

University of Sheffield

Navigating Multimodal Complexity: Advances in Model Design, Dataset Creation, and Evaluation Techniques



Peter George Jarvis Vickers

Supervisor: Prof. Nikolaos Aletras

Co-Supervisor: Dr. Loïc Barrault

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

in the

Department of Computer Science

September 2024

Declaration

I, Peter Vickers, hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted for any other degree or qualification in this, or any other university. All work presented is my own and does not contain work as a result of collaboration with others, except as specified in the text and Acknowledgements

Abstract

Ibn Sina, a philosopher of 11th-century Persia, wrote of a ‘Floating Man’. This man is floating through a void, without the use of his sight or touch or any of the senses which make us human. Yet as he has a human brain this man, according to Ibn Sina, is capable of imagining and reasoning with the capabilities of any other person.

With the development of Large Language Models the field of Artificial Intelligence has come close to making a ‘Floating Man’ - or at least making a ‘Floating Man’ with memories of more books than exist in the wildest dreams of the librarians of Alexandria or Oxford. In this thesis, we question if the ‘floating man’ of AI could benefit from more of his senses, reasoning that as humans a great deal of our experience is multimodal.

Our research aims to address the limitations of current NLP models that heavily rely on textual information, often at the expense of multimodal cues. Such errors highlight the critical need for multimodal approaches in many applications, of which we study Visual Question Answering, Citation Recommendation, and Eye-Tracking Prediction, where text alone can lead to biased, harmful, or simply incorrect outcomes, such as mistaking a metal table for one made of wood due to textual biases.

Through our research, we aim to show the potential for Multimodality in enriching the capabilities of Artificial Intelligence.

Acknowledgements

I am profoundly grateful to my supervisors, Dr. Loïc Barrault and Prof. Nikos Aletras, as well as my industry contact, Emilio Monti, for their invaluable guidance, support, and wisdom throughout this process. Their deep insights and the breadth of their knowledge have been foundational to my work.

My deepest, heartfelt thanks to Rosa Wainwright for her warmth, love, and strength, providing unwavering support throughout this journey. To Henry, our Schnauzer, whose instinctive companionship and timely reminders to take breaks have been more rewarding than I could express with mere beef tendons and walks.

I would also like to thank my friends and colleagues—Adam, Joe, Alanna, Seb, Matias, Shaun—and everyone at the Sheffield CDT in Speech and Language Technologies for the collaborative and friendly atmosphere they cultivated.

My gratitude extends to the CDT creators Prof. Rob Gaizauskas, Prof. Thomas Hain for making this research possible. I am likewise thankful for the tireless contribution of the support team: Stu, Rachael, and Lizzie. I am equally grateful to the administrative teams in the Department of Computer Science.

A special mention to my Amazon Internship colleagues: my manager Dr. Symeon Nikitidis, and colleagues Dr. Rusen Aktas and Dr. Mark Kiermayer. The experience was not only enjoyable but crucial in teaching me how to transform research for practical solutions.

I am deeply appreciative of the teams I worked with during my studies: the JSALT teams, ESPERANTO in 2022, and Better Together in 2023. Their collaboration was remarkable, and I am thankful for being part of it. Prof. Sanjeev Khudanpur deserves special thanks for his role in these workshops, creating an engaging and welcoming space, and Dr. Kenneth Church for his invaluable guidance and collaboration in the Better Together project.

This journey would not have been the same without my expedition companion, Roy. His readiness to explore the coldest and most inhospitable parts of Europe by foot or ski with provided much-needed glorious silliness whenever I needed to regroup. My gratitude also extends to the wonderful people I met in bothies, huts, and igloos along those travels, and the conversations and songs we shared.

Lastly, my deepest gratitude goes to my family—my grandparents Rachael and Graham, my Aunt Helen Ackroyd, and my brothers Robert and Eden. Their endless

support and love is the bedrock of my life.

This work is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1. The IT Services at The University of Sheffield for the provided High Performance Computing used in this research. This work was also supported by a grant from Amazon UK.

Contents

Acknowledgements	iii
1 Introduction	2
1.1 Background	2
1.2 Modalities	4
1.3 Knowledge-Density of Modalities	6
1.4 Research Directions	9
1.5 Description of Multimodal Tasks	10
1.5.1 External Knowledge Visual Question Answering	11
1.5.2 Eye-tracking Prediction	12
1.5.3 Citation Prediction	13
1.5.4 Task Modalities	14
1.6 Research Aims and Objectives	15
1.7 Thesis Overview: Publications and Contributions	17
2 Publication I: In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering	21
2.1 Introduction	21
2.2 Related Work	22
2.3 Methodology	24
2.3.1 Preprocessing Stage	24
2.3.2 Example KAVQA Samples with Facts	27
2.3.3 Reasoning Stage	27
2.4 Experimental Setup	28
2.5 Results	29
2.6 Analysis	30
2.6.1 Bias Studies	31
2.7 Conclusion and Future Work	33
3 Publication II: Blending Cognitively Inspired Features with Transformer- based Language Models for Predicting Eye Tracking Patterns	35
3.1 Introduction and Motivation	35

3.2	Task Description	36
3.3	Related Work	37
3.4	Experimental Design	38
3.4.1	Linguistic Features	38
3.4.2	Reading Specific Features	38
3.4.3	Type Summary Statistics from GECO	39
3.4.4	Multi-word Expression Features	39
3.4.5	XLNet	40
3.4.6	Regressors	41
3.5	Results	41
3.6	Analysis and Discussion	43
3.7	Conclusion and Future Work	44
3.8	Features Used	45
3.8.1	Model One Features	45
3.8.2	Model Two Features	45
3.9	Permutation Feature Importance	45
3.10	Description of features	46
4	Publication III: We Need to Talk About Classification Evaluation	
	Metrics in NLP	47
4.1	Introduction	47
4.2	Classification Evaluation Metrics	48
4.3	Experiment 1: Metric Evaluation on a Toy Setting	52
4.4	Experiment 2: Metric Evaluation on Natural Language Understanding Tasks	54
4.5	Experiment 3: Metric Evaluation in Visual Question Answering	55
4.5.1	GQA	56
4.5.2	KVQA	58
4.6	Experiment 4: Metric Evaluation on Formality Control for Spoken Language Translation	59
4.7	Discussion	60
4.7.1	Limitations of current metrics	60
4.7.2	Improving Evaluation of Classification Tasks in NLP	61
4.8	Conclusion	61
4.9	GQA Full Comparison	63

5	Publication IV: Comparing Edge-based and Node-based Methods on a Citation Prediction Task	65
5.1	Introduction	65
5.2	Related Work	68
5.3	Methodology	72
5.3.1	Forecasting Citation Prediction	72
5.3.2	Graph Partitioning	72
5.3.3	Dataset Balancing	74
5.3.4	Representation Models	74
5.3.5	Evaluation Task	76
5.3.6	Evaluation Implementation	77
5.4	Results and Analysis	78
5.5	Citation Prediction	79
5.5.1	Results with Informedness	80
5.5.2	Early and Late Bins	81
5.5.3	Comparisons and Combinations	82
5.6	Conclusion	86
5.7	Ethics	87
5.8	Limitations	87
5.9	ProNE-S Forecast Heatmap	89
5.10	ProNE-S Forecast Table	90
6	Publication V: SynthVQA: Towards Flexible External Knowledge VQA Dataset Creation	95
6.1	Introduction	95
6.2	Related Work	97
6.2.1	Visual Question Answering	97
6.2.2	External Knowledge VQA	98
6.2.3	Knowledge Base Question Answering	100
6.3	GRAVITY Framework	101
6.3.1	Knowledge Graph (KG)	101
6.3.2	Question Sampling	101
6.3.3	Image Collection and Entity Detection	103
6.3.4	KG Linking	103
6.3.5	Question Engine	104

6.3.6	Hard Negative Samples	104
6.3.7	Question Filtering.	104
6.4	SynthVQA Dataset	105
6.4.1	KG Linking	106
6.4.2	Question Engine	107
6.4.3	Image and Question Sampling	107
6.4.4	Question Forming	108
6.4.5	Question Filtering	108
6.4.6	Negatives for Multiple Choices Setting	109
6.5	Dataset Statistics	109
6.5.1	Creation Cost	111
6.6	Results	112
6.7	Results By Relation	113
6.8	Conclusion	115
6.9	LLM Prompts	118
6.9.1	Question Phrasing	118
6.9.2	Question Filtering	121
6.10	LLM Negative Sampling	123
6.11	LLM Guess	123
7	Conclusions	124
7.1	Summary	124
7.2	Research Questions	127
7.3	Impact of this Thesis	130
7.4	Future Directions	131

List of Figures

1.1	Relationship Between Modalities and Knowledge Density	9
1.2	SynthVQA Example from our Paper V. Question: What in this image was invented in the 1870s? Fact: <Metal Detector, time of discovery or invention, 1874> Answer: Metal Detector	12
1.3	Eye-tracking Example from the Task in our Paper II	13
1.4	Citation Prediction Sample from our Paper IV	14
1.5	Tasks and Modalities	14
1.6	Research Questions	15
2.1	Our Model	25
2.2	Exemplar KVQA Questions with Relevant Wikidata Facts and Totals.	28
3.1	XLNET Feature Prediction Model	40
3.2	Feature Importance by Target for Model 1 (Left) and Model 2 (Right).	42
4.1	Accuracy, Balanced Accuracy, F1-Macro, Informedness, MCC, and NIT of the same binary (top) or multi-class (bottom) classifier as a function of the class distribution and the model’s prediction capability from 0% (Random Guess) to 100% (Perfect).	53
4.2	Metrics on GQA Unbalanced (left) and Balanced (right) validation splits. Error-bars show the standard deviation across five runs. Numbers after the question category are (question count) and [answer class entropy]. .	56
4.3	Metrics on GQA Unbalanced. Questions are grouped by reasoning type annotation on the X axis and sorted by count. X axis labels gives the reasoning type, the number of samples, and the entropy of the answer class distribution	63
4.4	Metrics on GQA Balanced. Questions are grouped by reasoning type annotation on the X axis and sorted by count in GQA Unbalanced for comparison. X axis labels gives the reasoning type, the number of samples, and the entropy of the answer class distribution	64

5.1	Random Splits (top) vs Proposed Causal Split (bottom) for Table 5.1. Train split in green, test in blue. The bottom plot with the train-test cut-off in 2010 gives a temporally consistent split.	66
5.2	The literature doubles every nine years. Observations are denoted by circles and predictions by red lines.	67
5.3	There are many missing values. Many papers have abstracts (99M) and many have links in the citation graph (111M), but only 65M (31%) have both.	73
5.4	Histogram of <i>query</i> papers by bin. Specter is trained on triples: $\langle query, pos, neg \rangle$, where the <i>query</i> paper cites the <i>pos</i> . The distribution of <i>neg</i> (random negatives) is similar to <i>query</i> , though <i>pos</i> predates <i>query</i>	75
5.5	Results on ProNE-S ⁽⁵⁰⁾ (trained on $G^{(50)}$) and tested on $T^{(k)}$ for $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$; dashed lines compare with Specter.	77
5.6	Performance improves with larger training sets: ProNE-S ⁽⁶⁰⁾ > ProNE-S ⁽⁵⁰⁾ > ProNE-S ⁽⁴⁰⁾	78
5.7	ProNE-S Accuracy Across Forecast Horizons	81
5.8	ProNE-S–Specter Crossover: Metcalfe’s Law favors larger citation graphs (more than 82M papers).	82
5.9	Hybrid Accuracy Uplift from ϕ_p to ϕ_s across all ProNE-S versions	83
5.10	Hybrid Forecast Accuracy (with 5-point Moving Average) for $\phi_{\frac{P\tau}{S}}$ with ProNE-S ^{(0), (10), (30)} and Specter. The color bars indicate which line has the best accuracy. Differences between bars suggest the policy for combining text and context depends on the size of the training set (t) and forecasting horizon (h).	85
5.11	Full ProNE-S Cumulative Forecasting Results	89
6.1	SynthVQA samples generated automatically from our GRAVITY framework.	96
6.2	The GRAVITY framework. The visual entities extracted from the image are linked to the knowledge graph (Sec. 6.3.4), and then further processed by the question engine (Sec. 6.3.5) to generate question candidates. Hard negative answers are then created (Sec. 6.3.6). Those questions are then filtered according to several requirements (Sec. 6.3.7).	102
6.3	Answer Distribution.	109

6.4	Top: Relation Distribution (Unique Rel). Bottom: Relation Distribution (Intersection). We give the source of the relation first. WD=Wikidata, VGR=Visual Genome Relation, VGA=Visual Genome Attribute. . . .	111
6.5	Correlation of error rates across models for relations and tail entities in SynthVQA. Each scatter plot compares the error rates of two models, with each point representing a relation or tail entity. The diagonal shows model names. R-values indicate the strength of correlation between model errors.	116

List of Tables

1.1	Multimodal Tasks Considered in this Thesis	11
2.1	All facts retrieved from the KVQA Wikidata Release for the Samples in Figure 2.2. In the case of multiple tail values for a given head and relation, the values are numbered on subsequent lines.	27
2.2	Results in terms of % accuracy of the considered systems break down into question types along with the question types distribution (last column).	30
2.3	Ablation Study of Information. Q=Question, I=Image, F=Facts. Image refers to the Image feature stream. Results are expressed as % accuracy by question type.	31
2.4	Further Ablation and Adversarial Studies. *Adversarial Modality indicates that the sample from that modality was randomly assigned from the entire data split	32
3.1	Model MAE on Development Splits	41
3.2	Ranking on the CMCL Shared Task Test Data.	43
4.1	Classification Contingency Table	49
4.2	GLUE Results. See Wang et al. (2018) for tasks details and evaluation metrics. All values are scaled by 100. ‘All’ is a uniform weighted mean of the individual metric scores as in https://gluebenchmark.com/leaderboard	54
4.3	Model performance on KVQA across metrics.	57
4.4	Metric scores on Formality Control for Spoken Language Translation (En-De) between off-the-shelf and formality-aware MT systems.	60
5.1	1 cites 2, 2 cites 3,..., 5 cites 6	66
5.2	Accuracy of 10 ProNE-S models and Specter on citation prediction forecasting task. Lines indicate the train-test divide.	79
5.3	Accuracy and Informedness of 10 ProNE-S models on our Citation Prediction forecasting task. Lines indicate the train-test divide.	80
5.4	ProNE Cumulative Forecasting Results	94

6.1	Popular VQA datasets, their question count, and their creation method	100
6.2	Levels of Abstraction in Knowledge Graph Representations	100
6.3	LLM generated f-string Question Templates. The first row is a ‘Unique Relation’ graph isomorphism and the second is a ‘Unique Intersection’.	107
6.4	Linking statistics for SynthVQA. Linking is attempted in the order rows are shown.	110
6.5	Dataset statistics pre- and post-filtering with an LLM. Filtering removes ~70% of samples and ~50% of logical relations, although Entities/Literals remain diverse.	110
6.6	Results of text, image and multimodal models in SynthQA.	112
6.7	Stratified Success Rate and Question Count by Relation	113
6.8	Breakdown of relation sources (Visual Genome and Wikidata) and BLIP2-VQA accuracy on intersection questions.	113
6.9	The Relation and Tail Values with highest error rates for the Unique Relation SynthVQA Subsection	114
6.10	The Relation and Tail Values with highest error rates for the Unique Intersection SynthVQA Subsection	115

Nomenclature

CV	Computer Vision
EKVQA	External Knowledge Visual Question Answering
GNN	Graph Neural Network
KG	Knowledge Graph
ML	Machine Learning
NLP	Natural Language Processing
VQA	Visual Question Answering

Chapter 1

Introduction

Multimodal Artificial Intelligence refers to Artificial Intelligence (AI) approaches which use more than one modality. Modalities are defined as a distinct type of perception or experience (Baltrusaitis et al. 2019). Within AI, multimodal categories are typically bound to the human sensory modalities such *Vision*, *Sound*, *Touch* with the exception that *Natural Language* is seen as a distinct modality to *Vision* (Zhou and Shimada 2023).

In this thesis, we study how Multimodal Artificial Intelligence systems integrate information from different modalities. We study modalities across what we term the axis of *knowledge density*, the ratio of signal to noise inherent in each modality. We study this dynamic across three multimodal AI tasks: Visual Question Answering, Eye-Tracking Prediction, and Citation Prediction. Our process follows an ‘evaluation cycle’ (Fig 1.6): we design a new model, and to understand these models, we design datasets and a metric to better understand the model’s capabilities. We organise our Research Questions and findings to align with these stages: (1) Multimodal modelling, (2) Multimodal data, and (3) Evaluation of complex classification tasks.

1.1 Background

Multimodal AI dates back to MIT’s 1966 (not so) straightforward summer project of “spend[ing] the summer linking a camera to a computer and getting the computer to describe what it saw” (Papert 1966). Later work included incorporating visuals of lip movements in Speech Recognition (Yuhua et al. 1989), and Multimedia Retrieval (Yoshitaka and Ichikawa 1999). Recently, multimodal AI has had a renaissance due to the capabilities offered by scaling Neural Network models with self-supervised learning approaches across exponentially increasing computational and data resources (Nan 2023).

A focus on Language-only methods in AI has constituted a successful if roundabout pathway to success (Brown et al. 2020). At the time of this Thesis, remarkable

capabilities in text-based models have been achieved through neural network approaches, model and data scaling, and self-supervised learning (Ericsson et al. 2022). These models are revolutionising the workplace and the economy (The Economist 2023). The successes of text-based models are the paying off of a bet made on the primacy of the text modality which has existed throughout the history of the AI field. Even early ‘Good Old Fashioned AI’ systems defined by formal logic considered text-based conversational capacity to be the target to demonstrate high machine competence (Turing 1950). AI’s perennial focus on text-based capability itself continues a much older philosophical tradition which viewed language as a ‘primary modality’ (Harnad 1990; Grosz 2012).

This thesis seeks to gently push back against the AI-Philosophical consideration that ‘text is all you need’. In many real-world tasks, the use of non-textual data is necessary. In the Vision domain alone, these include Visual Question Answering (VQA) (Antol et al. 2015; Malinowski and Fritz 2014; Gao et al. 2015; Yu et al. 2015; Ren et al. 2015; Zhu et al. 2016; Wang et al. 2016; Johnson et al. 2016; Marino et al. 2019; Hudson and Manning 2019a; Shah et al. 2019; Wang et al. 2017b; Sampat et al. 2021; Schwenk et al. 2022), Visual Entailment (Xie et al. 2019), Image Captioning (Vinyals et al. 2015), Multimodal Sentiment Analysis (Zadeh et al. 2016), Multimodal Machine Translation (Elliott et al. 2015), and Cross-modal Retrieval (Srihari 1995).

To progress beyond this self-evident usefulness of other modalities, we ask: *In which tasks and sub-tasks is multimodal data useful, and to what degree, and is this constant?*

Answering this question requires the design and analysis of multimodal models, datasets, and metrics with a focus on diversity, difficulty, and diagnosis. To sort modalities and order our research, we introduce the concept of *knowledge-density*, the amount of data in a modality versus the useful information to be extracted from it, which we discuss in the Knowledge Density Section 1.3. We discuss these more throughout in our Research Questions Section 1.6.

We give a three-part introduction to our Research below. Firstly, we outline the modalities and modelling approaches thereof which we consider in this thesis: Images, Text, KG, Citations, and Expert Linguistic Features. Secondly, we define knowledge density and place the aforementioned modalities on this axis. Thirdly and finally, we introduce the three multimodal tasks of External Knowledge Visual Question Answering, Eye-Tracking Prediction, and Citation Recommendation which we develop in our study of integrating multimodal data.

1.2 Modalities

Images in a digital context are discretized representations of the Visual Modality (sight). In standard formats, images are represented as rectangular grids of pixels, where a pixel is a three-tuple of Red, Green, and Blue ‘channels’. The number of horizontal and vertical pixels, (width and height) specifies the resolution of an image. The range of intensities each of the Red, Green, and Blue channels can have is defined by the ‘bit depth’ of an image, with 8-bit (2^8) values typical. Therefore, the size of an image is defined by Resolution (= Width * Height) multiplied by the bit depth.

Knowledge Graphs are structured representations of Knowledge defined by an ontology and populated with nodes and edges representing entities and relations. The first proposed computational Knowledge Graph [KG], the Semantic Net (Richens 1958), was created to reduce semantic loss during translation. Recent KG such as Wikidata (Vrandečić and Krötzsch 2014) and DBpedia (Auer et al. 2007) represent World Knowledge in a format accessible to symbolic querying. By convention, KGs are node- and edge-attributed graphs where nodes are entities and edges are relations. Formally, they are defined by an ontology $O_{KG} \subseteq E \times R \times (E \cup L)$ (Hogan et al. 2021), where:

- E represents the set of entities,
- R is the set of relations (or properties) between entities,
- L is the set of literals, such as numbers, strings, or dates,
- $E \times R \times (E \cup L)$ signifies the possible triples formed by entities and relations, resulting in either another entity (E) or a literal (L).

Examples of $E \times R \times (E \cup L)$ include:

- An entity-to-entity relation: $\langle \text{J. R. Firth, Studied at, University of Leeds} \rangle$,
- An entity-to-literal relation: $\langle \text{J. R. Firth, Data of Birth, 17 June 1890 (Gregorian)} \rangle$.

Citation Graphs are structured representations of Citations between academic documents. They can be defined by the ontology $O_C \subseteq D \times C \times D$ where:

- D is the set of documents.

- C is the set of directed citations.

As citations can only be to other documents, D is both the origin and target.

We realise this ontology as a Citation Graph, where nodes are documents and directed edges reflect citations.

Linguistic Features This section details the various linguistic and reading-specific features.

- Part-of-speech (POS) information was incorporated, following the natural inclination of readers to fixate more on function words compared to open-class words. This POS tagging was accomplished using the Spacy library.
- Sentence Position Indicators: Binary indicators were used to mark words as either the first or last in their sentences.
- Frequency Measures: The analysis included raw and Zipf frequency measures of words
- Concreteness Norms: These features describe the abstractness of a word. We use the data from the human annotation in Brysbaert et al. (2014), mean, standard deviation, and the % of participants familiar enough with the word to accurately judge its concreteness.
- GECO Corpus Summary Statistics: type-level summary statistics for gaze features were generated and from the GECO eye-tracking corpus.
- Multi-Word Expression Features: An MWE lexicon and related metrics were created using the mwetoolkit annotations of the Wikitext-103 corpus (Cordeiro et al. 2016):
 - Binary indicators of MWE presence,
 - categorization of MWEs by syntactic pattern,
 - compositionality scores from MWEToolkit’s compositionality scoring function.
 - Skip-Gram embeddings generated from joining component words of MWEs in Wikitext-103 using underscores (i.e. climate change becomes climate_change) (Mikolov et al. 2013b).

1.3 Knowledge-Density of Modalities

The ratio of total information to useful information in each modality is not equal. In this section, we propose a definition of a way of quantifying this dimension which we term *knowledge density*, make calculations to quantify this property, and propose a grouping of the modalities we study according to this criteria.

Entropy is an information-theoretic measure of the uncertainty of a random variable. It is defined as:

$$H(X) = - \sum_i^n p(X = x) \log_2(p(X = x)) \quad (1.1)$$

In order to simplify our presentation, we assume a uniform probability over all values $p(X = x)$. We refer to this as the ‘uniformity assumption’. In practice, certain values for X will be more likely and the entropy of a modality will be lower.

Given $p(x)$ is uniform, the uniformity entropy which we term H_U can be written as:

$$H_U(X) = - \sum_i^n \frac{1}{n} \log_2\left(\frac{1}{n}\right) \quad (1.2)$$

$$= -n \cdot \frac{1}{n} \log_2\left(\frac{1}{n}\right) \quad (1.3)$$

$$= -\log_2\left(\frac{1}{n}\right) \quad (1.4)$$

$$= \log_2(n) \quad (1.5)$$

That is, the log of the number of possible states.

In most data structures Multimodal AI encounters, there are a number of ‘features’ in each sample S : pixels, words, nodes. Assuming each feature F is independent, and there are n possible states for each feature, the entropy over a complete sample is:

$$H_U(S) = \log_2(n^F) \quad (1.6)$$

$$= F \log_2(n) \quad (1.7)$$

That is, given a feature, we can estimate the entropy of that feature as the product of the number of features F (words, pixels, nodes) multiplied by the logarithm of the possible states for each feature n .

Image The number of features of an image is defined by the number of pixels multiplied by the number of channels C . The number of pixels is the product of the width W and height H in pixels. For colour images, there are typically three channels: Red, Green, and Blue. The possible states per feature is the bit depth which is conventionally expressed as a power of 2. A bit depth D of 8-bits = 2^8 values are typical for images.

Keeping our uniformity assumption, and taking the default Vision-Transformer (Dosovitskiy et al. 2020) input image size of 224x224, this gives a uniformity Shannon Entropy of:

$$\begin{aligned}
 H_U(\text{Image}) &= \log_2(D^{W \times H \times C}) \\
 &= W \times H \times C \times \log_2(D) \\
 &= 224 \times 224 \times 3 \times \log_2(2^8) \\
 &\approx 1,200,000
 \end{aligned} \tag{1.8}$$

1.2 million is an extremely high value for entropy, which reflects the fact that even a relatively small image has a large amount of information.

Text For text the uniformity assumption means that (a) all words are equally likely and (b) each word is independent of any preceding words, which is known as the unigram assumption. We let the sentence length L equal 10 and set the word vocabulary V be 10,000, meaning L^V possible sentences. In this case, a sentence has a uniformity Shannon Entropy of:

$$\begin{aligned}
 H_U(\text{Text}) &= \log_2(V^L) \\
 &= L \log_2(V) \\
 &= 10 \log_2(10,000) \\
 &\approx 133
 \end{aligned} \tag{1.9}$$

Knowledge Graph We consider a Knowledge Graph (KG) with 100k nodes N and 100 relation types R .

We consider a ‘Sample’ to be a triple of <Node, Edge, Node> taken from a KG. This may be considered as two nodes each with states drawn from N possible nodes and a connecting edge drawn from R possible states gives a uniformity Shannon Entropy of:

$$\begin{aligned}
H_U(\text{KG}) &= \log_2(N^2 \times R^1) \\
&= 2 \log_2(N) + \log_2(R) \\
&= 2 \log_2(100,000) + \log_2(100) \\
&= 2 \log_2(100,000) + \log_2(100) \\
&\approx 40
\end{aligned} \tag{1.10}$$

$H_U(\text{KG}) = 40$ is a lower value, reflecting the higher knowledge density of the KG. This presents an opportunity for multimodal systems to obtain high-quality and clear information without having to learn how to model exceptionally complex distributions seen in Images and Text.

This may be illustrated through quantifying the x saying ‘An image is worth x words’ (given our uniformity assumption). Taking the $H_U(\text{Image})$ of 1.2 million, we consider what length of sentence would be required to have the same uncertainty given the same uniform vocabulary of 10,000 as before:

$$\begin{aligned}
\log_2(V^L) &= 1,200,000 \\
L \log_2(10,000) &= 1,200,000 \\
L &= \frac{1,200,000}{\log_2(10,000)} \\
&\approx 90,000
\end{aligned} \tag{1.11}$$

Therefore an image is ‘worth’ 90,000 words in terms of Entropy under the uniformity assumption.

Next, we consider how large a Knowledge Graph must be for each triple to match the entropy of either a text sentence or an image. We do not include edge relations in our calculations as these make an insignificant contribution. First, we consider how many nodes are required for a triple from a KG to match a 10-word sentence with a uniform vocabulary of size 10,000:

$$\begin{aligned}
\log_2(N^2) &= 10,000 \\
2 \log_2(N) &= 10,000 \\
\log_2(N) &= 5,000 \\
N &= 2^{5,000} \approx 1.41 \times 10^{1,505}
\end{aligned} \tag{1.12}$$

Now we consider an image of the size, channels, and bit depth we discussed above:

$$\begin{aligned}
\log_2(N^2) &= 1,200,000 \\
2\log_2(N) &= 1,200,000 \\
\log_2(N) &= 600,000 \\
N &= 2^{600,000} \approx 9.94 \times 10^{180,618}
\end{aligned}
\tag{1.13}$$

These large numbers give an indication of the size of the change in complexities between modalities. This in turn suggests that approaches in multimodal AI should be sensitive to the challenges of highly-expressive data such as images and the opportunities of highly-structured data such as KG.

In real-world data, the entropies will be lower. By (A) assuming uniform distributions and (b) independence for all possible states within a data structure we have provided upper bounds on the entropy. However, the general trend of Image \gg Text \gg KG holds.

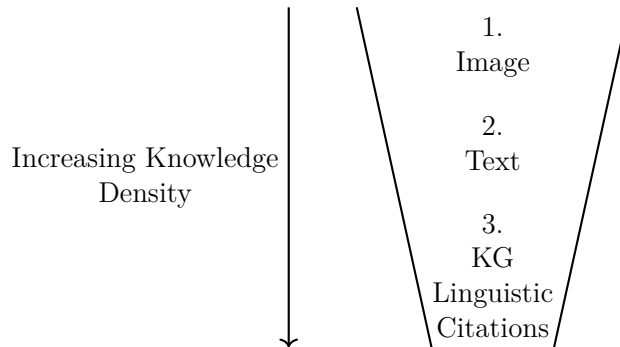


Figure 1.1: Relationship Between Modalities and Knowledge Density

Using the values obtained in these calculations, we group all of the modalities across our Research in Fig 1.1. Citation Graphs are similar to KG, where each node represents a paper and each edge represents a citation. Linguistic Features categorical variables representing certain properties of each word in a sentence, and may be considered as sentences with extremely restricted vocabulary ($V=100$).

1.4 Research Directions

In this thesis, we quantify the interaction of modalities across the knowledge-density dimension, from the *dense*: Knowledge Graphs (Schneider et al. 2022), Linguistic

Annotations (Chakrabarty et al. 2020), and Citations (Ostendorff et al. 2022), to the *medium*: Text, to the *sparse*: Images.

We consider these features in the context of real-world AI tasks. External Knowledge Visual Question Answering (Antol et al. 2015) (KG, Text, Images), Eye-Tracking Prediction (Hollenstein et al. 2021b) (Linguistic Features, Text), and Citation Recommendation (Färber and Jatowt 2020) (Citations, Text). We explore how text-based architectures and models perform on these tasks when additional modalities are provided as an additional signal.

The challenge is to make a model that can use very heterogeneous data where information is distributed across all modalities. In an ideal world, text could suffice for all tasks. However, many times it is inconvenient or even unreasonable to represent a task as text. Therefore, the ability to take information from less plentiful modalities is crucial to performance.

A crucial dimension in the development of AI systems with novel capabilities is their *evaluation*. Careful consideration of the triple factors: (I) dataset design (Sparck Jones 1994), (II) sample annotation (DeYoung et al. 2020), and (III) the evaluation metric (Blagec et al. 2022) are required to ensure a fair and informative test for a system under evaluation.

We are also concerned with how to quantify model performance. There has been research into the failings of Accuracy, (Ben-David 2007), and proposed possible alternatives without these issues (Brodersen et al. 2010; Valverde-Albacete and Peláez-Moreno 2014). Furthermore, we investigate how better datasets can support these better metrics to permit separable analysis of model capabilities.

1.5 Description of Multimodal Tasks

In this thesis, we study tasks which involve integrating structured knowledge to NLP systems. We target tasks which require at least two classes from Figure 1.1. We choose three tasks with activate research in Multimodal AI: External Knowledge Visual Question Answering, Eye-tracking Prediction, and Citation Prediction. We outline these tasks in Table 1.1 and discuss them further below.


Task	External Knowledge Visual Question Answering	Eye-Tracking Prediction	Citation Prediction
Input Modalities	Image (1), Text (2), KG (3)	Text (2), Linguistic Features (3)	Text (2), Citations (3)
Output	Answer/Classification	Eye Movements/Regression	Is Cited/Classification
Example	"What is the Alma Mater of the person wearing a Mortarboard?" +  + KG	What proportion of time did an average reader spend on the word "company" in the sentence "You shall know a word by the company it keeps"? + Linguistic Features	Does (Devlin et al. 2019b) cite (Firth 1957)? + Titles, Abstracts + Citation Graph

Table 1.1: Multimodal Tasks Considered in this Thesis

1.5.1 External Knowledge Visual Question Answering

Visual Question Answering (VQA) is the task of answering a text question about an image given a Question and Image pair (Antol et al. 2015). It has uses for accessibility (Gurari et al. 2018), education (Kembhavi et al. 2017), content moderation, and healthcare (Hasan et al. 2018).

Whilst this task is conceptually simple, a number of orthogonal factors increase the task complexity (Goyal et al. 2019; Shah et al. 2019). These are: Question Priors (Goyal et al. 2019; Hudson and Manning 2019a), Fact Compositionality (Johnson et al. 2016; Zhang et al. n.d.), Knowledge Obscurity (Wang et al. 2016; Wang et al. 2017b), and Image Comprehension (Thrush et al. 2022)

In this thesis, we consider the VQA subtask **External Knowledge Visual Question Answering** (EKVQA), which is Visual Question Answering which requires the use of knowledge which can not be learned from the training set (Marino et al. 2019).

We illustrate EKVQA with a sample from our SynthVQA dataset in Figure 1.2. Given the image, the task is to answer ‘What in this image was invented in the 1870s?’. The World Knowledge that the metal detector was invented in 1870s is not learnable from a general VQA train set (Shah et al. 2019), and so a source of External Knowledge is required to provide this fact. We illustrate this by providing the fact, although in practice a system will have to retrieve the relevant fact itself.



Figure 1.2: SynthVQA Example from our Paper V.

Question: What in this image was invented in the 1870s?

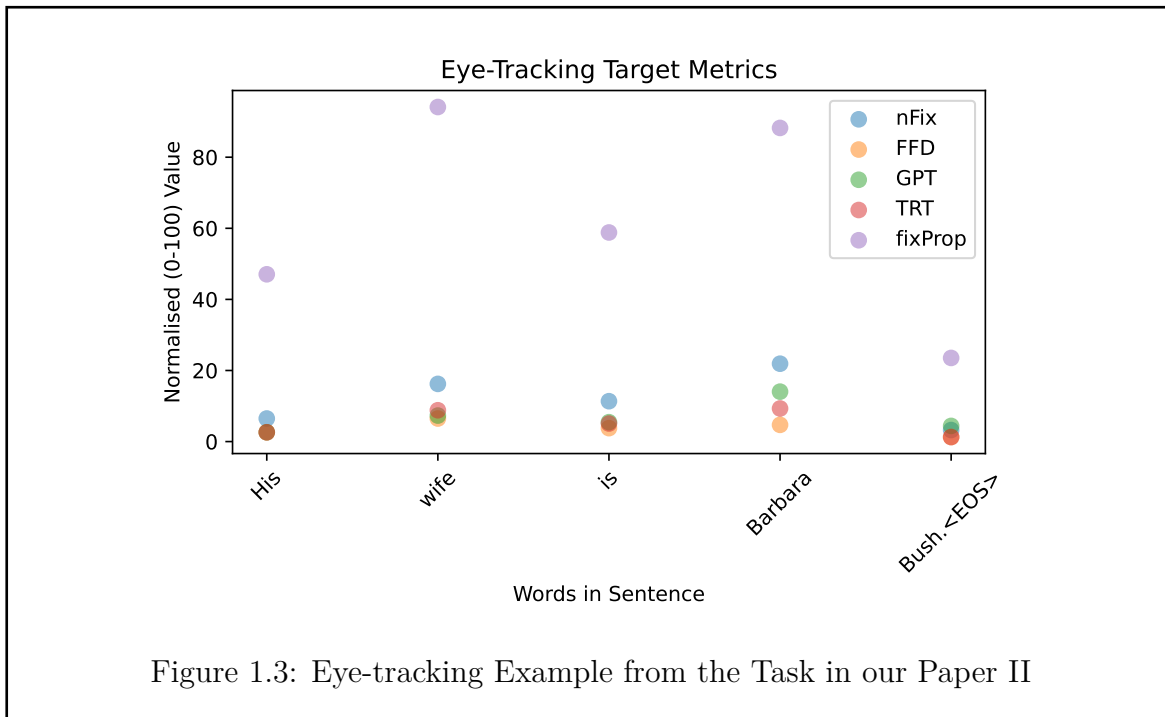
Fact: <Metal Detector, time of discovery or invention, 1874>

Answer: Metal Detector

Whilst models which can answer Vision-Language problems are a topic of intensive research (Chen et al. 2020; Lu et al. 2019; Radford et al. 2021; Li et al. 2023), the incorporation of Knowledge-Graph facts is far less studied and straightforward (Schwenk et al. 2022). Such systems use all three modalities: Vision, Text, and KG, and face the challenge of coordinating the information within these disparate data types to drive a prediction (Chen et al. 2021).

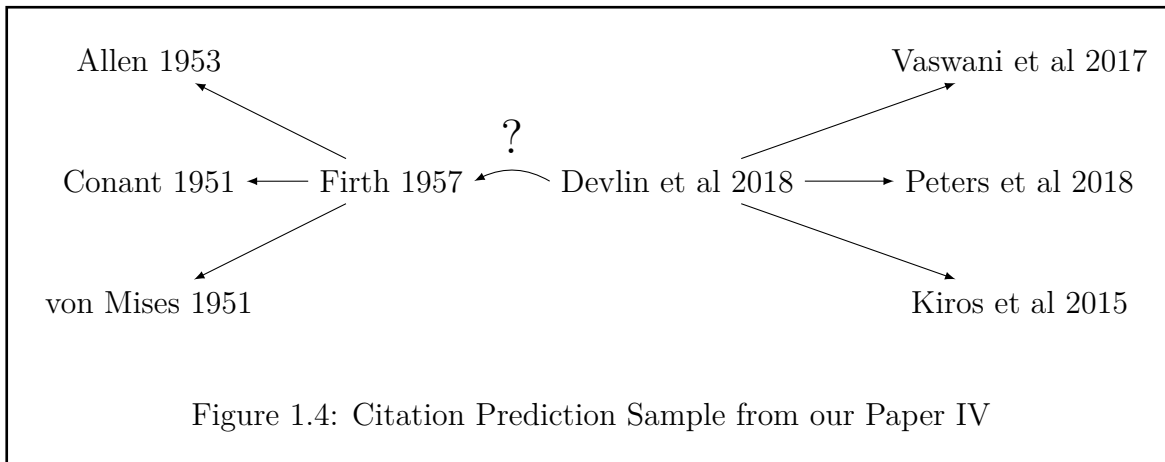
1.5.2 Eye-tracking Prediction

Eye-tracking Prediction is the task of predicting the path of the human gaze over the content of interest. Eye tracking data may be used either to understand human cognition or as an inductive bias for computational models (Hollenstein et al. 2021b). For computational models, they have been shown to enhance Named Entity Recognition (NER) (Hollenstein et al. 2019), Sarcasm Detection (Mishra et al. 2016), and Question Answering (Sood et al. 2020a). Studies have shown that human gaze duration is inversely correlated to the likelihood of a word given its context (Ehrlich and Rayner 1981).



1.5.3 Citation Prediction

Citation Prediction is the task of predicting whether a given source paper cites a given target paper (Färber and Jatowt 2020). It is designed to provide suggestions to paper authors on what to cite. In Global Citation Prediction, a model is provided with a singular representation of source and target papers, typically a subset of text, abstract, full-text, and other citations. Recent research has focused on models which input Title and Abstract text and optimise metric-based losses over samples of the citation graph. We also consider other citations.



1.5.4 Task Modalities

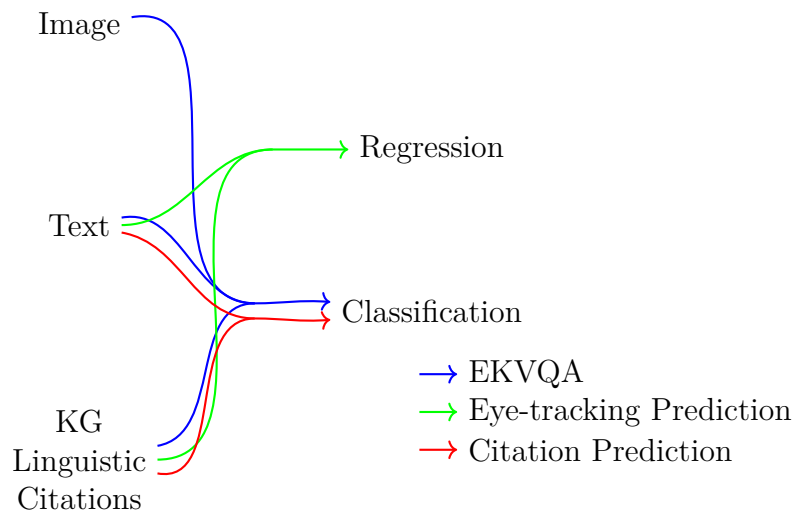


Figure 1.5: Tasks and Modalities

Each of these tasks requires the integration of different modalities across varying data scales and with unique targets. Importantly, each gives us a different view of the classes defined in Figure 1.1. We outline the modalities of each task and their densities in Figure 1.5. External Knowledge Visual Question Answering requires systems to reason over all Knowledge Densities: Images (level 1), text (level 2), and KG (level 3) to predict from a large set of answer classes. Eye Tracking Prediction uses Text (level 2) and expert Text features (level 3) to predict continuous features capturing human

gaze movements. Citation Prediction has text (level 2) and other citations (level 3) as input and targets a binary classification as to whether two papers cite each other.

1.6 Research Aims and Objectives

As discussed in Section 1.4 we conduct our study of multimodality in Artificial Intelligence through the VQA, Eye-tracking Prediction, and Citation Prediction tasks. We organise our Research Questions by the three dimensions of Model Development, Dataset Creation, and Evaluation Metrics. As Fig 1.6 indicates, these are mutually informing and reinforcing, and our work responds to limitations in one with novel research in another.

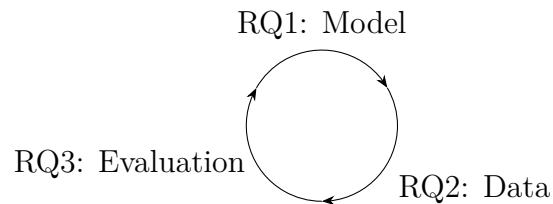


Figure 1.6: Research Questions

This thesis asks the following Research Questions:

1. **RQ1: How can we incorporate both knowledge-dense data structures (KG, Citations, Linguistic Annotations) and noisy modalities (Images) into text models?** In this Research Question, we ask what approaches can be used to build effective and robust systems which integrate multimodal data of different knowledge densities.

We address this question in Paper I with our Vision-Language-KG REUNITER model, a novel architecture for solving External Knowledge Visual Question Answering datasets. We show that this approach beats the prior state of the art on the Knowledge Aware VQA (KVQA) dataset by 19%. To gauge the contribution of each modality, we perform train and test time ablations of each modality. We are limited in our ability to fully categorise its reasoning performance by noise in the sub-task divisions the dataset.

We further address Modality Integration in Paper II where we train models for predicting human gaze patterns. We compare Text features from a fine-

tuned Language Model and a set of Linguistically motivated features through Permutation Feature Importance.

Finally, in Paper IV we evaluate the contribution of Text and Citation Graph features to the Citation Prediction task. We compare a state-of-the-art Text-based model with a Citation Graph-based model as the size of the Training Citation Graph grows. Finally, we consider strategies for combining Text and Citation features and find that this is dependent on the properties of both the train- and test-set.

2. **RQ2: How can we create multimodal datasets which are diverse, difficult, and diagnostic?** Our second research question explores the creation of novel multimodal datasets for probing the integration of modalities.

We address the sparsity of existing External Knowledge VQA (EKVQA) datasets by adapting techniques from Knowledge-Based Question Answering to EKVQA with our GRAVITY pipeline and the SynthVQA dataset it produces. We show that this approach generates diverse, difficult, and diagnostic samples which have a wide variety of underlying reasoning types and structures required to answer the questions. We benchmark several state-of-the-art VQA models on our SynthVQA and find they lack World Knowledge found in KG compared to text-only Language VQA models.

Furthermore, we create a novel Citation Prediction dataset to determine if there is a crossover point between knowledge-rich citation-based methods and text-based methods in this task. Furthermore, we speculate that there is a further task dynamic which has been elided in previous datasets which is the time dimension.

3. **RQ3: How can we improve the evaluation of classification models on datasets with diverse and imbalanced class distributions?** Many tasks in Multimodal AI are classification-based: given a set of possible options, a model must select the most plausible. Evaluation with ‘standard’ metrics such as Accuracy and F1-Macro does not report the true performance level of the systems and may lead to wrong interpretation of the results. Furthermore, if the number or distribution of answers changes, then it becomes even more challenging to compare capabilities. This is because the prior probability of a correct guess is a function of the distribution of the possible answers. At the same time, in reality, different question types have different answer distributions. For instance ‘What

is the table made of’ has fewer plausible answers than ‘How old is the person on the stage’? Furthermore, these distributions are uneven: ‘What is the table made of’ will be answered correctly by ‘wood’ more than ‘marble’ (Goyal et al. 2019). Therefore, we promote a debiased Informedness metric, to allow comparison between questions with different distributions, and use this to reevaluate our findings from **RQ1** and **RQ2**.

1.7 Thesis Overview: Publications and Contributions

Here we list the publications and datasets which resulted from the research undertaken in this thesis. The thesis itself is a Thesis by Publication, so these papers constitute the main body of the document where each publication is given a chapter. Here we list the publications in chronological order, giving an overview of how they interrogate the Research Questions and their high-level findings. In all papers, Peter Vickers was the sole First Author, except for paper II, where he was the joint First Author.

Publication 1:

In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering

In this work we develop a new approach for applying knowledge-graph features to Vision-Language models. Our contributions are as follows:

1. A new method for adding Knowledge Graph facts to Vision-Language models.
2. Training and evaluation on the Knowledge Visual Question Answering dataset, beating the state of the art by 19%.
3. The first full-scale ablation study over KVQA dataset, finding which reasoning types are problematic for our model.

This work was published in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* Vickers et al. (2021a).

The authors, in order as in the publication are: Peter Vickers, Dr. Nikolaos Aletras, Emilio Monti, Dr. Loïc Barrault

My contributions to this paper: research, idea development, methodology, model development, testing, and paper writing.

Publication 2:**Blending Cognitively Inspired Features with Transformer-based Language Models for Predicting Eye Tracking Patterns**

In this study, we examined the integration of cognitively and linguistically inspired features within transformer-based language models, specifically XLNet, to predict eye tracking patterns. Our key findings are as follows:

1. The investigation into the utility of linguistic and cognitive information, predicted by eye-tracking features, for the enhancement of eye-tracking prediction models.
2. The demonstration that a smaller pre-trained model (XLNet-base) can outperform a larger one (XLNet-large) in this context, challenging common assumptions about model size and performance.
3. The exploration of multi-word expressions (MWEs) in improving model predictions, finding limited benefits despite known cognitive processing advantages.
4. Detailed experimentation with a range of features, including word length, part-of-speech tags, and concreteness norms, revealing nuanced influences on prediction accuracy.
5. The employment of a Random Forest Regressor and ElasticNetCV for feature-rich and XLNet models respectively, with a comprehensive evaluation of feature importance and model performance.

This work was published in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* Vickers et al. (2021c).

The authors, in order as in the publication are: Peter Vickers, Rosa Wainwright, Dr. Harish Tayyar Madabushi, Prof Aline Villavicencio

My contributions to this paper: mentorship of Master's students, research, idea development, methodology, model development, testing, and paper writing.

Publication 3:**We Need to Talk About Classification Evaluation Metrics in NLP**

In this comprehensive study, we delve into the evaluation metrics utilized in Natural Language Processing (NLP) classification tasks, such as topic categorization and sentiment analysis. Our investigation reveals significant biases in widely-used metrics, prompting

a reevaluation of how model performance is measured. This work was motivated by our difficulties in Paper 1 with using existing classification metrics to compare across models and subtasks. Our key contributions include:

1. A critical comparison of standard classification metrics against more nuanced measures, advocating for the use of the Informedness metric as a more accurate baseline for evaluating task performance.
2. An extensive empirical analysis across a broad spectrum of NLP tasks, demonstrating that Informedness more effectively captures model generalizability and allows for fairer comparisons between models.
3. The release of first Python implementation of the Informedness and Normalized Information Transfer metrics, adhering to the SciKitLearn classifier format, to facilitate its adoption in future research.

This work was published in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* Vickers et al. (2023).

The authors, in order as in the publication are: Peter Vickers, Dr. Loïc Barrault, Emilio Monti, Prof Nikolaos Aletras

My contributions to this paper: metric development, research, idea development, methodology, metric implementation, data gathering, and paper writing.

Publication 4:

Comparing Edge-based and Node-based Methods on a Citation Prediction Task

In this dataset and benchmarking paper, we release a new citation prediction benchmark which tests models to perform on high scale data and against forecasting dynamics inherent in academic literature. Our contributions are as follows:

1. A new benchmark for citation prediction emphasizing scale and forecasting.
2. Empirical demonstration that larger graphs favor edge-based methods.
3. Empirical evidence that performance improves with t (larger training sets) and degrades with h (forecasting horizon).
4. We distribute the benchmark, evaluation code, and embeddings.

This work was a part of the 2023 JSALT Better Together Text+Content Summer Workshop and was submitted to *ACL Rolling Review (ARR)*.

The authors, in order as in the publication are: Peter Vickers, Dr. Kenneth W. Church.

My contributions to this paper: idea development, research, dataset development, model evaluation, and paper writing.

Publication 5:

SynthVQA: Automating Visual Question Answering Creation

In this work, we address the complexity and bias inherent in Visual Question Answering (VQA) datasets by introducing a novel VQA question generation pipeline. Our approach leverages deep question structure graph isomorphisms, making it highly expressive and relation-agnostic. Our contributions are as follows:

1. Development of a VQA Question generation pipeline that operates over deep question structure ‘graph isomorphisms,’ enabling the creation of highly expressive and relation-agnostic questions.
2. Creation of the SynthVQA dataset, a proof of concept dataset that is diverse, difficult, and diagnostic, allowing for in-depth analysis of what facts, relations, and isomorphisms are challenging for VQA models.
3. Demonstration that state-of-the-art VQA models lack factual knowledge from Wikidata compared to text-only models, highlighting the importance of external knowledge in improving VQA performance.

This work was submitted to *ACL Rolling Review (ARR)*.

The authors, in order as in the publication are: Peter Vickers, Dr. Loïc Barrault, Emilio Monti, Prof Nikolaos Aletras

My contributions to this paper: idea development, research, pipeline coding, dataset development, model testing, and paper writing.

Chapter 7 summarizes our research findings and indicates future research directions.

Publication I: In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering

2.1 Introduction

Visual Question Answering (VQA) is a popular multi-modal task of answering a question about an image. It tracks both inter-modal interactions and reasoning capabilities of models (Wang et al. 2017b; Marino et al. 2019). Recent studies have tested compositional reasoning (Johnson et al. 2016; Hudson and Manning 2019a) and the integration of external knowledge (Wang et al. 2017b; Wang et al. 2016; Shah et al. 2019; Marino et al. 2019) for VQA. In this paper, we address Knowledge-aware VQA (KVQA) (Shah et al. 2019)¹, defined as a VQA task where it is not reasonable to expect a model without access to a knowledge base to be able to answer the questions in the test set.

In a uni-modal textual context, both synthetic dataset (Kassner et al. 2020) and task-driven (Ding et al. 2020) studies of neural models have shown significant competence at symbolic reasoning. This is encouraging, as neural pretrained Language Models such as BERT (Devlin et al. 2019a) achieve state-of-the-art results in a wide range of natural language inference tasks and benchmarks such as Natural Language Inference (Bowman et al. 2015). (Rajani et al. 2019) uses pretraining on a domain-specific dataset to improve CommonsenseQA by 10% absolute accuracy. Tamborrino et al. (2020) develop an improved training objective to improve COPA by 10% absolute accuracy.

Bouraoui et al. (2020) find that BERT is capable of relational induction, whilst Broscheit (2019) and Petroni et al. (2020) find that BERT stores non-trivial world-knowledge.

Previous work has argued that restriction to a uni-modal context may itself impair reasoning performance (Barsalou 2008; Li et al. 2020). In a bi-modal Vision + Language (V+L) context, datasets such as CLEVR and GQA allow for the evaluation of both model reasoning and language grounding. Within this setting, Ding et al. (2020) and Lu et al. (2020) show that appropriate neural models trained on large quantities of

¹For data, examples, and licence information, please see <https://malllabisc.github.io/sources/kvqa/>

data can exhibit accurate reasoning.

In this paper, we propose a new method of applying a massively pretrained V+L BERT model (Chen et al. 2020) to the KVQA task (Shah et al. 2019). Our method is able to learn a set of reasoning types (confirming findings in Ding et al. (2020)) but can increase performance even more by incorporating external factual information. KVQA answers require attending to a knowledge base, allowing us to quantify the contribution of both explicit and implicit knowledge extracted from supervised training data. We also quantify the degree to which corpus bias makes certain question types harder, and outline how future datasets may be better balanced.

Our contributions are as follows:

- We perform factual integration into a V+L BERT-based model architecture VQA, leading to 19.1% accuracy improvement over previous baselines on KVQA.
- We evaluate our model’s reasoning capabilities through an ablation study, proposing explanations for poor performance on certain question types as well as highlighting our model’s strong preference for text and facts over the image modality.
- We conduct a bias study of the KVQA dataset, revealing both strengths and potential improvements for future VQA datasets.

2.2 Related Work

VQA tasks explicitly encourage grounded reasoning (Antol et al. 2015), with emphasis on a variety of sub-domains, such as commonsense (Zellers et al. 2019), compositionality and grounding (Suhr et al. 2020), factual reasoning (Wang et al. 2017b) or external knowledge reasoning (Wang et al. 2016; Marino et al. 2019; Shah et al. 2019). External knowledge reasoning modifies the VQA task by requiring various forms of symbolic inference across natural language, making the task more similar to that performed in Neural Reasoning Diagnostics.

This immediately raises concerns about bias, a known problem within VQA tasks (Goyal et al. 2019), which require active intervention from dataset designers to avoid (Hudson and Manning 2019a). Indeed, as supervised machine learning algorithms learn from annotated data only, if the data is heavily biased towards certain answer types, then answering less frequent question types is made even more complex. Whilst all of the four External Knowledge VQA datasets that we are aware of: FVQA (Wang et al.

2016), KB-VQA (Wang et al. 2017b), KVQA Shah et al. (2019), and OK-VQA (Wang et al. 2016) advertise intractability to current neural models, only FVQA evaluates both symbolic and neural systems, finding that the best neural system achieves 43.1%, whilst a hybrid system achieves 52.6%. Concerns around the out-of-domain robustness of neural models has led to a preference towards hybrid Neuro-Symbolic (Garcez and Lamb 2020) approaches (Hudson and Manning 2019b; Yi et al. n.d.). A paradigm has emerged where tasks needing explicit, compositional reasoning are best solved by Neuro-Symbolic systems, and those requiring implicit, commonsense reasoning are best solved by Neural systems (Zellers et al. 2019; Chen et al. 2020).

State-of-the-art systems for external knowledge VQA are based on Memory networks (MemNet, (Weston et al. 2014)). In Shah et al. (2019), the facts are extracted from the Knowledge Graph (KG) by considering the visual (from image) and eventually textual (from Wikipedia caption) entities. They are then embedded using a Bi-LSTM encoder and fed into the memory. After the question is embedded in a similar way, the resulting representation is used to query the memory by soft attention. Several stacked memory layers are used to better model multi-hop facts.

Wang et al. (2016) and Wang et al. (2017b) introduce two datasets, KB-VQA and FVQA respectively, and address the task with systems that perform searches in a visual knowledge graph formed from the image and a KB. The question is first mapped to a query of the form \langle visual object, relationship, answer source \rangle , which is then used to extract the supporting facts from the KB. They report improved results when compared to systems using LSTM, SVM and hierarchical co-attention (Lu et al. 2016).

In Marino et al. (2019), the OK-VQA is presented with some baseline results obtained with MUTAN (Ben-younes et al. 2017), a multimodal tensor-based Tucker decomposition which models interactions between visual (from CNN) and textual (from RNN) representations. Those systems exhibit rather low performance compared to those obtained on standard VQA, demonstrating that the corpus requires external knowledge to be solved correctly.

Similar datasets KB-VQA (Wang et al. 2017b), OK-VQA (Marino et al. 2019), FVQA (Wang et al. 2016) are smaller, put also present tasks requiring a knowledge base + describe systems (globally).

Recent work has introduced methods to incorporate visual information to create Vision+Language BERT models through joint multimodal embeddings (Chen et al. 2020; Su et al. 2019; Lu et al. 2019). First, image and text are embedded into the same

space, and then Transformer networks are applied as in the standard BERT model (Devlin et al. 2019a).

Our work is most similar to that of Shah et al. (2019) since the same preprocessing pipeline is used. However, our system does not use a memory network, and instead relies on on a BERT-based model (UNITER, see section 2.3) to model the relationship between question, facts, and image with self-attention layers.

2.3 Methodology

To answer KVQA with Neural models, we first take the V+L BERT model UNITER (Chen et al. 2020) with the highest score on the commonsense VQA task, VCR (Zellers et al. 2019).

In order to allow UNITER to accept external KG facts, we cast these facts to a textual form ‘Entity₁ Relation Entity₂’. To keep the input facts count small, we perform a *conditional search* of the KG. The KVQA task consists in finding a^* :

$$a^* = \underset{a \in A}{\operatorname{argmax}} p(a|q, i, K) \approx \underset{a \in A}{\operatorname{argmax}} p(a|q, i, k_{i,q}) \quad (2.1)$$

where a^* is the correct answer out of candidate set A ; and q , i , and K are a question, image and knowledge base, respectively. As shown, we may reduce the KG through a conditional search to find the relevant subset of facts $k_{i,q}$.

To define the subset $k_{i,q}$, we follow Shah et al. (2019) in extracting all facts from the knowledge base that are up to two hops from any entities detected by the textual entity linking or the face detection.

Our model, as presented in section 2.2 consists of two stages: preprocessing, which implements relevant fact extraction, and reasoning, which selects an answer from the question, facts, and image features.

2.3.1 Preprocessing Stage

For preprocessing and fact acquisition, we broadly reproduce the fact and feature extraction process used in Shah et al. (2019). We perform object detection with the Faster R-CNN network (Ren et al. 2017). A seven-dimensional normalised size and location vector is concatenated with the Faster R-CNN features.

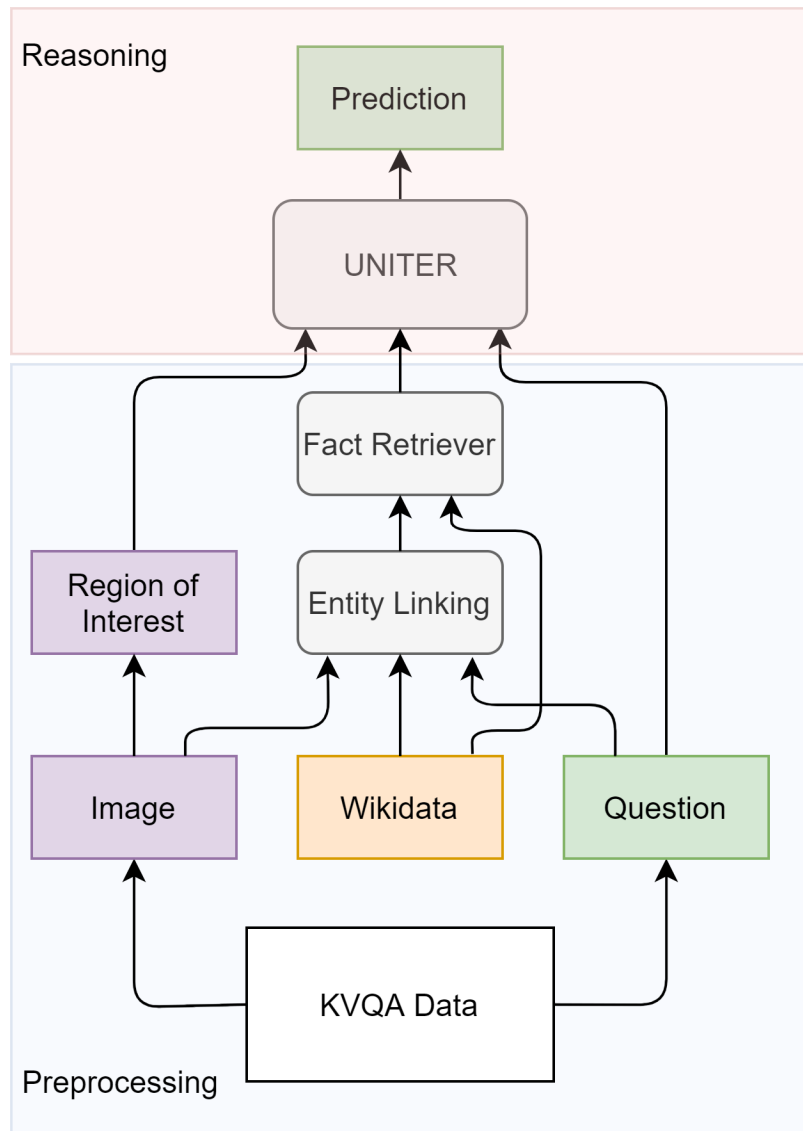


Figure 2.1: Our Model

For person detection, we use MTCNN (Zhang et al. 2016) and Facenet (Schroff et al. 2015) models, pretrained on the MS-celeb-1M (Guo et al. 2016) dataset, to generate 128-dimensional embeddings. We predict names by nearest-neighbour comparison with the KVQA reference dataset, which contains photos of celebrities linked to their entry in the Wikidata Knowledge Graph. Note, that we only consider identification and retrieval of facts about human entities in this task, as both the dataset and our method only consider features which identify humans. We treat the name identification as a multi-class classification problem, achieving a Micro-F1 of 0.539. We follow (Shah et al. 2019) them in also using the REL textual entity linker (Hulst et al. 2020) to predict persons from the image captions. Each image in KVQA is sourced from Wikipedia, and the community-sourced captions are retained in the final dataset. REL accepts a string and returns 0-n links to Wikidata entities. Since this is lower than reported in Shah et al. (2019), we follow them in applying a textual entity linker (Hulst et al. 2020) over supplied image descriptions. This setup increases our a per-image+caption Micro-F1 to 0.686.

We use the names of identified entities to query the (Shah et al. 2019)’s reduced Wikidata graph (Vrandečić and Krötzsch 2014) contained as part of the KVQA dataset. The reduced graph has a total of 18K Entities, and 164K Facts. The linked (human) entities are used as head entities to query one and two hop triples. Tail entities are the entity or qualification string of Wikidata. For instance, ‘Hillary Clinton, spouse, Bill Clinton’ has an entity as the tail, and this will lead to a further hop. Meanwhile ‘Hillary Clinton, date of birth, 1947-10-26’ has a qualifier as a tail, and will not lead to a further hop. The extracted facts are finally cast to the form ‘subject relation object’. We sort one-hop facts before two-hop ones in the context passed to the UNITER. For any individual entity in the reduced Wikidata graph, the mean number of facts is 6.9 with a standard deviation of 4.5. Considering retrieved facts per image in KVQA, the minimum number of KG facts retrieved is 0 (person not identified), whilst the maximum is 81 (6 people identified). Additionally, normalised image location facts are generated from these detections, such as ‘Barack Obama at 42 78’, which would indicate that the centre bounding box for Barack Obama is at normalised (0-100) position x=42, y=78 of the image. This adds one fact per detected person.

We do not consider predicates in the question: for instance the query ‘Who is the father of the person on the left?’ would be passed directly to our model along with the ‘image location facts’ and ‘entity facts’. We rely on the model’s own reasoning ability

Hop	1	2	1	2	1
Person	Hillary Clinton	Bill Clinton	Shahrukh Khan	Gauri Khan	Francis Condon
occupation	autobiographer (1) diplomat (2) lawyer (3) politician (4) research assistant (5) university teacher (6) writer (7)	politician (1) statesperson (2)	actor (1) film actor (2) film producer (3) presenter (4) screenwriter (5) singer (6) television presenter (7)	film producer	judge (1) lawyer (2) politician (3)
place of birth	Edgewater Hospital	Hope	Purna	New Delhi	Central Falls
date of birth	1947-10-26	1946-08-19	1965-11-02	1970-10-08	1891-11-11
alma mater	Maine East High School (1) Maine South High School (2) Wellesley College (3) Yale Law School (4)	Edmund A. Walsh School of Foreign Service (1) Georgetown University (2) Hot Springs High School (3) University College (4) Yale Law School (5)	Jamia Millia Islamia	University of Delhi	Georgetown University Law Center
spouse	Bill Clinton	Hillary Clinton	Gauri Khan	Shahrukh Khan	
sex	female	male	male	female	male
is member of	Republican Party	Democratic Party			Democratic Party
knows language	English	English (1) German (2)			English
native language	English	English			
religion	Methodism	Baptist	Islam	Hindu	
work started			1988-01-01		
date of death					1965-11-23
place of death					Boston

Table 2.1: All facts retrieved from the KVQA Wikidata Release for the Samples in Figure 2.2. In the case of multiple tail values for a given head and relation, the values are numbered on subsequent lines.

to select the correct facts.

2.3.2 Example KAVQA Samples with Facts

We illustrate two examples from the KVQA datasets in 2.2. The figure shows the Wikidata Image, the Question, and the facts retrieved from the KVQA Wikidata dump. In each case, we highlight the facts relevant to answering the question, and the total 1-hop and 2-hop facts. We detail all facts for these samples in 2.1. Note that we implicitly indicate which entities are present in the image through the location fact (see above). We mark image entities and their relevant facts in bold.

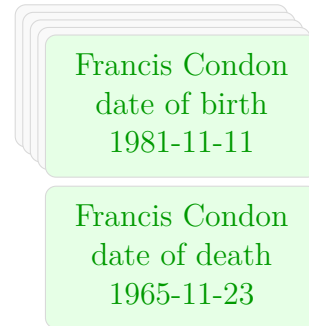
2.3.3 Reasoning Stage

The neural model we use, UNITER, is pretrained on MS COCO (Lin et al. 2014), Visual Genome (Krishna et al. 2016), Conceptual Captions (Sharma et al. 2018), and SBU Captions (Ordonez et al. 2011). It is a multi-task system that is trained on performing Masked Language Modeling, Image-Text Matching, and Masked Region Modeling (Chen et al. 2020).



Question: *For how many years did the person in the image live?*

1-Hop Facts 11



2-Hop Facts 0

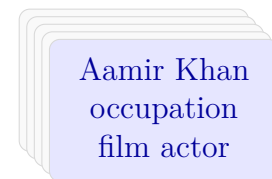
(a) Exemplar KVQA Question: Francis Condon's Lifespan



Question: *Do all the people in the image have a common occupation?*

1-Hop Facts 19

1-Hop Facts 14



2-Hop Facts 16
(Bill Clinton)

2-Hop Facts 11
(Gauri Khan)

(b) Exemplar KVQA Question: Common Occupation of Clinton and Khan

Figure 2.2: Exemplar KVQA Questions with Relevant Wikidata Facts and Totals.

2.4 Experimental Setup

We select the KVQA dataset for two reasons: to our knowledge, it is the largest external knowledge dataset (with 183k questions), and the questions are annotated with their reasoning types. We use accuracy as the evaluation metric and provide results over both

the entire dataset and also for each question type as provided in the KVQA dataset.

The baseline systems for KVQA are those presented in (Shah et al. 2019) and discussed in section 2.2. The first baseline is a stacked BLSTM encoder, operating over question and facts. This system has an overall accuracy of 48.0% . The second is the MemNet architecture and has the previously highest performing baseline accuracy at 50.2%.

We use the UNITER_BASE pretrained model available at the ChenRocks GitHub repository² with custom classification layers (MLP +softmax output layer). For task training, we merge retrieved facts with the question, dividing each statement with the ‘[SEP]’ token, following research that indicates that this token induces partitioning and pipelining of information across attention layers (Clark et al. 2019). The textual input stream is tokenised with the HuggingFace ‘bert-base-uncased’ tokeniser (Wolf et al. 2020). We set the maximum WordPiece sequences length to 412, the maximum visual objects count to 100, the learning rate to 8×10^{-5} and use AdamW (Loshchilov and Hutter 2017) as optimizer. Once preprocessing is completed, we train the UNITER model with the cross-entropy objective function for 80,000 iterations, which we empirically found to guarantee convergence.

2.5 Results

Table 2.2 shows the results of our system (UNITER), using a question label break-down similar to Shah et al. (2019). Overall, we observe that our system outperforms the previous baseline MemNet setting (see ‘World+WikiCap+ORG’ in Shah et al. (2019)) with an absolute improvement of 19%.

Our results show that UNITER is learning to perform reasoning more accurately than MemNet in all but two cases. In the question types involving multiple entities (‘Multi-Entity’, ‘Multi-Hop’, ‘Multi-Relation’), the increase is the greatest, suggesting that UNITER is able to robustly learn these reasoning here. We speculate that stacked self-attention layers in BERT are able to better attend to the many involved entities than MemNet.

We now discuss the performance of our model on its weakest categories, namely ‘Subtraction’ and ‘Spatial’. The poor performance on ‘Subtraction’ questions confirms previous results that BERT-like models require specialised pretraining for numerical

²<https://github.com/ChenRocks/UNITER>

Question Type	Model		Entropy (Base 2)
	MemNet	UNITER	
1-Hop	61.0	65.7	7.8
1-Hop Counting	-	78.0	1.4
1-Hop Subtraction	-	28.6	4.3
Boolean	75.1	94.6	1.1
Comparison	50.5	90.4	2.1
Counting	49.5	79.4	2.3
Intersection	72.5	79.4	1.2
Multi-Entity	43.5	77.1	3.3
Multi-Hop	53.2	87.9	3.7
Multi-Relation	45.2	75.2	7.1
Spatial	48.1	21.2	11.5
Subtraction	40.5	34.4	6.0
Overall	50.2	69.3	7.6

Table 2.2: Results in terms of % accuracy of the considered systems break down into question types along with the question types distribution (last column).

reasoning tasks (Geva et al. 2020). In the case of our model specifically, we note the lack of numerical reasoning tasks in UNITER’s pretraining regime. ‘Spatial’ is the model’s least accurate question type (21.4%) and the biggest absolute decrease from MemNet (-26.7%). This question type requires two-hop reasoning where the second hop is a numerical operation of the form $\underset{y}{\operatorname{argmin}}(x_i - y_i)$. Both of these have been shown to be problematic for BERT (Kassner et al. 2020; Geva et al. 2020).

2.6 Analysis

UNITER performs well at the reasoning tasks in general, with the most surprising result being that it apparently does better at multi-hop reasoning than one-hop. We believe that this can be explained by the presence of unbalanced distribution of answer types in the dataset perturbing the results (see Table 2.2). We discuss this in Section 2.6.1.

In order to better understand the reasoning capability of our model and the impact of each input modality, we perform an inference time ablation study, presented in Table 2.3.

Ablation of Image features (column ‘Q+F’) does not change the performance, suggesting that the model is not attending to image features. To confirm this hypothesis,

Question Type	Q+F+I	Q+F	Q+I	F+I	Q	F	I
1-Hop	65.7	65.7	32.4	3.9	32.4	3.8	4.5
1-Hop Counting	78.0	78.0	30.3	0.0	30.3	0.0	0.0
1-Hop Subtraction	28.9	28.6	28.8	0.8	30.3	0.6	6.5
Boolean	94.6	94.6	55.2	1.3	55.2	1.0	10.5
Comparison	90.4	90.4	38.7	1.0	38.7	0.9	10.7
Counting	79.4	79.4	66.1	0.6	65.9	0.4	1.4
Intersection	79.4	79.4	61.0	0.4	60.6	0.3	0.0
Multi-Entity	77.1	77.1	41.3	0.8	41.2	0.7	6.4
Multi-Hop	87.9	87.9	29.0	0.8	28.9	0.8	0.0
Multi-Relation	75.2	75.2	25.1	3.0	25.0	3.0	2.5
Spatial	21.2	21.2	0.0	13.0	0.0	13.0	0.0
Subtraction	34.4	34.4	1.3	1.0	0.9	0.7	0.0
Overall	69.3	69.3	31.6	3.1	31.5	3.0	3.6

Table 2.3: Ablation Study of Information. Q=Question, I=Image, F=Facts. Image refers to the Image feature stream. Results are expressed as % accuracy by question type.

we performed an experiment with adversarial images, obtaining very similar results for each question type and the same overall score (69.30%). We explain this behaviour by the fact that the preprocessing pipeline extracts all the required information as explicit facts which the model prefers over the more ambiguous visual features. We leave a deeper analysis for further work.

An interesting case is the ‘Spatial’ questions, where facts alone are able to correctly answer 13% of the questions. This is likely the result of the answers to this question type being entities present in the facts. Again, we observe that the model is not able to learn this information from the visual features.

2.6.1 Bias Studies

We briefly discuss the corpus bias, a well-known concern in VQA (Goyal et al. 2019). We consider question difficulty across three parameters: reasoning difficulty, task design, and corpus bias. Certain question types are inherently more complex, as discussed in Section 2.5. Additionally, the task may have different numbers of answer classes per task, effectively weakening any priors models might form (see Entropy column in Table 2.2). Finally, an unbalanced dataset may cause certain reasoning types to be underrepresented, making it harder for models to learn for them. ‘Spatial’ and

Question Type	Train Ablation		Adversarial Modality*	
	Q+I	Q	I	F
1-Hop	47.09	38.5	65.9	31.3
1-Hop Counting	66.1	61.5	75.2	50.5
1-Hop Subtraction	29.4	29.7	28.1	26.2
Boolean	83.9	67.3	94.1	57.5
Comparison	83.4	60.3	90.6	47.8
Counting	75.4	75.2	78.9	70.2
Intersection	67.6	67.9	76.8	61.2
Multi-Entity	69.4	57.2	76.4	47.6
Multi-Hop	56.5	50.2	87.9	38.4
Multi-Relation	47.3	38.9	75.2	28.3
Spatial	3.3	1.2	21.1	0.0
Subtraction	2.1	2.6	39.2	1.6
Overall	47.0	40.8	69.3	32.8

Table 2.4: Further Ablation and Adversarial Studies. *Adversarial Modality indicates that the sample from that modality was randomly assigned from the entire data split

‘Substraction’ questions are among the least represented in the training dataset, which increase their difficulty for the model.

Unseen answer classes are also an issue. For ‘Spatial’ questions, only 54.2% of the test answers (output classes) are actually seen during training, placing an upper bound on accuracy. We find 98.4% of ‘Spatial’ questions the model answered correctly and 95.7% of ‘Spatial’ question the model answered incorrectly were supplied with adequate facts by the preprocessing pipeline.

Training time ablation and adversarial experiments To further probe the task, we perform a training time ablation with first facts, and then facts and images removed (see Table 2.4). In this we seek to exhibit the capability of our model to leverage the available modalities and to compensate for the missing ones.

Through comparing the training time and inference time ablations, we can better understand the importance of a modality to solving the task.

Through comparing train and inference ablation of facts (‘Q+I’ column of Table 2.4 and of Table 2.3) we observe that when facts are unavailable at train time, the model attends to images to obtain 47.0% accuracy, which is 15.4% more than the 31.6% obtained by the corresponding inference time ablation. This indicates that the visual modality can provide useful information for this task.

We observe a similar trend in the fact and image ablation setting (‘Q’ column of Table 2.4 and of Table 2.3) that the model is able to greater leverage questions to make accurate predictions when additional modalities are never available.

We also perform adversarial checks, where random images or facts from the data split are presented at inference time. These align closely with the ablation study, with adversarial images (Column ‘I’ of Table 2.4) performing within 0.1% of blanked images (Column ‘Q+F’ of Table 2.4) and adversarial facts (Column ‘F’ of Table 2.4) performing within 1% of blanked facts (Column ‘Q+I’ of Table 2.4). These results confirm the importance of factual data and the unimportance of raw image features to a model trained on the full data.

2.7 Conclusion and Future Work

We evaluated our model and found that it improves on the previous state of the art by a substantial margin (19.1%). An ablation study revealed the specific strengths and weaknesses of our model on certain question categories when evaluated on the KVQA dataset. We show that the UNITER model is not actually using the visual input.³

In the future, we seek to create a large external knowledge dataset designed following KVQA with more entities besides persons to encourage grounded reasoning, and better calibration of answer types. We will also consider pretraining our model on closely related tasks. This will help to form a model capable of learning robust reasoning with a high degree of spatial specificity and entity discrimination.

Acknowledgements

Peter Vickers is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1.

Ethical Statement

This work is based on the open-source KVQA dataset, an English multimodal dataset, and the Wikidata knowledge base (also in English). No English-specific preprocessing

³We release code at <https://github.com/petervickers/Factuality-UNITER>

was used for this research and the UNITER model is language agnostic, which tends to suggest that this could generalize to other languages. We will make our code publicly available to ensure the reproducibility of our experiments.⁴

⁴<https://github.com/petervickers/UNITER-experiments>

Publication II: Blending Cognitively Inspired Features with Transformer-based Language Models for Predicting Eye Tracking Patterns

3.1 Introduction and Motivation

Many researchers now agree that eye movements during reading are not random (Rayner 1998); as a result, eye-tracking has been used to study a variety of linguistic phenomena, such as language acquisition (Blom and Unsworth 2010) and language comprehension (Tanenhaus 2007). Readers do not study every word in a sentence exactly once, so following patterns of fixations (pauses with the eyes focused on a word for processing) and regressions (returning to a previous word) provides a relatively non-intrusive method for capturing subconscious elements of subjects' cognitive processes.

Recently, cognitive signals like eye-tracking data have been put to use in a variety of NLP tasks, such as POS-tagging (Barrett et al. 2016), detecting multi-word expressions (Rohanian et al. 2017) and regularising attention mechanisms (Barrett et al. 2018): the majority of research utilising eye-tracking data has focused on its revealing linguistic qualities of the reading material and/or the cognitive processes involved in reading. The CMCL 2021 Shared Task of Predicting Human Reading Behaviour (Hollenstein et al. 2021a) asks a slightly different question: given the reading material, is it possible to predict eye-tracking behaviour?

Our ability to quantitatively describe linguistic phenomena has greatly increased since the first feature-based models of reading behaviour (i.e. Carpenter and Just (1983)). Informed by these traditional models, our first model tests 'simple' features that are informed by up-to-date expert linguistic knowledge. In particular, we investigate information about multi-word expressions (MWEs) as eye-tracking information has been used to detect MWEs in context (Rohanian et al. 2017; Yaneva et al. 2017), and empirically MWEs appear have processing advantages over non-formulaic language (Siyanova-Chanturia et al. 2017).

Our second model is motivated by evidence that Pre-trained Language Models

(PLMs) outperform feature based models in ways that do not correlate with identifiable cognitive processes (Sood et al. 2020b). Since many PLMs evolved from the study of human cognitive processes (Vaswani et al. 2017) but now perform in ways that do not correlate with human cognition, we wished to investigate how merging cognitively inspired features with PLMs may impact predictive behaviour. We felt this was a particularly pertinent question given that PLMs have been shown to contain information about crucial features for predicting eye tracking patterns such as parts of speech (Chrupała and Alishahi 2019; Tenney et al. 2019) and sentence length (Jawahar et al. 2019).

We therefore had the goals of providing a competitive Shared Task entry, and investigating the following hypotheses: A) Does linguistic/cognitive information that can be *predicted* by eye-tracking features prove useful for *predicting* eye-tracking features? B) Can adding cognitively inspired features to a model based on PLMs improve performance in predicting eye tracking features?

3.2 Task Description

The CMCL 2021 Shared Task of Predicting Reading Behaviour formulates predicting gaze features from the linguistic information in their associated sentences as a regression task. The data for the task consists of 991 sentences (800 training, 191 test) and their associated token-level gaze features from the Zurich Cognitive Language Processing Corpora (Hollenstein et al. 2018; Hollenstein et al. 2020). For each word, the following measures were averaged over the reading behaviour of the participants: FFD (*first fixation duration*, the length of the first fixation on the given word); TRT (*total reading time*, the sum of the lengths of all fixations on the given word); GPT (*go past time*, the time taken from the first fixation on the given word for the eyes to move to its right in the sentence); nFix (*number of fixations*, the total quantity of fixations on a word, regardless of fixation lengths) and fixProp (*fixation proportion*, the proportion of participants that fixated the word at least once). Solutions were evaluated using Mean Absolute Error (MAE). For more details about the Shared Task, see Hollenstein et al. (2021a).

3.3 Related Work

Transformer architectures Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019c) is a Language Representation model constructed from stacked Neural Network attention layers and ‘massively’ pre-trained on large Natural Language Corpora. In contrast with traditional language models, BERT is pre-trained in two settings: a ‘cloze’ task where a randomly masked word is to be predicted, and next sentence prediction. BERT or derivative models have been used to achieve state-of-the-art baselines on many NLP tasks (Devlin et al. 2019c; Yang et al. 2019). Analysis studies have shown that BERT learns complex, task-appropriate, multi-stage pipelines for reasoning over natural language, although there is evidence of model bias. XLNet (Yang et al. 2019) is an autoregressive formulation of BERT which trains on all possible permutations of contextual words, and removes the assumption that predicted tokens are independent of each other.

Similar studies To our knowledge, studies that attempt to *predict* cognitive signals using language models are fairly few and far between. Djokic et al. (2020) successfully used non-Transformer word embeddings to decode brain activity recorded during literal and metaphorical sentence disambiguation. Since RNNs may be considered more ‘cognitively plausible’ than Transformer based models, Merks and Frank (2020) compared how well these two types of language models predict different measures of human reading behaviour, finding that the Transformer models more accurately predicted self-paced reading times and EEG signals, but the RNNs were superior for predicting eye-tracking measures.

In a slightly different task, Sood et al. (2020b) compared LSTM, CNN, and XLNet attention weightings with human eye-tracking data on the MovieQA task (Tapaswi et al. 2016), finding significant evidence that LSTMs display similar patterns to humans when performing well. XLNet used a more accurate strategy for the task but was less similar to human reading.

Though these studies may indicate that Transformer models are not the most suited to eye-tracking prediction, they are still considered State of the Art in creating broad semantic representations and general linguistic competence (Devlin et al. 2019c). As such, we hoped they would allow us to investigate Carpenter and Just’s speculation that the dominance of word length and frequency for predicting eye-tracking behaviour may reduce “as the metrics improve for describing higher-level factors” like semantic

meaning (1983, p. 290).

3.4 Experimental Design

We pursued both feature engineering and deep learning approaches to the task; though both methods performed well independently, there was little improvement in predictive capability when combining their features (see Table 3.1). As such, we developed and submitted two models: Model 1 (Feature Rich) and Model 2 (XLNet). Additional details about the feature combinations used in our final models can be found in Appendices 3.8 and 3.10.¹

3.4.1 Linguistic Features

Each word in the training vocabulary was encoded as a one-hot vector. Since function words are more likely to be fixated than open class words (Carpenter and Just 1983), we included POS information generated by `Spacy` (Honnibal et al. 2020) (honouring the tokenisation in the training data). We included a binary indicator for whether a word was the first or last in its sentence to incorporate the knowledge that first and last fixations on a *line* are 5-7 letter spaces from the two respective ends (Rayner 1998). We generated raw frequencies (proportion per million words) and Zipf frequencies (Van Heuven et al. 2014).

Finally, concreteness norms (a measure of how ‘abstract’ a given word is) were included as features (mean, standard deviation, and the % of participants familiar enough with the word to accurately judge its concreteness; Brysbaert et al. (2014)). We specifically tested concreteness due to the unusually large coverage of the norms.

3.4.2 Reading Specific Features

Word length has been empirically demonstrated as a very good predictor of gaze features in many studies (i.e. Rayner and McConkie (1976) and Carpenter and Just (1983)). Duration of fixation is observed to increase for words that exceed the mean saccade length (7-9 letters), and probability of fixation is reduced for words shorter than half the mean saccade length (Rayner and McConkie 1976). Therefore, as features we included

¹For reproducibility purposes, our program code (including details of hyperparameters) is available here: <https://github.com/petervickers/CogNLP-Sheffield-CMCL-2021>

both the raw word lengths, and categorical variables representing word length as a proportion of a mean saccade length.

Since readers may store information about adjacent words (Rayner 1975; Rayner 1998; Barrett 2018), we also experimented with supplying features from previous and future words to each target word.

3.4.3 Type Summary Statistics from GECO

Following Barrett et al. (2016), we used the monolingual data from the GECO corpus (Cop et al. 2017) to generate type-level summary statistics for each word. Specifically, we averaged the gaze features across the 12 participants who completed the reading task, and normalised these features to reflect the normalisation of the Shared Trask training data. We then averaged these values again at the type (word) level. For words present in the task training data but not the GECO data, we estimated the values using means for words in the GECO data of a similar frequency (according to the `wordfreq`).

3.4.4 Multi-word Expression Features

We generated an MWE lexicon and summary metrics using the Wikitext-103 corpus (Merity et al. 2016) and `mwe toolkit` (Ramisch 2012). We chose Wikitext-103 since it provided a large variety of possible MWEs in a similar context to the ZuCo reading material (Hollenstein et al. 2020). We produced two indicator features for the presence of MWEs: a binary indicator, and a categorical variable summarising the syntactic pattern of the MWE, motivated by Yaneva et al.’s evidence that MWEs of different syntactic patterns display different eye-tracking characteristics (2017).

Following the method of Cordeiro et al. (2019), we joined component words of MWEs in Wikitext-103 using underscores (i.e. *climate change* became *climate_change*) and then generated Skip-gram word embeddings (Mikolov et al. 2013a) for all single words and MWEs identified in Wikitext-103. Using the `feat_comp` function in `mwe toolkit` (Ramisch 2012), these MWE embeddings were used to compute compositionality scores and weights (Cordeiro et al. 2019).²

MWEs identified in the training data were assigned MWE embeddings and compositionality information as features, and non-MWEs were assigned single word embeddings

²The score represents the degree to which the meaning of the MWE can be worked out from the meanings of its constituent words (i.e. ‘climate change’ has high compositionality, ‘cloud nine’ has low compositionality), and the weights estimate the semantic contribution of each word in the expression.

and zero values for compositionality.

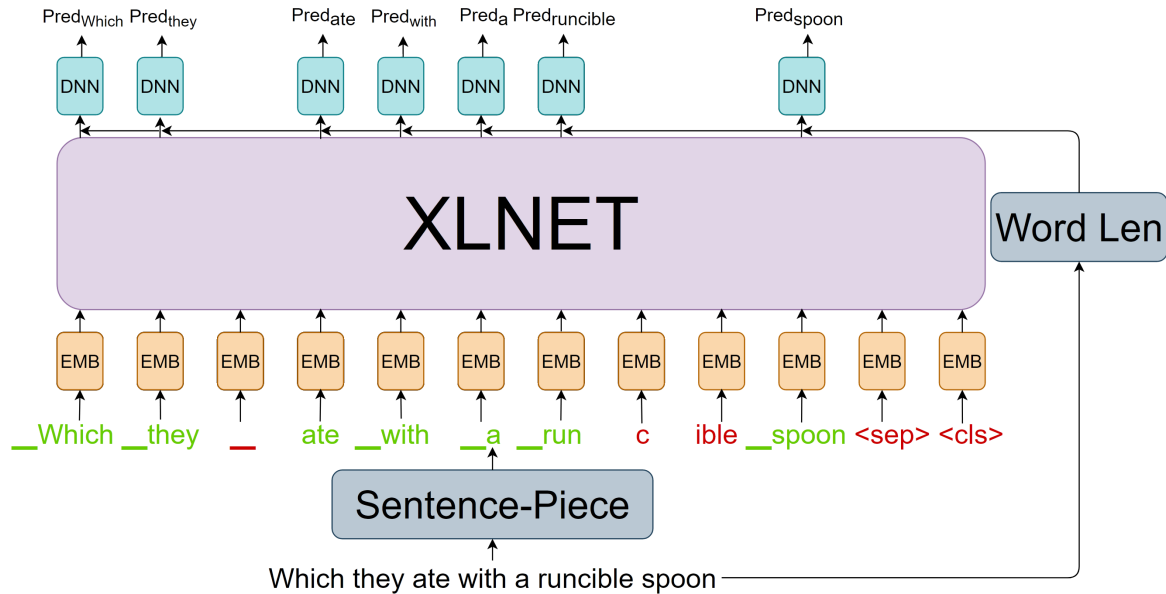


Figure 3.1: XLNET Feature Prediction Model

3.4.5 XLNet

In order to obtain Massively Pre-trained Language Model features we used XLNet. We finetuned a model that was pre-trained on BooksCorpus (Zhu et al. 2015), English Wikipedia, Giga5 (Courtney Napoles 2012), ClueWeb 2012-B (Callan et al. 2009), and Common Crawl text (Crawl 2019). For predictions, we took the final hidden representation of the first sub-word token encoding of each word. We concatenated this feature with an integer representing the total word length in characters to encourage the model to explicitly attend to word length. We tested the effectiveness of sub-word aggregation but found this reduced the model’s accuracy by an average of 0.04 MAE, which we speculate is due loss of information in the pooling operation whilst head sub-word units already contain contextual information. We then passed the concatenated sub-word and word-length features to a 3-layer dense Neural Network which was used to predict the Shared Task’s five target features. This 3-layer multi-feature Network was found to be optimal through experimentation. For stability, we used the Huber loss objective, which approximates L2 loss for small values and L1 loss for large values. We trained using the AdamW optimiser and with learning rates and training duration chosen through grid search across 3-fold cross-validation, obtaining an optimal learning

rate of 0.00001 and 800 epochs.

3.4.6 Regressors

To form predictions for the Feature Rich model we used a Random Forest Regressor implemented by `scikit-learn` (Pedregosa et al. 2011) with parameters [`max_depth = 7`, `n_estimators = 100`, `max_features = None`]. For the XLNet model, we collected the XLNet final state embeddings (identical to those fed into the DNN in Figure 3.1) along with the features [`word-len`, `CAT-pos`, `zipf-frequency`, `Is-EOS`, `Is-SOS`]. We then trained `scikit-learn`'s `ElasticNetCV` for 5-fold validation with parameters [`max_iter = 10000`, `l1_ratio=[0.1,0.3,0.5,0.7,1]`, `cv=5`].

3.5 Results

In Table 3.1 we present the MAE on validation splits of the training data. This information informed our choice of model submissions alongside a preference for models using more cognitive features.

Model/Split	1	2	3	Mean
ElasticNet(XLNet + ALL Features)	<u>3.918</u>	3.927	<u>3.891</u>	<u>3.912</u>
Feature Rich/Model 1	4.017	4.023	3.981	4.007
BERT-base-cased	4.030	4.045	3.977	4.012
ElasticNet(BERT-base-cased)	3.986	4.024	3.969	3.993
XLNet-base-cased	3.988	3.956	3.935	3.959
XLNet-base-cased (random init)	4.608	4.722	4.695	4.675
XLNet-large-cased	3.929	4.039	3.960	3.976
ElasticNet(XLNet-base-cased)/Model 2	3.921	3.924	3.896	3.914

Table 3.1: Model MAE on Development Splits

We submitted two sets of predictions from Model 2 (ElasticNet(XLNet-base-cased)) and one set of predictions from Model 1 (Feature Rich). Table 3.2 shows the ranking of Models 1 and 2 in the overall task. Our overall standing is shown to be 5th, with an MAE delta of 0.143 behind the best model. Whilst a prediction which combined Models 1 and 2 was slightly more accurate (see Table 3.1), we regard this improvement as within margin of error. We therefore focussed on Models 1 and 2 separately since this allowed for clearer comparisons between the two approaches.

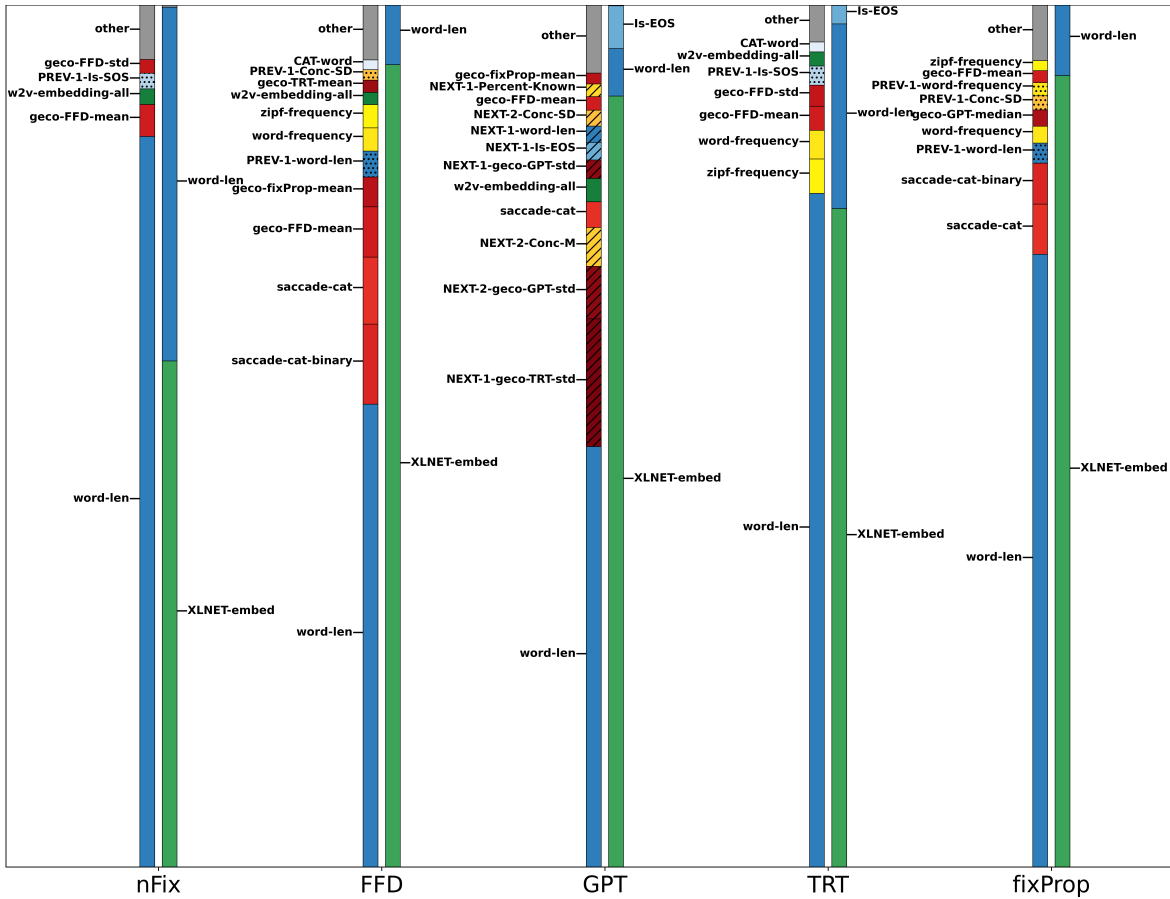


Figure 3.2: Feature Importance by Target for Model 1 (Left) and Model 2 (Right).

We give a brief outline of all approaches in Table 3.2. We use two finetuned language models, BERT and XLNet. BERT is a transformer-encoder model trained on a ‘cloze’ task and next sentence prediction. XLNet is trained on a permuted language modelling objective. The task is to predict a selected tokens given all possible combinations of previous tokens in the sentence. Both encoder models are trained for the CMCL task through the addition of three layer Feed-Forward Neural Networks to the final hidden state of each token. The final network layer has a hidden dimension of 5, which is trained with a Huber loss against the eye-tracking statistics for each word. As words may have multiple sub-tokens, we only train and infer values for the first token of each word. ElasticNet(XLNet + ALL Features) is a Linear Regression model with L1 and L2 regularization. Features from XLNet are the final hidden states of the first sub-word token of each word form the finetuned model described above. Feature Rich/Model 1 is SciKitLearn’s RandomForestRegressor over the features outlined in

Rank	Team (model)	MAE
1	LAST	3.8134
2	TALEP	3.8328
	...	
5	CogNLP@Sheffield (XLNet/Model 2)	3.9565
	...	
7	MTL782_IITD	4.0639
-	CogNLP@Sheffield (Feature Rich/Model 1)	4.0689
	...	
-	MEAN BASELINE	7.3699
13	IIIT_DWD	9.7615

Table 3.2: Ranking on the CMCL Shared Task Test Data.

Subsections 3.4.1-3.4.4 and fully defined in Section 3.10.

3.6 Analysis and Discussion

Our results (Table 3.1) support both our hypotheses introduced in Section 3.1.

We did not anticipate that XLNet-base would outperform XLNet-large, which had more pre-training data and layers. This is possibly due to the limited amount of training data specific to the task for fine-tuning, resulting in the larger model under-fitting. We are able to confirm that the knowledge XLNet learns through massive pre-training crucial to its performance in this arena - removal of this knowledge through weight randomisation increases MAE from 3.959 to 4.675. Hence we believe that both structure and pre-training of XLNet-base contribute to its success in this task.

We use normalised permutation feature importance (see Appendix 3.9) to better understand the value of different features and present it on a per-target basis for each model in Figure 3.2.

The most interesting outcome of our experiments was the fact that XLNet embeddings subsume information contained across most features except word length (especially in predicting nFix). It may be that the use of word-pieces obfuscate word length information thus requiring the explicit addition of that information. While the usefulness of features such as word length is consistent with the literature, we were surprised by the relative unimportance of MWE information given that many neurocognitive

studies have demonstrated differences in how they are processed (Siyanova-Chanturia et al. 2011; Siyanova-Chanturia et al. 2017; Cacciari and Tabossi 1988). An additional surprise is that even though the Skip-gram embeddings provide semantic information about single words as well as MWEs, the Feature Rich models make little use of them. Many of the Feature Rich models utilize the GECO features, which may be because they provide approximate guidance about the distributions of the various gaze features that would be difficult to learn directly given the sparsity of the training data.

3.7 Conclusion and Future Work

This work describes our submissions to the 2021 CMCL Shared Task: we contributed a Feature Rich model inspired by cognitive and linguistic information, and model predominantly based on contextual XLNet-base embeddings. We find that only a limited subset of the cognitive features (such as word length) are helpful in the XLNet model. To our surprise, neither XLNet-large embeddings nor MWE features provide performance improvements. However, we believe this indicates a need for further research into MWE representations as opposed to suggesting that MWEs are unimportant for creating effective cognitive models.

Acknowledgements

We are grateful to Cheng Cao, Elham Khodaei, Srivishnu Ethirajulu Krishnaraj and Ronan Ramdas Revadker for their help generating and testing the feature sets. PV and RW are supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1. RW is also supported by ZOO Digital. This work is also partially supported by the EPSRC grant EP/T02450X/1.

3.8 Features Used

We use the following features for each model. +N and +P indicate that associated data for the two next and two preceding words were included, respectively.

3.8.1 Model One Features

[CAT-pos+N+P, CAT-word+N+P, Conc-M+N+P, Conc-SD+N+P, Is-EOS+N+P, Is-SOS+N+P, Percent-Known+N+P, comp-score+N+P, comp-weights+N+P, geco-FFD-mean+N+P, geco-FFD-std+N+P, geco-GPT-median+N+P, geco-GPT-std+N+P, geco-TRT-mean+N+P, geco-fixProp-mean+N+P, geco-fixProp-std+N+P, geco-nFix-median+N+P, geco-nFix-std+N+P, is-mwe+N+P, is-strange+N+P, mwe-cat+N+P, saccade-cat+N+P, saccade-cat-binary+N+P, w2v-embedding+N+P, word-frequency+N+P, word-len+N+P, zipf-frequency+N+P]

3.8.2 Model Two Features

[XLNET-embed, CAT-pos, Is-EOS, Is-SOS, word-len, zipf-frequency]

3.9 Permutation Feature Importance

We use permutation feature importance (Breiman 2001) to better understand the impact of different features on each of the different models. This method measures the base error of the model against the error when one feature is randomly permuted, allowing for quantification of importance. That is for feature i :

$$FI_i = E_{base} - E_{perm_i}$$

We note that permutation methods have a tendency of attributing higher importance to correlated features (Nicodemus et al. 2010), whilst still being informative. Alternatives include per-feature retraining (Lei et al. 2016; Mentch and Hooker 2016) which was computationally intractable within the timeframe of the CMCL task duration.

3.10 Description of features

Feature (generated at the word-level unless specified)	Description	Data and tools used
CAT_word	One hot word encoding	
CAT_pos	Categorical encoding of Part-of-Speech tag	Honnibal et al. (2020)
Is_EOS	Binary variable indicating if word is the last in its sentence	
Is_SOS	Binary variable indicating if word is the first in its sentence	
Conc_M	Mean concreteness norm assigned to the lemmatized form of the word. Words not covered by the dataset of norms were given a 'neutral' score of 3 (concreteness rated on a Likert scale from 1-5)	Brysbaert et al. (2014)
Conc_SD	Standard deviation of concreteness values assigned to lemmatized form of word. Words not covered by the dataset of norms were assigned the mean of Conc_SD for all other words	Brysbaert et al. (2014)
Percent_Known	Proportion of participants asked to estimate concreteness norms that were familiar enough with the word to judge its concreteness. Words not covered by the dataset of norms were assigned a value of 1	Brysbaert et al. (2014)
word_len	Number of characters in the word	
saccade_cat	Categorical representation of number of characters in relation to average saccade length (categories were 1-3, 4-7, 8-10 and 11+ letters)	
saccade_cat_binary	Binary categorical representation of number of characters in relation to average saccade length (categories were 1-3 letters and 4+ letters)	
word_frequency	Frequency of word per million words	Speer et al. (2018)
zipf_frequency	Frequency of word per million words on the zipf scale	Speer et al. (2018)
NEXT_n_FEAT	Attaches FEAT for the next n words to the current word (i.e. NEXT_1_Is_EOS attaches Is_EOS for the next word to the current word)	
PREV_n_FEAT	Attaches FEAT for the previous n words to the current word	
geco_FEAT_mean	Mean average of all measurements of FEAT for this word in GECO. If the word was not present in GECO, the mean of means for words with comparable frequency in natural language was used	Cop et al. (2017)
geco_FEAT_median	Median average of all measurements of FEAT for this word GECO. If the word was not present in GECO, the mean of medians for words with comparable frequency was used	Cop et al. (2017)
geco_FEAT_std	Standard deviation of all measurements of FEAT for this word in GECO. If the word was not present in GECO, mean of standard deviations for words with comparable frequency was used	Cop et al. (2017)
is_mwe	Binary indicator showing if word is part of an MWE in this context	Ramisch (2012)
mwe_cat	Categorical representation of whether the word is part of an MWE in this context, where categories are based on syntactic patterns (i.e. adjective noun compound, verb + preposition phrase)	Ramisch (2012) Loper and Bird (2002)
w2v_embedding	300 dimensional Skip-gram embedding for the word or MWE. If the word is part of an MWE in this context, the Skip-gram embedding trained for the MWE is used instead. Embeddings are trained using the Wikitext-103 corpus, where multiword expressions are reformatted to be concatenated using underscores (i.e. <i>multiword_expression</i>)	Ramisch (2012) Mikolov et al. (2013a) Rehurek and Sojka (2011) Merity et al. (2016)
comp_score	Compositionality score for the MWE calculated using <code>mwetoolkit</code> . Words not part of MWEs are assigned a value of 0	Ramisch (2012) Cordeiro et al. (2019)
comp_weights	Weights used for each word to calculate the <code>comp_score</code> for the MWE (certain words may contribute more semantic meaning to an MWE than others). Words not part of MWEs are assigned a value of 0	Ramisch (2012) Cordeiro et al. (2019)
is_strange	Binary indicator of non-standard formatting or non-alphanumeric characters in the current word (generated using regular expressions)	

Publication III: We Need to Talk About Classification Evaluation Metrics in NLP

4.1 Introduction

Some of the most widely used classification metrics for measuring classifier performance in NLP tasks are *Accuracy*, *F1-Measure* and the *Area Under the Curve - Receiver Operating Characteristics (AUC-ROC)*. For example, seven out of nine tasks of popular NLP benchmark GLUE (Wang et al. 2018) use either Accuracy or F1.

Such metrics reduce the full collection of true classes y and predicted classes \hat{y} to a single scalar value. For instance accuracy, the most common classification metric, is equal to the proportion of predicted classes which match true classes. Whilst capturing all the qualities of a classifier in any single scalar value is rather impossible (Chicco et al. 2021), the quality of the heuristic rule (Valverde-Albacete et al. 2013) influences both the overall ranking of models and the intra-task understanding of model capability.

It is difficult to evaluate true model ability with Accuracy due to the ‘Accuracy Paradox’ (Ben-David 2007): simply guessing the most common class can reward a score equal to that class’s prevalence in the test set. We expand this paradox into two phenomena: (1) the reward given to models that predict more classes which appear more often (are more prevalent) (Lafferty et al. 2001); and (2) the probabilistic lower bound for accuracy being much greater than zero for random guessing models in most realistic scenarios, a phenomenon we term *baseline credit* (Youden 1950).

F1-Measure (Manning and Schütze 1999) is the harmonic mean of precision and recall and so represents a balance of two desirable characteristics of classifiers. F1 is defined against a single class, and so within even a binary classification case its value changes if the classes are reversed. Additionally, the weighting of precision and recall is a function of the model itself (Hand and Christen 2018), making it a poor metric for ranking models. In order to handle the multi-class case, macro- and micro- averaging strategies have been proposed. In the single-label case we consider, micro averaging is reduced to Accuracy, whilst macro-averaging is equivalent to averaging the F1 score across all classes. Therefore, F1-Macro retains both the biases of F1 in the single class

case and introduces a further heuristic in weighting all classes equally regardless of class prevalence.

An alternative to the F-Measure, the Receiver Operating Characteristic (ROC) curve visually presents the trade-off between Recall and Precision as a function of the decision threshold. The Area Underneath the ROC Curve (AUC) is a metric which integrates the ROC curve to return a scalar value. As Hand (2009) has shown, AUC is effectively applying a cost function dependent on the False Positive Rate of the specific classifier, so systems cannot be compared if they have different False Positive Rates.

In this paper we perform an extensive empirical analysis of various classification metrics in synthetic and real settings. We advocate for using **Informedness**, an unbiased and cognitively plausible multi-class classification metric (Powers 2003; Powers 2013) for comparing classification performance of different models instead of common metrics such as accuracy and F1. This metric avoids crediting models exhibiting guessing or bias which distort the comparability of mainstream classification models. Informedness reports the proportion of the time a classifier makes an informed decision; that is, a decision better than bias exploitation strategies. Finally, it allows comparison between tasks of different bias or complexity, and negates the need for dataset re-balancing to ‘fit the metric’.

Our main contributions are as follows:

- A definition of Informedness as a classification metric suited to NLP applications
- Synthetic and real task comparisons of Informedness against an extensive list of classification metrics
- An in-depth analysis on how the use of different metrics can affect model ranking and within task understanding of model capabilities
- Python implementation of Informedness and Normalised Information Transfer to encourage further study within the community

4.2 Classification Evaluation Metrics

We begin by defining various classification metrics and discussing their strengths and limitations.

Metrics operate over a set of classifications, where a true class y and a predicted class \hat{y} are the two elements in each classification. Both y and \hat{y} are indications of a class from out of a set of classes C . The full classification output

$$\left(\begin{array}{c} \{y = C_0, \hat{y} = C_0\}, \\ \{y = C_1, \hat{y} = C_0\}, \\ \{y = C_1, \hat{y} = C_1\}, \\ \vdots \end{array} \right)$$

is unwieldy, so a metric is used to reduce the set more compact form, typically a single scalar value. First, the set of classifications may be considered as a Confusion Matrix (or contingency table), which is an $N \times N$ matrix with the columns by convention indicating the true class and the rows indicating the predicted class. Cells are assigned the number of classification events for the given actual and predicted class. In most NLP cases, creating a classification matrix is a non-destructive operation as the only information lost is the order of the classifications.

As part of our definitions, we introduce the per-class contingency table:

	Class of Interest c	Other Class	Real Class
Class of Interest c	TP $_c$		FP $_c$
Other Class	FN $_c$		TN $_c$
Predicted Class			

Table 4.1: Classification Contingency Table

We define this table for a class of interest c . In the binary case, this would be one of two classes and hence two tables could be created, each the 180° rotation of the other. In the multi-class case, there will be c such matrices.

From this table we also introduce Class Prevalence: the proportion of all samples which have a given real class, and Class Bias: the proportion of all samples which have a given predicted class. Prevalence is $(TP+FN)/(TP+FN+TN+FN)$. Prediction Bias is $(TP+FP)/(TP+FN+TN+FN)$.

Since an $N \times N$ is considered too complex to compare models, a further simplification is often used to produce a single scalar value. As this reduction is an information-destructive operation (Chicco et al. 2021), the heuristic rule (Valverde-Albacete et al. 2013) which the metric applies to obtain a single value will determine what that metric considers be a ‘good’ model.

Accuracy: It is defined as the proportion of correctly identified samples out of a total set of evaluation samples. Accuracy encodes the heuristic that the best model will have the most correctly predicted instances. This prior allows for the ‘accuracy paradox’ where an uninformed model may guess the most common class artificially overestimating the generalizability score.

$$\text{Accuracy} = \frac{1}{S} \sum_{c=0}^C TP_c \quad (4.1)$$

where C is the number of classes, TP_c is the number True Positives for class c and S is the total number of samples.

Balanced Accuracy: This is a variant designed to counteract the class-frequency-weighted nature of accuracy (Brodersen et al. 2010). As shown by Chicco et al. (2021), the binary case is equivalent to a re-scaled Informedness (see below).

F-Measure: This metric is defined as the harmonic mean of the Precision and Recall of a binary classifier.

$$\text{F1-Macro} = \frac{1}{C} \sum_{c=0}^C \frac{TP_c}{TP_c + \frac{1}{2}(FP_c + FN_c)} \quad (4.2)$$

where TP_c, FP_c, FN_c denote True Positives, False Positives and False Negatives for each class c . In the multi-class case (3+ classes), those are computed for each class in turn. F1-Macro encodes the heuristic that the average of F1-Measure for all classes is a good representation of model performance. However, this has no intuitive interpretation. Additionally, as the number of negative samples increases, the number of samples which are misclassified as positive will also increase. As F1 is independent of the total number of samples, it ignores this important component of model assessment. F-Measure may be generalised to multi-class classification through micro or macro averaging. Micro-averaging sums the True Positives, False Positives, and False Negatives when calculating Precision and Recall, and is equivalent to accuracy in the uni-label case. Macro-averaging takes the arithmetic mean over Precision and Recall for every class.

Kappa: This is a family of metrics which calculate the inter-annotator reliability between annotators, rather than the performance of a classifier on a task. However, they account for the probability of chance agreement. Given annotators a_0, a_1 they take the general form:

$$k = \frac{\text{Accuracy}(a_0, a_1) - \text{Chance Agreement}(a_0, a_1)}{1 - \text{Chance Agreement}(a_0, a_1)} \quad (4.3)$$

Kappa metrics differ in how they estimate from how the chance agreement is calculated (Cohen 1960). It is possible to use Kappa as a metric for classification systems by defining the system and the true labels as annotators (Ben-David 2007). However, Powers (2012) has shown that Kappa is unfair to models in cases where the rates of true classes and predicted classes are unequal.

Informedness: This metric treats classification evaluation as an ‘odds game’, where a model with no predictive capability is unable to gain any credit through either **label bias** or **baseline credit**. It was first proposed in the binary case as Youden’s J-statistic (Youden 1950) and was generalised to the multi-class case in Powers (2003). Informedness is defined as the proportion of samples for which the model guesses better than random chance. The expected value of a model which is always correct is 1, and the expected value of a model which predicts correctly $x\%$ of the time, and guesses from the prevalence $100-x\%$ of the time is x .

For a class with an empirical probability (prevalence) of $p(y = c)$, the gain (or loss) i for a single prediction is computed as:

$$i(y, \hat{y}) = \begin{cases} \frac{1}{p(y = c)} & \text{if } \hat{y} = c \\ -\frac{1}{1 - p(y = c)} & \text{if } \hat{y} \neq c \end{cases} \quad (4.4)$$

where $p(y = c)$ is the empirical probability of class c , calculated from the test set. Scores are aggregated across the whole classification set as:

$$I = \sum_{c=0}^C \frac{p(\hat{y} = c)}{N} \sum_y \mathbb{1}(y = c) i(y, \hat{y}) \quad (4.5)$$

Where $\mathbb{1}(y = c)$ is an indicator function which takes 1 when $y = c$ and 0 otherwise.

Mathew’s Correlation Coefficient (MCC): MCC is a measure of the correlation of the predicted classes \hat{y} with the true classes y . Whilst its definition ensures that random guessing will score 0, for any model better than random guessing, it will not report the possibility of random chance. MCC is dependent on the relative frequencies of classes in the test set, which makes comparison between models evaluated on different datasets impossible (Chicco et al. 2021). Formally, MCC is defined as:

$$\text{MCC} = \frac{\text{Cov}(\hat{y}, y)}{\sigma_{\hat{y}} \cdot \sigma_y} \quad (4.6)$$

Normalized Information Transfer (NIT): This information-theoretic measure reports the degree to which the classifier reduces the uncertainty of the input distribution by considering the information transfer through the classifier. It was introduced by Valverde-Albacete et al. (2013). Formally, NIT is defined as:

$$\text{NIT} = 2^{\text{MI}_{\hat{y},y} - H_{U_y}} \quad (4.7)$$

Where $\text{MI}_{\hat{y},y}$ is the Mutual Information of the Real and Predicted Classes, whilst H_{U_y} is the Entropy of the Real Classes if they come from a uniform distribution.

As with Informedness, NIT considers prevalence, forcing classifiers to add Shannon Information, that is, to correctly classify samples, in order to increase the metric score.

4.3 Experiment 1: Metric Evaluation on a Toy Setting

We first compare the metrics outlined in Section 4.2 on a toy setting, aiming to unveil the main differences between them. We assume a simulated model as follows:

- First, we sample from a uniform distribution $[0,1]$ and then pick the correct label if the sample is smaller than model predictive power;
- Otherwise, we randomly sample from the class-prevalence weighted output distribution.
- We score a simulated model with a fixed probability of making a correct classification

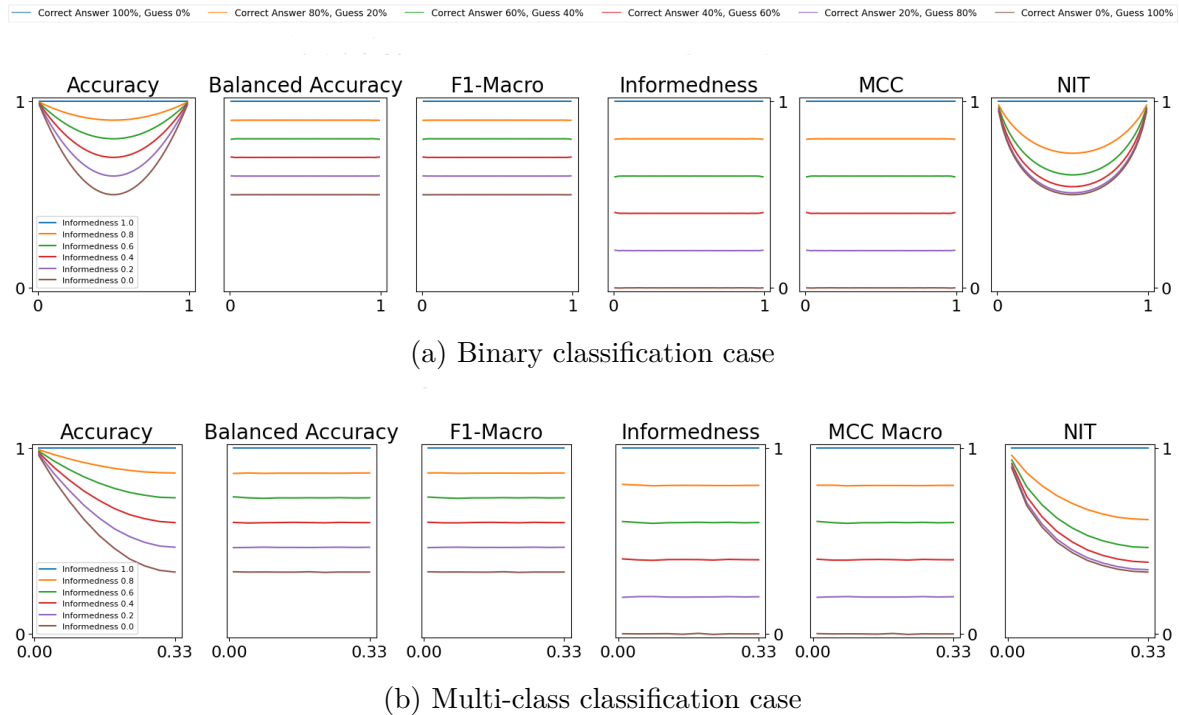


Figure 4.1: Accuracy, Balanced Accuracy, F1-Macro, Informedness, MCC, and NIT of the same binary (top) or multi-class (bottom) classifier as a function of the class distribution and the model’s predictive capacity from 0% (Random Guess) to 100% (Perfect).

We believe this is an acceptable representation of how a reasonably designed and trained neural network would behave.

Figure 4.1 shows the performance of a binary (top) and multi-class (bottom) classifier as a function of the class distribution and the model’s predictive capacity from random guess to perfect.

In the binary case, we first observe that Accuracy becomes more distorted as the prevalence of either class increases. On the other hand, Balanced Accuracy and F1-Macro score are robust against prevalence, but are susceptible to random chance exploitation. Surprisingly, the NIT is superficially similar to accuracy. This can be explained by the fact that when one class is far more probable than the others, the Mutual Information between a random distribution sampled from the same prior is high.

In both binary and multi-class cases MCC-Macro appears to behave exactly as Informedness. This only holds in the case where the *classification ability of the model*

Model (Metric)	Single Sentence		Similarity and Paraphrase			Natural Language Inference					All
	CoLA	SST-2	MRPC	QQP	STS-B	MNLI-M	MNLI-MM	QNLI	RTE	WNLI	
DistillBERT (Acc.)	79.7	90.5	84.2	77.4	51.8	81.4	81.6	88.6	57.6	56.3	74.9
DistillBERT (Inform.)	57.0	81.0	69.4	77.4	41.6	72.1	72.5	77.2	14.7	-43.1	52.0
Random Guess (Acc.)	58.1	51.4	56.7	53.5	18.3	33.5	33.6	50.0	49.9	51.8	45.7
Random Guess (Inform.)	01.2	02.8	-01.1	00.0	01.0	00.1	00.5	00.0	-00.3	02.0	00.6
Δ Accuracy	21.6	39.1	27.5	23.9	33.5	47.9	48.0	38.6	07.7	04.5	29.2
Δ Informedness	55.8	78.2	70.5	77.4	40.6	72.0	72.0	77.2	15.0	-45.1	51.4

Table 4.2: GLUE Results. See Wang et al. (2018) for tasks details and evaluation metrics. All values are scaled by 100. ‘All’ is a uniform weighted mean of the individual metric scores as in <https://gluebenchmark.com/leaderboard>.

is constant across classes (Chicco et al. 2021). We simulate model ability as a function of prevalence, so our figures do not capture this dynamic of the MCC-Macro. However, we do show that in this case Informedness correctly identifies the underlying probability of the model making an informed decision.

4.4 Experiment 2: Metric Evaluation on Natural Language Understanding Tasks

Next, we compare metrics across a range of NLU tasks and show that the metric choice affects the model ranking. First, we test on the **GLUE** Multi-Task Natural Language Understanding Benchmark. GLUE is a suite of nine NLP tasks representing a range of domains, biases, and difficulties (Wang et al. 2018). Interestingly the GLUE employs different metrics across tasks, i.e. Accuracy, MCC, Pearson Correlation and Spearman’s Correlation. MCC is a discretised version of the Pearson correlation and Spearman’s Correlation is the Pearson Correlation calculated on the Rank transformation of the values. To make the continuous $[0, 5]$ STS-B task values tractable for classification metrics, we discretize into $[0, 5] \cap \mathbb{Z}$ by rounding to the nearest integer.

We experiment with following two approaches:

- **Random Guess:** A ‘most likely’ guesser, which chooses the most common class from training;
- **DistilBERT:** We also finetune DistilBERT (Sanh et al. 2019) for five epochs on each sub-task.

Table 4.2 shows model performance across models, metrics and tasks. For the sake of clarity, the last two lines show the difference between DistilBERT and Random Guess

scores. The ‘All’ column is a uniform-weighted mean of the metric scores across the GLUE tasks. In the case of informedness, it represents the average probability of an informed decision across all nine tasks. The use of Informedness across the GLUE tasks allows for direct comparison with the knowledge that bias is discounted.

First, we note that sampling classes according to their prior probability (see *Guess* rows) produces high accuracy scores for many tasks whilst Informedness remains very close to 0. This fact makes it clear that Informedness provides a more interpretable metric when it comes to evaluating model capability. For all tasks, we observe a lower Informedness than Accuracy. This is expected due to the properties of the metrics shown in Figure 4.1. For unbalanced tasks (CoLA, MRPC, WNLI), the gap between accuracy and Informedness is increased as Informedness removes the *label bias* gain. In the three-class tasks (MNLI-M and MNLI-MM), the delta between accuracy and Informedness is reduced but still pronounced.

WNLI is the most interesting result. DistilBERT accuracy (56.3) is a small amount (4.5) larger than random guessing which suggests a weakly predictive model. However, Informedness is strongly negative (-43.1), which suggests that the model is *underperforming* the prior class distribution to a large degree. We hypothesise this is because the WNLI task is **adversarial**. We quote the GLUE authors: ‘Due to a data quirk, the development set is adversarial: hypotheses are sometimes shared between training and development examples, so if a model memorizes the training examples, they will predict the wrong label on corresponding development set example.’ (Wang et al. 2018) Here accuracy suggests a weak model, whilst Informedness reports the real behaviour.

Another advantage of Informedness is the possibility of direct comparison between tasks with varying bias (e.g. CoLA and SST-2) and varying classes (e.g. CoLA and MNLI) without the need to correct for prevalence. Because MCC gives each class equal weight, it cannot be used to compare across tasks with varying class distributions (Chicco et al. 2021). Informedness and NIT support comparison between tasks, but NIT may be confusing for task comparison as it awards credit for guessing.

4.5 Experiment 3: Metric Evaluation in Visual Question Answering

Visual Question Answering (VQA) is the task of answering a question about an image and is often cast as a classification task which requires selecting a correct answer from

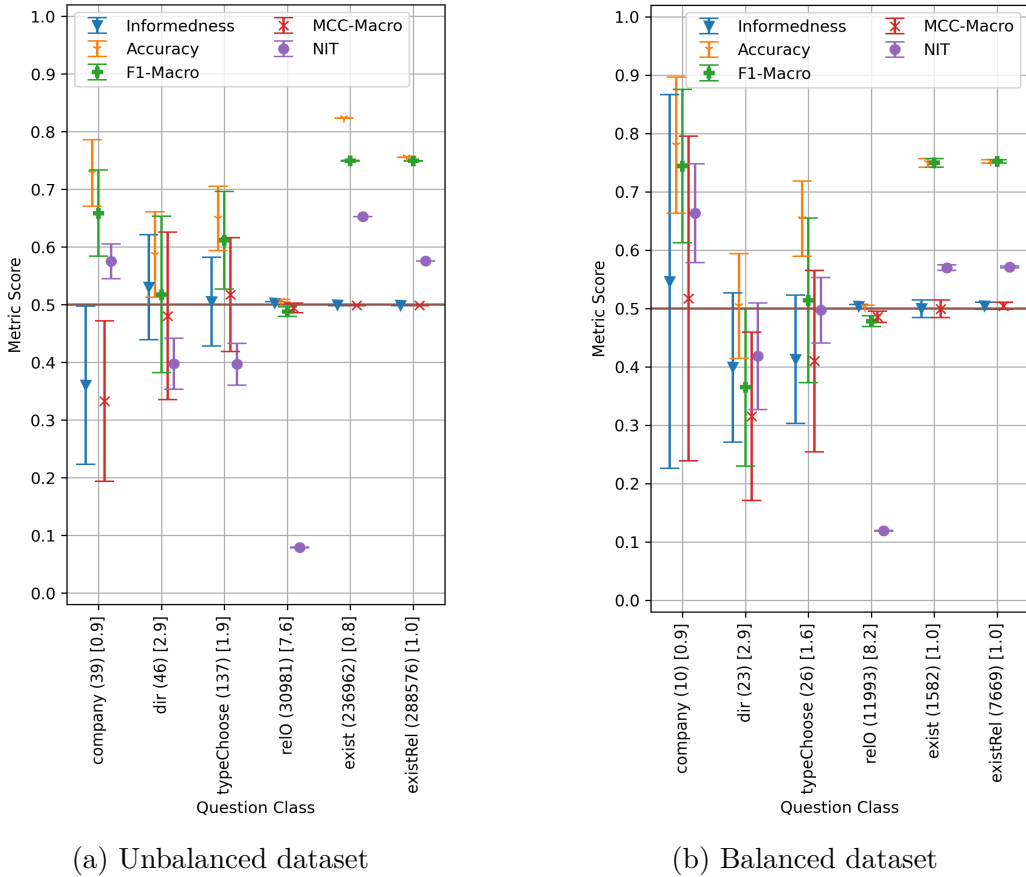


Figure 4.2: Metrics on GQA Unbalanced (left) and Balanced (right) validation splits. Error-bars show the standard deviation across five runs. Numbers after the question category are (question count) and [answer class entropy].

a large set of candidate classes (Antol et al. 2015). Due to the real-world imbalances (for instance, more tables are made of wood than marble), VQA datasets have high tendencies to inherent biases, making accuracy a poor metric to use.

In this work, we consider two VQA datasets: (1) GQA (Hudson and Manning 2019a) and (2) KVQA (Shah et al. 2019)

4.5.1 GQA

We select GQA for the high variance in class count and prevalence across question types. It provides ‘unbalanced’ and ‘balanced’ versions. ‘Unbalanced’ is the default dataset and features a strong prevalence skew due to real world biases towards certain classes. ‘Balanced’ is a resampled version of dataset where the class distributions have

Question type	Dataset		Metric				
	Classes	Entropy	Accuracy	F1-Macro	Informedness	MCC-Macro	NIT
1-Hop	5336	7.4	66.9	10.8	64.6	10.8	25.8
1-Hop Count.	5	1.1	79.3	38.9	58.1	31.5	58.1
1-Hop Subtr.	66	4.1	26.5	03.0	18.8	02.9	17.3
Boolean	2	1.0	94.9	63.2	89.7	89.7	81.9
Comparison	11	2.1	91.1	37.0	90.2	47.3	84.9
Counting	9	2.1	80.9	56.1	75.4	56.2	61.2
Intersect.	2	1.0	79.5	78.5	56.3	59.5	62.1
Multi-Ent.	81	3.2	78.0	10.8	76.1	12.0	56.5
Multi-Hop	119	3.6	87.9	34.8	87.0	43.9	68.9
Multi-Relat.	4104	6.8	75.4	11.7	73.7	12.1	38.1
Spatial	1260	10.0	19.9	07.4	18.6	09.2	16.3
Subtract.	93	5.9	39.8	36.6	45.9	34.3	08.6

Table 4.3: Model performance on KVQA across metrics.

been resampled to reduce the class imbalance.

With GQA, we perform an intra-dataset comparison. Such a comparison is a common step in model and dataset analysis when researchers wish to compare the relative capabilities of a model on different sub-tasks. We provide a model with a predictable behaviour by simulating a 50% probability of choosing the correct answer and a 50% probability of sampling from the class prevalence within a question type. For clarity, we only examine the low-frequency categories ‘company’, ‘dir’ and ‘typeChoose’ and the high-frequency categories ‘relO’, ‘exist’, and ‘existRel’. Results for a representative sub-set of the question types are shown in Figure 4.2. Refer to Appendix 4.9 for the full dataset results.

First, we have many cases where Accuracy, Balanced Accuracy and F1-Macro are 75% on binary questions. This baseline credit makes it hard to compare between model performance, which is calibrated to be uniform, across dataset sub-tasks. Practically, we are not able to use Accuracy, F1-Macro, or NIT to look at ‘typeChoose’ questions and see if the model is as strong as on ‘existRel’. Meanwhile, MCC-Macro and Informedness converge on the correct value (0.5) even with the 46 samples in ‘dir’ question type. The ‘dir’ case demonstrates how the deletion of samples to create a more uniform prevalence is not required with sophisticated metrics. That is, Informedness and MCC are closer to the true value for ‘dir’ with the unbalanced sample than with the balanced one. Meanwhile, the balanced dataset has only a minor effect on accuracy and F1-score, with ‘dir’ and ‘typeChoose’ questions being slightly closer to an unbiased score. This reinforces our hypothesis that dataset balancing is not the correct approach to

evaluation.

For the questions with many samples (‘reLO’, ‘exist’, and ‘existRel’), all metrics have low variance. For ‘exist’, and ‘existRel’, F1-Macro and Accuracy converge on 0.75, which reflects correctly predicting a binary task half the time, and randomly guessing the other half. For the ‘reLO’ question class, Accuracy and F1-Macro tend to the true proportion of the time the model is predicting the correct answer, but this can be attributed to the higher entropy for this class of questions. The same behaviour can be observed for additional question types in section 4.9.

These experiments show that that Informedness automatically accounts for prevalence imbalance and provides a better assessment of the model capability. Whilst MCC appears similar, it over-punishes classifiers which have variable per-class performance (Chicco et al. 2021), which we do not believe is in line with desired characteristics of classifiers in NLP.

4.5.2 KVQA

Having established metric characteristics through controlling model performance, we now move to model evaluation in the wild. First, the KVQA dataset (Shah et al. 2019) provides multiple question type attributes for each question. The task requires reasoning over retrieved knowledge graph facts as well as arithmetical operations. For modelling, we select ‘REUNITER’, a simple yet effective transformer based model (Vickers et al. 2021b), and re-evaluate it with informedness.

We are interested in this case for the opportunity to have a metric which allows comparison *within* a dataset *between* subsections with different class distributions. We present results across unbiased metrics Informedness, MCC-Macro and NIT (Powers 2003; Chicco et al. 2021; Valverde-Albacete and Peláez-Moreno 2014) along with accuracy grouped by question type in Table 4.3.

The ‘1-Hop’ category is a superset of many question types requiring a single KG fact to answer. This question type is scored very differently across all metrics but the difference between Informedness (64.6%), MCC-Macro (10.8%) and NIT (25.8%) is especially striking given the agreement between Informedness and NIT in the synthetic case from Section 4.5.1. This range indicates the model is doing well in general: if it were guessing from a prior, it would have an Informedness of zero. The difference can be explained by the different dynamics of Informedness and MCC raised above. The model is much better than random chance at predicting certain popular classes, but

struggles with low-frequency obscure classes. This is supported by a high accuracy at the same time as a low F1-Macro (12.9). In this case, F1-Macro, MCC, and NIT harshly and unfairly penalize the model.

Looking at the ‘Intersection’ type, we see the opposite behaviour. Accuracy and F1-Macro are all fairly high (78.5 and above) while Informedness is rather low (56.3). This means that Accuracy and F1-Macro exaggerate the predictive power of the model for this type of question. The similar score of MCC-Macro (59.5%) to Informedness indicates that the model has even performance across classes.

Interestingly, accuracy reports that the model is poor at ‘subtraction’ questions, which Informedness is much higher (45.9). We hypothesise this is because (1) transformer models are not good at arithmetic without extensive task-specific pretraining and (2) the high number of output labels will have lower *baseline credit*.

Through the use of Informedness, we come to a different conclusion of the relative strengths of the model. We find that the model has better mathematical ability than accuracy indicated, whilst the ability to reason over intersectional facts is much poorer than accuracy reports. For example, this could lead to focus on improving this sub-task in the future.

Meanwhile, we have the issue that both Informedness and NIT are proposed as suitable metrics for reporting the cross-task capability of different classifiers, but they report divergent scores and sub-task rankings. This is because both metrics target different criteria: NIT the transmission of information from the true labels to the predicted labels, and Informedness the probability of an informed decision. We propose that Informedness is a more intuitive measure for NLP, and refer to Section 4.3 for a toy example demonstration.

4.6 Experiment 4: Metric Evaluation on Formality Control for Spoken Language Translation

In the last set of experiments, we consider a contextual task involving machine translation (MT). The *Special Task on Formality Control for Spoken Language Translation* (Anastasopoulos et al. 2022) evaluates an MT model to correctly express the desired formality (either *formal* or *informal*) in its translation hypotheses. Focusing on the English-to-German language pair, we use the winning system proposed by Vincent et al. (2022). The model is trained to recognise a formality token to generate adequate

	Off-the-shelf MT	Formality-aware MT
Accuracy	50.0	95.4
Balanced Accuracy	50.0	95.3
F1-Macro	49.2	95.4
Informedness	00.0	91.8

Table 4.4: Metric scores on Formality Control for Spoken Language Translation (En-De) between off-the-shelf and formality-aware MT systems.

translations, and an off-the-shelf formality-unaware MT model on the test set provided by the organisers. We report accuracy, Balanced accuracy, F1-Macro and Informedness on the English-to-German test set.

Table 4.4 displays metric scores between off-the-shelf and formality-aware MT systems. We see that the model with no knowledge of the formality is still able to achieve accuracy and F1-score of around 0.5, which seems to mean that the model is able to correctly produce a translation with correct formality 50% of the time. Meanwhile, Informedness drops to zero. As the dataset is balanced, this is a product of Informedness removing baseline credit making it a more suitable choice as an evaluation metric.

Overall, Informedness provides a better and more interpretable measure of the system capability to model the task. This demonstrates that Informedness can be used as an effective tool for comparing two different systems.

4.7 Discussion

4.7.1 Limitations of current metrics

The results obtained across all experiments highlight that widely-used metrics (e.g. Accuracy, F1-Macro) for classification evaluation in NLP feature biases which suggest higher performance than either intuitive reasoning or information theory support. Importantly, this bias makes comparing classifiers across tasks with different class distributions impossible.

Additionally, through the analysis of a real model on the KVQA task, we showed that traditional metrics are not suited to intra-dataset analysis when evaluating a single model’s performance across various sub-tasks. This is highly problematic, as knowing if a model is better at a particular sub-task such as the sub-tasks of addition or syntactic

parsing is crucial for model analysis.

4.7.2 Improving Evaluation of Classification Tasks in NLP

Across all experiments, we found that Informedness better captures model generalizability than all other metrics. Given this finding and the main limitation of popular metrics such as Accuracy and F1 across different NLP tasks, *we encourage the community and practitioners to consider reporting Informedness alongside metrics such as Accuracy and F1 in future experiments and analyses.*¹

4.8 Conclusion

We have presented an extensive empirical analysis of various classification metrics across a wide range of tasks including NLU, VQA and MT with controlled formality. Our experiments demonstrated that the use of a class-invariant metric, Informedness, allows for a fairer ranking and understanding of model generalization capacity.

Whilst we find that Informedness is the most intuitive metric, we also found that it is also the fairest in driving inter and intra-model comparisons.

Finally, we provide `sklearn.metrics` style implementations of both NIT and Informedness, previously unavailable in Python

We hope that our work is the first step towards rethinking the way NLP classification systems are evaluated in the future and will raise awareness to the community.

¹For a discussion of the limitations of Informedness, see Limitations section.

Limitations

Informedness cannot fully represent all of the characteristics of a classification system within a single scalar value. It assumes that the distribution of classes in the training and test set are identical. This assumption is used to determine the loss and gain for a particular class according to the distribution in the test set. However, we allow for train class distributions to be passed to our implementation of Informedness.

In this work, we further assume that an uninformed model will reproduce the training distribution. In the case that models are poorly parameterised, or the testing set is very small, this may not be the case. This could lead to models which are not using the input data to have Informedness scores other than zero. Likewise, systems which use strategies such as ‘guess the most common’ may have Informedness scores other than zero.

Informedness is sensitive to the number of evaluation samples, which may result in less stable estimation of model’s performance in situations with low numbers (< 50) of examples. We consider that all metrics are subject to this and that it is reasonable to expect that evaluation is performed on sizeable test sets.

4.9 GQA Full Comparison

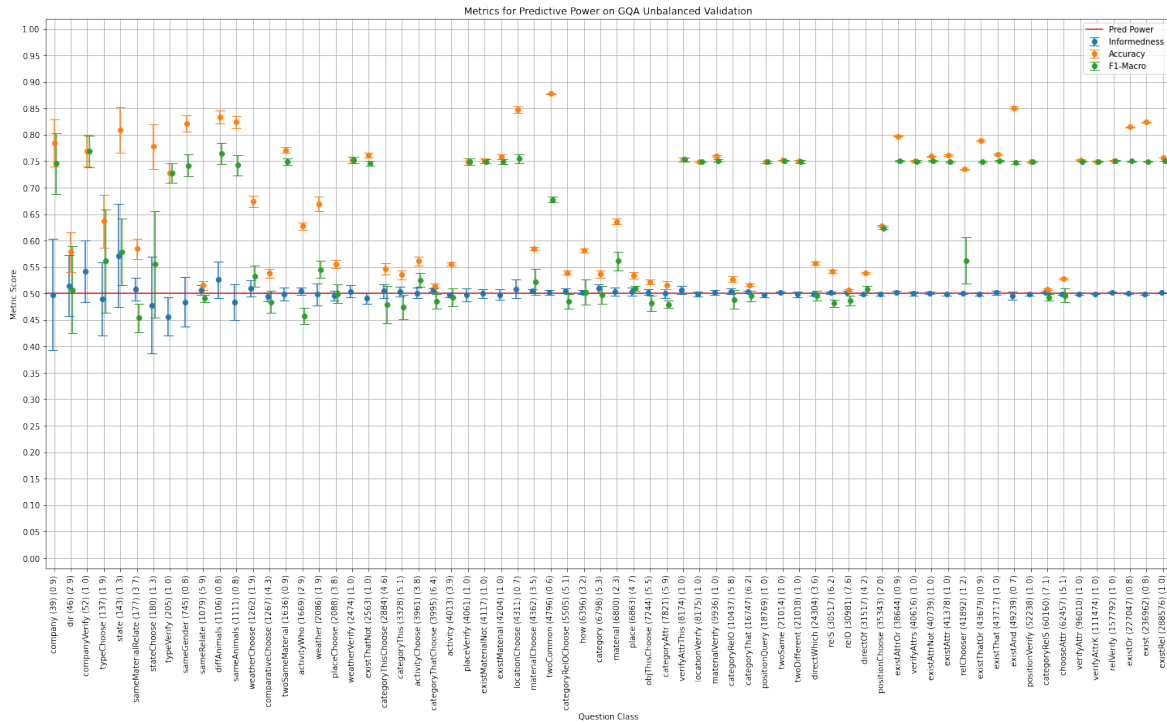


Figure 4.3: Metrics on GQA Unbalanced. Questions are grouped by reasoning type annotation on the X axis and sorted by count. X axis labels gives the reasoning type, the number of samples, and the entropy of the answer class distribution

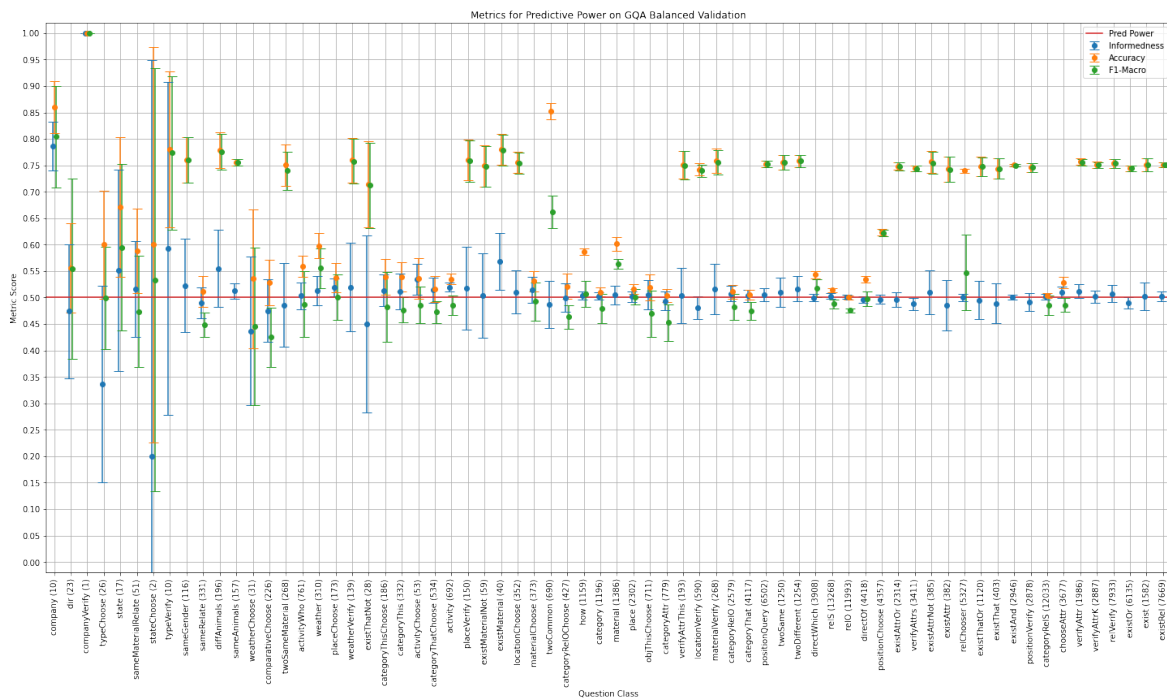


Figure 4.4: Metrics on GQA Balanced. Questions are grouped by reasoning type annotation on the X axis and sorted by count in GQA Unbalanced for comparison. X axis labels gives the reasoning type, the number of samples, and the entropy of the answer class distribution

Publication IV: Comparing Edge-based and Node-based Methods on a Citation Prediction Task

5.1 Introduction

Citation Prediction is the task of predicting whether a given paper cites a target paper (Färber and Jatowt 2020). Imagine the scenario where an author is writing a paper and is open to suggestions about what to cite. Consider a recommender system which will suggest papers in Semantic Scholar (S2)¹ (Ammar et al. 2018), a collection of 200 million academic papers from many fields.² Recommendations can be based on whatever is available in the input draft, including both text and references.

In order to make progress toward this ambitious goal, we introduce a new Citation Prediction task with an emphasis on the time dimension, and evaluate both a node-based model and an edge-based model on this task. The node-based model focuses on titles and abstracts, and the edge-based model focuses on citations.

These models have not been previously compared with one another on graphs of different sizes, especially in a forecasting scenario. Standard benchmarks such as Open Graph Benchmark (OGB)³ (Hu et al. 2020; Hu et al. 2021) and SciRepEval (Singh et al. 2023) evaluate models such as Graph Neural Networks (GNNs)⁴ (Scarselli et al. 2009; Zhou et al. 2018; Wu et al. 2019) and Specter (Cohan et al. 2020) on various academic document modelling tasks including citation prediction.

Unfortunately, most benchmarks are too small to see the region where edge-based methods overtake node-based methods. We expect citations (edges) to outperform text (nodes) when the graph is large enough because of network effects.

These scaling and forecasting issues are important for citation tasks because the

¹<https://www.semanticscholar.org/product/api>

²Medicine (45M), Chemistry (13M), Computer Science (13M), Biology (13M), Materials Science (10M), Engineering (8M), Physics (7M), Psychology (7M), Mathematics (5M), Political Science (4M), Business (4M), Sociology (3M), Geography (3M), Economics (3M), Environmental Science (3M), Geology (3M), History (2M), Art (2M), Philosophy (1M).

³<https://ogb.stanford.edu/docs/lsc/>

⁴<https://web.stanford.edu/class/cs224w/>

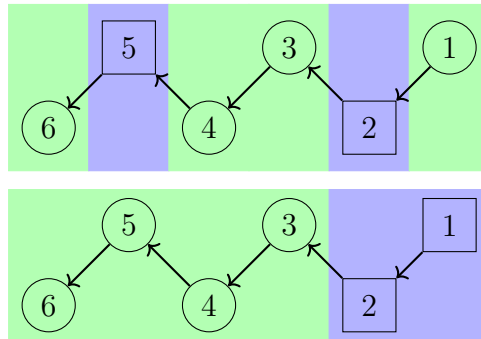


Figure 5.1: Random Splits (top) vs Proposed Causal Split (bottom) for Table 5.1. Train split in green, test in blue. The bottom plot with the train-test cut-off in 2010 gives a temporally consistent split.

Paper	Year	Title
1	2018	[...] Photogramment imaging
2	2016	Convenient probe of $S(1D2)$ [...]
3	2005	Megapixel ion imaging [...]
4	2003	Direct current slide imaging [...]
5	1995	profiles of $CI(2Pj)$ photofragments [...]
6	1988	Adiabatic dissociation of [...]

Table 5.1: 1 cites 2, 2 cites 3,..., 5 cites 6

literature is growing exponentially, doubling every 9 years⁵ (Wade 2022; Kinney et al. 2023). This growth rate is shown in Figure 5.2 for a collection of more than 200 million papers in Semantic Scholar (S2).

The scaling properties mentioned above are somewhat similar to Metcalfe’s Law (Metcalfe 2013). Metcalfe’s Law applies when benefits scale with edges (n^2) and costs scale with vertices (n). In a telephone network, costs scale with subscribers (n), and benefits scale with connections between subscribers (n^2). These network effects are often cited⁶ for the success of businesses such as telephones (AT&T), web search (Google) and social media (Facebook).

Our contributions are as follows:

1. A new benchmark for citation prediction emphasizing scale and forecasting.
2. Empirical demonstration that larger graphs favor edge-based methods.

⁵<https://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html>

⁶https://en.wikipedia.org/wiki/Metcalfe%27s_law

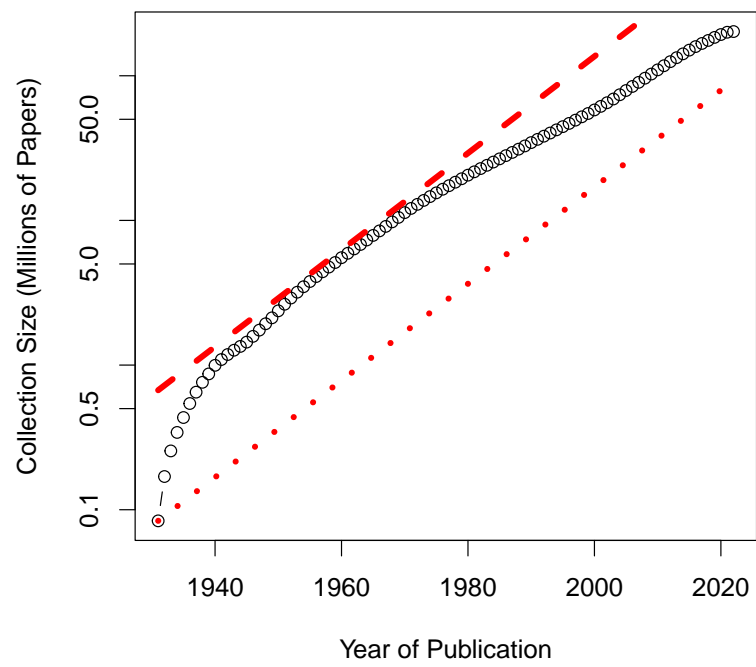


Figure 5.2: The literature doubles every nine years. Observations are denoted by circles and predictions by red lines.

3. Empirical evidence that performance improves with t (larger training sets) and degrades with h (forecasting horizon).
4. We will distribute the benchmark, evaluation code and embeddings.⁷

5.2 Related Work

Citation Prediction is situated within the broader context of recommendation systems. Recommendation systems may model the content ‘Content-based’ (Bhagavatula et al. 2018a) or the preferences of other users ‘Collaborative’ (Resnick and Varian 1997). Citation Prediction the feature of study (citations) may be regarded as both content-based (Caragea et al. 2013), or collaborative (McNee et al. 2002). This is because citations are both significant features of academic documents (content-based) and a representation of the preferences of the document authors (collaborative), so systems may take either approach (Liang and Lee 2023).

In their study of the related topic of Paper Recommendation Systems (Beel et al. 2016) found that 55% of approaches are content-based, whilst collaborative filtering applied to 18%, and graph-based to only 16%. The remaining approaches were hybrids of there or expert systems.

When creating a Citation Prediction model, prior citations may be either be researched as a tool to meet the information need Wilson (1997) of researchers Strohman et al. (2007) and Bethard and Jurafsky (2010a), or as a feature to improve academic document representations in general Cohan et al. (2020) and Yasunaga et al. (2022).

Methodologies Citation Recommendation is subdivided into two sub-categories: Global Citation Recommendation and Local Citation Recommendation. Global Citation Recommendation identifies papers which are cited by a paper given a general or ‘global’ representation of a paper such as the title and abstract. Meanwhile, Local Citation Recommendation identifies papers based on a local text passage such as ‘Citing Sentence’ (Färber and Jatowt 2020). In this paper, we study Global Citation Recommendation.

Within the Content-Based approach for Global Citation Recommendation there are two high-level methodologies: Text-based and Citation-based (Liang and Lee 2023). Text-based methods are Content-Based establish a measure between two papers based on a (sub)set of their textual content (Bhagavatula et al. 2018a). Citation-based

⁷<https://anonymous.4open.science/r/nacl-forecasting-DD3C/>

methods establish a measure between two papers based on a (sub)set of their citations (McNee et al. 2002; Liang and Lee 2023).

Datasets Datasets in Citation Prediction are derived from Paper Repository sources such as the ACL Anthology or Open Academic Graph. Landmark papers in the field have released static datasets based on sub-sampled snapshots of these repositories, and these have formed benchmarking targets. However, as we discuss, these datasets are (a) small, domain-specific, and not time-attributed. We therefore follow Färber and Jatowt (2020) in detailing potential repositories for deriving Citation Prediction datasets which meet these requirements.

The CiteSeer dataset Nallapati et al. (2008) contains 3,312 scientific publications with 4,732 citation links from the Citeseer collection. The CiteSeerX dataset contains 2M papers Caragea et al. (2014), but contained many duplicated papers, and a cleaned version was released Wu et al. (2017).

ACL Anthology Network (AAN) dataset offers a contains 10,921 papers from ACL Computational Linguistics venue (Bethard and Jurafsky 2010b).

The ogbl-citation2 dataset (Roy 2017), a medium-sized collection of 2.9 million nodes and 30.5 million edges, is designed specifically for link prediction tasks in citation networks.

Which not a formal citation recommendation dataset, a 1M subset of SemanticScholar has been used to train and evaluate the Specter and SciNCL document representation models Cohan et al. (2020) and Ostendorff et al. (2022). The evaluation portion is termed SciDocsCohan et al. (2020).

The SciRepEval Singh et al. (2023) benchmark contains an expanded 6M citation recommendation data alongside additional tasks, and was used to train the Specter2 model.

Resources The ACM Digital Library contains 2.4 million publications and 9.7 million citations datasets derived from ACM. Datasets derived from this resource accounted for 22% of such citation prediction studies according to (Beel et al. 2016).

DBLP, another comprehensive computer science bibliography, encompasses over 3 million papers and 25.2 million citation relationships, and is likewise highly popular for citation recommendation experiments, accounting for 33% of citation datasets according to (Beel et al. 2016).

As potential resources, the Open Academic Graph (OAG), an integration of the

Microsoft Academic Graph (Sinha et al. 2015) and AMiner (Tang et al. 2008), contains over 200 million papers and their citations.

The CORE database, while not yet widely used in recommendation systems (Färber and Jatowt 2020), contains 306 million scholarly resources.

The Semantic Scholar (S2) project provides 200M papers with 1.2B Citations. They make their data available through an API or a dump. Semantic Scholar is an academic paper recommendation tool. The owners make the underlying paper repository available, which includes identifiers and metadata, including citations.

Evaluation Citation Recommendation are almost always evaluated offline against a gold standard dataset, which are held-out citations from the network (Beel et al. 2016; Färber and Jatowt 2020). We are not aware of Online Studies, although it would be possible for Paper Recommendation services such as Google Scholar or S2 to conduct them. User Studies are encountered in twice in the literature to our knowledge (McNee et al. 2002; Gori and Pucci 2006).

Offline evaluation involves the creation of a labelled evaluation dataset ahead of time. In the context of Citation Recommendation, this is simply a separate split of the Citation Network data used to train the model, and may be termed ‘citation re-prediction’ Färber and Jatowt (2020). There are two settings for evaluating the offline datasets. First is as a classification task between cited and uncited papers, where precision, recall, F1 score are used. Second is as a retrieval problem of ranking the cited paper(s) above uncited ones, where the Mean Averaged Precision (MAP), Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain (nDGC) are used. nDGC is proposed in cases where the score may be other than 0 (for non-cited papers) and 1 (for cited papers), such as with (He et al. 2010) where co-cited papers are given some credit.

Applications Citation Prediction has a number of downstream applications. Firstly, it aids authors in identifying relevant works to cite, enhancing the quality of their references. For researchers, it facilitates the discovery of relevant literature (Beel et al. 2015; Steinert 2017). It has been used to assist matching reviewers to papers for conference reviewing based on their expertise and publication history (Dumais and Nielsen 1992; Yarowsky and Florian 1999; Mimno and McCallum 2007; Zhang et al. 2023). Similarly, it can be used in expert identification (Yimam-Seid and Kobsa 2003; Maybury 2006; Tu et al. 2010).

Existing Approaches As we discuss above, the literature covers a diverse range of modeling techniques, from collaborative filtering and hand-crafted feature-based models to graph-based methods and deep learning approaches (Jiang et al. 2018). This diversity allows for different tradeoffs in terms of complexity, interpretability, and performance. Simpler feature-based models and some graph-based methods, have shown promise in handling large-scale citation networks and document collections (Brack et al. 2021).

However, these approaches face several limitations.

1. Evaluation challenges persist, as most evaluations use a 'citation re-prediction' approach with offline metrics, which may not fully capture the quality or usefulness of recommendations (Färber and Jatowt 2020).
2. There is a significant lack of user studies and online evaluations with real users (Beel et al. 2016; Färber and Jatowt 2020; Pillai and Deepthi 2022).
3. Furthermore, many methods struggle with the cold start problem, recommending newly published papers with few or no citations (Bhagavatula et al. 2018b).
4. Temporal aspects are often overlooked, as citation patterns change over time Hall et al. (2008), but most models do not explicitly account for the temporal dynamics of citation networks. The only approach we are aware is Local Citation Recommendation, not Global, and divides data into two-year periods from 2007-2016, creating just five time segments for analysis (He and Chen 2018).
5. Most methods focus on relevance, potentially leading to echo chambers, with only a few approaches explicitly considering recommendation diversity (Noordeh et al. 2020).
6. Cross-lingual and cross-discipline limitations are evident, as the majority of work focuses on English-language papers and specific fields like computer science (Jiang et al. 2018).
7. Data limitations, such as limited access to full-text papers, often force systems to rely only on metadata or abstracts (Färber and Jatowt 2020; Pillai and Deepthi 2022).
8. Ethical considerations, such as the potential for automated systems to perpetuate or introduce new biases in citation practices, are not fully addressed in current research (Liang and Lee 2023).

Our approach seeks to specifically address (4): the temporal nature of academic literature, and, by extension, citations. We argue that due to the temporal nature of citations, the partitioning needs to be conducted along time dimension, and importantly, that the time duration t of the training data and the forecast horizon h must be considered. We implicitly address (3), as we evaluate the performance of models on recent forecasting horizons (low h). Our ProNE method addresses (6) by being language-agnostic.

5.3 Methodology

5.3.1 Forecasting Citation Prediction

The Citation Prediction Task is simple: predicting whether paper v_k and v_l cite one another i.e. $(v_k, v_l) \in E$. We define the distance $d(v_k, v_l)$ as the length of the shortest path between vertices v_k and v_l in the citation graph. To make the task harder, we sample relatively challenging negatives, where v_k and v_l are 2-4 hops from one another, i.e., $2 \leq d(v_k, v_l) \leq 4$.

Given the full Semantic Scholar citation graph G , we begin by taking random walks of up to 11 hops. These walks are then filtered with BFS to retain only those walks where $1 \leq d(v_k, v_l) \leq 4$. Verified walks are added to our evaluation dataset, structured as: $\langle v_k, v_l, d(v_k, v_l), \text{bin} \rangle$, where v_k and v_l are two papers in a verified walk, and bin is $\max(\text{bin}(v_k), \text{bin}(v_l))$, the bin of the more recent of the two papers. Binning is discussed in more detail in the next section.

5.3.2 Graph Partitioning

We construct a citation graph, G , based on data from S2. S2 maintains a dataset of around 200 million academic documents dating from 1684 CE up to the current time. Each entry has a primary document id. Document ids are often associated with a title, abstract and citations, though these values can be missing (and incorrect). Figure 5.3 shows that many papers have abstracts, A , and many papers have links in the citation graph, L , but relatively few have both. By construction, random walks are based on L and therefore, the proposed benchmark is a subset of L (papers with links).

To build a causal forecasting task we require a view of the graph which respects the time-dynamics of academic literature (Kuhn 1962; Hall et al. 2008). To do this, we

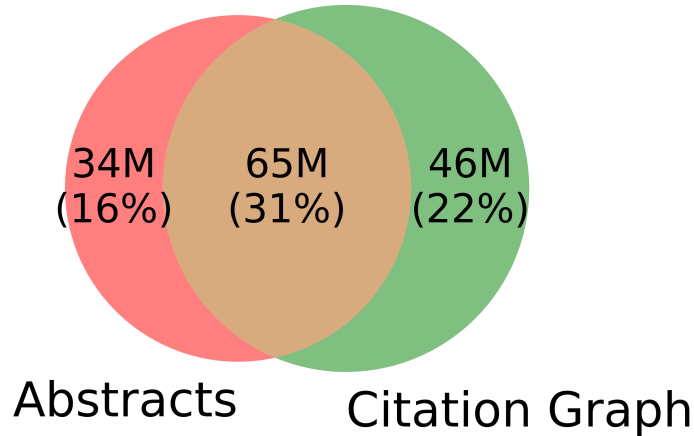


Figure 5.3: There are many missing values. Many papers have abstracts (99M) and many have links in the citation graph (111M), but only 65M (31%) have both.

split the citation graph evolution into 100 chronological sub-graphs.

First, we construct a citation graph $G = (V, E)$, where:

- V is the set of vertices, $\{v_1, v_2, \dots\}$, where v_i is a document id
- E is the set of edges. An edge is a pair of document ids, (v_i, v_j) , where document v_i cites document v_j .

Each vertex, $v_i \in V$, has a publication date.⁸ We use these dates to partition the 200 million total documents in V into the 100 equal-sized bins: $V_0, V_1, V_2, \dots, V_{99}$. Each bin contains approximately two million documents. Let $bin(v_k)$ indicate the bin for paper v_k . That is, if $v_k \in V_b$, then $bin(v_k)$ is b . $bin(v_k)$ is a number between 0 and 99. The bin of an edge, $bin((v_i, v_j))$, is $max(bin(v_i), bin(v_j))$, which is usually $bin(v_i)$ since edges are usually causal. That is, papers typically cite papers in the past, and rarely cite papers in the future.

Due to the exponential pace of paper publications in Figure 5.2, the first bin, V_0 , encompasses papers from 1684 to 1936 CE, while the final bin, V_{99} , encapsulates papers from 2022 to 2023. More details are presented in Appendix 5.9.

For each bin, V_i , we construct a subgraph $G_\tau \subset G$, where $G_\tau = (V_\tau, E_\tau)$. τ is a bin (the max bin of nodes and edges in G_τ). That is, G_τ consists of nodes, V_τ , and edges, E_τ , where V_τ are the documents in bin τ , and E_τ are citations from papers in V_τ to

⁸In fact, there are many missing values (and incorrect values). The set of papers, V , is limited to papers with (non-missing) publication dates.

papers in $V_{i \leq \tau}$. In other words, we allow citations from papers in the current bin (bin τ) to papers in the current bin or in previous bins.

With this partitioning we may set the train-test splits to be between any two subsequent bins. Firstly, this allows us to adjust the size of the training graph to study the effects of scale. Secondly, it allows us to associate test predictions with a time bin. This allows for analysis of test performance as a forecasting task with a forecasting horizon of h . That is, if we train a model on bins up to t , how does the performance of the model on bin $t + 1$ compare with the performance on bin $t + 5$? In general, the task becomes more difficult with larger horizons h .

Let $G^{(\tau)}$ be a cumulative graph:

$$G^{(\tau)} = \sum_{i=0}^{\tau} G_i \quad (5.1)$$

That is $G^{(\tau)} = (V_{i \leq \tau}, E_{i \leq \tau})$, where $V_{i \leq \tau}$ are documents in the current bin or previous bins, and $E_{i \leq \tau}$ are citations from these papers to papers in the current bin or previous bins.

Thus, $G^{(\tau)}$ is the union of all single-bin subgraphs up to and including G_{τ} . We will refer to the vertices and edges in $G^{(\tau)}$ as $V^{(\tau)}$ and $E^{(\tau)}$, respectively.

5.3.3 Dataset Balancing

Within our raw Citation Prediction dataset, we find a non-constant ratio of 1-hop to [2,4] hop labels across different bins. On average, 1-hop labels constitute 28.9%. The accuracy paradox (Valverde-Albacete and Peláez-Moreno 2014) means that bins with a higher prevalence of [2,4] hops will be easier to obtain higher scores on. To rectify this, we down-sample the more-than-average prevalence class until the 28.9% rate is achieved. This adjustment results in the deletion of 7.6% samples. More details can be found in Appendix 5.9.

5.3.4 Representation Models

We use our dataset to compare both node-text and edge-citation representation models.

Text: Specter is an academic document model designed to accept paper titles and abstracts as input. Specter is initialized to the SciBERT model (Beltagy et al. 2019), a variant of BERT trained through masked text denoising of academic documents. Specter

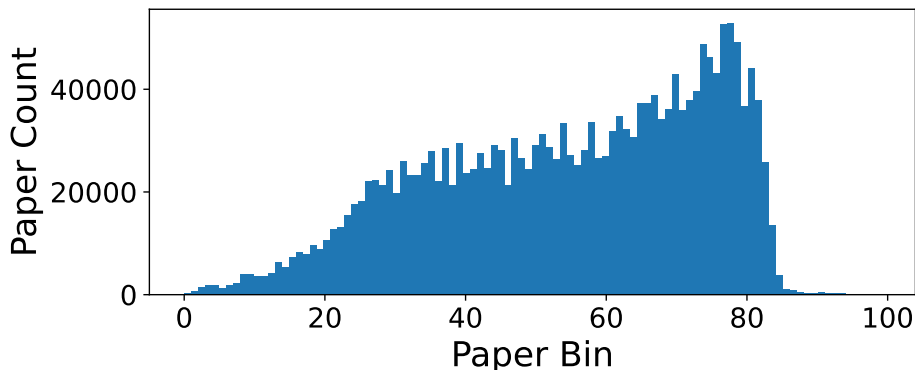


Figure 5.4: Histogram of *query* papers by bin. Specter is trained on triples: $\langle query, pos, neg \rangle$, where the *query* paper cites the *pos*. The distribution of *neg* (random negatives) is similar to *query*, though *pos* predates *query*.

is further trained to minimise a triplet loss across a query paper, a positive paper, and a negative paper. The positive paper is cited by the query, whilst the negative paper is not. The underlying training goal is to ensure the [CLS] token representation of the positive paper is closer (in L2 distance) than that of the negative paper. We use the more recent Specter2 model release.⁹ Specter2 is trained over an approximately 2M paper portion of the S2 corpus. Figure 5.4 shows the distribution of Specter’s training data across our time bins. Discounting the 0.7% of papers for which the publication date is unavailable, 99.8% of Specter2 training papers appear in bins [0-85], which span 1684-2019. We therefore call the Specter release Specter⁽⁸⁵⁾.

Citations: There are a number of methods such as Node2Vec (Grover and Leskovec 2016), DeepWalk (Zhuoren et al. 2014) and ProNE (Zhang et al. 2019b) that take the citation graph as input and apply techniques such as spectral clustering to return an embedding for each paper (vertex). As discussed in section 4.1.1. of (Cai et al. 2018), these methods produce embeddings where cosines can be interpreted in terms of the input graph. If two vectors have a large cosine, then the corresponding nodes are relatively close to one another in the input graph, though details depend on methods and hyperparameters.

We used the `nodevector`¹⁰ implementation of ProNE to generate embeddings. We refer to the ProNE model trained on the S2 citation graph as: ProNE-S. In our experiments, the bottleneck is the SVD of the citation graph. Time and space requirements

⁹<https://huggingface.co/allenai/specter2>

¹⁰<https://github.com/VHRanger/nodevectors>

for SVD grow non-linearly with the size of the graph. Our SLURM cluster allows us to request 2 TBs of RAM and 5 days of runtime per job. The larger graphs consumed about half of these resources. We will need to replace the SVD with an approximation if the literature grows faster than our cluster.

ProNE is a transductive model, meaning it generates embeddings only for documents in the training set, but not for other documents. We introduce a *centroid* approximation to estimate vectors for other documents. The centroid approximation is:

$$vec(v_k) \approx \sum_{v_l \in fanout(v_k)} vec(v_l) \quad (5.2)$$

where $vec(v)$ is the embedding for document v , and $fanout(v)$ is the set of papers that are reachable in one step from v (i.e. cited and citing papers). When evaluating ProNE-S embeddings, we prefer the original transductive embeddings and fallback with the centroid assumption when necessary.

We are interested evaluating two scaling effects:

1. The effect of graph scale on representation quality (size of τ)
2. The impact of time duration between train data and evaluation data (forecast horizon)

To evaluate these scaling effects, we train ProNE on each cumulative subgraph, $G^{(\tau)}$, $\tau \in [0, 99]$, resulting in 100 ProNE-S $^{(\tau)}$ models. ProNE-S $^{(\tau)}$ is trained on $G^{(\tau)}$, and maps documents in $V^{(\tau)}$ to vectors.

5.3.5 Evaluation Task

We evaluate both ProNE-S and Specter models on our Forecasting Citation Prediction dataset. Similar to cumulative graphs, we use the notation $M^{(\tau)}$ to indicate the maximum graph partition which a model is trained on. We report results for all bins, but like (Färber and Jatowt 2020), we are particularly interested in predictions for papers published after the training set: bins $> \tau$. Results for bin $\leq \tau$ are less interesting because those bins were used for training.

As discussed in Section 5.1, we are interested in measuring trends in forecasting capability: how does accuracy depend on the interval between training time and evaluation time? Consistent with research in other domains (and common sense), we

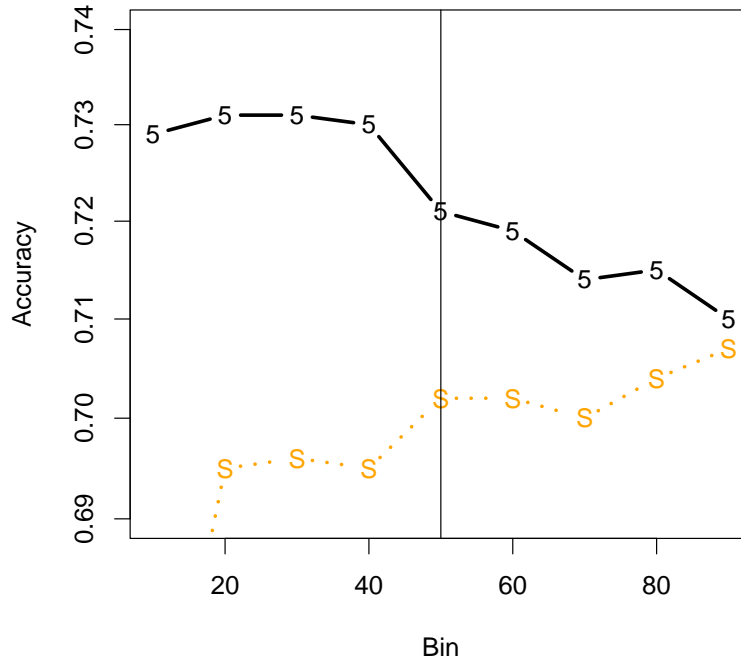


Figure 5.5: Results on ProNE-S⁽⁵⁰⁾ (trained on $G^{(50)}$) and tested on $T^{(k)}$ for $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$; dashed lines compare with Specter.

expect the task to become more challenging when the evaluation is based on papers that are published well after the papers in the training split.

5.3.6 Evaluation Implementation

We perform classification by taking the cosine similarity of model’s representations of papers A and B and evaluating against a learn-able threshold. For each model, we use the first 1/6th of the data as a validation split to find this threshold, and evaluate on the remaining 5/6th of papers.

Missing Values: In the case of missing values for either paper, we predict by sampling from a Bernoulli distribution parameterised by the train distribution of the overall rate of 1 vs [2,4] hops (28.9%).

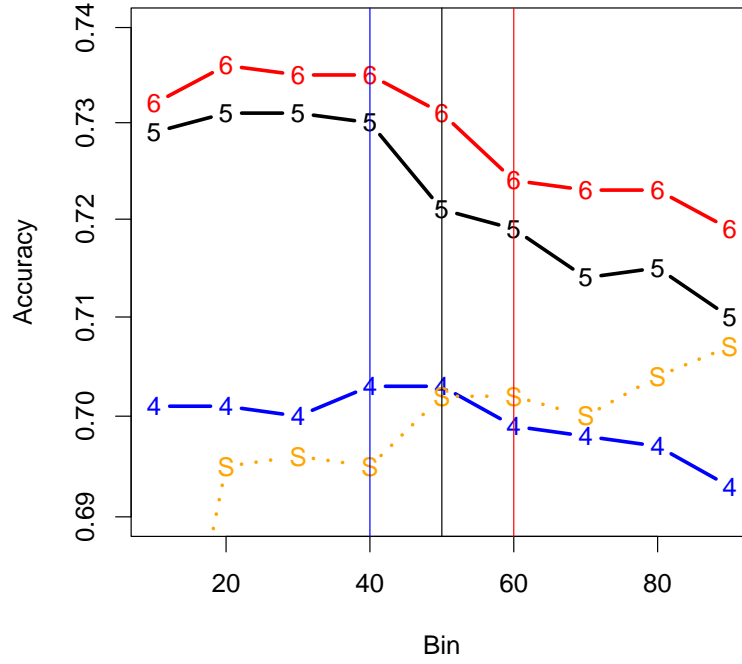


Figure 5.6: Performance improves with larger training sets: $\text{ProNE-S}^{(60)} > \text{ProNE-S}^{(50)} > \text{ProNE-S}^{(40)}$

5.4 Results and Analysis

Figure 5.5 shows the performance of $\text{ProNE-S}^{(50)}$ (labeled ‘5’) on every tenth evaluation bin. Accuracy is better over the training set (left of the vertical line). Accuracy suddenly decreases moving into the first forecasting bin ($h=1$) and then slowly decreases further into the future. These results confirm our Forecast Dynamics assumption, that as time into the future increases, the model’s performance degrades. We plot the Specter⁽⁸⁵⁾ (labeled ‘S’) for comparison. Despite having more recent training data, Specter underperforms $\text{ProNE-S}^{(50)}$.

Figure 5.6 is like Figure 5.5, but Figure 5.6 shows performance of three ProNE-S models, trained on $\text{ProNE-S}^{(40)}$ (blue/4), $\text{ProNE-S}^{(50)}$ (black/5) and $\text{ProNE-S}^{(60)}$ (red/6), respectively. Note that the red line is consistently above the black line, and the black line is consistently above the blue line ($\text{ProNE-S}^{(60)} > \text{ProNE-S}^{(50)} > \text{ProNE-S}^{(40)}$) because training on more bins is better than training on fewer bins. Figures 5.5-5.6

ProNE-S Train Bins	Test Bin										Mean
	0	10	20	30	40	50	60	70	80	90	
0-0	0.543	0.573	0.557	0.559	0.566	0.563	0.561	0.562	0.561	0.558	0.560
0-10	0.701	0.706	0.662	0.670	0.663	0.663	0.665	0.659	0.659	0.659	0.663
0-20	0.736	0.724	0.725	0.709	0.703	0.700	0.698	0.693	0.690	0.686	0.699
0-30	0.701	0.701	0.701	0.700	0.703	0.703	0.699	0.698	0.697	0.693	0.703
0-40	0.751	0.729	0.731	0.731	0.730	0.721	0.719	0.714	0.715	0.710	0.724
0-50	0.772	0.732	0.736	0.735	0.735	0.731	0.724	0.723	0.723	0.719	0.733
0-60	0.756	0.732	0.732	0.736	0.735	0.734	0.732	0.726	0.724	0.725	0.738
0-70	0.726	0.726	0.733	0.734	0.734	0.735	0.735	0.737	0.726	0.730	0.743
0-80	0.731	0.732	0.736	0.736	0.737	0.733	0.734	0.735	0.733	0.730	0.745
0-90	0.736	0.731	0.737	0.733	0.736	0.736	0.735	0.740	0.739	0.740	0.750
Specter	0.569	0.661	0.695	0.696	0.695	0.702	0.702	0.700	0.704	0.707	0.701

Table 5.2: Accuracy of 10 ProNE-S models and Specter on citation prediction forecasting task. Lines indicate the train-test divide.

show that accuracy is better when tested on the training set, and declines the more we predict into the future.

Figures 5.5-5.6 are based on Table 5.2, which reports results for every tenth ProNE-S model and for Specter⁽⁸⁵⁾ in Table 5.2. These expanded results confirm the two results of note for ProNE-S model:

1. Accuracy degrades with h (forecast horizon), as shown in Figure 5.7. This observation is validated by OLS regression analysis, where accuracy drops by 0.0009 (coefficient: -0.0009, t-value: -41.246, p-value: <0.0001).
2. Accuracy improves with t (size of training set). For every additional training bin, accuracy improves by 0.0009 (coefficient: 0.0009, t-value: 12.280, p-value: <0.0001). This supports the use of very large training graphs.

Appendix 5.9 shows the full results for all cumulative 100 ProNE-S models and results on each 100 evaluation bins.

5.5 Citation Prediction

We rerun our experiments on Citation Prediction from Section 5.4 using the Informedness metric from Paper III. As stated in Section 5.3.2 we segment the evaluation set by the maximum time bin of either citing or cited paper. The exact ratio of non-citing to citing pairs was not constant: mean 0.288, but a standard deviation of 0.084. In order to avoid the Accuracy Paradox from introducing noise into our results, we first

Accuracy Train Bins	Test Bin										Mean
	0	10	20	30	40	50	60	70	80	90	
0-0	0.782	0.528	0.530	0.544	0.553	0.556	0.557	0.558	0.562	0.570	0.574
0-10	0.709	0.701	0.623	0.627	0.639	0.651	0.657	0.665	0.670	0.679	0.662
0-20	0.405	0.611	0.648	0.646	0.672	0.683	0.685	0.696	0.705	0.712	0.646
0-30	0.304	0.553	0.602	0.624	0.664	0.680	0.687	0.702	0.713	0.723	0.625
0-40	0.434	0.629	0.667	0.679	0.703	0.705	0.708	0.717	0.728	0.737	0.671
0-50	0.491	0.653	0.683	0.693	0.712	0.718	0.716	0.726	0.733	0.743	0.687
0-60	0.507	0.664	0.687	0.698	0.715	0.722	0.725	0.730	0.733	0.745	0.692
0-70	0.520	0.667	0.693	0.700	0.716	0.724	0.728	0.740	0.735	0.748	0.697
0-80	0.471	0.650	0.681	0.692	0.713	0.721	0.726	0.738	0.744	0.751	0.689
0-90	0.533	0.669	0.696	0.699	0.717	0.726	0.728	0.742	0.748	0.756	0.701
Informedness	0	10	20	30	40	50	60	70	80	90	Mean
0-0	0.328	0.033	0.015	0.029	0.030	0.023	0.024	0.014	0.018	0.024	0.054
0-10	0.390	0.399	0.167	0.124	0.102	0.093	0.089	0.080	0.072	0.070	0.159
0-20	0.125	0.140	0.140	0.083	0.076	0.063	0.052	0.043	0.042	0.036	0.080
0-30	0.000	0.000	0.000	0.000	0.011	0.013	0.016	0.013	0.014	0.012	0.008
0-40	0.162	0.187	0.197	0.185	0.181	0.140	0.128	0.113	0.109	0.088	0.149
0-50	0.236	0.250	0.256	0.245	0.240	0.219	0.195	0.185	0.173	0.149	0.215
0-60	0.250	0.280	0.277	0.270	0.266	0.250	0.241	0.229	0.210	0.189	0.246
0-70	0.249	0.290	0.301	0.288	0.283	0.272	0.265	0.271	0.241	0.232	0.269
0-80	0.196	0.239	0.251	0.240	0.240	0.226	0.222	0.221	0.212	0.196	0.224
0-90	0.284	0.294	0.305	0.288	0.285	0.277	0.265	0.276	0.268	0.259	0.280

Table 5.3: Accuracy and Informedness of 10 ProNE-S models on our Citation Prediction forecasting task. Lines indicate the train-test divide.

calculated the ratio of non-citing to citing pairs across the entire dataset. We then ensured this ratio was found in each of the 100 time-segmented bins by down-sampling the more-than-globally prevalent class. This resulted in the exclusion of 215,000 samples, which is 7.56% of the dataset.

5.5.1 Results with Informedness

We rerun the experiment with no down-sampling but the Informedness metric introduced in III.

The results over the full data in Table 5.3 confirm those in Paper III. We note the following differences: The lack of citations in early bins allows models to report higher accuracy scores without being any better at prediction. This is observed through the accuracy score of 0.782 for the model trained only on bin 0 evaluated only on bin 0. Meanwhile, the Informedness for the same setting is 0.328 - much lower and reflecting the lower power of this model. We are also able to discern that the model trained on bins 0-30 is broken, and is merely sampling the prior probability of classes in each bin.

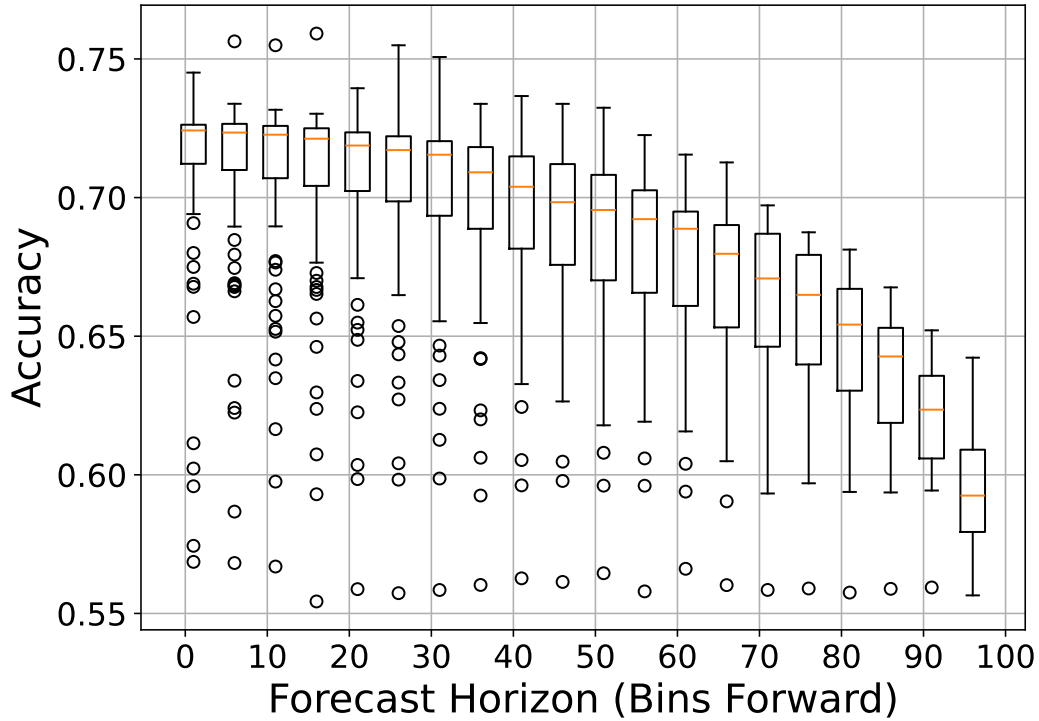


Figure 5.7: ProNE-S Accuracy Across Forecast Horizons

This is very hard to identify with accuracy as the rate of prevalence of each classes changes through time. Informedness, however, is in the range 0.000-0.016 for all bins, which makes analysis much more straightforward. Finally, we note that the best scores are around 0.27 Informedness, which is quite low. This suggests a much larger headroom for future improvements than the accuracy score indicates.

5.5.2 Early and Late Bins

Early bins (Bins 0-5) and late bins (Bins 95-99) produce outliers in our evaluation with both Specter and ProNE-S models, although the effect is most noticeable with Specter. We speculate that two effects are in play: (1) Specter’s training set is skewed towards more recent papers (see Figure 5.4) and (2) older papers and newer papers have relatively noisy metadata due to issues such as OCR noise and preprints. OCR is more common for older papers and preprints are more common for newer papers.

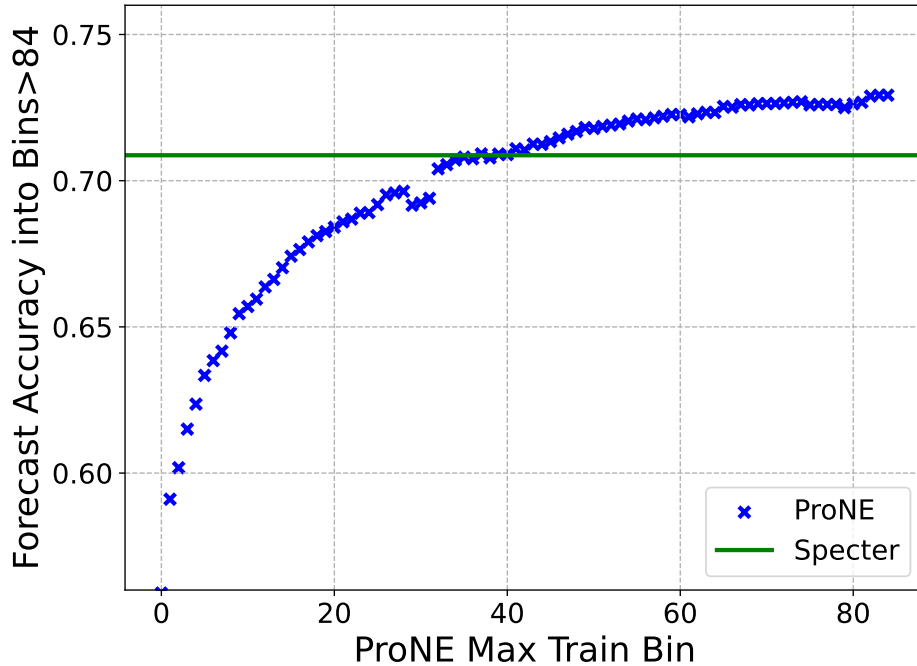


Figure 5.8: ProNE-S–Specter Crossover: Metcalfe’s Law favors larger citation graphs (more than 82M papers).

5.5.3 Comparisons and Combinations

The next two subsections will discuss:

1. **Comparisons:** As suggested above, larger t (training data) favors ProNE-S, but where is the cross-over point?
2. **Combinations:** Ensembles of ProNE-S and Specter can be better than either by itself.

Comparing Text and Context (Citations)

As previously discussed, we train 100 ProNE-S models times on increasing cumulative graphs from G^0 to G^{99} . There is only a single Specter model, trained by S2 on papers published up to 2019, corresponding to our bin 85.¹¹ In Table 5.2 ProNE-S overtakes Specter at around bin 40, but it is not possible to make a precise determination from looking at every tenth bin.

¹¹<https://github.com/allenai/specter>

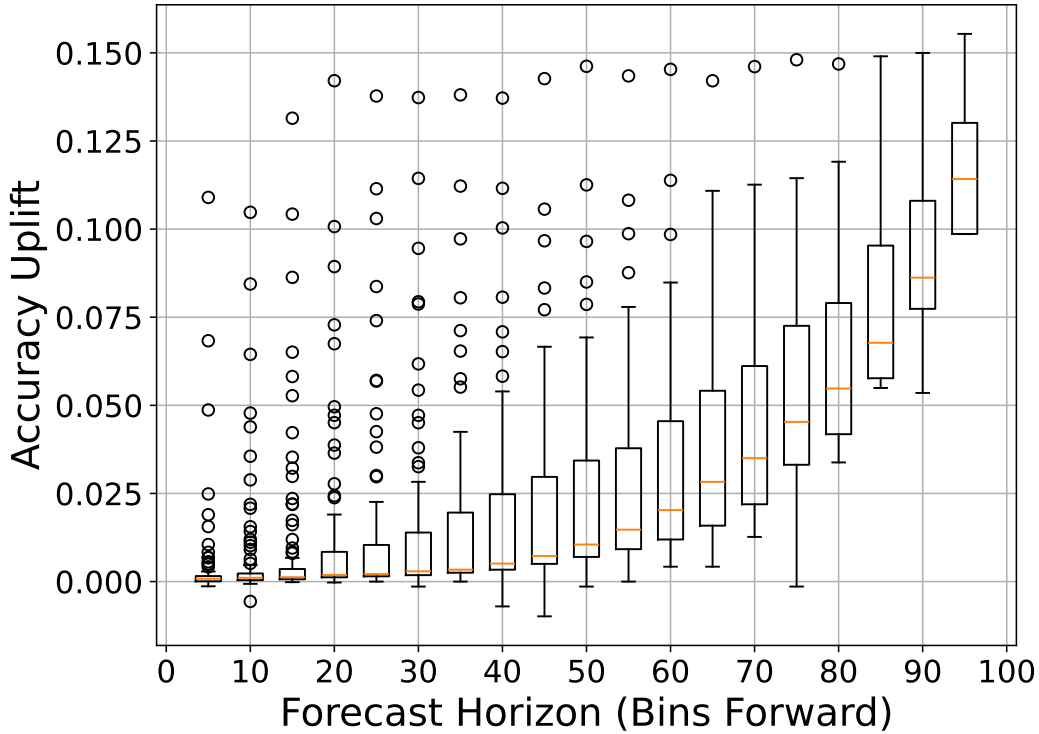


Figure 5.9: Hybrid Accuracy Uplift from ϕ_p to ϕ_z across all ProNE-S versions

According to our time binning, Specter2 is Specter⁽⁸⁵⁾, so we evaluate both ProNE-S and Specter on bins 86 to 99. For ProNE-S^(τ), we use all models where $\tau \leq 85$ as later iterations can access the test the citations during training. As we seek to compare models, we average the accuracy over all forecasting bins rather than considering them individually. Figure 5.8 shows the averaged forecasting accuracy across for each model across bins 86 to 99. Specter outperforms the smaller ProNE-S models (left side of plot), whilst ProNE-S is better for larger graphs (right side of plot). The crossover point is around ProNE-S⁽⁴¹⁾, which has about 82 million papers.

It is remarkable that ProNE-S⁽⁴¹⁾ has comparable accuracy with Specter⁽⁸⁵⁾, given the large difference in t (training data). ProNE-S⁽⁴¹⁾'s training data ends at bin 41 (2007), 11 years before the beginning of the test set (2018).

Combining Text and Context

We have evaluated the forecast capabilities of individual models and compared two different types of models, text and citation. A further area of evaluation is combining the predictions of the two different model types. Certain document representation models

such as GNNs use both text and citations (nodes and edges) as input features. However, these methods may not apply to the case of missing values, which Figure 5.3 shows is more than half of S2. Additionally, we have shown that text- and citation-based models have different forecast characteristics. In a real world application, it may be desirable to optimise performance for papers published this year, given models trained on much older papers. This is somewhat equivalent to optimising for a particular evaluation bin. We show that under these conditions the optimal combination policy will change over time. To show this, we evaluate several strategies for ensembling Specter and ProNE-S models.

More formally: consider a scenario with N models M_i , each providing a forecast $F_i(t)$ at time t . The objective is to devise a combination policy ϕ that maximizes the forecast accuracy $A_i(t)$ for a given evaluation set. We will show that the optimal combination policy ϕ is not consistent across:

1. Different forecast horizons
2. Model variants with different train data (bins)

To make our study as clear as possible, we pick a straightforward hybrid system, which is to use ProNE-S citation embeddings when available, and otherwise, fall back to Specter text embeddings. We term this $\phi_{\frac{P\tau}{S}}$, (where τ indicates the max train bin of the ProNE-S model). We run this policy across all ProNE-S $^{(\tau)}$ models, recording results as distances from the test-train split (i.e. for ProNE-S $^{(n)}$, predictions on bin $n + 3$ count as a 3 bin forecast). Results are shown in Figure 5.9. This hybrid system improves performance, especially on extreme-range forecasting for the undertrained bin ProNE-S $^{(\tau)}$ models where $\tau < 40$.

In Section 5.5.3, we observed that the relative performance of ProNE-S and Specter depend on t (size of training set) and h (forecast horizon). We therefore compare $\phi_{\frac{P\tau}{S}}$ to two ‘no-op’ baselines: (1) ProNE-S only: $\phi_{P\tau}$ and (2) Specter only: ϕ_S .

We evaluate all three policies on three ProNE-S variants: ProNE-S $^{(0)}$, ProNE-S $^{(10)}$, and ProNE-S $^{(30)}$. We choose these small τ s to explore the region where Specter and ProNE-S have similar accuracy. The miss rate for ProNE-S $^{(0)}$ is 89.5%, for ProNE-S $^{(10)}$ 40.0%, and for ProNE-S $^{(30)}$, it is 8.8%. We show that no single policy dominates over all forecasting horizons h . That is, there are regions where $\phi_{\frac{P\tau}{S}}$ is best, and other regions where $\phi_{P\tau}$ is best, and regions where ϕ_S is best, as indicated by the color bars in Figure 5.10. The color bars also show that ensembling is often effective. Note that there

is more green (ensembling) in Figure 5.10 than red (Specter only) and blue (ProNE-S only).

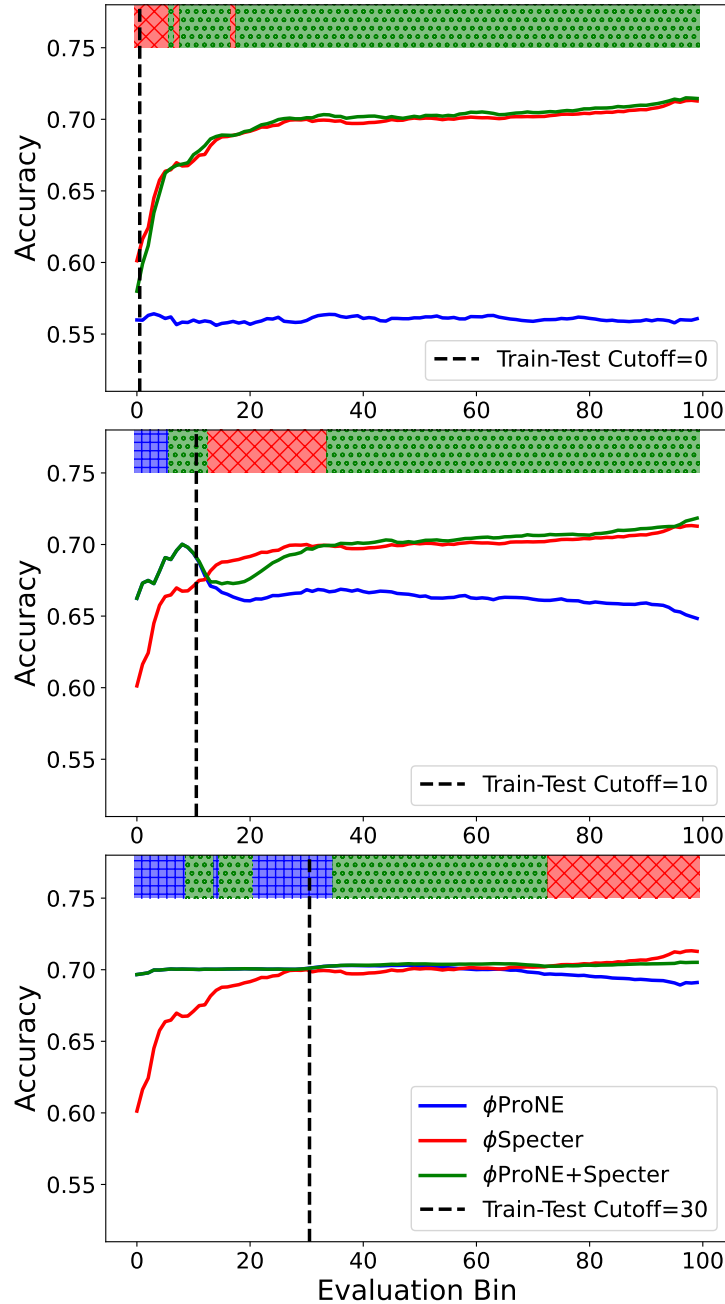


Figure 5.10: Hybrid Forecast Accuracy (with 5-point Moving Average) for $\phi_{\frac{P_T}{S}}$ with ProNE-S^{(0), (10), (30)} and Specter. The color bars indicate which line has the best accuracy. Differences between bars suggest the policy for combining text and context depends on the size of the training set (t) and forecasting horizon (h).

We start with the least trained ProNE-S⁽⁰⁾ : $\phi_{\frac{P_0}{S}}$. This ensemble uses ProNE-S when ProNE-S is able to form embeddings, and falls back to Specter otherwise. We plot the $\phi_{\frac{P_0}{S}}$ vs ϕ_{P_0} vs ϕ_S in the top plot of Figure 5.10. To highlight which ϕ is performing best at a given time-period, we shade above the graph the color of the best policy.

We observe that with ProNE-S⁽⁰⁾, ϕ_S outperforms the hybrid system $\phi_{\frac{P_0}{S}}$ for the first 8 bins, before converging for later forecast horizons. Convergence over latter bins is due to Specter dominating the hybrid system as ProNE-S has little coverage so far out from the training data. In $\phi_{\frac{P_{10}}{S}}$ (middle plot) we see again that the under-trained ProNE-S system adds noise to short-range forecasts (bins 11-30), where ϕ_S – the Specter system alone – performs best. Finally the bottom plot of $\phi_{\frac{P_{30}}{S}}$ shows that $\phi_{\frac{P_{30}}{S}}$ becomes the optimal policy for forecasting near term (bins 31-70), whilst Specter remains more accurate for long range forecasting (bins 71-80). In short:

1. The optimal policy ϕ varies over models and time horizons. This can be seen through the color of best ϕ changing across each version of ProNE-S we plot in Figure 5.10.
2. By modelling Text and Citations separately, we are able to change the feature combination policy ϕ for different forecast horizons, which produces higher accuracy overall.

The hybrid system increases coverage from 96.0% to 99.0% for bin 50. We also find an uplift in prediction performance of 3.91% on average in forecasting bins, however, this varies by forecast distance and number of ProNE-S training bins. Further, we find that there is more opportunity for ensembling when there are more missing bins. Benefits for ensembling decrease when there is more training data.

5.6 Conclusion

We produced a link-prediction benchmark based on Semantic Scholar (S2) to measure forecasting capability of document models at scale. We then evaluated 100 ProNE-S models and Specter on data binned by time. Forecasting is important because science evolves over time. We found:

1. Performance improves as t increases (more training data over more time), and
2. Performance degrades as h increases (predictions further and further into the future).

Metcalf’s Law suggests edge-based methods such as ProNE-S should be relatively effective for larger graphs. We found that that was correct, with a cross-over point around 82 million papers, much larger than the training set for Specter.

Since ProNE is transductive (and does not generate embeddings for novel papers), we introduced the centroid assumption, so ProNE-S can generate embeddings for papers that cite known papers. This extended version of ProNE-S performed well on our forecasting benchmark, especially when trained on larger graphs.

We further investigated the relationship of text and context models through an analysis of model combination strategies. First, we found that simply using Specter as a fallback for missing values in ProNE-S boosts long-horizon predictions significantly. Secondly, through evaluation over 100 cumulative ProNE-S models, we found the optimal model combination policy for combining Specter and ProNE-S depends on forecast horizon, h .

In practical applications, $h \gg 0$, since we typically train models once, and then use them much later for inference. Retraining more often will reduce h , and improve predictions at inference time. If we plan to combine models such as Specter and ProNE-S, the combination should be reevaluated more often since combinations also depend on h .

Finally, we will make our benchmark, embeddings and scripts available for further research and experimentation.

5.7 Ethics

It is good for society to make the scientific literature more accessible. There is no sensitive data in this work. All of data we use for creating our benchmark is freely available through Semantic Scholar. The Specter2 model is available on Huggingface. We make the ProNE-S and Specter embeddings available on Globus. We release our benchmark on our project GitHub along with our evaluation code. Both embeddings and code are available under the MIT license.

5.8 Limitations

Scientific literature is growing quickly, and our benchmark will need to be updated frequently to stay relevant. Training ProNE-S over a 200 million paper benchmark

requires significant compute resources (2+ days on a 2TB RAM HPC machine). Citations are not recorded as accurately for non-English language papers. Specter will likely not perform well for non-English language text and abstracts.

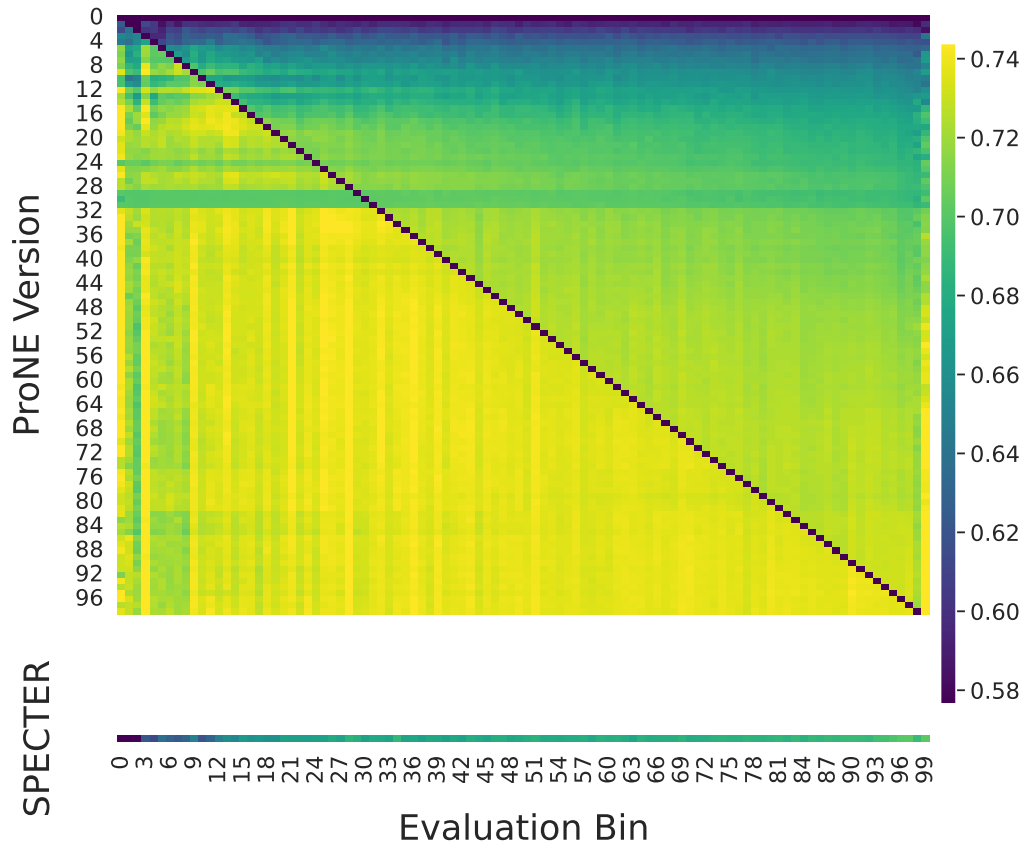


Figure 5.11: Full ProNE-S Cumulative Forecasting Results

5.9 ProNE-S Forecast Heatmap

5.10. PRONE-S FORECAST TABLE

Table with columns: Max Train Bin (0-98), 40-59. Contains a grid of numerical forecast values.

Publication V: SynthVQA: Towards Flexible External Knowledge VQA Dataset Creation

6.1 Introduction

Visual Question Answering (VQA) is the task of automatically answering questions given corresponding images. It rose to prominence as a benchmark for vision-language models with the publication of the first VQA dataset by Antol et al. (2015). Since then, many similar datasets have been published. These can be categorized into three main groups: traditional, external knowledge, and relational reasoning.

In traditional VQA, questions are sourced from human annotators, with variations in images or annotator prompts accounting for their diversity (Goyal et al. 2019). In External Knowledge VQA (EKVQA) datasets (Wang et al. 2016), questions require knowledge from outside the question-image pair to answer. EKVQA datasets source their external knowledge from either human annotators (Marino et al. 2019; Schwenk et al. 2022) or Knowledge Graphs (KG) (Wang et al. 2017b; Shah et al. 2019). Relational reasoning datasets probe a model’s ability to resolve spatial questions over an image. Questions are generated with a ‘question engine’ which operates over an abstracted representation of object locations and relations (Johnson et al. 2016; Hudson and Manning 2019a). The engine picks a specific type of predefined question template, and then grounds this template against an abstract graph representation of objects in the image, the ‘scene graph’.

Most VQA datasets suffer from biases. Human annotators tend to ask similar questions with highly skewed answer distributions, the ‘How many X? 2’ problem (Goyal et al. 2019). Real-world distributions are biased, the ‘What is the table made of? Wood’ problem (Hudson and Manning 2019a). Moreover, EKVQA datasets typically consist of few samples (around 10k per dataset). The AOKVQA dataset is the largest human annotated dataset at 25k samples (Schwenk et al. 2022). The cost of human annotation required for such datasets constrains their size and the accessibility of creating them. This is troubling due to the over-concentration of resources on small static distributions in applied Machine Learning (Church and Kordoni 2022).

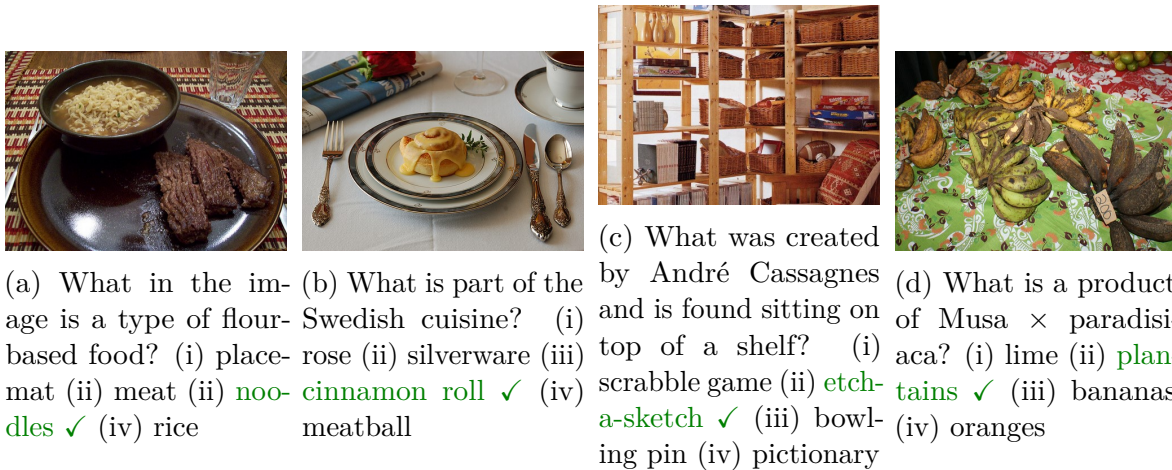


Figure 6.1: SynthVQA samples generated automatically from our GRAVITY framework.

Synthetic data creation has been proposed as to mitigate these concerns (Qian et al. 2023). The text-only analogue of EKVQA is Knowledge Base Question Answering (KBQA). KBQA research has focused on how to sample more expressive questions that static templates allow (Su et al. 2016; Gu et al. 2021).

Whilst Large Language Models (LLMs) are capable of learning a great number of facts, they fail to apply these facts to new linguistic contexts (Berglund et al. 2023). We speculate that EKVQA with straightforward yet diverse facts provided by an expressive graph sampling method will challenge state-of-the-art Vision-Language models such as BLIP-2 (Li et al. 2023).

In this paper, we propose a new framework for generating EKVQA samples. Similar to Relational Reasoning VQA, our system samples facts from graphs which represent images (scenegraphs). Our approach is novel in that (a) the graph is composed of relations from an external Knowledge Graph (b) instead of using templates which are bound to specific relations, sample from *structures* of edges and nodes. This approach is used to generate expressive and diverse KBVQ datasets (Dutt et al. 2023). We provide examples of our dataset in Figure 6.1. Our contributions are:

1. The Graph-based Reasoning for Automated Visual Intelligence Test Yield (GRAVITY) framework which automates creation of diverse, flexible, grounded VQA datasets.
2. A new dataset named SynthVQA, generated by applying GRAVITY over 10k images from Visual Genome (Krishna et al. 2016) and facts from Wikidata.
3. Evaluation of state-of-the-art models on our dataset, finding that SynthVQA is

challenging (50.2% accuracy for BLIP-2-AOKVQA).

This work aims to democratize the creation of EKVQA datasets by dramatically reducing the costs of annotation.

6.2 Related Work

Visual Question Answering (VQA) rose to prominence as a benchmark for Vision-Language models with the publication of the eponymous Visual Question Answering dataset 2015 (Antol et al. 2015). Since then, many other tasks fitting the paradigm of text questions over images have been published.

6.2.1 Visual Question Answering

We categorize Visual Question Answering datasets into three high-level groups:

Group 1 is ‘traditional’ VQA, where questions are elicited from human annotators, with variations in images or annotator prompts accounting for their diversity (Goyal et al. 2019). These include VQA (V1), VQA V2, Visual7W, FM-IQA, Visual Genome, which use human annotators to generate questions over images Antol et al. (2015), Zhu et al. (2016), Malinowski and Fritz (2014), Gao et al. (2015), and Krishna et al. (2016). Notably, VQA V2 adjusted away from overly informative priors by incorporating images with anti-prior answers Goyal et al. (2019). VQA-CP v1/2 focused on changing priors to enhance model robustness Agrawal et al. (2018). R-VQA filters VG Questions, keeping only those where the question-answer has a high semantic similarity with the underlying fact Lu et al. (2018).

Group 2 is ‘Knowledge Base’ or ‘External Knowledge’ VQA Wang et al. (2016). These datasets are designed to be more challenging than traditional VQA due to the inclusion of Real-World or Commonsense Knowledge into questions. Whilst a typical VQA question might be ‘What is the chrome object?’, an External Knowledge VQA question might be ‘In which year was the chrome object invented?’ This external knowledge may come from either a Knowledge Graph or human annotators.

6.2.2 External Knowledge VQA

Formally, EKVQA is defined as:

$$a^* = \underset{a \in A}{\operatorname{argmax}} p(a|q, i, K) \quad (6.1)$$

where a^* is the predicted answer, A is the set of all possible answers and q, i, K are a text question, an image, and a Knowledge Graph respectively. An additional constraint is that all of q, i , and K are required to solve the dataset, i.e. the question cannot be answered with only partial information.

Fact-based VQA (FVQA) Wang et al. (2016) contains 2190 images with 5826 questions and 4216 facts. Each question is paired with a supporting fact and an image. Questions are generated by requiring human annotators to ask questions that both an visual and external commonsense reasoning to answer. Through human annotation they find that 97.6% of questions require common sense knowledge to answer, and 99% of the supporting-facts provided represent this common sense knowledge. Samples are annotated with a supporting fact and a visual concept. FVQA samples are annotated with a required fact for answering and the location of the answer (‘image’ or ‘KG’).

Knowledge-Base VQA (KB-VQA) Wang et al. (2017a) contains 700 images with 2402 questions which require external knowledge, and their system uses DBpedia Auer et al. (2007). Questions are generated by asking human participants to instantiate question templates with features present in an image and concepts present in the DBpedia knowledge base. The authors partition their data into ‘Visual’, ‘Common-sense’ and ‘KB-knowledge’. ‘Visual’ may be answered by visual concepts learned from relations in the training data: ‘Is there a table in this image?’; ‘Common-sense’ questions should be hard to solve purely from relations in the training data, but easy for adult humans for instance: ‘How many mammals are in this image?’; ‘KB-knowledge’ require the average adult to refer to an external data-source to answer. Samples are annotated with a question structure category. KB-VQA samples are annotated with a ‘template-type’ which correspond to a question template.

Knowledge-aware Visual Question Answering (KVQA) Shah et al. (2019) contains 24k images with 183k questions which require external knowledge, and a ‘closed-world’ subset of the Wikidata knowledge base. Questions are generated by asking human annotators to create questions which require external knowledge from pre-defined templates. Ground truth answers are then found through SPARQL queries

of the Wikidata knowledge base. They place an explicit focus on ‘KB-knowledge’ (questions which would require the average adult to refer to an external data-source to answer). They restrict query entities to famous people who feature in the Wikidata knowledge base. Samples are annotated with reasoning type labels. KVQA samples are annotated with multiple labels indicating the kind of KG retrieval and the kind of reasoning required.

Outside Knowledge VQA (OK-VQA) Marino et al. (2019) contains 14k images with 14k questions annotated by humans to be ‘hard for robots’. They do not supply or recommend a specific knowledge base, although their best-performing system utilises Wikipedia. They categorise questions into knowledge domains, such as ‘Transportation’ and ‘Cooking’. OKVQA samples have semantic question type annotations.

Augmented OK-VQA (A-OKVQA) Schwenk et al. (2022) is a follow-up work to OK-VQA, expanding on its approach while addressing some of its limitations. A-OKVQA contains approximately 25k questions paired with images from the COCO image dataset. A-OKVQA required human annotators to provide rationales for the underlying reasoning for each question. Additionally, they perform more rigorous question filtering process, which removes 60% of initially questions. Finally, due to the filtering to remove common questions, A-OKVQA exhibits a long-tail distribution of answers, with many answers appearing infrequently, challenging models to handle rare or unseen answers.

Group 3 is ‘Relational Reasoning.’ These datasets probe a models ability to resolve spatial questions over an image. Questions are generated with a ‘question engine’ which operates over an abstracted representation of object locations and relations Johnson et al. (2016) and Hudson and Manning (2019a). The engine will pick a specific type of predefined question template, and then find a valid realisation of this across the image scene graph. CLEVR creates synthetic scenes and runs scene graph engines over them Johnson et al. (2016). GQA implements a question engine over real paired images and scenegraphs from Visual Genome to outputs compositional Relational Reasoning questions Hudson and Manning (2019a). RAVEN uses Raven’s non verbal reasoning ‘Progressive Matrices’ to power a question engine Zhang et al. (2019a).

Dataset	Questions	Question Gen	Languages	Images	KG
VQA	614K	Human	English	205K	No
VQA v2.0	1.1M	Human	English	205K	No
VQA-CP v1	245K	Human	English	118K	No
VQA-CP v2	658K	Human	English	219K	No
Visual7W	328K	Human	English	47K	No
FM-IQA	316K	Human	Chinese, English	158K	No
CLEVR	1.0M	Engine	English	100K	No
RAVEN	7.0M	Engine	English	1.1M	No
R-VQA	335K	Human	English	123K	No
GQA	2.2M	Engine	English	113K	No
KBVQA	2.4K	Human	English	700	Yes
FVQA	5.8K	Human	English	2.1K	Yes
KVQA	183K	Engine*	English	24K	Yes
OK-VQA	13K	Human	English	14K	No
A-OKVQA	25K	Human	English	24K	No
Synth-VQA	7.7k	Engine	English	7.7K	Yes

Table 6.1: Popular VQA datasets, their question count, and their creation method


Type	Example	Abstraction
Graph Isomorphism		1
Logical Template	<ENTITY><POWERED BY><ENTITY>	2
Logical Form	<TRAIN><POWERED BY><ELECTRICITY>	3
Question	What is powered by electricity? Train	4

Table 6.2: Levels of Abstraction in Knowledge Graph Representations

6.2.3 Knowledge Base Question Answering

Here we outline relevant prior work in KBQA. Note that KBQA is text-only. Gu et al. (2021) create a Question Answering dataset from Knowledge Bases through human annotated mappings. See Table 6.2 for examples of the levels of abstraction used in KBQA. They first sample exemplar sets of connected entities ‘Logical Forms’ (level 3) which they then manually construct templates for mapping into Natural Language questions (level 4). They then sample further sets of entities (level 3) with the same template (level 2) and reground the template to generate questions (level 4). This work is further generalised by Dutt et al. (2023) to graph isomorphisms (level 1). Graph isomorphisms lose the specific attributes of the node or edges and retain only the structure (Lan and Jiang 2020; Li and Ji 2022). These forms can then be used to seek patterns in the graph to sample suitable questions. However when isomorphism

are used to sample from a graph, they will generate multiple logical templates. Each logical template requires a specific question template to map from its logical forms to questions.

6.3 GRAVITY Framework

In this section, we define our Graph-based Reasoning for Automated Visual Intelligence Test Yield (GRAVITY) framework for automatic generation of VQA datasets. Then, in Section 6.4, we discuss how we implement this framework over Visual Genome and Wikidata. A full overview of the framework is presented in Figure 6.2.

6.3.1 Knowledge Graph (KG)

A KG is defined by an ontology $O \subseteq E \times R \times (E \cup L)$, where:

- E represents the set of entities,
- R is the set of relations (or properties) between entities,
- L is the set of literals, such as numbers, strings, or dates,
- $E \times R \times (E \cup L)$ signifies the possible triples formed by entities and relations, resulting in either another entity (E) or a literal (L).

Examples of triples included in the KG:

- an entity-to-entity relation: `<'Barack Obama'; 'alma mater'; 'Harvard Law School'>`
- an entity-to-literal relation: `<'Barack Obama'; 'height'; '1.87m'>`.

6.3.2 Question Sampling

We follow Dutt et al. (2023) in sampling questions from as graph isomorphisms. These are simply distinct patterns of linked nodes and edges in the Knowledge Graph. Sampling a graph with isomorphisms returns a wide variety of logical forms L_q . For instance, the simplest identity isomorphism could generate logical forms as diverse as `<'Columbia College', 'alma mater of', 'Barack Obama'>` and `<'Stonehenge', 'age', '4,000 years'>`.

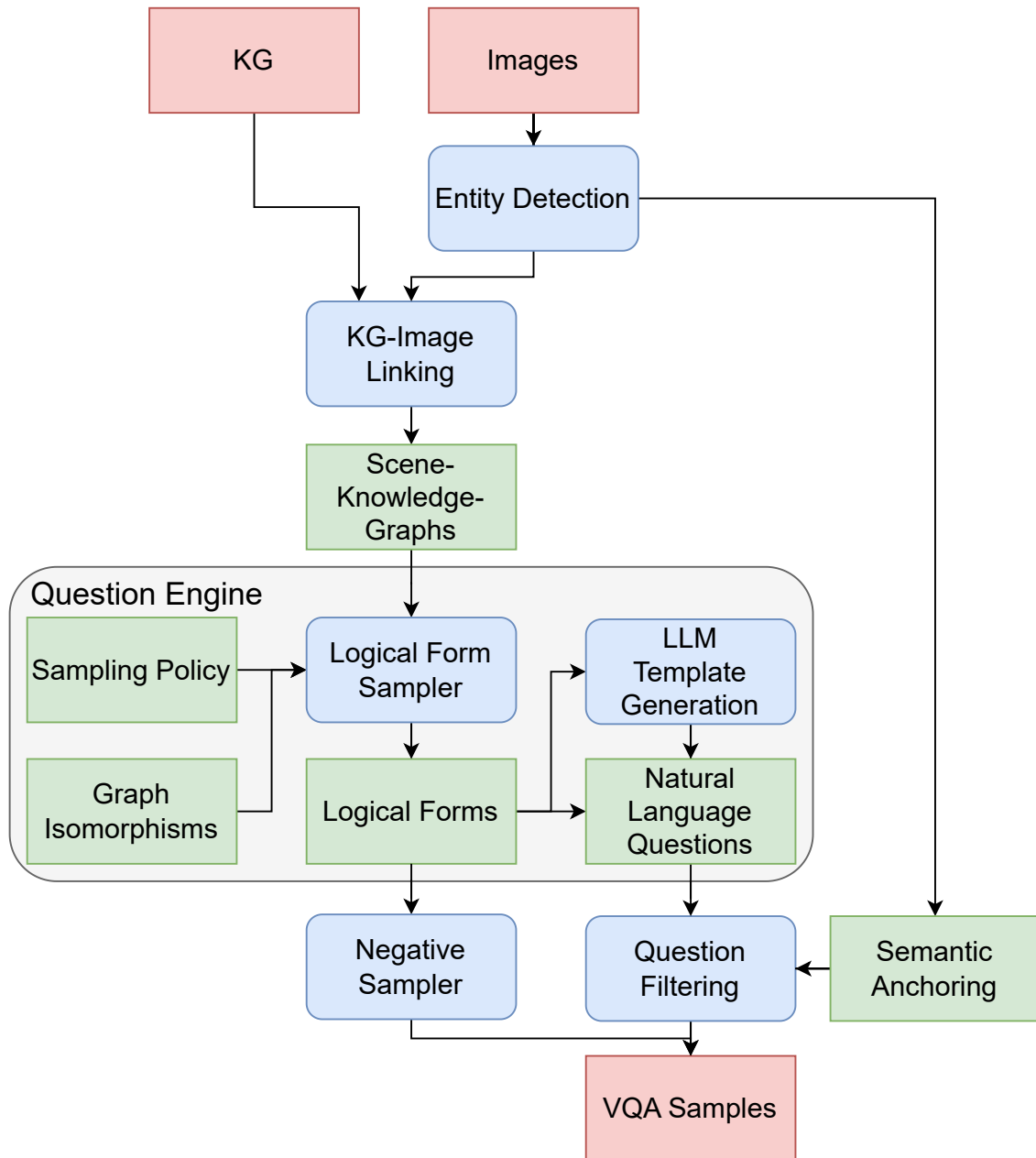


Figure 6.2: The GRAVITY framework. The visual entities extracted from the image are linked to the knowledge graph (Sec. 6.3.4), and then further processed by the question engine (Sec. 6.3.5) to generate question candidates. Hard negative answers are then created (Sec. 6.3.6). Those questions are then filtered according to several requirements (Sec. 6.3.7).

6.3.3 Image Collection and Entity Detection

A set of images, denoted as $I = \{i_1, i_2, \dots, i_n\}$, forms the visual dataset for the VQA tasks. As Equation 6.1 states, the image must be required for reasoning in EKVQA tasks. The entity detection process involves identifying significant objects, concepts, or regions within an image, denoted as $e_i \in D$, where D represents the detected entities in an image.

6.3.4 KG Linking

Unambiguous questions require a total function mapping M that associates each detected entity e_i in D with a corresponding entity E_k in the Knowledge Graph K . This linking process is critical for integrating visual data with structured knowledge.

We consider each image I_j linked to a KG as new KG, which we call a ‘Scene Knowledge Graph’ (SKG), $I_j \circ K = S_j$. This is because in VQA, questions are only over single images. The formal definition is given by:

- A detected entity set $D = \{e_1, e_2, \dots, e_m\}$ within an image,
- A mapping $M: D \rightarrow E$, associating each detected entity e_i with an entity E_k in the KG,
- The original Knowledge Graph K modeled as $O \subseteq E \times R \times (E \cup L)$,

The Scene Knowledge Graph SKG is constructed as the image and the one-hop KG locality as follows:

Initial Node Set: Start with $N_{SKG} = \{E_k | e_i \in D, M\}$, the set of entities in K obtained by mapping the detected entities e_i with M .

Expansion Process: For each entity $E_k \in N_{SKG}$, we augment N_{SKG} with all entities E' and literals L that are directly connected to E_k in K , along with the relations R that connect them. The SKG is then the sub-graph of K induced by N_{SKG} , containing all logical forms, i.e. triplets of entities, literals and their relation, relevant to the detected entities D in the image.

6.3.5 Question Engine

Given graph isomorphisms and one or more *SKG*, we generate logical forms by sampling from each. We represent logical forms as graph triples `<`pizza';`made with';`pineapple'>`. In order to function as EKVQA questions, these representations need to be converted to a natural language form such as ‘What is made with pineapple?’. With a restricted set of logical templates, human authoring of manual mappings is a practical, but not scalable, solution.

We want to allow (a) drop-in use of any KG and (b) diversity of logical templates which fit graph isomorphisms. As previously stated, each logical template requires a custom question template to map its grounded logical form to a natural language question. `<`pizza';`made with';`pineapple'>` \rightarrow ‘What is made with pineapple?’. We propose the use of a Large Language Models (LLM) for generating templates to natural language questions.

6.3.6 Hard Negative Samples

VQA is a multiple choice task. To be challenging we must find three convincing (but not correct) hard negative answers for each question. For the first two distractors the answers with the highest Pointwise Mutual Information (PMI) between the relation plus the grounding entity and and all answers across all Scene Knowledge Graphs.

Given a logical form L and a candidate answer A , the Pointwise Mutual Information (PMI) is defined as $\text{PMI}(L, A) = \log\left(\frac{P(L,A)}{P(L)P(A)}\right)$, where $P(L, A)$ is the joint probability of observing both L and A together, $P(L)$ is the probability of observing L , and $P(A)$ is the probability of observing A .

For the third distractor we call an LLM with the question and a prompt requesting a plausible answer that is distinct from the answer and PMI distractors.

6.3.7 Question Filtering.

The previous steps generate questions with variable quality. We filter out samples failing the following requirements:

1. *Require reasoning over the image.*
2. *Require reasoning over the text.*
3. *Is not ambiguous.*

4. *Is well formed.*
5. *Derive from valid image-KG linking.*

Point 1 is fulfilled by design. Entities from the images are used to query the KG. Human (inter)annotation is typically used to evaluate samples against these criteria. We use an LLM to detect issues [2-5]. As LLM cannot ‘see’ images, we use the twenty most popular entities in the image as a semantic signal for the filtering operation. We provide the prompt in Appendix 6.9.2.

6.4 SynthVQA Dataset

This section presents the setup we used to create the SynthVQA dataset following the methodology presented in the previous section.

For our KG, we use the Wikidata world knowledge graph (Vrandečić and Krötzsch 2014). Wikidata is licensed under Creative Commons Zero license, allowing data reuse for any purpose. Wikidata satisfies ontology O we defined in Section 6.3.1. It has $\langle E \times R \times E \rangle$ relations of the form $\langle \text{'Earth'; 'diameter'; '12,742 km'} \rangle$. We denote the Wikidata KG as K_w .

For our Image set I we use Visual Genome (VG) (Krishna et al. 2016). Visual Genome is licensed under Creative Commons BY license, allowing data reuse given attribution. These images are human annotated with entities and also provide scene graphs. VG contains 113K images along with human-annotated objects, relations, and attributes. These annotations constitute a scene graph which fulfills our definition of a KG. However, these ‘KG’ are unlike large scale World or Commonsense KG in that they do not make the universality assumption: the grounding of ‘largest mammal’ in Wikidata is ‘Blue Whale’, whilst in an image it may be a person, or a dog, or have no answer. We denote this locality with K^l . We term VG specific scene graphs as $K_{v_i}^l$, where i is the image index within VG. We enforce that the question must include knowledge from the KG. E.g. ‘What is red and on a sidewalk in this image?’ is rejected because is entirely relational reasoning, whilst ‘What is red in this image and invented in 19C?’ is included because it contains K_w knowledge.

6.4.1 KG Linking

In order to use the question engine over Wikidata for VQA, we need to link entities in K_v^l to K_w . The Commonsense Knowledge Graph (CSKG) project provides mappings from WordNet synsets used in VG to Wikidata IDs (Ilievski et al. 2021). Table 6.4 shows that CSKG covers 53% of all entities and 39% of all unique entities. We link the remaining unmapped entities through a two stage process:

1. We query Wikidata with the English language name of the entity. If this returns multiple results, we disambiguate by picking the entity with the most site links (Wikipedia pages in different languages).
2. If (1) returns no results, we search Wikipedia for the entity and return the Wikidata ID associated with the first result page.

Coverage statistics are given in Table 6.4. We end up with 68.6K linked unique entities.

As discussed previously, each set of entities $E_j \in K_{v_i}^l$ will have different logical forms. Therefore, we need to sample questions at the *image* level rather than at the Wikidata level. Practically we achieve this by expanding $K_{v_i}^l$ with localities from K_w around linked entities:

$$\text{OneHop}(e_k) = \{(e_k, r, x) \in K_w \mid e_k \in E_j, r \in R, x \in E_j \cup L\}$$

$$K_{wv_i}^l = K_{v_i}^l \cup \bigcup_{e_k \in E_j} \text{OneHop}(e_k)$$

where E_j is the set of all entities in $K_{v_i}^l$.

We gather these one-hop relations through the Wikidata SPARQL API. As most entities are seen repeatedly across images in VG, we cache all $\text{OneHop}(e_k)$ relations locally to save time and API usage. Note that as we only expand one-hop localities around linked entities, we restrict graph isomorphisms to depths of one. We term these unions of KG and Image Scenegraps K_{wv}^l ‘Scene Knowledge Graphs’.

We implement Scene Knowledge Graphs with the Python NetworkX package (Hagberg et al. 2008), with entities as nodes and literals relations as edges. Both the

Logical Forms (Triples)	Question Template
<canvas, fabrication method, plain weave> <donuts, fabrication method, deep frying>	What in the image was made using {fact[2]} as its {fact[1]}?
<bottle, in front of, computer>, <bottle, subclass of, container> <stool, in front of, couch>, <stool, subclass of, seat>	What is located in front of a {fact[0][2]} and is a type of {fact[1][2]}?

Table 6.3: LLM generated f-string Question Templates. The first row is a ‘Unique Relation’ graph isomorphism and the second is a ‘Unique Intersection’.

values <`Earth`;`made of`;`42`> and the source <`Wikidata`;`VG relation`;`VG attribute`> are stored as feature attributes.

6.4.2 Question Engine

Our question engine uses two graph isomorphisms to sample logical forms for questions: *One Hop Unique* and *One Hop Intersection*.

One Hop Unique. This is a tuple which uniquely identifies one entity within the Scene Knowledge Graph. If a Scene Knowledge Graph has the <E, R, E> triple <car; has_part; combustion_engine> and no other entity E_o fulfills < E_o ;`has_part`;`combustion_engine`> then this is a valid question.

This is the simplest form of graph isomorphism found in reasoning datasets and is often the most popular (77.9% of GrailQA (Gu et al. 2021; Dutt et al. 2023), 60.2% of KVQA). We only include Wikidata logical forms for this setting.

One Hop Intersection. Questions pair two 3-tuples which together uniquely identify one entity within the Scene Knowledge Graph. If a Scene Knowledge Graph has the <E, R, E> triple <`train`;`powered by`;`electricity`> and <`train`;`invented in`;`19th century`> and no other entity E_o fulfills < E_o ;`has powered by`;`electricity`> AND < E_o , ‘invented in’, ‘19th century’>, but at least one other entity fulfils each individually, this is a valid logical form. These questions are comparatively rarer (3.8% of GrailQA, 44.2% of KVQA (Gu et al. 2021; Dutt et al. 2023)).

6.4.3 Image and Question Sampling

Diversity is crucial for a VQA dataset to be challenging. Image diversity ensures that models are evaluated on a range of different objects and scenes. Fact diversity ensures models are challenged to reason across a range of image types.

We set a desired number of questions Q equal to 10,000. We then sample one logical form from each image. For each isomorphism, we pick the form which contains the least recently seen relation. We fall back to a random pick if multiple relations are unseen.

6.4.4 Question Forming

As explained in Section 6.3.5, each logical template requires its own question template to map to a natural language question. As a scalable alternative to human template annotation, we experiment with LLMs.

Our task is of the form: *Given triple(s) t , return a f -string into which we can insert the relation and grounding (final) entity to make a question for which the answer is the first entity.*

We experiment with the LLama 2 70B (Touvron et al. 2023) and Phi 2 (Javaheripi and Bubeck 2023) LLMs for this task. We find that they are unable to map forms to templates: they often incorporate the answer into the question and lose the semantics of the logical form. We find it interesting that such a relatively simple task is not possible for Open LLM and leave this as a future research area.

We use GPT-4 to form questions. We find that prompting the model to first copy the answer entity E_a before forming the question from $\langle R, E_r/L \rangle$ reduces the rate of questions containing the answer. Examples of model output is given in Table 6.3. The prompt is provided in Appendix 6.9.2.

6.4.5 Question Filtering

Quality filtering has been explored in previous work in synthetic corpus generation. Gardent et al. (2017) use *human* annotation to check against the following three criteria: *Does the text sound fluent and natural?*, *Does the text contain all and only the information from the data?*, *Is the text good English (no spelling or grammatical mistakes)?* Agarwal et al. (2021) explore *automatic* filtering through fine-tuning BERT (Devlin et al. 2019a) on the WebNLG 2017 human assessment task of semantics and fluency (Gardent et al. 2017).

In a similar fashion, we design an LLM filtering step to remove bad questions. We instruct the model to remove questions which are ambiguous or nonsensical. We use the twenty most popular entities in the image as a semantic signal for the filtering operation. We provide the prompt and example input and output in Appendix 6.9.2.

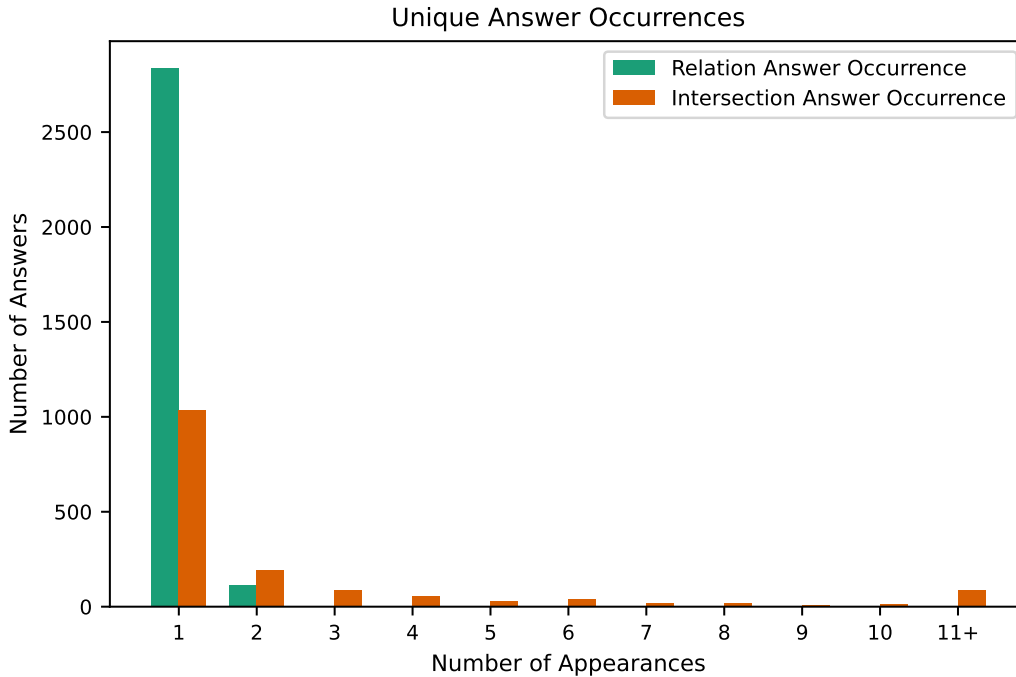


Figure 6.3: Answer Distribution.

Again, we find that GPT-4 is more accurate than other open LLMs.

We run this prompt over the generated questions, finding that 30.8% of Unique Relation questions and 47.0% of Unique Intersection questions are suitable. We give statistics on pre- and post-filtering SynthVQA in Table 6.5.

6.4.6 Negatives for Multiple Choices Setting

For unique relation questions, we pick the two negatives from the image sorted by PMI with the logical form. For unique intersection questions, we pick one other entity which fulfils the relation and one LLM negative. In both cases we call an LLM to sample the third negative. The LLM prompt is included in Appendix 6.10.

6.5 Dataset Statistics

There are 3.8M total and 68.8K unique entities in Visual Genome. We link 68.6K of these unique entities, which is 99.3%. At the Logical Form stage, we find 1.6M candidate triples for Unique Relations. This includes 78K unique answers and 491K

Linking Method	Total	Unique
CSKG	52.30%	39.13%
Wikidata	41.16%	39.38%
Wikipedia	6.48%	21.24%
None	0.05%	0.25%

Table 6.4: Linking statistics for SynthVQA. Linking is attempted in the order rows are shown.

unique relations when considered as (relation, entity/qualifier) 2-tuples. We provide statistics in Table 6.5. Full logical forms are $\langle R, E/L \rangle$ for Unique Relation and $\langle R_1, E/L_1, R_2, E/L_2 \rangle$ for Unique Intersection.

Filtering	Dataset Element	Unique Relation	Unique Intersection
Pre	Images	10,000	10,000
	LF/Questions	6,927	9,999
	Answers	9,238	2,899
	Logical Relations	585	683
	Logical Entities/Literals	6,231	3,955
Post	Images	3,089	4,703
	LF/Questions	2,171	1,581
	Answers	2,958	4,703
	Logical Relations	169	397
	Logical Entities/Literals	1971	2,326

Table 6.5: Dataset statistics pre- and post-filtering with an LLM. Filtering removes $\sim 70\%$ of samples and $\sim 50\%$ of logical relations, although Entities/Literals remain diverse.

Our sampling approach naturally samples from all types of relations in the source KG which match the structural pattern. This is made possible by the graph isomorphism sampler. Furthermore Table 6.5 shows that we have 196 relation types in Unique Relation and 397 in Unique Intersection.

We plot all relations in Figure 6.4. The figure visually represents the high degree of diversity in our dataset. A high diversity of relations is good because it increases the range of knowledge which models are evaluated and trained on. This compares favourably to other EKVQA datasets where the variety is restricted by fixed question templates. In the case of datasets with hand-crafted KG templates, the relations vocabulary is

restricted to the number of hand-authored templates. For human-generated questions, the reasoning type annotation is restricted by the difficulty of identifying the underlying facts.

6.5.1 Creation Cost

The cost of creating datasets determines how much data is going to be available for the community, and who is able to create it. Both traditional and EKVQA datasets use

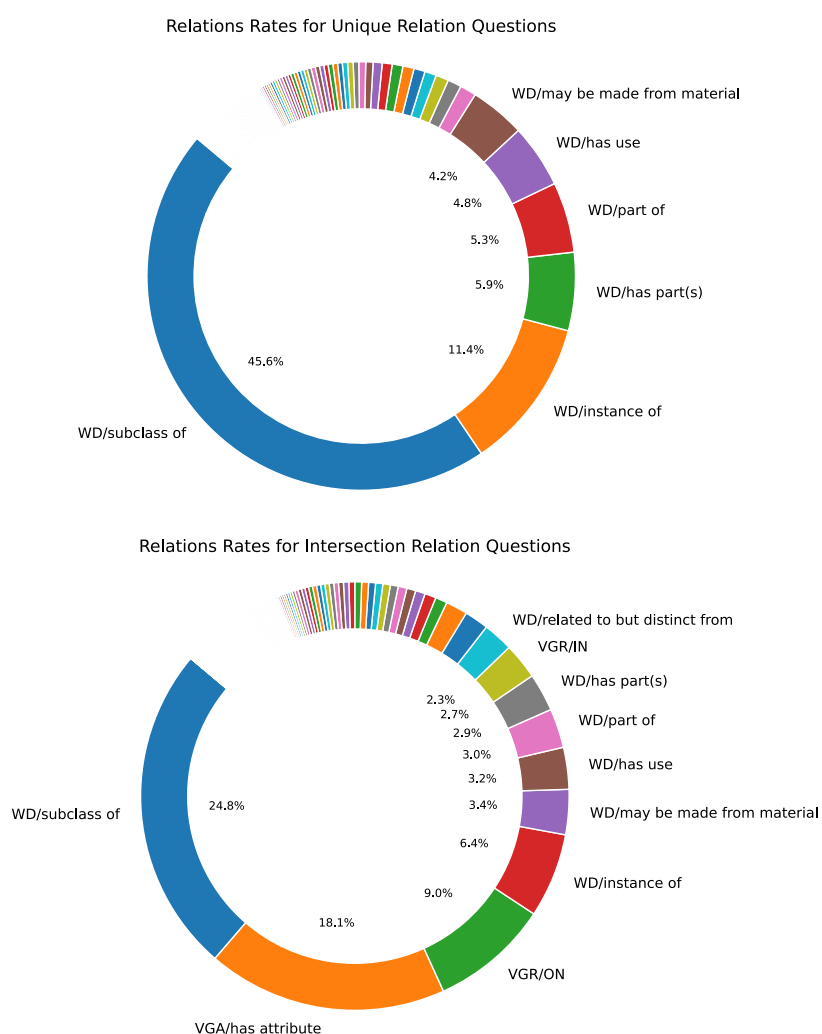


Figure 6.4: Top: Relation Distribution (Unique Rel). Bottom: Relation Distribution (Intersection). We give the source of the relation first. WD=Wikidata, VGR=Visual Genome Relation, VGA=Visual Genome Attribute.

Method	Relation	Intersection	All
Guess			
Random	0.246	0.255	0.251
Random Weighted	0.177	0.282	0.240
Most Common	0.180	0.325	0.267
Text			
T5	0.584	0.343	0.439
GPT-4	0.493	0.432	0.456
Image			
CLIP B-32	0.291	0.290	0.290
Text & Image			
BLIP-2-AOKVQA	0.522	0.489	0.502

Table 6.6: Results of text, image and multimodal models in SynthQA.

human annotations. Authors do not often release the pay or duration of annotation contracts. All of the experiments in this paper cost less than \$100 in GPT API costs. Our work enables researchers without large financial or annotation resources to create diverse VQA datasets. Through modifying the KG Linking and Querying code, any KG may be used to source facts. This work democratizes VQA dataset creation.

6.6 Results

We benchmark SynthVQA across different models and show the results in Table 6.6. We perform the evaluation with a similar set of models to AOKVQA (Schwenk et al. 2022).

We provide the results obtained with three random baselines (top-part **Guess**). First, we randomly sample one of the four answers with equal chance (“Random”). Second, we randomly sample one of the four answers with weighting proportional to the overall probability of appearance of that answer in the corpus (“Random Weighted”). Finally, we pick the answer candidate which appears the most frequently across the whole corpus (“Most Common”).

We also report the results obtained with text only QA systems. We use the T5 model Khashabi et al. (2020) and the GPT-4 Model. We provide the prompts in Appendix 6.11. For the image only setting, we use the CLIP B-32 model (Radford et al. 2021). We take the answer whose text encoding has the highest cosine similarity with the image encoding.

For the multimodal Question & Image setting, we use the BLIP-2 VQA model (Li et al. 2023). We test the AOKVQA fine-tuned variant. First, we observe that even a well performing model such as BLIP-2 struggles in providing high performance in this test dataset. Next, we note that the text only models remain very strong, in particular the T5 model marginally outperforms (+6.2%) BLIP-2-AOKVQA for ‘Relation’ questions. We speculate that the reason for this is that first, real-world priors which negate the importance of reasoning over the image (Goyal et al. 2019). This indicates that there is still some room to improve the question generation so that they heavily rely on the visual input. Second, VQA models may have insufficient factual knowledge to discriminate the correct answer from the two image distractors.

Next we analyse the kinds of questions that models are successful or unsuccessful on. BLIP2-model, and OpenCLIP H-14. Because we have access to the diverse underlying logical forms, we stratify results by accuracy for the most common relation types. These are presented in Table 6.7.

6.7 Results By Relation

Relation	Question Count	BLIP-2-AOKVQA (%)	CLIP B-32 (%)	T5 (%)
subclass of	2406	40.40	24.65	71.99
instance of	578	43.25	32.87	66.61
has part(s)	356	39.33	30.34	57.58
has use	343	39.94	27.41	68.22
part of	333	34.53	29.73	75.68
may be made from material	246	27.24	18.29	62.20
has quality	98	30.61	25.51	44.90
uses	92	40.22	31.52	61.96
used by	72	50.00	22.22	66.67
facet of	59	30.51	47.46	71.19

Table 6.7: Stratified Success Rate and Question Count by Relation

Intersection Relation Sources	Count	Accuracy
VG Relationship	2812	0.504
VG Attribute	1703	0.324
Wikidata	188	0.490

Table 6.8: Breakdown of relation sources (Visual Genome and Wikidata) and BLIP2-VQA accuracy on intersection questions.

Finally, we consider the effect of sampling Intersection questions from both K_w (Wikidata) and K_v^l (Visual Genome). To keep our dataset as EKVQA, we enforce that at least one relation always comes from K_w . In Table 6.8 we show the rates at which the second relation comes from either Wikidata, Visual Genome relations, or Visual Genome attributes. We find that BLIP-2-AOKVQA is worst (32.4% accuracy) over questions which require reasoning over both Visual Attributes and Wikidata Facts.

Error Analysis

A key advantage of SynthVQA is access to the underlying facts. We use this to perform an error analysis. For each model, we report the highest rate of incorrectly answered questions for every relation and every tail entity. We define error rate as (times seen and answer wrong/times seen). In the case of a tie, we break the tie by the overall *times seen and answer wrong*. We report the top errors for Relation Questions in Table 6.9 and for Intersection Questions in 6.10.

Model	Value	Relation			Value	Tail Entity		
		Error Rate	Error Count	Total Count		Error Rate	Error Count	Total Count
T5	has characteristic	1.0	12	12	text	1.0	12	12
	has effect	1.0	12	12	food ingredient	1.0	8	8
	original combination	1.0	8	8	steel	1.0	8	8
	name in kana	1.0	8	8	biomaterial	1.0	8	8
	significant event	1.0	8	8	vexillology	1.0	8	8
GPT-4	field of work	1.0	8	8	paper	1.0	20	20
	sex or gender	1.0	12	12	advertising	1.0	20	20
	course	1.0	8	8	shore	1.0	20	20
	contributing factor of	1.0	12	12	costume accessory	1.0	32	32
	followed by	1.0	12	12	trousers	1.0	24	24
CLIP B-32	part of	1.0	656	656	container	1.0	92	92
	has use	1.0	592	592	architectural element	1.0	84	84
	may be made from material	1.0	516	516	particular anatomical entity	1.0	72	72
	uses	1.0	152	152	product category	1.0	68	68
	parent taxon	1.0	108	108	class of anatomical entity	1.0	60	60
BLIP-2-AOKVQA	found in taxon	1.0	12	12	ungulate	1.0	16	16
	female form of label	1.0	24	24	infrastructure	1.0	20	20
	follows	1.0	20	20	human	1.0	20	20
	has characteristic	1.0	12	12	protection	1.0	20	20
	name	1.0	12	12	synthetic fiber	1.0	16	16

Table 6.9: The Relation and Tail Values with highest error rates for the Unique Relation SynthVQA Subsection

Interestingly, these results show that models fails completely for certain less frequent relations and tail entities. Crucially, these vary by model. This suggests that lack of access to all modalities makes answering certain question types impossible. Unsurprisingly, the Image only CLIP model has errors with much more common relations such as the relation *part of* (656 questions) or the tail entity container (92 questions).

This leads us to model the correlation of error rates across models. We record the error rate for all relations and all tail entities across both Relation and Intersection

Model	Type	Model			Tail			
		Error Rate	Error Count	Total Count	Error Rate	Error Count	Total Count	
T5	has characteristic	1.0	12	12	text	1.0	12	12
T5	has effect	1.0	12	12	food ingredient	1.0	8	8
T5	original combination	1.0	8	8	steel	1.0	8	8
T5	name in kana	1.0	8	8	biomaterial	1.0	8	8
T5	significant event	1.0	8	8	vexillology	1.0	8	8
GPT-4	field of work	1.0	8	8	paper	1.0	20	20
GPT-4	sex or gender	1.0	12	12	advertising	1.0	20	20
GPT-4	course	1.0	8	8	shore	1.0	20	20
GPT-4	contributing factor of	1.0	12	12	costume accessory	1.0	32	32
GPT-4	followed by	1.0	12	12	trousers	1.0	24	24
CLIP B-32	part of	1.0	656	656	container	1.0	92	92
CLIP B-32	has use	1.0	592	592	architectural element	1.0	84	84
CLIP B-32	may be made from material	1.0	516	516	particular anatomical entity	1.0	72	72
CLIP B-32	uses	1.0	152	152	product category	1.0	68	68
CLIP B-32	parent taxon	1.0	108	108	class of anatomical entity	1.0	60	60
BLIP-2-AOKVQA	found in taxon	1.0	12	12	ungulate	1.0	16	16
BLIP-2-AOKVQA	female form of label	1.0	24	24	infrastructure	1.0	20	20
BLIP-2-AOKVQA	follows	1.0	20	20	human	1.0	20	20
BLIP-2-AOKVQA	has characteristic	1.0	12	12	protection	1.0	20	20
BLIP-2-AOKVQA	name	1.0	12	12	synthetic fiber	1.0	16	16

Table 6.10: The Relation and Tail Values with highest error rates for the Unique Intersection SynthVQA Subsection

datasets and plot the correlation of error rates in Fig 6.5. These plots confirm that the modality is very important in determining what is answered correctly. The highest correlation (0.32) is between T5 and GPT-4, which are both text-only in our system. Meanwhile, the correlation between text-only and image-only models is 0.02 for both T5/CLIP and GPT-4/CLIP, suggesting that the difference in modalities contributes to the systematic differences in the errors of the systems. At the same time, the very low correlation in error rates between all of the systems suggests the potential performance gains of an ensemble model.

6.8 Conclusion

This paper presented a new methodology for creating EKVQA datasets with an automated framework we term **GRAVITY**. Our method is cheap and does not require human annotations for generating challenging datasets. We also release SynthVQA, a dataset created through **GRAVITY** applied on 10k images from Visual Genome obtained with a very limited budget. Our SynthVQA dataset is challenging, as demonstrated by the results obtained with several state-of-the-art systems showing that they struggle to reach high accuracy. Furthermore, questions retain their underlying logical forms, allowing for diagnostic analysis of VQA models.

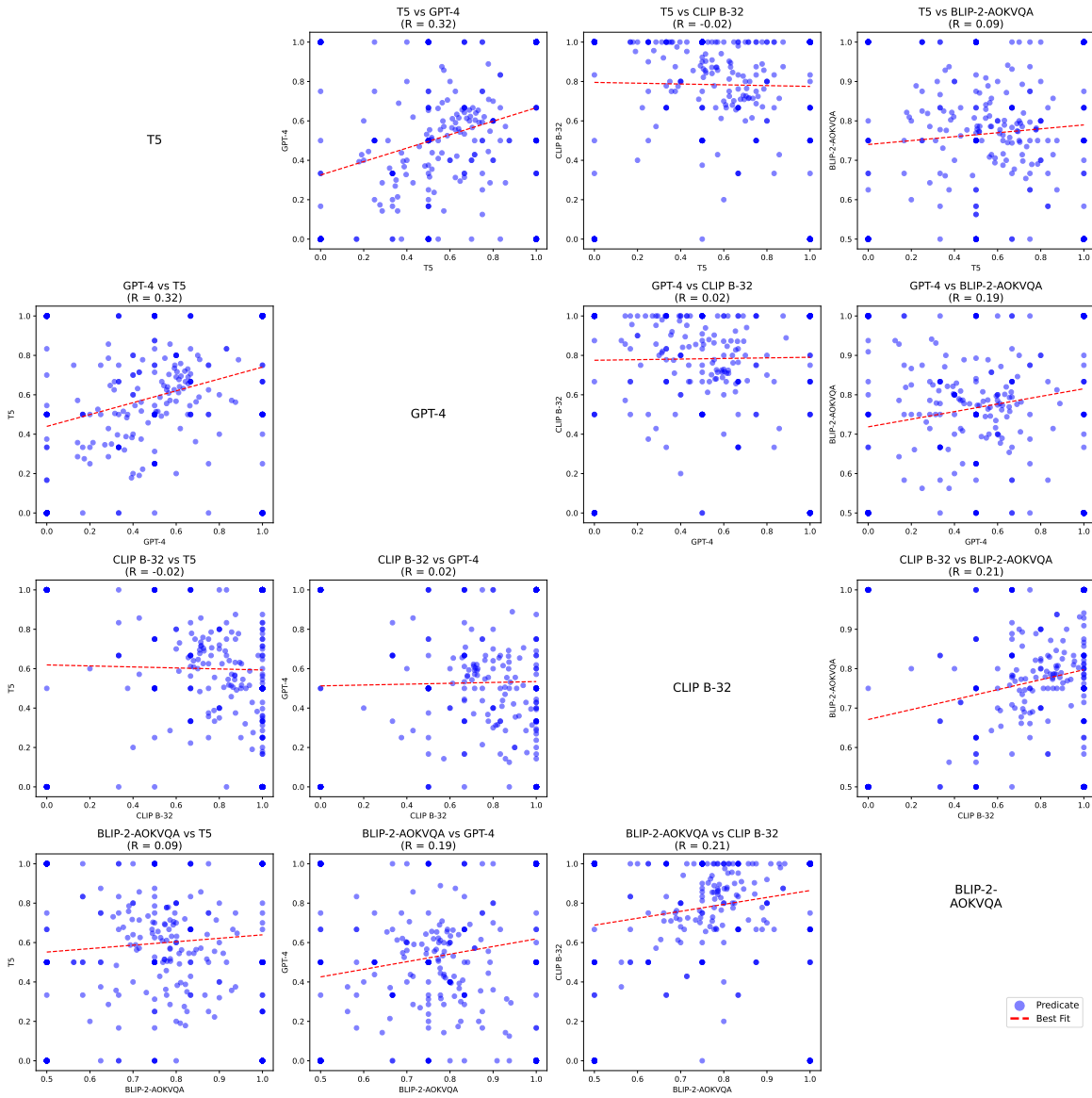


Figure 6.5: Correlation of error rates across models for relations and tail entities in SynthVQA. Each scatter plot compares the error rates of two models, with each point representing a relation or tail entity. The diagonal shows model names. R-values indicate the strength of correlation between model errors.

Limitations

Our framework relies on existing knowledge graphs (KGs) and scene graph annotations from Visual Genome. This reliance may limit the diversity and depth of knowledge represented in the SynthVQA dataset, as the Wikidata KG does not cover all knowledge

domains. Furthermore, the accuracy and completeness of scene graph annotations will impact the quality of generated questions and answers.

Our automated generation pipeline significantly reduces the need for human annotation, which is a major advantage in terms of scalability and cost. However, this approach may miss subtle nuances and complexities in visual scenes and questions that human annotators could capture. The absence of human validation in the question and answer generation process could lead to inaccuracies or unrealistic question-answer pairs.

Whilst we manually review all samples in SynthVQA, we cannot guarantee that all samples are bias free or unoffensive for all.

Ethics

Generation of questions from on KGs and Scenegraphs raises ethical considerations regarding the potential inclusion of harmful or sensitive topics within the data. Additionally, the automated nature of the process may inadvertently propagate biases present in the underlying data sources.

6.9 LLM Prompts

6.9.1 Question Phrasing

One Hop Unique

You are tasked with creating natural language templates from logical forms for a Visual

Question Answering (VQA) task. The forms are given as:

'ANSWER'=fact[0], 'RELATION'=fact[1], 'QUALIFIER'=fact[2].

Your role involves translating examples of logical forms into f-string templates that generate

English questions. These questions should:

Be fluent and natural-sounding.

Include all necessary information from the input without adding extraneous details.

Be grammatically correct and free of spelling errors.

It's permissible to substitute original relations with better words or phrases that preserve

the original meaning but enhance naturalness and clarity.

Must include the 'QUALIFIER'=fact[2], may include the 'RELATION'=fact[1] or rephrase it,

but NEVER include the 'ANSWER'=fact[0].

EXAMPLES:

INPUT:

'loose straw', 'by-product of', 'grain production'
'dung', 'by-product of', 'animal husbandry'

OUTPUT:

What here is a {fact[1]} of {fact[2]}?

INPUT:

'redshirt', 'inspired by', 'Star Trek: The Original Series'
'uncle sam', 'inspired by', 'Samuel Wilson'

OUTPUT:

What here was {fact[1]} {fact[2]}?

INPUT:

'redshirt', 'inspired by', 'Star Trek: The Original Series'

'uncle sam', 'inspired by', 'Samuel Wilson'

OUTPUT:

What here was {fact[1]} {fact[2]}?

INPUT:

'overcast', 'does not have quality', 'precipitation'

'pirate', 'does not have quality', 'credit'

'go cart', 'does not have quality', 'street legality'

OUTPUT:

What here can't be said to have {fact[2]}?

INPUT:

'newspaper' 'time of discovery or invention' '1605-01-01T00:00:00Z'

'jacket zipper' 'time of discovery or invention' '1893-01-01T00:00:00Z'

'bulb' 'time of discovery or invention' '1834-01-01T00:00:00Z'

'tubing' 'time of discovery or invention' '1904-01-01T00:00:00Z'

'tarmac' 'time of discovery or invention' '1902-01-01T00:00:00Z'

OUTPUT:

What in the image had a {fact[1]} of {fact[2]}?

END OF EXAMPLES

One Hop Intersection

You are tasked with creating natural language templates from logical forms for a Visual Question Answering (VQA) task. You will be given a number of examples which share relations.

The forms are given as:

'ANSWER'=fact[0][0], 'RELATION'=fact[0][1], 'QUALIFIER'=fact[0][2]. 'ANSWER
'=fact[1][0], 'RELATION'=fact[1][1], 'QUALIFIER'=fact[1][2].

Both of these facts are needed to uniquely identify the ANSWER.

Your role involves translating examples of logical forms into a single f-string template that generate English questions when applied to all the examples. These English questions should:

Be fluent and natural-sounding.

Include all necessary information from the input without adding extraneous details.

Be grammatically correct and free of spelling errors.

It's permissible to substitute original relations with better words or phrases that preserve the original meaning but enhance naturalness and clarity.

Must include the 'QUALIFIERS'=fact[0][2] and fact[1][2], may include the 'RELATIONS'=fact[0][1] and fact[1][1] or rephrase them, but NEVER include the 'ANSWER'=fact[0][0],fact[1][0].

EXAMPLES:

INPUT:

```
(('air', 'has use', 'lifting gas'), ('air', 'subclass of', 'mixture'))'
'(('stairs', 'has use', 'transport'), ('stairs', 'subclass of', 'thoroughfare'))'
'(('food', 'has use', 'eating'), ('food', 'subclass of', 'disposable product'))'
'(('cardboard', 'has use', 'mulch'), ('cardboard', 'subclass of', 'material'))'
'(('hair brush', 'has use', 'hairdressing'), ('hair brush', 'subclass of', 'personal hygiene item'))'
```

OUTPUT:

What has both a use for {fact[0][2]} and is a type of {fact[0][2]}?

INPUT:

```
('pointer finger', 'anatomical location', 'hand'),
('pointer finger', 'venous drainage', 'palmar digital veins'))
(('thumb', 'anatomical location', 'hand'), ('thumb', 'venous drainage', 'Dorsal venous network of hand'))
```

OUTPUT:

What has the {fact[0][1]} of the {fact[0][2]} and exhibits {fact[0][1]} into the {fact[0][2]}?

INPUT:

```
('apples', 'color', 'yellow'), ('apples', 'color', 'red'))
(('rainbow', 'color', 'green'), ('rainbow', 'color', 'blue'))
(('apples', 'color', 'yellow'), ('apples', 'color', 'green'))
(('bananas', 'color', 'yellow'), ('bananas', 'color', 'brown'))
(('apple', 'color', 'pink'), ('apple', 'color', 'green'))
```

OUTPUT:

What has both a {fact[0][2]} and {fact[1][2]} {fact[0][1]}?

INPUT:

```
(('ball', 'shape', 'sphere'), ('ball', 'subclass of', 'toy'))
((('sheet cake', 'shape', 'rectangular cuboid'), ('sheet cake', 'subclass of',
    'cake'))
(('napkin', 'shape', 'rectangle'), ('napkin', 'subclass of', 'linens'))
(('mug', 'shape', 'cylinder'), ('mug', 'subclass of', 'cup'))
((('globe', 'shape', 'sphere'), ('globe', 'subclass of', 'physical model'))
OUTPUT:
What has a {fact[0][2]} {fact[0][1]} and is a type of {fact[0][2]}?
END OF EXAMPLES
```

6.9.2 Question Filtering

You are an annotator of a Visual Question Answering dataset.
Your task is to review a single sample and review if the question is valid.

Reasons to reject:

- Entity is incorrectly linked
- Question is nonsensical
- Question is badly formatted
- Question is ambiguous

To assist you, we provide 20 ground truth entities that are in the image:

Note: The explanations are only for illustrative purposes.

Respond only 'Valid' or 'Invalid' responses for the actual input samples.

EXAMPLES:

Image Objects: window, tree, car, building, street light, walk sign,
backpack, man, road,
crosswalk, sidewalk, sign, sneakers, bike, walk, trees, pole, lights
Fact: (has use, track cycling), A: bike, Q: What is used in track cycling?
Response: Valid
Explanation: bikes are used for track cycling.

Image Objects: inbox tray, computer keyboard, paper, composition book, water
bottle,
surge protector, desk, computer speaker, computer monitor, sticky note,
office chair,

armrest, spots, cpu, telephone, wall, pen, mouse, cup

Fact: (related to but distinct from, scale degree), A: sticky note,

Q: What in the image is related to but distinct from a scale degree?

Response: Invalid

Explanation: Bad linking: 'sticky note' is not related to 'scale degree'

Image Objects: desk, picture, photo, pen, telephone, baby, wall, scissors,
book, keyboard, orange cloth, chair, pens, cup, monitor, mouse, pad,
computer, calendar, floor

Fact: (subclass of, container), A: tray, Q: What in the image is a subclass
of a container?

Response: Valid

Explanation: Trays are containers

Image Objects: ['flower' 'building' 'driveway' 'window' 'lamp post' 'roof' '
tree']

Fact: (located in the administrative territorial entity, Springfield), A:
entrance way,

Q: Where in the administrative territorial entity of Springfield is located
?\nResponse"}]

Response: Invalid

Explanation: bad linking: 'entrance way' is not specifically in Springfield

Image Objects: ['book' 'cord' 'shelf' 'stapler' 'top' 'base' 'ground' '
monitor']

Fact: (has part(s) of the class, element), A: support,

Q: What part(s) of the class element is present?\nResponse"}]

Response: Invalid

Explanation: bad question: nonsense

Image Objects: ['tree' 'leaves' 'car' 'bridge' 'sign' 'greenleaves' 'mirror' '
'highway']

Fact: (subclass of, road traffic control device), A: streetsign,

Q: What in the image is a subclass of road traffic control device?

Response: Valid

Explanation: streetsign the only road traffic control here

END OF EXAMPLES

6.10 LLM Negative Sampling

You are assisting in the creation of a VQA dataset.

You are to help provide a convincing distractor answer.

Please give a likely one or two word answer without a definite article.

INPUT:

True Answer: <CORRECT_ANSWER>

Distractors: <PMI_DISTRACTOR_1><PMI_DISTRACTOR_2>

Question: <LLM_TEMPLATED_QUESTION>

Other Distractors:

6.11 LLM Guess

This the template prompt used to query GPT4.0.

You are answering questions in a VQA dataset.

You are not provided the image.

Please pick the most likely answer given the question and the options.

Give the string of the answer, not its ordinal letter. E.g., Output 'Dog'
not '(b)'.

INPUT:

Question: <LLM_TEMPLATED_QUESTION>

Options: <CORRECT_ANSWER><PMI_DISTRACTOR_1><PMI_DISTRACTOR_2><LLM_DISTRACTOR
>

OUTPUT:

Chapter 7

Conclusions

This thesis presented work on improving and understanding the use of multiple Modalities for AI systems across three real-world tasks. To categorise Multimodal data we introduced the concept of **Knowledge Density**, the ratio of a modality’s entropy to useful information, in Section 1.3, where we claim that Images are knowledge-sparse, text is knowledge-rich, and KG, citations, and expert features are knowledge-dense. We then designed systems for three Multimodal Tasks: External Knowledge Visual Question Answering (images, text, KG), Eye Tracking Prediction (linguistic features, text), and Citation Prediction (citation graphs, text). In the case of EKVQA and Citation Prediction, our models outperformed prior baselines in the task-specific metrics. Feature analyses across all tasks were critical in evaluating the contribution of each modality. Faced with the limitations of existing datasets, we designed new datasets with a focus on underlying knowledge facts and reasoning types for EKVQA and for graph scale and forecasting for Citation Prediction. These datasets address significant challenges and diagnostic gaps in the field.

Furthermore, we have introduced, evaluated, and applied the Informedness metric for a more rigorous comparison of models on Classification tasks, notably EKVQA and Citation Prediction. This metric permits a fairer assessment of model performance across and within datasets and advances the evaluation of Multimodal AI models.

7.1 Summary

Publication I: In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering introduced a new method to enhance the reasoning capabilities

of a Vision-Language model for External Knowledge Visual Question Answering by integrating knowledge-dense facts extracted from a Knowledge Graph. When evaluated on the KVQA dataset, our method outperformed the previous baseline by 19%. We also performed an extensive analysis highlighting the limitations of our best-performing model through an ablation study.

Our REUNITER model demonstrated an overall absolute improvement of 19% over the previous State-of-the-Art model, with considerable gains in question types involving reasoning over multiple entities, and KG triples in the ‘Boolean’, ‘Comparison’, and ‘Multi-Hop’ categories, achieving accuracy of over 85% in all. However, it struggled with ‘Subtraction’ and ‘Spatial’ questions, which we attributed to BERT-like models’ limitations in numerical reasoning and spatial reasoning tasks, respectively. Further, we noted that whilst certain question types are inherently more complex, the unbalanced nature of the target answer classes complicated performance measuring.

The KVQA dataset aimed to minimize bias by using strict templates for question generation, yet it introduces an answer distribution bias due to the real-world priors associated with each template. This design makes it relatively easy for models to predict answers based on the question type, setting a baseline accuracy that models must exceed to demonstrate genuine understanding. We quantified this bias by reporting the random guess performance and answer distribution entropy per subtask. However, its unbalanced distribution of reasoning and answer types complicates the assessment of model weaknesses. We theorized that addressing these imbalances in future datasets and metrics could provide clearer insights into models’ true reasoning strengths and limitations.

Publication II: Blending Cognitively Inspired Features with Transformer-based Language Models for Predicting Eye Tracking Patterns evaluated the value of a variety of Knowledge-Dense cognitively- and linguistically-motivated features for predicting eye-tracking patterns over text. We considered these features as both standalone model inputs and supplements to contextual word embeddings from a finetuned version of the Auto-Regressive Masked-Denoising Language XLNET model. Contrary to Paper I, where the KG features were helpful, we found that only a limited subset of the most simple Linguistic features contributed to our best-performing model.

Publication III: We Need to Talk About Classification Evaluation Metrics in NLP addressed the downsides of applying simplistic classification metrics to our

prior research in Paper I. We started by outlining the issues with Accuracy, F-1 Macro, and Balanced Accuracy. We showed that the random-guess-normalised ‘Informedness’ metric describes useful model properties in its scoring. Informedness makes the scoring of datasets with very low answer-distribution entropy more intuitive to understand. Moreover, it enables comparison between question types with very different entropy in their answer distributions.

We re-assessed the results of REUNITER on KVQA from Paper I and found more convincing evidence that the model is strong due to above-zero Informedness scores across the board. We also make sub-task level reevaluations, such as the model having a moderate performance at ‘subtraction’ which accuracy reports that the model is poor at (39.8), due to Informedness being much higher (45.9). Meanwhile the Informedness score (56.3) for reasoning over ‘intersectional’ KG facts is much poorer than accuracy (79.5) reports. This suggests areas to focus on for model development in the future.

Publication IV: Comparing Edge-based and Node-based Methods on a Citation Prediction Task considered the third multimodal task of this thesis: Citation Prediction. We were motivated by the fact that Paper I found that knowledge-dense features are helpful for EKVQA, whilst Paper II found them unhelpful for Eye-tracking Prediction. We speculated that there exists a cross-over point where a sufficient scale of knowledge-rich features can outperform text. To test this hypothesis, we designed a new benchmark for Citation Prediction with a focus on graph scale. By training and evaluating a series of graph-embedding models on successively increasing time-segmented sub-graphs, we found a point at which citation graph-based features outperform text-based features for Citation Prediction at 82 million papers. Our largest model outperforms the text model by a considerable margin: 4.9%. Furthermore, the time-based stratification of our dataset allowed us to empirically demonstrate that long-term citation prediction is harder than short-term predictions. Finally, we found that the best policy for combining text and citation-based models depends on both the size of the train citation graph and the forecast horizon of interest. This means that *the optimal ensemble of multimodal features depends on the characteristics of both the train and test sets.*

Publication V: SynthVQA: Towards Flexible External Knowledge VQA Dataset Creation implemented an automated pipeline for generating VQA samples from ‘graph isomorphisms’. This pipeline is highly expressive due to being agnostic

to the underlying logical forms. Our approach therefore overcomes the limited topic diversity of previous methods. We create a diverse, difficult, and diagnostic dataset, SynthVQA, allowing us to analyse which facts, relations, and isomorphisms are hard for any VQA model. Our dataset shows that state-of-the-art VQA lack factual knowledge from Wikidata compared to text-only Question Answering models.

7.2 Research Questions

Here we discuss the research findings for our Research Questions outlined in Section 1.6.

RQ1: How can we incorporate modalities with Knowledge-rich (KG, Linguistic Annotations, Citations) and Knowledge-sparse (Images) to text models?

We investigated the integration of modalities with varying knowledge densities—images (knowledge-sparse), text (knowledge-rich), Knowledge Graphs (KG), citations, and expert features (knowledge-dense)—across three distinct tasks. Our findings suggest that the usefulness of adding knowledge-rich modalities to text models is dependent upon three factors: the presence of a knowledge gap, the detail provided by the new modality, and the model’s ability to utilize this information effectively.

Publication I: In Factuality demonstrated the effectiveness of KG features and our integration strategy for enhancing a Vision-Language model’s reasoning capabilities for External Knowledge Visual Question Answering, achieving a considerable improvement over the baseline. In this case, there was a knowledge gap due to questions being designed around external data, and we were able to locate the relevant information straightforwardly. However, our model’s performance varied across question types, indicating that the specific sub-task (reasoning over multiple facts, numerical operations, spatial reasoning) the impact varied across question types, indicating the specific nature of the information gap and its ability to interpret KG data.

Publication II: Blending Cognitively Inspired Features revealed that in the context of eye-tracking prediction, the integration of expert linguistic features did not provide improvement over a fine-tuned language model, suggesting that the added knowledge-rich features were not sufficient to fill a relevant knowledge gap or were redundant given the information already captured by the language model.

Publication IV: Comparing Edge-based and Node-based Methods found that for Citation Prediction, the efficacy of knowledge-rich citation features surpassed that of

text-based features once the citation graph used for training reached a certain scale. This highlighted the importance of the availability of lots of multimodal data for models to train on in determining the value of integrating additional modalities.

Ultimately, we found that adding knowledge-rich modalities is useful only when there is (a) a knowledge gap to be filled (b) sufficient detail in the new modality to assist the system, and (c) capability within the model to make use of this information.

RQ2: How can we create multimodal datasets which are diverse, difficult, and diagnostic?

In Paper V we release SynthVQA. Existing EKVQA datasets which challenge models to do multimodal reasoning with external support are not large. The largest in KVQA at 183K, but this relies on a restricted set of logical templates to produce formulaic questions with similar reasoning types. The second largest is A-OKVQA at 25K samples, which has higher diversity due to questions coming from human annotators without the restriction of logical templates to align to. We used the method of Graph Isomorphism sampling from Knowledge-Based Question Answering to generate maximally diverse EKVQA questions from a Knowledge Graph. This method ensures a broad range of reasoning types are represented, moving beyond the limitations of fixed templates and mimicking the variety of real-world questions. We defined graph patterns and then searched across linked Image/Knowledge Graphs for patterns which fitted these isomorphisms. We created a new approach for mapping sampled facts to Natural Language Questions with LLMs and used these to create diverse questions. Using LLMs allows for natural phrasing of the vast variety of questions we sample, removing the restriction on question types seen in prior EKVQA work. SynthVQA, with its emphasis on linked Image/Knowledge Graphs, presents unique challenges that test models' ability to integrate and reason with external knowledge, providing clear diagnostics on where models need improvement. Text-only Question Answering models perform extremely well on our system, even when adversarial distractors are introduced. This indicates that the state-of-the-art Visual Question Answering models are lacking in world knowledge.

In Paper IV we release the Citation Forecasting dataset . Citation Forecasting allows for the comparison of text and citation-based document representation models, as well as hybrid versions of them. Our work demonstrated that citation-based approaches outperform text-based when the knowledge-rich citation graph has enough scale.

RQ3: How can we improve the evaluation of classification models on datasets with diverse and imbalanced class distributions? Paper I presented

a per-question type breakdown based on the question annotations provided in Shah et al. (2019). However, these were hard to interpret as the answer-class distribution was not constant across question types. In Paper III, we underscored the limitations of conventional metrics like Accuracy, F-Measure, and AUC-ROC, particularly in the presence of class imbalance and varied class distributions. Through extensive experiments across a spectrum of NLP tasks, we made the case for the adoption of Informedness, a metric that evaluates the ability of models to make decisions better than random guessing, adjusting for class prevalence.

Our investigation spans several NLP domains, including EKVQA, Natural Language Understanding, and Machine Translation, revealing the impact of metric choice on reported model performance. In the GLUE benchmark, for instance, we demonstrate that while models may achieve high Accuracy, their Informedness scores reveal a more modest capability, particularly in tasks with considerable class imbalance. This discrepancy emphasizes the advantages Informedness offers in capturing a model’s genuine performance by discounting the advantage gained from label bias.

For VQA tasks, our exploration of GQA and KVQA again highlighted the pitfalls of relying on Accuracy and F-Measure in environments with inherent biases and real-world class distributions. Through controlled experiments, we show that Informedness provides a more stable and accurate reflection of a model’s capability across different question types and datasets, permitting a more nuanced understanding of model strengths and weaknesses.

Furthermore, our analysis extended to the formality control in Machine Translation, where we contrasted the performance of formality-aware and unaware MT systems. Again, Informedness eliminated the baseline credit granted by Accuracy and F1, offering a more straightforward view of a system’s true capacity to handle formality nuances in translation (0 for unaware systems).

In our rerun of the Citation Prediction experiments, we used Informedness to address the potential biases introduced by the Accuracy Paradox, particularly in light of a variable ratio of non-citing to citing pairs across our dataset. We removed the consistent ratio of non-citing to citing pairs in our time-segmented evaluation bins, and then adopted Informedness as our primary evaluation metric to gain clearer insights into our model’s performance.

The results reveal insights that were not apparent through the original accuracy-based results. For instance, in the early bins where citation occurrences are less frequent,

models appeared to perform better in terms of Accuracy due to the predominant class being non-citations. Specifically, a model trained solely on the earliest bin (0-0) achieved an Accuracy of 0.782 on its training bin but only an Informedness of 0.328, due to the inflated sense of performance that Accuracy can portray in such skewed datasets.

Informedness also played a diagnostic role for our model training when we discovered that a model trained on bins 0-30, which should theoretically perform well across time bins, showed minimal Informedness across all time-horizons, hovering around 0.000 to 0.016. This contrasted with its high Accuracy scores. This results highlighted how Informedness can unmask models that effectively predict no better than random chance.

Finally, our analysis indicates that the highest Informedness scores achieved across all models and bins was approximately 0.27, suggesting considerable room for improvement in Citation Prediction Forecasting, whilst the Accuracy scores of 0.70 might otherwise suggest nearing a performance ceiling.

Our findings advocate for the broader adoption of Informedness in evaluating citation prediction models, especially in the context of time-segmented data. By providing a more accurate reflection of a model’s predictive power and its ability to make informed decisions beyond mere prevalence bias, Informedness enables a more nuanced understanding of model performance and potential areas for enhancement.

7.3 Impact of this Thesis

The contributions of this thesis may be placed into two categories, (1) incremental work that contributes to the direction of the field, and (2) work which opens new research directions.

In class (1) are Papers I, II, and the dataset part of V. Paper I released the ReUNITER architecture for adding Knowledge Graph facts to the Vision-Language BERT architecture. This sits within both the research direction of expanding the modal input field of Transformer models and of Retrieval-Augmented Generation. Paper II explored the use of expert linguistic features along with fine-tuned Masked Denoising Language models. This is a type of Linguistic feature analysis which is useful for low-resource languages. Paper V released a Visual Question Answering dataset with External Knowledge which is challenging, cheap, and clearly annotated.

In class (2) are papers III, IV, and the pipeline part of V. The metric from Paper III can be used to draw greater insights into model performance on classification tasks

with low question-distribution entropy. The forecasting approach from paper IV opens a new direction in the Citation-prediction literature, and our dataset allows researchers to benchmark their systems against this criterion easily and fairly. Paper V’s **GRAVITY** pipeline opens a new approach for rapidly generating fact-grounded Visual Question Answering samples from Knowledge-Intensive groundings.

7.4 Future Directions

Enhancing Model Capabilities Paper I explored models that retrieve facts from Knowledge Graphs for use in Vision-Language tasks. Given advancements in such methods, such as Retrieval Augmented Generation in Lin and Byrne (2022), reconsideration with the latest Vision-Language models and Retrieval techniques is warranted. Additionally, refining question tagging by reasoning type and underlying fact in more datasets, as we developed in our SynthVQA dataset in Publication V, could further improve model diagnostics.

Paper II highlighted the potential of incorporating linguistic features into eye-tracking prediction. A pretrained English-language model outperformed expert features, but this may not be the case for low-resource languages. Future research could identify the threshold at which Knowledge-Rich features surpass Language Models across languages, as we did for Citation Prediction in Paper IV.

Expanding Dataset Utility Papers IV and V reported novel dataset creation, with IV introducing a citation-prediction dataset and V releasing the SynthVQA dataset for Knowledge-Enhanced Visual Question Answering. The next steps involve scaling these datasets in scale, diversity, and complexity. For SynthVQA, expanding to include the full range of facts within the Visual Genome/Wikidata linked collection could enrich the dataset’s complexity. Additionally, employing more intricate fact patterns, such as ‘two-hop unique’ fact chains, and integrating alternative or multiple Knowledge Graphs, like Commonsense ConceptNet, could diversify question types and enhance the dataset’s utility for testing AI models’ reasoning capabilities. For Citation Prediction, we seek to expand our approach to Local Citation Recommendation, where the task is to predict a cited paper given a citing sentence. The task will have an increased reliance on text due to the citing sentence being required to identify the cited paper, so this offers an intriguing opportunity to study the interactions of Citation Graphs and

Citing Language modalities across time.

Advancing Evaluation Metrics Paper III introduced the Informedness metric to provide a fairer analysis of classification performance, especially in tasks with low question-distribution entropy. Future research should delve into metrics like Normalised Information Transfer and explore methodologies for fair comparisons across diverse class distributions. Addressing the sensitivity of Informedness to uncommon classes and calibrating this metric to accommodate non-equal train-test distributions are critical areas for development. Furthermore, community acceptance of novel metrics is crucial to their adaptation. Future should present researchers with applications of different metrics and record their ability to understand the relative performance of models given a set of metrics under study.

Moreover, examining how to effectively average across tasks in multi-task benchmarks to accurately reflect a model’s overall performance remains an open question. This exploration could lead to more nuanced evaluation frameworks that better capture the complexities of model performance across tasks.

Bibliography

- Agarwal, Oshin, Heming Ge, Siamak Shakeri, and Rami Al-Rfou (2021). Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, pp. 3554–3565.
- Agrawal, Aishwarya, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi (2018). Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ammar, Waleed, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine Van Zuylen, and Oren Etzioni (2018). Construction of the Literature Graph in Semantic Scholar. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 3*, pp. 84–91.
- Anastasopoulos, Antonios, Loïc Barrault, Luisa Bentivogli, Marceley Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri,

- Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe (2022). Findings of the IWSLT 2022 Evaluation Campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Ed. by Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà. Dublin, Ireland (in-person and online): Association for Computational Linguistics, pp. 98–157.
- Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh (2015). VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2015 Inter, pp. 2425–2433. ISBN: 9781467383912.
- Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives (2007). DBpedia: A nucleus for a Web of open data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 4825 LNCS. Springer, Berlin, Heidelberg, pp. 722–735. ISBN: 3540762973.
- Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis Philippe Morency (2019). Multimodal Machine Learning: A Survey and Taxonomy.
- Barrett, Maria (2018). Improving natural language processing with human data: Eye tracking and other data sources reflecting cognitive text processing. PhD thesis.
- Barrett, Maria, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard (2018). Sequence Classification with Human Attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Ed. by Anna Korhonen and Ivan Titov. Brussels, Belgium: Association for Computational Linguistics, pp. 302–312.
- Barrett, Maria, Joachim Bingel, Frank Keller, and Anders Søgaard (2016). Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data. In *Proceedings of ACL 2016: Short Papers*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 579–584.
- Barsalou, Lawrence W (2008). Grounded Cognition. In *Annual Review of Psychology* 59, pp. 617–645. ISSN: 00664308.

- Beel, Joeran, Bela Gipp, Stefan Langer, and Corinna Breitinger (2015). Research-paper recommender systems: a literature survey. In *International Journal on Digital Libraries* 17, pp. 305–338.
- Beel, Joeran, Bela Gipp, Stefan Langer, and Corinna Breitinger (2016). Research-paper recommender systems: a literature survey. In *Int. J. Digit. Libr.* 17.4, pp. 305–338. ISSN: 1432-5012.
- Beltagy, Iz, Kyle Lo, and Arman Cohan (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620.
- Ben-David, Arie (2007). A lot of randomness is hiding in accuracy. In *Engineering Applications of Artificial Intelligence* 20.7, pp. 875–885. ISSN: 0952-1976.
- Ben-younes, Hedi, Rémi Cadene, Matthieu Cord, and Nicolas Thome (2017). MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision 2017-October*, pp. 2631–2639.
- Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans (2023). The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". In.
- Bethard, Steven and Dan Jurafsky (2010a). Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*.
- Bethard, Steven and Dan Jurafsky (2010b). Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10*. Toronto, ON, Canada: Association for Computing Machinery, pp. 609–618. ISBN: 9781450300995.
- Bhagavatula, Chandra, Sergey Feldman, Russell Power, and Waleed Ammar (2018a). Content-Based Citation Recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker,

- Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 238–251.
- Bhagavatula, Chandra, Sergey Feldman, Russell Power, and Waleed Ammar (2018b). Content-Based Citation Recommendation. In *ArXiv* abs/1802.08301.
- Blagec, Kathrin, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald (2022). A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*. Ed. by Tatiana Shavrina, Vladislav Mikhailov, Valentin Malykh, Ekaterina Artemova, Oleg Serikov, and Vitaly Protasov. Dublin, Ireland: Association for Computational Linguistics, pp. 52–63.
- Blom, E. and S. Unsworth (2010). *Experimental Methods in Language Acquisition Research*. Language learning and language teaching. John Benjamins Pub. Company. ISBN: 9789027219961.
- Bouraoui, Zied, Jose Camacho-Collados, and Steven Schockaert (2020). Inducing Relational Knowledge from BERT. In *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 7456–7463.
- Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (2015). A large annotated corpus for learning natural language inference. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (ACL), pp. 632–642. ISBN: 9781941643327.
- Brack, Arthur, Anett Hoppe, and Ralph Ewerth (2021). Citation Recommendation for Research Papers via Knowledge Graphs. arXiv: 2106.05633 [cs.DL].
- Breiman, Leo (2001). Random forests. In *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125.
- Brodersen, Kay H., Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann (2010). The balanced accuracy and its posterior distribution. In *Proceedings - International Conference on Pattern Recognition*, pp. 3121–3124. ISSN: 10514651.
- Broscheit, Samuel (2019). Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*. Association for Computational Linguistics, pp. 677–685. ISBN: 9781950737727.

- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 2020-December*. ISSN: 10495258.
- Brysbart, Marc, Amy Beth Warriner, and Victor Kuperman (2014). Concreteness ratings for 40 thousand generally known English word lemmas. In *Behavior Research Methods* 46.3, pp. 904–911. ISSN: 15543528.
- Cacciari, Cristina and Patrizia Tabossi (1988). The comprehension of idioms. In *Journal of Memory and Language* 27, pp. 668–683. ISSN: 00906905.
- Cai, Hongyun, Vincent W Zheng, and Kevin Chen-Chuan Chang (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. In *IEEE transactions on knowledge and data engineering* 30.9, pp. 1616–1637.
- Callan, Jamie, Mark Hoy, Changkuk Yoo, and Le Zhao (2009). Clueweb09 data set.
- Caragea, Cornelia, Adrian Silvescu, Prasenjit Mitra, and C. Lee Giles (2013). Can't see the forest for the trees? a citation recommendation system. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '13. Indianapolis, Indiana, USA: Association for Computing Machinery, pp. 111–114. ISBN: 9781450320771.
- Caragea, Cornelia, Jian Wu, Alina Maria Ciobanu, Kyle Williams, Juan Pablo Fernández Ramírez, Hung-Hsuan Chen, Zhaohui Wu, and Colin Giles (2014). CiteSeer x : A Scholarly Big Dataset. In *European Conference on Information Retrieval*.
- Carpenter, P. A. and M. A. Just (1983). What your eyes do while your mind is reading. In *Eye movements in reading: Perceptual and language processes*. Ed. by Keith Rayner. New York: Academic Press., pp. 275–307.
- Chakrabarty, Abhisek, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita (2020). Improving Low-Resource NMT through Relevance Based Linguistic Features Incorporation. In *Proceedings of the 28th International Conference on Computational*

- Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4263–4274.
- Chen, Xiangning, Cho-Jui Hsieh, and Boqing Gong (2021). When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. In *arXiv preprint arXiv:2106.01548*.
- Chen, Yen Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu (2020). UNITER: UNiversal Image-TExt Representation Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12375 LNCS, pp. 104–120. ISSN: 16113349.
- Chicco, Davide, Niklas Tötsch, and Giuseppe Jurman (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. In *BioData Mining* 14.1, p. 13. ISSN: 1756-0381.
- Chrupała, Grzegorz and Afra Alishahi (2019). Correlating Neural and Symbolic Representations of Language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2952–2962.
- Church, Kenneth Ward and Valia Kordoni (2022). Emerging Trends: SOTA-Chasing. In *Natural Language Engineering* 28.2, pp. 249–269. ISSN: 1351-3249.
- Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning (2019). What Does BERT Look at? An Analysis of BERT’s Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 276–286.
- Cohan, Arman, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld (2020). SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 2270–2282.
- Cohen, Jacob (1960). A Coefficient of Agreement for Nominal Scales. In *Educational and Psychological Measurement* 20.1, pp. 37–46.

- Cop, Uschi, Nicolas Dirix, Denis Drieghe, and Wouter Duyck (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. In *Behavior Research Methods* 49.2, pp. 602–615. ISSN: 15543528.
- Cordeiro, Silvio, Carlos Ramisch, and Aline Villavicencio (2016). mwetoolkit+sem: Integrating Word Embeddings in the mwetoolkit for Semantic MWE Processing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 1221–1225.
- Cordeiro, Silvio, Aline Villavicencio, Marco Idiart, and Carlos Ramisch (2019). Un-supervised compositionality prediction of nominal compounds. In *Computational Linguistics* 45.1, pp. 1–57. ISSN: 15309312.
- Courtney Napoles Matthew R. Gormley, Benjamin Van Durme (2012). Annotated English Gigaword. In *Linguistic Data Consortium*.
- Crawl, Common (2019). Common Crawl.
- Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova (2019a). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1, pp. 4171–4186. ISBN: 9781950737130.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019c). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

- Short Papers*). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.
- DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace (2020). ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4443–4458.
- Ding, David, Felix Hill, Adam Santoro, and Matt Botvinick (2020). Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures. In *CoRR*.
- Djokic, Vesna G, Jean Maillard, Luana Bulat, and Ekaterina Shutova (2020). Decoding Brain Activity Associated with Literal and Metaphoric Sentence Comprehension Using Distributional Semantic Models. In *Transactions of the Association for Computational Linguistics*.
- Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In.
- Dumais, S. and J. Nielsen (1992). Automating the assignment of submitted manuscripts to reviewers. In pp. 233–244.
- Dutt, Ritam, Sopan Khosla, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiah (2023). GrailQA++: A Challenging Zero-Shot Benchmark for Knowledge Base Question Answering. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi. Nusa Dua, Bali: Association for Computational Linguistics, pp. 897–909.
- Ehrlich, Susan F. and Keith Rayner (1981). Contextual effects on word perception and eye movements during reading. In *Journal of Verbal Learning and Verbal Behavior* 20.6, pp. 641–655. ISSN: 0022-5371.

- Elliott, Desmond, Stella Frank, and Eva Hasler (2015). Multilingual Image Description with Neural Sequence Models. In.
- Ericsson, Linus, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales (2022). Self-Supervised Representation Learning: Introduction, advances, and challenges. In *IEEE Signal Processing Magazine* 39.3, pp. 42–62. ISSN: 15580792.
- Färber, Michael and · Adam Jatowt (2020). Citation recommendation: approaches and datasets. In *International Journal on Digital Libraries* 21, pp. 375–405.
- Firth, J (1957). A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford.
- Gao, Haoyuan, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu (2015). Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question. In *Advances in Neural Information Processing Systems* 28.
- Garcez, Artur d’Avila and Luis C Lamb (2020). Neurosymbolic AI: The 3rd Wave.
- Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini (2017). Creating Training Corpora for NLG Micro-Planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 179–188.
- Geva, Mor, Ankit Gupta, and Jonathan Berant (2020). Injecting Numerical Reasoning Skills into Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 946–958.
- Gori, Marco and Augusto Pucci (2006). Research Paper Recommender Systems: A Random-Walk Based Approach. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI’06)*, pp. 778–781.
- Goyal, Yash, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh (2019). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *International Journal of Computer Vision* 127.4, pp. 398–414. ISSN: 15731405.
- Grosz, Barbara J. (2012). What question would turing pose today? In *AI Magazine* 33.4, pp. 73–81. ISSN: 07384602.

- Grover, Aditya and Jure Leskovec (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864.
- Gu, Yu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su (2021). Beyond I.I.D.: Three levels of generalization for question answering on knowledge bases. In *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, pp. 3477–3488.
- Guo, Yandong, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao (2016). MS-celeb-1M: A dataset and benchmark for large-scale face recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9907 LNCS, pp. 87–102. ISBN: 9783319464862.
- Gurari, Danna, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham (2018). VizWiz Grand Challenge: Answering Visual Questions From Blind People. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE Computer Society, pp. 3608–3617.
- Hagberg, Aric, Pieter Swart, and Daniel S Chult (2008). Exploring network structure, dynamics, and function using NetworkX. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
- Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008). Studying the History of Ideas Using Topic Models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Ed. by Mirella Lapata and Hwee Tou Ng. Honolulu, Hawaii: Association for Computational Linguistics, pp. 363–371.
- Hand, David and Peter Christen (2018). A note on using the F-measure for evaluating record linkage algorithms. In *Statistics and Computing* 28.3, pp. 539–547. ISSN: 15731375.
- Hand, David J (2009). Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve. In *Mach. Learn.* 77.1, pp. 103–123. ISSN: 0885-6125.
- Harnad, Stevan (1990). The symbol grounding problem. In *Machine Intelligence: Perspectives on the Computational Model*, pp. 89–100. ISBN: 9781136525049.

- Hasan, Sadid A, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P Lungren (2018). Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. In *Conference and Labs of the Evaluation Forum*.
- He, Jianguan and Chaomei Chen (2018). Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications. In *Frontiers in Research Metrics and Analytics* 3. ISSN: 2504-0537.
- He, Qi, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles (2010). context-aware Citation Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: ACM, pp. 421–430. ISBN: 978-1-60558-799-8.
- Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia D'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann (2021). Knowledge Graphs. In *ACM Computing Surveys (CSUR)* 54.4. ISSN: 15577341.
- Hollenstein, Nora, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (2021a). CMCL 2021 Shared Task on Eye-Tracking Prediction. In *Proceedings of the Workshop on Cognitive Modelling and Computational Linguistics*.
- Hollenstein, Nora, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (2021b). CMCL 2021 Shared Task on Eye-Tracking Prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Ed. by Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. Online: Association for Computational Linguistics, pp. 72–78.
- Hollenstein, Nora, Jonathan Rotsztein, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. In *Scientific Data* 5.
- Hollenstein, Nora, Marius Troendle, Ce Zhang, and Nicolas Langer (2020). ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. Tech. rep., pp. 11–16.

- Hollenstein, Nora, Eth Zurich, Maria Barrett, Marius Troendle, Francesco Bigioli, Nicolas Langer, and Ce Zhang (2019). Advancing NLP with Cognitive Language Processing Signals. In.
- Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020). spaCy: Industrial-strength Natural Language Processing in Python.
- Hu, Weihua, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec (2021). OGB-LSC: A large-scale challenge for machine learning on graphs. In *arXiv preprint arXiv:2103.09430*.
- Hu, Weihua, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec (2020). Open graph benchmark: Datasets for machine learning on graphs. In *Advances in neural information processing systems* 33, pp. 22118–22133.
- Hudson, Drew A and Christopher D Manning (2019a). GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2019-June, pp. 6693–6702. ISBN: 9781728132938.
- Hudson, Drew A and Christopher D Manning (2019b). Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems*. Vol. 32.
- Hulst, Johannes M. van, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries (2020). REL: An Entity Linker Standing on the Shoulders of Giants. In *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2197–2200.
- Ilievski, Filip, Pedro Szekely, and Bin Zhang (2021). CSKG: The CommonSense Knowledge Graph. In *Extended Semantic Web Conference (ESWC)*.
- Javaheripi, Mojan and Sébastien Bubeck (2023). Phi-2: The Surprising Power of Small Language Models. Accessed: 01/2024.
- Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah (2019). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3651–3657.

- Jiang, Zhuoren, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu (2018). Cross-language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph. In *CoRR* abs/1812.11709. arXiv: 1812.11709.
- Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick (2016). CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua, pp. 1988–1997.
- Kassner, Nora, Benno Krojer, and Hinrich Schütze (2020). Are Pretrained Language Models Symbolic Reasoners over Knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 552–564.
- Kembhavi, Aniruddha, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi (2017). Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5376–5384.
- Khashabi, Daniel, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi (2020). UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 1896–1907.
- Kinney, Rodney Michael, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld (2023). The Semantic Scholar Open Data Platform. In *ArXiv* abs/2301.10140.

- Krishna, Ranjay, Justin Johnson, Yannis Kalantidis Yahoo, David Ayman Shamma, Yuke Zhu, Oliver Groth, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei (2016). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations Human trajectory forecasting View project hybrid intrusion detection systems View project Visual Genome Connecting Language and Vision Using Crowdsourced Dense Image A. In *Article in International Journal of Computer Vision* 123.1, pp. 32–73.
- Kuhn, T S (1962). *The structure of scientific revolutions*. Chicago.
- Lafferty, John D, Andrew McCallum, and Fernando Pereira (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*.
- Lan, Yunshi and Jing Jiang (2020). Query Graph Generation for Answering Multi-hop Complex Questions from Knowledge Bases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 969–974.
- Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman (2016). Distribution-Free Predictive Inference For Regression. In *Journal of the American Statistical Association* 113.523, pp. 1094–1111.
- Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of Machine Learning Research* 202, pp. 20351–20383. ISSN: 26403498.
- Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang (2020). What Does BERT with Vision Look At? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5265–5275.
- Li, Mingchen and Shihao Ji (2022). Semantic Structure Based Query Graph Prediction for Question Answering over Knowledge Graph. In *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia

- Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 1569–1579.
- Liang, Yicong and Lap-Kei Lee (2023). A Systematic Review of Citation Recommendation Over the Past Two Decades. In *International Journal on Semantic Web and Information Systems* 19, pp. 1–22.
- Lin, Tsung Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8693 LNCS. PART 5. Springer Verlag, pp. 740–755.
- Lin, Weizhe and Bill Byrne (2022). Retrieval Augmented Visual Question Answering with Outside Knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 11238–11254.
- Loper, Edward and Steven Bird (2002). NLTK: The Natural Language Toolkit. In *CoRR*.
- Loshchilov, Ilya and Frank Hutter (2017). Fixing Weight Decay Regularization in Adam. In *CoRR* abs/1711.0.
- Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*. Ed. by H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett. Vol. 32. Curran Associates, Inc., pp. 13–23.
- Lu, Jiasen, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee (2020). 12-in-1: Multi-Task Vision and Language Representation Learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh (2016). Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Advances in Neural Information Processing Systems*. Ed. by D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett. Vol. 29. Curran Associates, Inc.

- Lu, Pan, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang (2018). R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. In *CoRR* abs/1805.09701. arXiv: 1805.09701.
- Malinowski, Mateusz and Mario Fritz (2014). A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems* 27.
- Manning, Christopher D and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.
- Marino, Kenneth, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi (2019). OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June, pp. 3190–3199.
- Maybury, Mark T. (2006). Expert Finding Systems. https://www.mitre.org/sites/default/files/pdf/06_1115.pdf.
- McNee, Sean M., Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl (2002). On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*. CSCW '02. New Orleans, Louisiana, USA: Association for Computing Machinery, pp. 116–125. ISBN: 1581135602.
- Mentch, Lucas and Giles Hooker (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. In *Journal of Machine Learning Research* 17.26, pp. 1–41.
- Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher (2016). Pointer Sentinel Mixture Models. In *Proceedings of ICLR 2017*. arXiv: 1609.07843.
- Merkx, Danny and Stefan L. Frank (2020). Comparing Transformers and RNNs on predicting human sentence processing data. arXiv: 2005.09471 [cs.CL].
- Metcalf, Bob (2013). Metcalfe’s law after 40 years of Ethernet. In *Computer* 46.12, pp. 26–31.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). Efficient Estimation of Word Representations in Vector Space. Tech. rep.

- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., pp. 3111–3119.
- Mimno, David and Andrew McCallum (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 500–509.
- Mishra, Abhijit, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya (2016). Harnessing Cognitive Features for Sarcasm Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 1095–1104.
- Nallapati, Ramesh, Amr Ahmed, Eric P. Xing, and William W. Cohen (2008). Joint latent topic models for text and citations. In *Knowledge Discovery and Data Mining*.
- Nan, Duan (2023). Frontier Review of Multimodal AI. English. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*. Ed. by Jiajun Zhang. Harbin, China: Chinese Information Processing Society of China, pp. 110–118.
- Nicodemus, Kristin K., James D. Malley, Carolin Strobl, and Andreas Ziegler (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. In *BMC Bioinformatics* 11.1, p. 110. ISSN: 14712105.
- Noordeh, Emil, Roman Levin, Ruochen Jiang, and Harris Shadmany (2020). Echo Chambers in Collaborative Filtering Based Recommendation Systems. In *CoRR* abs/2011.03890. arXiv: 2011.03890.
- Ordonez, Vicente, Girish Kulkarni, and Tamara Berg (2011). Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*. Ed. by J Shawe-Taylor, R Zemel, P Bartlett, F Pereira, and K Q Weinberger. Vol. 24. Curran Associates, Inc., pp. 1143–1151.
- Ostendorff, Malte, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm (2022). Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In *Proceedings of the 2022 Conference on Empirical*

- Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 11670–11688.
- Papert, Seymour A. (1966). The Summer Vision Project. In.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Petroni, Fabio, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel (2020). Language models as knowledge bases? In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, pp. 2463–2473. ISBN: 9781950737901.
- Pillai, Reshma S and LR Deepthi (2022). A Survey on Citation Recommendation System. In *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*. IEEE, pp. 423–429.
- Powers, David (2003). Recall and Precision versus the Bookmaker. In *Cognitive Science - COGSCI*, pp. 529–534.
- Powers, David M W (2012). The Problem with Kappa. In *European Chapter of the Association for Computational Linguistics* 13, pp. 345–355.
- Powers, David M W (2013). A Computationally and Cognitively Plausible Model of Supervised and Unsupervised Learning. In *Advances in Brain Inspired Cognitive Systems*, pp. 145–156. ISBN: 978-3-642-38786-9.
- Qian, Zhaozhi, Bogdan-Constantin Cebere, and Mihaela van der Schaar (2023). Synthcity: facilitating innovative use cases of synthetic data in different data modalities. In *arXiv:2301.07573*.
- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by Marina

- Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8748–8763.
- Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher (2019). Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 4932–4942.
- Ramisch, Carlos (2012). A generic and open framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of the ACL 2012 Student Research Workshop*. September. Association for Computational Linguistics, pp. 61–66. ISBN: 978-3-319-09206-5.
- Rayner, Keith (1975). The perceptual span and peripheral cues in reading. In *Cognitive Psychology* 7.1, pp. 65–81. ISSN: 00100285.
- Rayner, Keith (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. In *Psychological Bulletin* 124.3, pp. 372–422.
- Rayner, Keith and George W. McConkie (1976). What guides a reader’s eye movements? In *Vision Research* 16.8, pp. 829–837. ISSN: 00426989.
- Rehurek, Radim and Petr Sojka (2011). Gensim–python framework for vector space modelling. In *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2.
- Ren, Mengye, Ryan Kiros, and Richard S Zemel (2015). Exploring Models and Data for Image Question Answering. In *Advances in Neural Information Processing Systems* 28.
- Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6, pp. 1137–1149. ISSN: 01628828.
- Resnick, Paul and Hal R Varian (1997). Recommender systems. In *Communications of the ACM* 40.3, pp. 56–58.
- Richens, R H (1958). Interlingual Machine Translation. In *The Computer Journal* 1.3, pp. 144–147. ISSN: 0010-4620.

- Rohanian, Omid, Shiva Taslimipour, Victoria Yaneva, and Le An Ha (2017). Using Gaze Data to Predict Multiword Expressions. In *Proceedings of RANLP 2017*, pp. 601–609.
- Roy, Dwaipayan (2017). An improved test collection and baselines for bibliographic citation recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2271–2274.
- Sampat, Shailaja Keyur, Akshay Kumar, Yezhou Yang, and Chitta Baral (2021). CLEVR_HYP: A Challenge Dataset and Baselines for Visual Question Answering with Hypothetical Actions over Images. In *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 3692–3709.
- Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *ArXiv* abs/1910.0.
- Scarselli, F., M. Gori, A. Tsoi, M. Hagenbuchner, and G. Monfardini (2009). The Graph Neural Network Model. In *IEEE Transactions on Neural Networks* 20, pp. 61–80.
- Schneider, Phillip, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes (2022). A Decade of Knowledge Graphs in Natural Language Processing: A Survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang. Online only: Association for Computational Linguistics, pp. 601–614.
- Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 07-12-June, pp. 815–823. ISBN: 9781467369640.
- Schwenk, Dustin, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi (2022). A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. In.
- Shah, Sanket, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar (2019). KVQA: Knowledge-Aware Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 8876–8884. ISSN: 2159-5399.

- Sharma, Piyush, Nan Ding, Sebastian Goodman, and Radu Soricut (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 1. Association for Computational Linguistics (ACL), pp. 2556–2565. ISBN: 9781948087322.
- Singh, Amanpreet, Mike D’Arcy, Arman Cohan, Doug Downey, and Sergey Feldman (2023). SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 5548–5566.
- Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Paul Hsu, and Kuansan Wang (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*.
- Siyanova-Chanturia, Anna, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter J.B. van Heuven (2017). Representation and processing of multi-word expressions in the brain. In *Brain and Language* 175, pp. 111–122. ISSN: 10902155.
- Siyanova-Chanturia, Anna, Kathy Conklin, and Norbert Schmitt (2011). Adding more fuel to the fire: an eye-tracking study of idiom processing by native and non-native speakers. In *Second Language Research* 27.2, pp. 251–272.
- Sood, Ekta, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu (2020a). Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 12–25.
- Sood, Ekta, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu (2020b). Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 12–25.

- Sparck Jones, Karen (1994). Towards Better NLP System Evaluation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Speer, Robyn, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan (2018). LuminosoInsight/wordfreq: v2.2.
- Srihari, R K (1995). Automatic indexing and content-based retrieval of captioned images. In *Computer* 28.9, pp. 49–56.
- Steinert, Laura (2017). Beyond Similarity and Accuracy - A New Take on Automating Scientific Paper Recommendations. In pp. 1–161.
- Strohman, Trevor, W. Bruce Croft, and David D. Jensen (2007). Recommending citations for academic papers. In pp. 705–706.
- Su, Weijie, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai (2019). VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *arXiv*.
- Su, Yu, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan (2016). On Generating Characteristic-rich Question Sets for QA Evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, pp. 562–572.
- Suhr, Alane, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi (2020). A corpus for reasoning about natural language grounded in photographs. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 6418–6428. ISBN: 9781950737482.
- Tamborrino, Alexandre, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin (2020). Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 3878–3887.
- Tanenhaus, Michael K (2007). Spoken language comprehension: Insights from eye movements. In *The oxford handbook of psycholinguistics*.

- Tang, Jie, Jing Zhang, Limin Yao, Juan-Zi Li, Li Zhang, and Zhong Su (2008). Arnet-Miner: extraction and mining of academic social networks. In *Knowledge Discovery and Data Mining*.
- Tapaswi, Makarand, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler (2016). MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.
- The Economist (2023). Large, creative AI models will transform lives and labour markets.
- Thrush, Tristan, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross (2022). Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2022-June*, pp. 5228–5238. ISSN: 10636919.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael, Smith Ranjan, Subramanian Xiaoqing, Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey

- Edunov, and Thomas Scialom (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. In.
- Tu, Yuancheng, Nikhil Johri, Dan Roth, and Julia Hockenmaier (2010). Citation Author Topic Model in Expert Search. In *Coling 2010: Posters*. Ed. by Chu-Ren Huang and Dan Jurafsky. Beijing, China: Coling 2010 Organizing Committee, pp. 1265–1273.
- Turing, Alan M (1950). Computing Machinery and Intelligence. In *Mind* LIX.236, pp. 433–460. ISSN: 0026-4423.
- Valverde-Albacete, Francisco J. and Carmen Peláez-Moreno (2014). 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. In *PLOS ONE* 9.1, e84217. ISSN: 1932-6203.
- Valverde-Albacete, Francisco José, Jorge Carrillo-de-Albornoz, and Carmen Peláez-Moreno (2013). A Proposal for New Evaluation Metrics and Result Visualization Technique for Sentiment Analysis Tasks. eng. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. Vol. 8138. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 41–52. ISBN: 9783642408014.
- Van Heuven, Walter J B, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert (2014). SUBTLEX-UK: A new and improved word frequency database for British English. In *The Quarterly Journal of Experimental Psychology* 67.6, pp. 1176–1190.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.
- Vickers, Peter, Nikolaos Aletras, Emilio Monti, and Loïc Barrault (2021a). In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 468–475.

- Vickers, Peter, Nikolaos Aletras, Emilio Monti, and Loïc Loïc Barrault (2021b). In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering. In pp. 468–475.
- Vickers, Peter, Loïc Barrault, Emilio Monti, and Nikolaos Aletras (2023). We Need to Talk About Classification Evaluation Metrics in NLP.
- Vickers, Peter, Rosa Wainwright, Harish Tayyar Madabushi, and Aline Villavicencio (2021c). CogNLP-Sheffield at CMCL 2021 Shared Task: Blending Cognitively Inspired Features with Transformer-based Language Models for Predicting Eye Tracking Patterns. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Ed. by Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. Online: Association for Computational Linguistics, pp. 125–133.
- Vincent, Sebastian, Loïc Barrault, and Carolina Scarton (2022). Controlling Formality in Low-Resource NMT with Domain Adaptation and Re-Ranking: SLT-CDT-UoS at IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Ed. by Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà. Dublin, Ireland (in-person and online): Association for Computational Linguistics, pp. 341–350.
- Vinyals, O, A Toshev, S Bengio, and D Erhan (2015). Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 3156–3164.
- Vrandečić, Denny and Markus Krötzsch (2014). Wikidata: a free collaborative knowledgebase. In *Communications of the ACM* 57.10, pp. 78–85. ISSN: 0001-0782.
- Wade, Alex D (2022). The Semantic Scholar Academic Graph (S2AG). In *Companion Proceedings of the Web Conference 2022*.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355.

- Wang, Peng, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel (2017a). Explicit Knowledge-based Reasoning for Visual Question Answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, {IJCAI-17}*, pp. 1290–1296.
- Wang, Peng, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel (2017b). Explicit knowledge-based reasoning for visual question answering. In *IJCAI International Joint Conference on Artificial Intelligence*. Vol. 0. International Joint Conferences on Artificial Intelligence, pp. 1290–1296. ISBN: 9780999241103.
- Wang, Peng, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick (2016). FVQA: Fact-based Visual Question Answering. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.10, pp. 2413–2427.
- Weston, Jason, Sumit Chopra, and Antoine Bordes (2014). Memory Networks. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.
- Wilson, T.D. (1997). Information behaviour: An interdisciplinary perspective. In *Information Processing and Management* 33.4, pp. 551–572. ISSN: 0306-4573.
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, pp. 38–45.
- Wu, Jian, Athar Sefid, Allen C. Ge, and C. Lee Giles (2017). A Supervised Learning Approach To Entity Matching Between Scholarly Big Datasets. In *Proceedings of the 9th Knowledge Capture Conference*.
- Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu (2019). A Comprehensive Survey on Graph Neural Networks. In *IEEE Transactions on Neural Networks and Learning Systems* 32, pp. 4–24.
- Xie, Ning, Farley Lai, Derek Doran, and Asim Kadav (2019). Visual Entailment: A Novel Task for Fine-Grained Image Understanding. In.

- Yaneva, Victoria, Shiva Taslimipoor, Omid Rohanian, and Le An Ha (2017). Cognitive Processing of Multiword Expressions in Native and Non-native Speakers of English: Evidence from Gaze Data. Tech. rep.
- Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In.
- Yarowsky, David and Radu Florian (1999). Taking the load off the conference chairs—towards a digital paper-routing assistant. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.
- Yasunaga, Michihiro, Jure Leskovec, and Percy Liang (2022). LinkBERT: Pretraining Language Models with Document Links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 8003–8016.
- Yi, Kexin, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli Deepmind, and Joshua B Tenenbaum (n.d.). Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Tech. rep.
- Yimam-Seid, Dawit and Alfred Kobsa (2003). Expert-finding systems for organizations: Problem and domain analysis and the DEMOIR approach. In *Journal of Organizational Computing and Electronic Commerce* 13.1, pp. 1–24.
- Yoshitaka, Atsuo and Tadao Ichikawa (1999). A survey on content-based retrieval for multimedia databases. In *IEEE Transactions on Knowledge and Data Engineering* 11.1, pp. 81–93. ISSN: 10414347.
- Youden, W J (1950). Index for Rating Diagnostic Tests. In *Cancer* 3.1, pp. 32–35. ISSN: 10970142.
- Yu, Licheng, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg (2015). Visual Madlibs: Fill in the blank Image Generation and Question Answering. In.
- Yuhas, Ben P., Moise H. Goldstein, and Terrence J. Sejnowski (1989). Integration of acoustic and visual speech signals using neural networks. In *IEEE Communications Magazine* 27.11, pp. 65–71. ISSN: 01636804.

- Zadeh, Amir, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency (2016). MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. In *CoRR* abs/1606.06259.
- Zellers, Rowan, Yonatan Bisk, Ali Farhadi, and Yejin Choi (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2019-June, pp. 6713–6724. ISBN: 9781728132938.
- Zhang, Chi, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu (2019a). RAVEN: A Dataset for Relational and Analogical Visual Reasoning. In *CoRR* abs/1903.02741. arXiv: 1903.02741.
- Zhang, Chi, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu (n.d.). RAVEN: A Dataset for Relational and Analogical Visual Reasoning. Tech. rep.
- Zhang, Jie, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding (2019b). ProNE: Fast and Scalable Network Representation Learning. In *IJCAI*. Vol. 19, pp. 4278–4284.
- Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao (2016). Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. In *IEEE Signal Processing Letters* 23.10, pp. 1499–1503.
- Zhang, Yu, Yan-Jun Shen, Xiusi Chen, Bowen Jin, and Jiawei Han (2023). "Why Should I Review This Paper?" Unifying Semantic, Topic, and Citation Factors for Paper-Reviewer Matching. In *ArXiv* abs/2310.14483.
- Zhou, Jie, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun (2018). Graph Neural Networks: A Review of Methods and Applications. In *ArXiv* abs/1812.08434.
- Zhou, Yutong and Nobutaka Shimada (2023). Vision + Language Applications: A Survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 826–842.
- Zhu, Y, R Kiros, R Zemel, R Salakhutdinov, R Urtasun, A Torralba, and S Fidler (2015). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27.

- Zhu, Yuke, Oliver Groth, Michael Bernstein, and Li Fei-Fei (2016). Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December*, pp. 4995–5004. ISSN: 10636919.
- Zhuoren, Jiang, Z. Xiaozhong Liu, and Liangcai Gao (2014). Dynamic topic/citation influence modeling for chronological citation recommendation. In *International Conference on Information and Knowledge Management, Proceedings 2014-November (November)*, pp. 15–18.