at minimising the error between predicted and actual ordinal values. The combined metric of accuracy and correlation also favoured STL, with a score of 0.8163 compared to 0.7908 for BA-MTL. Overall, the single-task learning approach outperformed the balanced attentive multi-task learning approach in all evaluated metrics for the ordinal classification task.

Method	Accuracy	MAE	ACC+Corr
STL	0.8716	0.1283	0.8163
BA-MTL	0.8497	0.1503	0.7908

Table 6.6: Comparing performance of the ordinal classifier in Single-Task and Balanced Attentive Multi-Task Learning

¹Our FullFusion Model presented in Chapter 5

Table 6.7 demonstrates the performance of the multi-label classification task for the Single-Task Learning (STL) and Balanced Attentive Multi-Task Learning (BA-MTL) approaches. The BA-MTL approach outperformed the STL method across all evaluated metrics. Additionally, BA-MTL exhibited lower Hamming Loss (0.1562 vs. 0.1630) compared to STL, indicating better overall performance in the multi-label classification task.

Method	Precision	Recall	F1 Score	Hamming Loss	Exact Match Ratio
STL	0.6410	0.6188	0.6108	0.1630	0.8370
BA-MTL	0.6482	0.6398	0.6278	0.1562	0.8438

Table 6.7: Comparing performance of the multi-label classifier in Single-Task and Balanced Attentive Multi-Task Learning

It is noteworthy that the classification tasks were successfully trained without encountering convergence issues noted in the previous approach. Although the BA-MTL approach did not yield improved results for the ordinal classification task, it considerably enhanced performance in both report generation and multi-label classification tasks compared to single-task learning baselines. These outcomes highlight the effectiveness of our approach in improving task-specific performance within a multi-task learning framework.

Qualitative Results

To provide a qualitative assessment of the generated reports, Figure 6.4 presents a visual comparison between the input image, ground truth labels, and reports generated by the STL and BA-MTL approaches. Corresponding statements across the generated reports and ground truth have been colour-coded for ease of comparison, enabling the identification of accurate predictions and discrepancies from the ground truth (GT).

In the first sample, the STL model’s output of ”tissue on the frontal view” and ”the lateral view is limited due to patient” does not match with the ground truth and shows hallucination by the model. Additionally, the STL model fails to capture all findings and fails in grammar and sentence structure. It generated incomplete and incoherent sentences such as ”the lateral view is limited due to the patient has been interval placement of a right picc line has been removed.” In contrast, the BA-MTL model’s output is more coherent and closely aligned with the ground truth. The BA-MTL output correctly identifies all the findings in the ground truth except ”heart size remains mildly enlarged.” The repetition of ”the right hemidiaphragm is seen” is an error, but it is less severe compared to the hallucinations seen in the STL model.

The BA-MTL model also shows better performance on grammatical rules and produces more complete sentences, although the redundancy suggests a minor mistake in the model’s ability to generate entirely natural text.


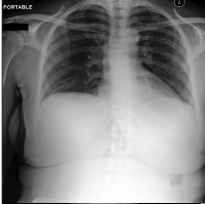
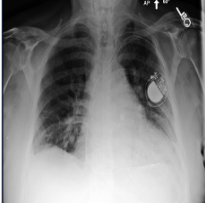
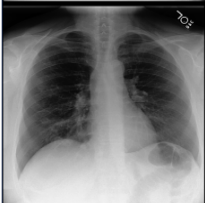

Input Image	Ground Truth	STL	BA-MTL
	heart size remains mildly enlarged mediastinal and hilar contours are normal pulmonary vasculature is normal lungs are clear no pneumothorax or pleural effusion is present no acute osseous abnormalities seen no subdiaphragmatic free air is present	tissue on the frontal view the lateral view is limited due to patient has been interval placement of a right picc line has been removed there is no pleural effusion or pneumothorax the cardiac and mediastinal silhouettes are unremarkable no acute	cardiac mediastinal and hilar contours are normal pulmonary vasculature is normal lungs are clear no pleural effusion or pneumothorax is present no acute osseous abnormalities identified no free air below the right hemidiaphragm is seen the right hemidiaphragm is seen there
	the cardiac mediastinal and hilar contours are normal lung volumes are low no focal consolidation pleural effusion or pneumothorax is visualized there are no acute osseous abnormalities no free air is demonstrated under the diaphragms	tooth is new since the prior study is not clearly visible on the frontal view though this is not substantiated on the frontal view there is not excluded there is no pulmonary edema no pleural effusion or pneumothorax is identified no	the lungs are clear without focal consolidation no pleural effusion or pneumothorax is seen the cardiac and mediastinal silhouettes are stable and hilar contours are unremarkable there is no acute osseous abnormalities are detected on the lateral view no free air
	transvenous right pacer lead follows the expected course into the right ventricle moderate to severe cardiomegaly is unchanged prominence of the pulmonary vasculature is unchanged and compatible with mild vascular congestion there is no focal lung consolidation there is no pleural effusion or pneumothorax	and right chest wall portacath is again noted with tip in the right atrium the lungs are clear without focal consolidation pleural effusion or pneumothorax the cardiac silhouette is enlarged there is a right upper lobe retrocardiac opacity is again seen	with mild to moderate interstitial edema the heart is mildly enlarged the mediastinal and hilar contours are normal pulmonary vascular congestion is not engorged the lungs are otherwise clear without focal consolidation no pleural effusion or pneumothorax is seen there is
	chest the lungs are clear without focal opacities pulmonary edema pleural effusion or pneumothorax the cardiac and mediastinal contours are normal there is no free air beneath the right hemidiaphragm	on the lateral view the the lateral view is a the chest ct chest the lungs are clear there is no pleural effusion or pneumothorax the cardiomeastinal silhouette is normal no acute osseous abnormalities are identified there is no free air	this study is clear no focal consolidation effusion or pneumothorax the cardiomeastinal silhouette is normal imaged osseous structures are intact no free air below the right hemidiaphragm is seen the right hemidiaphragm is seen the right hemidiaphragm is seen the right
	streaky bibasilar opacities are seen most suggestive of atelectasis the lungs are otherwise clear noting a slightly the cardiomeastinal silhouette is normal tortuous descending thoracic aorta	rotation to low lung volumes cause bronchovascular crowding no focal consolidation pleural effusion or pneumothorax heart size is normal mediastinal contour is normal imaged osseous structures are intact no free air below the right hemidiaphragm is seen diaphragms and lower thoracic	projecting over the right chest wall with the heart size is normal the mediastinal and hilar contours are unremarkable there is no pleural effusion or pneumothorax the pulmonary edema the pulmonary edema there is no pneumothorax is seen within the right

Figure 6.4: Illustrative comparison of generated reports from different learning approaches with ground truth

In the second sample, the STL model again demonstrates strong hallucinations, incomplete and disjointed sentence structure, and fails to capture important details. The report generated by the BA-MTL model aligns more closely with the ground truth but it still shows some omissions and grammatical errors.

The results from the third sample also show that the STL model includes hallucinations, such as mentioning a “right chest wall portacath”, and uses poor grammar and structure, like “there

is a right upper lobe retrocardiac opacity is again seen.” It also incorrectly states the cardiac silhouette is enlarged. BA-MTL model correctly stated cardiomegaly, however, stated severity level of cardiomegaly does not align with the ground truth. It also includes “the mediastinal and hilar contours are normal” which is not presented in GT, but it is important to note that the generated statement is clinically correct. BA-MTL model’s report is more coherent, however, it still has some grammar issues.

In the fourth sample, the STL model’s report contains repetitions and non-logical phrasing like “on the lateral view the the lateral view is a the chest ct chest,”. Despite this, it correctly identifies clear lungs, no pleural effusion, pneumothorax, or free air, and normal cardiac and mediastinal contours. The BA-MTL model’s report also includes these findings, however, it also includes repetitions like “the right hemidiaphragm is seen”. In this sample, we observed that both models avoid strong hallucinations but suffer from the repetition of irrelevant statements.

The STL model’s report in the last sample discusses “rotation to low lung volumes” and “bronchovascular crowding,” which are not mentioned in the GT or exist in the input image. It correctly identifies no pleural effusion or pneumothorax and normal heart size, however, it fails to mention the abnormalities noted in the GT. Additionally, it ends with incomplete sentences like “diaphragms and lower thoracic.” The BA- MTL model’s report is more concise and better in grammar structure, however, it still has repetitions like “the pulmonary edema the pulmonary edema”. Most importantly, both models fail to detect the crucial findings presented in the input image as well as ground truth.

6.5 Discussion and Conclusion

This chapter proposed a novel framework for report generation by leveraging multi-modal data and a multi-task learning approach. Our proposed models aimed to improve the representation learning capabilities by optimising the relevant tasks. Ultimately, it bridges the gap between image understanding and natural language generation by enhancing the quality and coherence of generated medical reports.

The problem is approached using two architectural designs and all models are trained using single and multi-task learning strategies. Additionally, multi-modal learning is employed in both approaches by integrating 10 additional features along with visual data. The first framework,

CM-MTL, was designed to aim to prioritise the report generation task whereas the second framework, BA-MTL, employed a balanced approach to improve the performance of all tasks.

The results from CM-MTL demonstrate the benefits of multi-task learning, particularly with proper task weighting, for improving text generation quality. The multi-task model with text generation prioritisation (CM-MTL-TP) outperformed single-task learning baselines across all language generation metrics. While single-task learning achieved better performance on the auxiliary tasks of ordinal and multi-label classification, it suffered from severe overfitting after only a few epochs. In contrast, multi-task learning helped prevent overfitting for these tasks and train for additional epochs.

Qualitative analysis of CM-MTL models also showed that CM-MTL-TP generated more coherent narratives that better captured logical relationships between medical findings. The generated reports exhibited good identification of key findings from the images.

However, the convergence problem of the classification tasks encountered in the CM-MTL models may be leading to imbalanced learning and preventing better representation learning for data modalities. BA-MTL model addressed this problem by enhancing the classification decoders with attention mechanisms and applying gradient clipping. These improvements allow us to adjust each task's converge and manage the learning process dynamically instead of trying to assign the weights manually.

BA-MTL results show that multi-task learning significantly enhances the performance of report generation and multi-label classifier on all metrics used, while the ordinal classifier demonstrates a reduction in performance. Visual results from the BA-MTL model also proved that multi-task learning helps considerably to capture key findings and generate consistent and coherent text. It also tends to less hallucinate compared to STL models and is better at detecting abnormalities, however, repeated phrases still remain a challenge in both approaches.

Our proposed approach and findings can inform future research on combining computer vision and natural language processing for medical applications. Furthermore, our approach's success in the medical domain suggests the potential for its generalisation to other domains. The principles of the integrated learning approach, as demonstrated in our research, can be applied to a wide range of applications beyond medical image analysis. For example, in image captioning (Sirisha and Sai Chandana 2022), our model's ability to understand the visual con-

tent and generate coherent textual descriptions could significantly enhance the accessibility and interpretability of images in various fields, including social media, e-commerce, and more.

Moreover, in grounded story generation (Hong et al. 2023), our approach can be extended to create compelling and contextually relevant narratives based on visual cues, making it suitable for content generation in the entertainment and creative industries. The integrated learning approaches, which proved effective in our medical domain application, could play an important role in advancing these related domains, improving the quality and relevance of content generated from visual inputs.

Finally, the synergies between vision and language training could lead to more contextual, logical, and human-like computer-generated text. Our model, however, still has limitations in optimising content selection and flow that provide opportunities for improvement. However, the success demonstrated in this medical application underscores the potential of multi-modal, multi-task learning for a wide range of domains.

Chapter 7

CONCLUSIONS AND FUTURE OUTLOOK

7.1 Overview

In this thesis, we investigate the automated generation of integrated and structured radiology reports by leveraging clinically relevant information and tasks to establish a more robust pipeline. This research is framed by two central research questions. The first research question—how robust frameworks can be established to enhance the utilisation of multiple modalities and ensure seamless integration and joint processing of diverse data— is addressed by proposing two frameworks that are presented in Chapter 4 and Chapter 5. Both frameworks address this question by introducing innovative approaches to handle multiple inputs simultaneously and integrate various data types. These frameworks demonstrate how pre-processing pipelines and input-specific encoders could effectively prepare diverse data for further processing, resulting in a unified embedding that optimally represents the integrated features from all data sources.

The second research question—how representation learning capabilities of neural network models can be improved through joint optimisation for relevant tasks, and the impact on generalisation and performance of radiology report generation— is addressed by the models presented in Chapter 6. This framework investigated the potential of multi-task learning paradigms to improve the model’s ability to capture complex relationships between different modalities and tasks, ultimately leading to more accurate and contextually rich radiology reports.

7.2 Significance and Implications

This research offers a multifaceted contribution to the field of medical report generation. We introduced a multi-input end-to-end network by incorporating medical ontology information to enrich image representation. This framework demonstrates how supplementary non-imaging data can improve the network’s ability to interpret image inputs and the importance of domain knowledge in image representation.

One of the key findings of this approach is highlighting the importance of case-specific information in assisting neural networks to fully comprehend and interpret image inputs. As most improvements are observed in normal cases, this finding suggests that the network requires more explicit and domain-specific knowledge beyond imaging data and information already included (utilised) in the final report.

The development of the FullFusion method represents a substantial improvement forward in multi-modal data fusion techniques. It simultaneously processes and combines 11 distinct non-imaging features which are traditionally not included in the final report. These features comprise information embedded in the images and radiologists’ working patterns. During the typical report generation process, radiologists reference many details observed in the images, which are not explicitly documented in the reports. Our approach incorporates these previously non-utilised data points, representing an innovative step not explored in existing literature. FullFusion method outperforms both single and group-based baselines across all evaluated metrics (as presented in Table 5.2), demonstrating the advancement and efficacy of this feature integration strategy.

The FullFusion approach is further extended and refined in the cross-task framework presented in Chapter 6. This framework enhances the joint optimisation process by incorporating multi-modal fusion strategies. As a result, it achieves a ROUGE-L score of 0.33 which demonstrates a significant improvement over the typical range of 0.26 to 0.28 reported in the comparative literature as shown in Table 6.5. An approximate 22% increase underscores the efficacy of our proposed methodology in advancing the state-of-the-art in this domain.

Importantly, the cross-task framework demonstrates improved coherence in narrative generation. As illustrated in Figure 6.4, this approach reduces hallucination compared to singular task baselines, indicating enhanced representation learning and logical relationship capture.

Furthermore, the strength of our approaches is demonstrated through the extensive ablation studies and systematic analyses detailed in this thesis. These include (1) a component-wise analysis demonstrating the impact of different views of the image and various encoder-decoder variants, (2) an assessment of the contribution of individual modalities and modality groups to the generated text and (3) an examination of the effect of different weighting approaches on learned features across various tasks.

From a clinical perspective, this thesis highlights the critical importance of incorporating supplementary data for automated chest X-ray (CXR) report generation. While recent literature has explored the integration of non-imaging or additional data sources in report generation, resulting in multi-modal data utilisation emerging as a prominent research direction, our work stands out in its comprehensive approach to data integration and joint optimisation. To the best of our knowledge, this study remains the only one to incorporate such a wide array of disparate data types and optimise complex vision tasks in radiology report generation, proposing a unique architecture specifically designed to handle this multi-modal complexity.

Overall, the proposed approaches not only improve text generation accuracy but also present a versatile framework applicable to various multi-modal data scenarios. By quantifying the performance of individual modalities and demonstrating the synergistic effects of multi-modal fusion, this thesis offers valuable insights and a roadmap for future developments in multi-modal and cross-modal AI systems.

7.3 Nuances in Multi-Modal Learning

In thoroughly assessing the overall results, this thesis also has yielded several additional findings that have significant implications for multi-modal learning. Comprehending the problem domain's boundaries, challenges, and framework limitations within the application area was crucial for identifying the optimal approach. To start with, evaluating dataset constraints is an important step in domain-specific modality translation tasks because the network components necessitate sufficient capacity to process the input and generate meaningful outputs. When input images are similar and their corresponding reports consist of long sequences rather than brief indications of normality, the decoder must fully utilise the image representation throughout the decoding process to understand the context. The selection of the decoder, as well as utilising the hybrid image representation, helps the model reduce hallucinations in the generated

reports.

Secondly, in medical applications, particularly in radiology, the approach cannot be a simple adaptation of the relevant methodology or data handling; the specialised vocabulary used in reports, the diversity of the report’s language and the critical nature of accuracy in medical diagnosis all require specialised consideration. It is important to handle the task in accordance with the working field and the original methodology of the radiologists. We prioritised alignment with real-world practices over maximising dataset size or NLG performance metrics. This approach involved excluding duplicated image views and reports that contain historical notes that cannot be extracted from the current image, ensuring temporal alignment between inputs and target reports, and balancing report types to prevent bias towards normal cases. While this resulted in a smaller dataset compared to existing literature, it enhances the clinical relevance and reliability of our results.

Furthermore, some data that may seem insignificant or did not improve results on its own can become valuable when combined with other data types. For example, a patient’s complaint or symptoms alone might not be very informative, but when combined with imaging data and relevant clinical data, it could provide crucial context for diagnosis. For instance, a patient reporting chest pain might not offer enough information for a conclusive diagnosis. However, when this symptom is combined with high blood pressure and imaging data, it provides more context that can lead to a more accurate diagnosis of conditions like heart disease or pneumonia. Conversely, data that shows strong statistical significance individually might still reduce model accuracy if not properly integrated with other types or if it introduces conflicting information. This is exemplified in Table 5.3, which presents the individual contributions of each data point. Notably, the ICD title achieved better performance in a singular approach compared to the text-based approach. However, its performance was still lower than the FullFusion approach. The ICD title and text-based data may provide conflicting information, and the model might struggle to reconcile these differences, leading to reduced accuracy. By aggregating diverse data types, the model can leverage additional context and evidence to make more informed decisions. This approach helps in compensating for discrepancies between individual data sources, leading to improved accuracy and robustness. This observation demonstrates the importance of holistic data integration and the potential limitations of relying on individual data points in isolation, regardless of their apparent statistical strength.

Moreover, effectively managing data multi-modality introduces network complexity even for individual tasks. Balancing this complexity is important when dealing with both multi-modality and multi-tasking. Networks that are too simple may fail to capture the intricacies of multiple tasks, while overly complex networks can be difficult to train efficiently. Additionally, a static approach to task prioritisation, such as always prioritising the task with the highest accuracy, may not be optimal in all cases. As demonstrated in Section 6.4.4, a dynamic approach that adjusts task priorities during training resulted in better performance across both clinical and computational metrics.

7.4 Limitations and Challenges

The research presented in this thesis is subject to several limitations that not only highlight areas for future research but also demonstrate an understanding of the challenges inherent in the field of computer-aided radiology report generation.

A primary challenge lies in the inherent variability of radiologists' language use. Despite our efforts to normalise and standardise the text data, the diverse terminology used to describe similar findings (e.g., "within normal limits", "appears to be normal", "seems normal" and "normal") introduces ambiguity. This linguistic diversity, while reflective of real-world practices, complicates the training process and leads to inconsistencies in the generated reports.

This language variability issue is closely tied to another critical limitation: the inadequacy of current evaluation metrics for radiology report generation. Standard metrics fail to capture the nuanced differences between generated and ground truth reports, particularly in cases where small changes in wording can have significant clinical implications, such as the distinction between "no cardiomegaly" and "cardiomegaly."

Some researchers have employed the CheXpert automated labelling tool to categorise both machine-generated and reference reports into 14 different categories related to thoracic diseases and support devices to tackle this problem. By comparing the CheXpert-labelled outcomes of the generated reports with those of the ground truths, they evaluated model performance using precision, recall, and F1-score. In our efforts to address this problem, we leveraged the BioClinical BERT model for contextual understanding and semantic similarity, which represents a step forward. While these advancements are promising, there remains a need for more sophisticated

evaluation methods that can assess content selection, contextual similarity, clinical relevance, and sentence structure in a manner that aligns with radiological expertise.

The challenges of language diversity and metric inadequacy are further compounded by inconsistencies in reporting practices among radiologists. Not all normalities are explicitly stated in every report, presenting a challenge in evaluating the model’s performance. When the model generates accurate predictions that are not present in the original report, it becomes difficult to distinguish between correct inferences and potential errors. This limitation underscores the need for a more comprehensive approach to ground truth data collection and annotation, possibly involving multiple expert reviews to create more complete, structured reference reports.

The research is further constrained by the lack of additional, comparable datasets that include all the relevant multi-modal data, limiting the validation of the proposed methods’ generalisability. We created balanced subsets from the MIMIC database for each approach. However, since all the data still originates from a single hospital, this limitation remains unaddressed.

Finally, the computational demands of processing vast amounts of multi-modal data simultaneously present a practical limitation to the research. Filtering and limiting data size, due to restricted access to high-performance computing resources, may have impacted the model’s potential performance and the scope of experiments conducted.

7.5 Future Directions

The limitations identified in this research open up several avenues for future work, which could significantly advance the field of computer-aided radiology report generation.

1. Expanding Model Generalisation in Cross-Domain Applicability

Our empirical evaluation provides robust evidence supporting the superiority of our multi-modal strategy over traditional image-only or single-modal approaches. The successful implementation of multi-modal and multi-task learning frameworks in radiology not only advances the field but also suggests promising avenues for application in other medical domains that necessitate the integration of multi-modal data.

Our network is designed with adaptability in mind, allowing customisation and reproducibility. The principles demonstrated in this thesis can enhance the quality and coherence of generated content in different domains. This cross-domain applicability highlights

the impact of the proposed methodologies, suggesting a wide range of future research directions and practical applications. Importantly, the methodologies are not confined to the specific setup and can be applied more broadly, fostering innovation beyond our context. They have the potential to enhance existing image-to-text generation models like BLIP (J. Li, D. Li, et al. 2022) and LLaVA (H. Liu et al. 2024). This underscores that our research not only stands independently but also holds the potential to advance current state-of-the-art systems in this field.

2. Enhancing Contextual Adaptability and Reducing Hallucination

Exploring patient history over time and comparing it with previous studies presents opportunities to refine our model’s capabilities. By incorporating longitudinal data and patient-specific factors—such as chronic conditions like smoking history—we can tailor our approach to better account for relevant clinical features. This direction not only aims to improve the accuracy of diagnostic assessments but also holds the potential to address the persistent issue of hallucination in language generation tasks.

Hallucination remains the most prevalent and significant problem in language generation tasks. While the proposed approaches have mitigated this issue to some extent, its complete elimination remains unsolved. This limitation is particularly concerning in real-life medical applications, where false information could have serious consequences.

Adopting adversarial training techniques alongside knowledge distillation methods is also promising to mitigate hallucination. Adversarial training involves training a discriminator network alongside the generator to distinguish between real and generated outputs, penalising hallucinated content by aligning generated data closer to real-world patterns. Knowledge distillation guides models to mimic outputs from more reliable sources or ensembles, reducing the generation of unrealistic or improbable outputs.

3. Optimising Large Language Models with Parameter-Efficient Fine-Tuning

Another direction is the integration of Large Language Models (LLMs) as decoders in our network, coupled with parameter-efficient fine-tuning techniques and soft-prompt engineering. This approach could potentially address some of the current limitations in language generation and adaptability.

LLMs, pre-trained on vast amounts of text data, have demonstrated remarkable capabil-

ities in understanding and generating human-like text across various domains. By using an LLM as a decoder, we can leverage its broad knowledge base and linguistic capabilities to generate more coherent and contextually appropriate radiology reports. Parameter-efficient fine-tuning techniques, such as LoRA (Low-Rank Adaptation) or Adapter layers, allow us to adapt these large models to our specific task without the need to update all parameters. This approach reduces computational requirements and mitigates the risk of catastrophic forgetting, where the model loses its general language understanding while adapting to a specific task.

Soft-prompt engineering involves prepending learnable continuous vectors to the input of the LLM. These "soft prompts" can be optimised during fine-tuning to guide the model towards generating task-specific outputs. In our context, we will design soft prompts that encode radiological expertise and domain-specific knowledge, potentially improving the accuracy and relevance of generated reports. This combination of techniques could enable us to create a more flexible and powerful report generation system that can adapt to the nuances of radiological language while maintaining the ability to generate coherent and contextually appropriate text.

4. Exploring Interpretability and Causality for Real-World Deployment

Given the critical nature of medical applications, enhancing the interpretability and causal understanding of our models is crucial for real-world deployment. Future work will focus on developing techniques to provide clear explanations for the model's decisions and outputs. We plan to investigate methods such as attention visualisation, saliency mapping, and concept-based explanations to make the model's decision-making process more transparent. This could involve highlighting which parts of the input images or clinical data most influenced the generated report, helping radiologists understand and verify the model's reasoning.

Furthermore, we aim to incorporate causal inference techniques to better understand the relationships between different input features and the generated outputs. This could involve developing causal models that capture the underlying mechanisms of disease progression and manifestation in medical images, leading to more robust and reliable report generation. These advancements in interpretability and causality will not only improve the trustworthiness of the system but also potentially provide new insights into radio-

logical practice, supporting both the model's adoption and the advancement of medical knowledge.

5. Expanding to Different Data Sources: Ultrasound Scans in Pregnancy

To further validate and extend our methodology, we are initiating a collaboration with a local hospital to collect fetal ultrasound scans, particularly focusing on the 12- and 20-week screening scans during pregnancy. This new data source presents an exciting opportunity to apply our improved methodology to a different but equally critical area of medical imaging.

The 20-week scan, in particular, involves the sonographer looking for 11 rare conditions. Our goal is to adapt our multi-modal, multi-task learning approach to this new context. We aim to generate comprehensive reports for these ultrasound scans, classify the baby's development as normal or abnormal, indicate the likelihood of the 11 rare conditions, and predict the baby's gender.

This expansion to ultrasound scans will allow us to test the generalisability of our approach across different imaging modalities and medical contexts. It also presents new challenges, such as incorporating time-series data from fetal development. By successfully adapting our methodology to this new domain, we could potentially develop a powerful tool to assist in prenatal care, providing more accurate and comprehensive information to healthcare providers and expectant parents.

6. Integrating Computational Models in Clinical Workflows

Acknowledging that Large Language Models (LLMs) have revolutionised text generation capabilities in recent years, and have increasingly incorporated multi-modal data integration, generation, and explanations. However, current models like GPT-4V, despite its prominence, cannot yet generate radiology reports accurately, even though it has far more parameters than our best model. Recent research (Jiang et al. 2024) challenges the practicality of employing GPT-4V in radiology workflows, highlighting its current limitations in understanding and generating meaningful radiology reports. Furthermore, it is important to consider the practical constraints faced by researchers in the healthcare domain from a computational perspective. Despite the availability of advanced multi-modal LLMs, the resources required for training or even fine-tuning these models are often prohibitively large, rendering their practical application challenging in many research projects. From

a clinical standpoint, and considering the existing infrastructure in most hospitals in the UK and other developed countries, there is a lack of support for these systems to generate radiology reports. Effective cloud integration and access to user-friendly cloud resources within hospitals are essential, but such infrastructure is not yet commonplace. Although there is significant potential for the use of LLMs in radiology report generation, addressing this challenge requires tackling the more complex issue of integrating large-scale cloud-based systems to manage data and train these models effectively. This highlights the need for future research and development to focus on creating feasible solutions for integrating advanced computational models within the existing healthcare infrastructure.

These directions represent opportunities to advance further in the field of computer-aided medical imaging analysis and report generation. These efforts aim not only to improve the technical capabilities of our systems but also to enhance their practical utility and trustworthiness in real-world clinical settings.

7.6 Final Remarks

In this thesis, we present a comprehensive investigation of deep neural architectures for vision-language modelling, specifically for the task of CXR report generation. We explore the incorporation of multi-modal data to analyse the value that imaging data brings to radiology report generation. Additionally, we address the real-world challenge of multi-label and ordinal classification for CXR interpretation and report generation through multi-modal cross-task learning. The findings revealed by this research, combined with methodological contributions in vision-language modelling, are expected to constitute a road map for future developments in multi-modal medical reporting and image interpretation systems.

References

- Akhter, Yasmeena, Singh, Richa, and Vatsa, Mayank (2023). “AI-based radiodiagnosis using chest X-rays: A review”. In: *Frontiers in Big Data* 6, p. 1120989.
- Aksoy, Nurbanu, Ravikumar, Nishant, and Frangi, Alejandro F (2023). “Radiology report generation using transformers conditioned with non-imaging data”. In: *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 12469. SPIE, pp. 146–154.
- Alfarghaly, Omar, Khaled, Rana, Elkorany, Abeer, Helal, Maha, and Fahmy, Aly (2021). “Automated radiology report generation using conditioned transformers”. In: *Informatics in Medicine Unlocked* 24, p. 100557.
- Anderson, Peter, He, Xiaodong, Buehler, Chris, Teney, Damien, Johnson, Mark, Gould, Stephen, and Zhang, Lei (2018). “Bottom-up and top-down attention for image captioning and visual question answering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086.
- Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua (2014). “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473*.
- Banerjee, Satanjeev and Lavie, Alon (2005). “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72.

- Bayoudh, Khaled, Knani, Raja, Hamdaoui, Fayçal, and Mtibaa, Abdellatif (2021). “A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets”. In: *The Visual Computer*, pp. 1–32.
- Bhattacharjee, Deblina, Zhang, Tong, Süssstrunk, Sabine, and Salzmann, Mathieu (2022). “Mult: An end-to-end multitask learning transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12031–12041.
- Chan, Heang-Ping, Samala, Ravi K, Hadjiiski, Lubomir M, and Zhou, Chuan (2020). “Deep learning in medical image analysis”. In: *Deep learning in medical image analysis: challenges and applications*, pp. 3–21.
- Chauhan, Geeticka, Liao, Ruizhi, Wells, William, Andreas, Jacob, Wang, Xin, Berkowitz, Seth, Horng, Steven, Szolovits, Peter, and Golland, Polina (2020). “Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer, pp. 529–539.
- Chen, Xinlei and Lawrence Zitnick, C (2015). “Mind’s eye: A recurrent visual representation for image caption generation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2422–2431.
- Chen, Zhihong, Song, Yan, Chang, Tsung-Hui, and Wan, Xiang (2020). “Generating radiology reports via memory-driven transformer”. In: *arXiv preprint arXiv:2010.16056*.
- Cornia, Marcella, Stefanini, Matteo, Baraldi, Lorenzo, and Cucchiara, Rita (2020). “Meshed-memory transformer for image captioning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10578–10587.
- Demner-Fushman, Dina, Kohli, Marc D, Rosenman, Marc B, Shooshan, Sonya E, Rodriguez, Laritza, Antani, Sameer, Thoma, George R, and McDonald, Clement J (2016). “Preparing a collection of radiology examinations for distribution and retrieval”. In: *Journal of the American Medical Informatics Association* 23.2, pp. 304–310.
- Donahue, Jeffrey, Anne Hendricks, Lisa, Guadarrama, Sergio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor (2015). “Long-term recurrent convolutional

- networks for visual recognition and description”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634.
- Gan, Zhe, Gan, Chuang, He, Xiaodong, Pu, Yunchen, Tran, Kenneth, Gao, Jianfeng, Carin, Lawrence, and Deng, Li (2017). “Semantic compositional networks for visual captioning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5630–5639.
- Gao, Jing, Li, Peng, Chen, Zhikui, and Zhang, Jianing (2020). “A survey on deep learning for multimodal data fusion”. In: *Neural Computation* 32.5, pp. 829–864.
- Hayat, Nasir, Geras, Krzysztof J, and Shamout, Farah E (2022). “MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images”. In: *arXiv preprint arXiv:2207.07027*.
- Hodosh, Micah, Young, Peter, and Hockenmaier, Julia (2013). “Framing image description as a ranking task: Data, models and evaluation metrics”. In: *Journal of Artificial Intelligence Research* 47, pp. 853–899.
- Hong, Xudong, Mehra, Khushboo, Sayeed, Asad, and Demberg, Vera (Sept. 2023). “Visually Grounded Story Generation Challenge”. In: *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*. Prague, Czechia: Association for Computational Linguistics, pp. 17–22. URL: <https://aclanthology.org/2023.inlg-genchal.3>.
- Huang, Shih-Cheng, Pareek, Anuj, Seyyedi, Saeed, Banerjee, Imon, and Lungren, Matthew P (2020). “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines”. In: *NPJ digital medicine* 3.1, p. 136.
- Irvin, Jeremy, Rajpurkar, Pranav, Ko, Michael, Yu, Yifan, Ciurea-Ilcus, Silviana, Chute, Chris, Marklund, Henrik, Haghgoo, Behzad, Ball, Robyn, Shpanskaya, Katie, et al. (2019). “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 590–597.
- Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross, Guadarrama, Sergio, and Darrell, Trevor (2014). “Caffe: Convolutional architecture

- for fast feature embedding”. In: *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675–678.
- Jiang, Yuyang, Chen, Chacha, Nguyen, Dang, Mervak, Benjamin M, and Tan, Chenhao (2024). “GPT-4V Cannot Generate Radiology Reports Yet”. In: *arXiv preprint arXiv:2407.12176*.
- Jin, Haibo, Che, Haoxuan, Lin, Yi, and Chen, Hao (2024). “Promptmrg: Diagnosis-driven prompts for medical report generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 3, pp. 2607–2615.
- Jing, Baoyu, Xie, Pengtao, and Xing, Eric (2017). “On the automatic generation of medical imaging reports”. In: *arXiv preprint arXiv:1711.08195*.
- Johnson, Alistair EW, Bulgarelli, Lucas, Shen, Lu, Gayles, Alvin, Shammout, Ayad, Horng, Steven, Pollard, Tom J, Hao, Sicheng, Moody, Benjamin, Gow, Brian, et al. (2023). “MIMIC-IV, a freely accessible electronic health record dataset”. In: *Scientific data* 10.1, p. 1.
- Johnson, Alistair EW, Pollard, Tom J, Greenbaum, Nathaniel R, Lungren, Matthew P, Deng, Chih-ying, Peng, Yifan, Lu, Zhiyong, Mark, Roger G, Berkowitz, Seth J, and Horng, Steven (2019). “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs”. In: *arXiv preprint arXiv:1901.07042*.
- Karpathy, Andrej and Fei-Fei, Li (2015). “Deep visual-semantic alignments for generating image descriptions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137.
- Kokkinos, Iasonas (2017). “Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6129–6138.
- LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick (1998). “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.
- Lewis, Mike, Liu, Yinhan, Goyal, Naman, Ghazvininejad, Marjan, Mohamed, Abdelrahman, Levy, Omer, Stoyanov, Ves, and Zettlemoyer, Luke (2019). “Bart: Denoising sequence-to-

- sequence pre-training for natural language generation, translation, and comprehension”. In: *arXiv preprint arXiv:1910.13461*.
- Li, Guang, Zhu, Linchao, Liu, Ping, and Yang, Yi (2019). “Entangled transformer for image captioning”. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8928–8937.
- Li, Junnan, Li, Dongxu, Xiong, Caiming, and Hoi, Steven (2022). “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation”. In: *International conference on machine learning*. PMLR, pp. 12888–12900.
- Li, Junnan, Selvaraju, Ramprasaath, Gotmare, Akhilesh, Joty, Shafiq, Xiong, Caiming, and Hoi, Steven Chu Hong (2021). “Align before fuse: Vision and language representation learning with momentum distillation”. In: *Advances in neural information processing systems* 34, pp. 9694–9705.
- Lin, Chin-Yew (2004). “Rouge: A package for automatic evaluation of summaries”. In: *Text summarization branches out*, pp. 74–81.
- Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence (2014). “Microsoft coco: Common objects in context”. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, pp. 740–755.
- Liu, Fenglin, Ge, Shen, Zou, Yuexian, and Wu, Xian (2022). “Competence-based multimodal curriculum learning for medical report generation”. In: *arXiv preprint arXiv:2206.14579*.
- Liu, Haotian, Li, Chunyuan, Wu, Qingyang, and Lee, Yong Jae (2024). “Visual instruction tuning”. In: *Advances in neural information processing systems* 36.
- Liu, Wei, Chen, Sihan, Guo, Longteng, Zhu, Xinxin, and Liu, Jing (2021). “Cptr: Full transformer network for image captioning”. In: *arXiv preprint arXiv:2101.10804*.
- Lu, Jiasen, Xiong, Caiming, Parikh, Devi, and Socher, Richard (2017). “Knowing when to look: Adaptive attention via a visual sentinel for image captioning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383.

- Mao, Junhua, Wei, Xu, Yang, Yi, Wang, Jiang, Huang, Zhiheng, and Yuille, Alan L (2015). “Learning like a child: Fast novel visual concept learning from sentence descriptions of images”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2533–2541.
- Marques, Gonçalo, Agarwal, Deevyankar, and De la Torre Díez, Isabel (2020). “Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network”. In: *Applied soft computing* 96, p. 106691.
- Montagnon, Emmanuel, Cerny, Milena, Cadrin-Chênevert, Alexandre, Hamilton, Vincent, Derennes, Thomas, Ilinca, André, Vandenbroucke-Menu, Franck, Turcotte, Simon, Kadoury, Samuel, and Tang, An (2020). “Deep learning workflow in radiology: a primer”. In: *Insights into imaging* 11, pp. 1–15.
- Moon, Jong Hak, Lee, Hyungyung, Shin, Woncheol, Kim, Young-Hak, and Choi, Edward (2022). “Multi-modal understanding and generation for medical images and text via vision-language pre-training”. In: *IEEE Journal of Biomedical and Health Informatics* 26.12, pp. 6070–6080.
- Nooralahzadeh, Farhad, Gonzalez, Nicolas Perez, Frauenfelder, Thomas, Fujimoto, Koji, and Krauthammer, Michael (2021). “Progressive transformer-based generation of radiology reports”. In: *arXiv preprint arXiv:2102.09777*.
- Pang, Ting, Li, Peigao, and Zhao, Lijie (2023). “A survey on automatic generation of medical imaging reports based on deep learning”. In: *BioMedical Engineering OnLine* 22.1, pp. 1–16.
- Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing (2002). “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.
- Pavlopoulos, John, Kougia, Vasiliki, Androutsopoulos, Ion, and Papamichail, Dimitris (2022). “Diagnostic captioning: a survey”. In: *Knowledge and Information Systems* 64.7, pp. 1691–1722.
- Pei, Xiangdong, Zuo, Ke, Li, Yuan, and Pang, Zhengbin (2023). “A Review of the Application of Multi-modal Deep Learning in Medicine: Bibliometrics and Future Directions”. In: *International Journal of Computational Intelligence Systems* 16.1, p. 44.

- Phi, M. (Sept. 2018). *Illustrated Guide to LSTM's and GRU's: A step by step explanation*. Published on 24 September. Towards Data Science. Towards Data Science. URL: <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21> (visited on 04/20/2020).
- Rajpurkar, Pranav, Irvin, Jeremy, Zhu, Kaylie, Yang, Brandon, Mehta, Hershel, Duan, Tony, Ding, Daisy, Bagul, Aarti, Langlotz, Curtis, Shpanskaya, Katie, et al. (2017). “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning”. In: *arXiv preprint arXiv:1711.05225*.
- Ramirez-Alonso, Graciela, Prieto-Ordaz, Olanda, López-Santillan, Roberto, and Montes-Y-Gómez, Manuel (2022). “Medical report generation through radiology images: an Overview”. In: *IEEE Latin America Transactions* 20.6, pp. 986–999.
- Safitra, Muhammad Fakhrol, Lubis, Muharman, Kusumasari, Tien Fabrianti, and Putri, Deyana Prastika (2024). “Advancements in Artificial Intelligence and Data Science: Models, Applications, and Challenges”. In: *Procedia Computer Science* 234, pp. 381–388.
- Sandstede, J, Lipke, C, Beer, M, Hofmann, S, Pabst, T, Kenn, W, Neubauer, S, and Hahn, D (2000). “Age-and gender-specific differences in left and right ventricular cardiac function and mass determined by cine magnetic resonance imaging”. In: *European radiology* 10, pp. 438–442.
- Sarvamangala, DR and Kulkarni, Raghavendra V (2022). “Convolutional neural networks in medical image understanding: a survey”. In: *Evolutionary intelligence* 15.1, pp. 1–22.
- Simonyan, Karen and Zisserman, Andrew (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556*.
- Singh, Sonit, Karimi, Sarvnaz, Ho-Shon, Kevin, and Hamey, Len (2021). “Show, tell and summarise: learning to generate and summarise radiology findings from medical images”. In: *Neural Computing and Applications* 33.13, pp. 7441–7465.
- Sirisha, Uddagiri and Sai Chandana, Bolem (2022). “Semantic interdisciplinary evaluation of image captioning models”. In: *Cogent Engineering* 9.1, p. 2104333.

- Sirshar, Mehreen, Paracha, Muhammad Faheem Khalil, Akram, Muhammad Usman, Alghamdi, Norah Saleh, Zaidi, Syeda Zainab Yousuf, and Fatima, Tatheer (2022). “Attention based automated radiology report generation using CNN and LSTM”. In: *Plos one* 17.1, e0262209.
- SS, Roshan Adhithya, Priyadharshini, M, and Kalinathan, Lekshmi (2023). “Image Caption Generation For Blind Users Of Social Media Websites”. In.
- Stahlschmidt, Sören Richard, Ulfenborg, Benjamin, and Synnergren, Jane (2022). “Multimodal deep learning for biomedical data fusion: a review”. In: *Briefings in Bioinformatics* 23.2, bbab569.
- Stefanini, Matteo, Cornia, Marcella, Baraldi, Lorenzo, Cascianelli, Silvia, Fiameni, Giuseppe, and Cucchiara, Rita (2022). “From show to tell: A survey on deep learning-based image captioning”. In: *IEEE transactions on pattern analysis and machine intelligence* 45.1, pp. 539–559.
- Su, Yixuan, Shu, Lei, Mansimov, Elman, Gupta, Arshit, Cai, Deng, Lai, Yi-An, and Zhang, Yi (2021). “Multi-task pre-training for plug-and-play task-oriented dialogue system”. In: *arXiv preprint arXiv:2109.14739*.
- Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew (2015). “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.
- Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, and Wojna, Zbigniew (2016). “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.
- Tan, Mingxing and Le, Quoc (2019). “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR, pp. 6105–6114.
- Tanida, Tim, Müller, Philip, Kaissis, Georgios, and Rueckert, Daniel (2023). “Interactive and Explainable Region-guided Radiology Report Generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7433–7442.

- Vandenhende, Simon, Georgoulis, Stamatios, Van Gansbeke, Wouter, Proesmans, Marc, Dai, Dengxin, and Van Gool, Luc (2021). “Multi-task learning for dense prediction tasks: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.7, pp. 3614–3633.
- Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Lukasz, and Polosukhin, Illia (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.
- Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru (2015). “Show and tell: A neural image caption generator”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164.
- Wang, Cheng, Yang, Haojin, Bartz, Christian, and Meinel, Christoph (2016). “Image captioning with deep bidirectional LSTMs”. In: *Proceedings of the 24th ACM international conference on Multimedia*, pp. 988–997.
- Wang, Lin, Ning, Munan, Lu, Donghuan, Wei, Dong, Zheng, Yefeng, and Chen, Jie (2022). “An inclusive task-aware framework for radiology report generation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 568–577.
- Wang, Peng, Yang, An, Men, Rui, Lin, Junyang, Bai, Shuai, Li, Zhikang, Ma, Jianxin, Zhou, Chang, Zhou, Jingren, and Yang, Hongxia (2022). “Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework”. In: *International conference on machine learning*. PMLR, pp. 23318–23340.
- Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, and Summers, Ronald M (2018). “Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058.
- Wang, Zhanyu, Han, Hongwei, Wang, Lei, Li, Xiu, and Zhou, Luping (2022). “Automated radiographic report generation purely on transformer: A multicriteria supervised approach”. In: *IEEE Transactions on Medical Imaging* 41.10, pp. 2803–2813.

- Wei, Jason, Bosma, Maarten, Zhao, Vincent Y, Guu, Kelvin, Yu, Adams Wei, Lester, Brian, Du, Nan, Dai, Andrew M, and Le, Quoc V (2021). “Finetuned language models are zero-shot learners”. In: *arXiv preprint arXiv:2109.01652*.
- Wu, Xing, Li, Jingwen, Wang, Jianjia, and Qian, Quan (2023). “Multimodal contrastive learning for radiology report generation”. In: *Journal of Ambient Intelligence and Humanized Computing* 14.8, pp. 11185–11194.
- Xie, Saining, Girshick, Ross, Dollár, Piotr, Tu, Zhuowen, and He, Kaiming (2017). “Aggregated residual transformations for deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500.
- Xiong, Yuxuan, Du, Bo, and Yan, Pingkun (2019). “Reinforced transformer for medical image captioning”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer, pp. 673–680.
- Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron, Salakhudinov, Ruslan, Zemel, Rich, and Bengio, Yoshua (2015). “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR, pp. 2048–2057.
- Xue, Yuan, Xu, Tao, Rodney Long, L, Xue, Zhiyun, Antani, Sameer, Thoma, George R, and Huang, Xiaolei (2018). “Multimodal recurrent model with attention for automated radiology report generation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 457–466.
- Yang, Shaokang, Niu, Jianwei, Wu, Jiyan, and Liu, Xuefeng (2020). “Automatic medical image report generation with multi-view and multi-modal attention mechanism”. In: *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, pp. 687–699.
- Yang, Shuxin, Wu, Xian, Ge, Shen, Zhou, S Kevin, and Xiao, Li (2022). “Knowledge matters: Chest radiology report generation with general and specific knowledge”. In: *Medical image analysis* 80, p. 102510.

- Yang, Zhilin, Yuan, Ye, Wu, Yuexin, Cohen, William W, and Salakhutdinov, Russ R (2016). “Review networks for caption generation”. In: *Advances in neural information processing systems* 29.
- You, Quanzeng, Jin, Hailin, Wang, Zhaowen, Fang, Chen, and Luo, Jiebo (2016). “Image captioning with semantic attention”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659.
- Yuan, Jianbo, Liao, Haofu, Luo, Rui, and Luo, Jiebo (2019). “Automatic radiology report generation based on multi-view image fusion and medical concept enrichment”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 721–729.
- Zhang, Tianyi, Kishore, Varsha, Wu, Felix, Weinberger, Kilian Q, and Artzi, Yoav (2019). “Bertscore: Evaluating text generation with bert”. In: *arXiv preprint arXiv:1904.09675*.
- Zhang, Wei, Nie, Wenbo, Li, Xinle, and Yu, Yao (2019). “Image caption generation with adaptive transformer”. In: *2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, pp. 521–526.
- Zhang, Yizhe, Sun, Siqu, Galley, Michel, Chen, Yen-Chun, Brockett, Chris, Gao, Xiang, Gao, Jianfeng, Liu, Jingjing, and Dolan, Bill (2019). “Dialogpt: Large-scale generative pre-training for conversational response generation”. In: *arXiv preprint arXiv:1911.00536*.
- Zhao, Guosheng, Zhao, Zijian, Gong, Wuxian, and Li, Feng (2023). “Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification”. In: *Artificial Intelligence in Medicine* 146, p. 102714.
- Zhu, Xinxin, Li, Lixiang, Liu, Jing, Peng, Haipeng, and Niu, Xinxin (2018). “Captioning transformer with stacked attention modules”. In: *Applied Sciences* 8.5, p. 739.

Appendix A

Integrative Attention Network: Natural Image Application

A.1 Introduction

In order to validate the proposed concept of enhancing results by providing additional context and methodology, the experiments were conducted on a natural image dataset. A topic for each image has been extracted from the entire dataset including test and validation sets using an unsupervised topic modelling algorithm. The topics extracted were a set of words that are already placed in the annotations of the given image; therefore, these word collections have not been used for the experiments. Instead, all topics were manually tagged based on the keywords they contained, and a descriptive tag that was never seen by the network for each data was generated. These tags are used along with the image representations in the text generation process. This section provides the details of our approach to generating text from an image. Respectively, it briefly introduces the datasets, demonstrates the process of preparing text and image data for experiments and analyses the data used, provides the architectural design of the transformer-based network with implementational details, and finally presents results.

A.1.1 Natural Image Datasets

Flickr8K and MS-COCO benchmark datasets are used to demonstrate the effectiveness of the proposed architecture. The Flickr8K dataset (Hodosh et al. 2013) contains over 8,000 images with a different shape in JPEG format that is each paired with five different captions at the

sentence level. Each image has a different shape and format and different sources of descriptions for the photographs. MS-COCO (Microsoft Common Objects in Context) is a large-scale dataset (T.-Y. Lin et al. 2014) that consists of over 80,000 images, and each image has five different sentence-level annotations. The data format used for this task is similar for all three datasets, and an example image with five corresponding captions from the Flickr8K Dataset is shown in Figure A.1.



1. A child in a pink dress is climbing up a set of stairs in an entry way.
2. A girl going into a wooden building.
3. A little girl climbing into a wooden playhouse.
4. A little girl climbing the stairs to her playhouse.
5. A little girl in a pink dress going into a wooden cabin.

Figure A.1: An example image with different captions from Flickr8K dataset

A.1.2 Data Pre-processing and Dataset Formation

In the text data preparation, firstly, all characters were converted into lowercase, and punctuation, tokens with a number and stop words were removed. As mentioned in Section 3.3.1, each image is described with five different captions, the second step was to merge all captions into a five-sentence-long description for each image. The descriptions obtained were processed for the LDA algorithm by applying the following steps, respectively, 1) lemmatize the words, 2) create a dictionary and 3) create a bag-of-words corpus. After, the coherence values of the topic numbers from 1 to 100 at a step size of 10 were calculated to find the optimal topic number. After a series of experiments, the optimal number of topics was determined to be 21.

Image Name	1000268201_693b08cb0e.jpg
Descriptions	A child in a pink dress is climbing up a set of stairs in an entryway. A girl going into a wooden building. A little girl climbing into a wooden playhouse. A little girl climbing the stairs to her playhouse. A little girl in a pink dress going into a wooden cabin.
Lemmatized	child, pink, dress, climbing, set, stair, entry, girl, go, wooden, building, little, girl, climb, wooden, playhouse, little, girl, climb, stair, playhouse, little, girl, pink, dress, go, wooden, cabin

Table A.1: A sample of pre-processed text data for LDA model

In the dataset, each data point consists of the image ID, caption, associated topic and keywords. It is important to note that, the concatenated captions (Table A.1) were only used in topic modelling to ensure that all captions associated with an image are assigned the same topic. After creating the data points, we divide the dataset into training, validation, and testing sets with the ratio of 80%, 10% and 10%. The image, corresponding captions and topic are used as input during the training phase. Since we extracted a topic for each image in the entire dataset before creating the data points, the topic assigned is also used along with the image during inference for prediction. The details of topic extraction are explained in the Network Design section. For text parsing, 1) texts are converted to tokens, 2) ;start_i and ;end_i tokens are added to the captions to be decoded, and 3) out of vocabulary, words are replaced with ;junk_i . Next, each text data is padded to a fixed length. In image processing, all images are resized to 299 x 299, and position and colour augmentation are applied for each image in the training set. Finally, a data loader object is created, and the data is fetched from the dataset and fed into the model in batches.

A.1.3 Network Design

The current network architecture was built to demonstrate the feasibility of the proposed method. The caption generation model has four main components: the convolutional model, LDA model, stack of transformer-encoder blocks and stack of transformer-decoder blocks. The image is first processed by the convolutional model and the extracted features are fed into the transformer-encoder. A new representation of images is calculated by the self-attention layer and output is sent to the following encoder blocks as an input. Meanwhile, the topic model assigns a topic number and five keywords for each caption. These keywords are used in the manual topic tagging process to define a descriptive name for each topic, and 21 manual tags are formed as a result. To establish an implicit connection between image and captions through the tags generated, the decoder is conditioned with tags by concatenating the output of the first encoder block with tag embeddings. On the other hand, each of the five captions corresponding to the given image is converted into vector representation with positional encoding and fed to the transformer-decoder one by one along with the semantically-enhanced image representation.

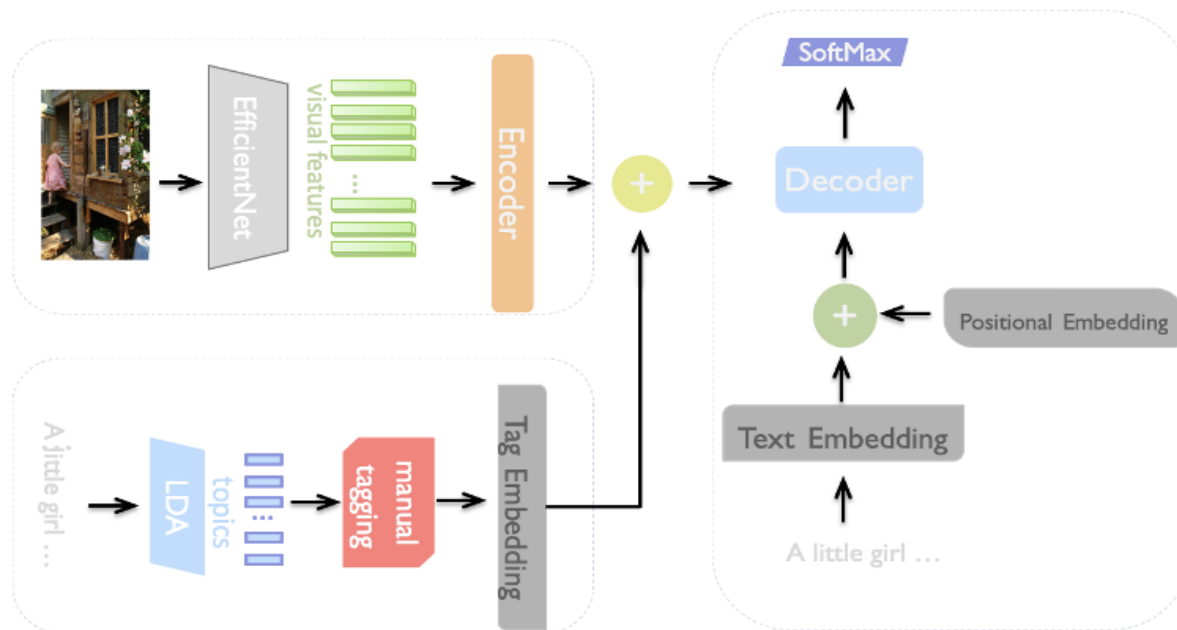


Figure A.2: An overall framework of using additional context

Implementation Details

In the encoding phase, EfficientNet has been used as a base model to extract visual features and all parameters are left non-trainable. The pre-trained model receives resized (299x299) images and generates a 1280-length visual feature vector. After applying layer normalisation, the vector is passed through the dense layer with the ReLU activation function. The output of the dense layer is sent to the multi-head self-attention layer followed by the normalisation layer. Meanwhile, the LDA model was used to extract the topics and keywords of each caption. LDA, however, generates different topics every time, even when the model is trained on the same corpus. Therefore, several experiments have been carried out in order to determine the optimal number of topics and the most relevant set of keywords. As a result, the ideal number was set to 21 and the extracted topics are represented by numbers from 0 to 20. Then, each topic number was manually assigned a label to describe the scene. For example, the topic that consist of the keywords “tent, fire, people, cold, play” was given the tag “camping”. The objective is here to allow the model to better grasp the content of the image by presenting previously unseen information to the network, as well as to establish an implicit connection to alleviate the impacts of the semantic gap. The tags are converted to a vector representation using the embedding layer. A self-attention block that is presented in the original transformer architecture has keys, queries, and values as demonstrated in Equation 2.4, where softmax is defined in Equation 2.2.

In our proposed model, the tag embedding has been used as a semantic feature along with the visual features as shown in Equation A.1 where SSAT stands for Semantic Self-Attention Tag.

$$\text{SSAT}(Q, [K_{\text{enc}}; K_{\text{tag}}], [V_{\text{enc}}; V_{\text{tag}}]) = \text{softmax} \left(\frac{Q[K_{\text{enc}}; K_{\text{tag}}]^{\top}}{\sqrt{d_k}} \right) [V_{\text{enc}}; V_{\text{tag}}] \in \mathbb{R}^{n \times (d_{v_{\text{enc}}} + d_{v_{\text{tag}}})} \quad (\text{A.1})$$

Where: K_{enc} is the key matrix from the encoder output

K_{tag} is the key matrix from the tag vector

V_{enc} is the value matrix from the encoder output

V_{tag} is the value matrix from the tag vector

[;] represents the concatenation operation

The concatenated $[K_{\text{enc}}; K_{\text{tag}}]$ and $[V_{\text{enc}}; V_{\text{tag}}]$ matrices are used to compute attention weights and the output.

Furthermore, the captions corresponding to the given image were passed through the embedding layer. However, the transformer does not implement a loop and all inputs are processed in parallel. Although this is one of its main advantages over recurrence models, the notion of the sequence order is lost during this operation. Therefore, positional information is injected into the output of the embedding layer with the positional encoding (Equation 2.5) of the input sequence and it is sent to the first layer of the decoder which is a masked multi-head self-attention layer.

Decoder begins with <start>token, generates the words one by one, and stops the decoding process when <end>token is generated. Due to its auto-regressive nature, it takes the information from the previous iteration to predict the next word. Different from self-attention in the encoder, the masking method is used here to ensure that the attention mechanism does not share any information about future tokens. In this way, each token only has access to information regarding itself and the previously generated values. The MHA layer is followed by the normalisation layer and encoder-decoder self-attention layer, respectively.

The second MHA layer is known as the encoder-decoder self-attention layer since it incorporates information from both the encoder and the decoder. When calculating self-attention, the queries matrix is created from the previous layer, and the semantically enhanced visual features are used for the keys and values matrix. In previous self-attention calculations, the sequence was paying attention to itself, however, the concept here is to ensure the model pays attention to the input

sequence when generating the target sequence. The output goes through the normalisation layer followed by the dense layer with the ReLU activation function and the last dense layer takes the outputs and functions as a classifier. The final output, then, passes through the SoftMax layer, which calculates probability scores between 0 and 1, and returns the word with the highest probability score as a predicted word.

The model was trained using the Adam optimiser with a learning rate of $1e-4$ and a customised SparseCategoricalCrossentropy loss function. A batch size of 64, a maximum caption length of 24, a maximum tags length of 3, a caption vocabulary size of 9000, a tags vocabulary size of 33, a dense dimension of 512, and an embedding dimension of 512 were used during training and validation.

A.1.4 Results and Discussion

This section first describes the inferences of experiments and presents the results of the models. The effectiveness of the proposed model has been demonstrated on two benchmark datasets: Flickr8K, and MS COCO. The results presented below have been conducted on the Flickr8K dataset due to limited access to GPU. After assigning the optimal components and parameters, the final model has been evaluated on both Flickr8K and MS COCO datasets. In order to evaluate the results of the models, BLEU Score and BERTScore were used. Finally, the limitations of the proposed model are discussed.

To begin with, as indicated in earlier sections, the adoption of transfer learning improves the model performance, especially when the dataset size is limited. In this context, the pre-trained models InceptionV3, VGG-19 and EfficientNet were used to analyse the impact of the different pre-trained models on the results. We have evaluated the models on 1020 samples of the unseen test set randomly selected from the Flickr8K dataset. Table A.2 displays the average Bleu -1 to -4 Score and BERTScore of the three-fold cross-validation results for the different pre-trained models. EfficientNet yielded better results than InceptionV3 and VGG19 in most metrics, therefore, EfficientNet was used in the final design to extract visual features from the given image.

In order to prove that providing additional context to the neural network has a beneficial impact on model performance, the comparison experiments were conducted with and without additional context. The identical train-test split was used during training in both models to

CNN Model	B_1	B_2	B_3	B_4	BS_P	BS_R	BS_F1
EfficientNet	0.55	0.41	0.42	0.48	0.52	0.50	0.50
InceptionV3	0.46	0.37	0.42	0.49	0.51	0.45	0.47
VGG-19	0.48	0.39	0.40	0.48	0.51	0.46	0.48

Table A.2: Performance Summary of EfficientNet, InceptionV3, and VGG-19, with 'B_n' for BLEU-n Scores and denoting BERTScore as BS with P for Precision, R for Recall, and F1 for F1 Score

avoid any bias caused by the dataset distribution. Table A.3 and Table A.4 demonstrate the average scores obtained from 5-set of experiments conducted by both models. Furthermore, we calculated probability values with a significance level of 0.05 for each metric, and as shown in Table A.5 the results are statistically significant.

Experiments	B_1	B_2	B_3	B_4	BS_P	BS_R	BS_F1
1-set	0.55	0.42	0.42	0.47	0.54	0.50	0.51
2-set	0.54	0.41	0.42	0.47	0.54	0.49	0.50
3-set	0.55	0.43	0.44	0.49	0.55	0.49	0.50
4-set	0.53	0.41	0.42	0.47	0.54	0.48	0.50
5-set	0.53	0.42	0.41	0.46	0.53	0.49	0.51

Table A.3: Base Captioning Model Performance Across 5 Sets of Experiments

Experiments	B_1	B_2	B_3	B_4	BS_P	BS_R	BS_F1
1-set	0.56	0.43	0.44	0.49	0.56	0.51	0.52
2-set	0.57	0.44	0.45	0.50	0.55	0.50	0.51
3-set	0.57	0.43	0.45	0.51	0.57	0.51	0.52
4-set	0.58	0.44	0.46	0.52	0.57	0.52	0.53
5-set	0.57	0.44	0.45	0.51	0.56	0.51	0.52

Table A.4: Semantically Enhanced Captioning Model Performance Across 5 Sets of Experiments

Metrics	B_1	B_2	B_3	B_4	BS_P	BS_R	BS_F1
p_value	0.001	0.005	0.002	0.001	0.002	0.002	0.005

Table A.5: Statistical Significance (p-values) Comparing Evaluation Metrics for Captioning Models

After concluding that the results generated are not likely to occur randomly, and our proposed approach provides better results, the semantically-enhanced model was also tested on the MS COCO dataset. Since the size of the MS COCO dataset is over ten times larger than Flickr8k, the model was able to learn better and provide higher accuracy (Table A.6).

Dataset	B_1	B_2	B_3	B_4	BS_P	BS_R	BS_F1
Flickr8K	0.57	0.43	0.45	0.50	0.56	0.51	0.52
MS-COCO	0.61	0.51	0.52	0.55	0.62	0.57	0.55

Table A.6: Image Captioning Performance Analysis: BLEU and BERTscore on Flickr8K and MSCOCO datasets



A black and white dog is swimming in the water.



A black dog is running through the the water in the ocean.



A man in a green shirt and helmet is riding a bicycle.

Figure A.3: The captions generated from our semantically-enhanced captioning model

Although the current model generates considerable results (Figure A.3), semantic addition in this experiment did not exactly reflect our proposed approach. Tags assigned may not give accurate information for each image as they were manually named according to the keywords they contain. However, this limitation has been tackled in the medical dataset as context has been extracted from patient records for each medical image.

Appendix B

Comparative Results of CM-MTL Model in Classification

Method	Accuracy	MAE	ACC+Corr
STL	0.8790	0.1210	0.8197
CM-MTL_EQ	0.8553	0.1447	0.7907
CM-MTL_TP	0.8640	0.1360	0.8005

Table B.1: Comparing performance of the ordinal classifier in Single-Task, Task-Prioritised Multi-Task and Equal-Weight Multi-Task Learning

Method	Precision	Recall	F1 Score	Hamming Loss	Exact Match Ratio
STL	0.7520	0.6552	0.7005	0.1466	0.8534
CM-MTL_EQ	0.6502	0.6641	0.6562	0.1618	0.8382
CM-MTL_TP	0.6603	0.6775	0.6694	0.1598	0.8402

Table B.2: Comparing performance of the multi-label classifier in Single-Task, Task-Prioritised Multi-Task and Equal-Weight Multi-Task Learning