

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/145157>

Copyright and reuse:

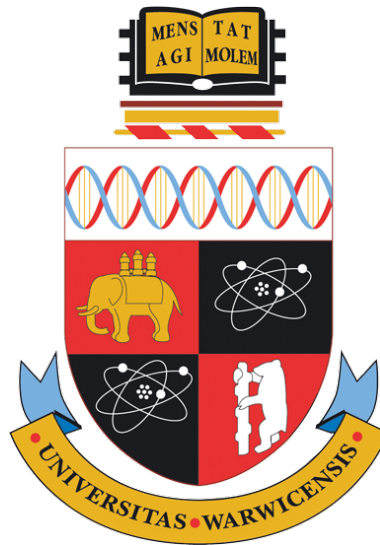
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk



Neural Models for Stepwise Text Illustration

by

Vishwash Batra

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Computer Science

June 2020

Contents

List of Tables	v
List of Figures	vii
Acknowledgments	ix
Declarations	x
Abstract	xii
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Relevant prior work	4
1.3 Research Questions and Objectives	5
1.4 Technical Challenges	7
1.5 Contributions	7
1.5.1 Neural Caption Generation for News Images	7
1.5.2 GutenStories and Stepwise Recipe datasets	7
1.5.3 Variational Recurrent Sequence to Sequence Retrieval for Stepwise Illustration	7
1.5.4 Context-Dependent Text Illustration and Description Retrieval	8
1.6 Thesis Outline	8
Chapter 2 Background	10
2.1 Prerequisites in Natural Language Processing	10
2.1.1 Representation	11
2.1.2 N-gram	11
2.1.3 TF-IDF	12
2.2 Deep neural-networks based architectures	12
2.2.1 Convolutional Neural Networks (CNN)	12
2.2.2 Recurrent Neural Networks (RNN)	12
2.2.3 Encoder-Decoder Architecture	13

2.2.4	Neural language methods with attention	15
2.3	Multimodal representation learning	15
2.3.1	Image-Text correspondence	16
2.3.2	Hinge loss	18
2.3.3	Order-embeddings	18
2.4	Missing modality inference	19
2.4.1	Image Captioning	19
2.4.2	Two-stage architecture	19
2.4.3	Image Synthesis	22
2.5	Cross-Modal Retrieval	23
2.5.1	Atomic Cross-Modal Retrieval	23
2.5.2	Sequential Cross-Modal Retrieval	23
2.5.3	Common Embedding Space Learning	26
2.6	Other related multimodal tasks	27
2.6.1	Image-sentence Ranking	28
2.6.2	Visual Question-Answering	28
2.6.3	Multimodal Summarisation	28
2.7	Critical Review	30
Chapter 3 Datasets and Evaluation		31
3.1	Cooking Recipe Datasets	31
3.2	BBC News Dataset	32
3.3	GutenStories	34
3.3.1	Construction	34
3.3.2	mini-GutenStories	37
3.4	StepwiseRecipeDataset	38
3.5	Evaluation Measures	40
3.5.1	Precision	40
3.5.2	Recall	41
3.5.3	Recall@K or R@K	41
3.5.4	Story Recall@K or StR@K	41
3.5.5	Visual Saliency Recall	41
3.5.6	Textual Saliency Recall	42
3.5.7	Inception Score	42
3.5.8	BLEU	42
Chapter 4 Neural Caption Generation for News Images		43
4.1	Introduction	43
4.2	Problem Formulation	46
4.2.1	Dataset	46
4.2.2	Existing Methods	47

4.3	Proposed Methodology	47
4.3.1	Text and Image Representation	48
4.3.2	LSTM Training	49
4.3.3	Variant Architecture	52
4.4	Experiments	52
4.4.1	Methods	54
4.4.2	Evaluation Metrics	55
4.4.3	Results	55
4.5	Error Analysis	57
4.6	Conclusion	59

Chapter 5 Variational Recurrent Sequence-to-Sequence Retrieval for Stepwise Illustration 61

5.1	Introduction	61
5.2	Variational Recurrent Seq2seq (VRSS) Retrieval Model	64
5.2.1	Problem Formulation	64
5.2.2	Text Encoder	67
5.2.3	Image Encoder	67
5.2.4	Incorporating Context	67
5.2.5	Latent Topic Modeling	68
5.2.6	Image Retrieval	68
5.2.7	Overall Objective Function	69
5.2.8	Parameter Configuration	69
5.3	Experimental Setup	71
5.3.1	Models for Comparison	71
5.3.2	Evaluation methods	73
5.4	Results and Discussion	75
5.4.1	Automatic Evaluation	75
5.4.2	Human Evaluation	77
5.5	Error Analysis	78
5.6	Embedding Analysis	79
5.7	Conclusion	83

Chapter 6 Context-Dependent Text Illustration and Description Retrieval 84

6.1	Introduction	84
6.2	Challenges	86
6.3	Problem Formulation	87
6.4	Methods	89
6.4.1	LDA-based	90
6.4.2	Joint embedding learning	90

6.4.3	Event Representation Embeddings	93
6.4.4	Stacked Cross Attention Network (SCAN)	93
6.4.5	Context-based Models	94
6.5	Experimental Setup – mini-GutenStories	94
6.5.1	mixed-LDA training	95
6.5.2	Results and Discussion	95
6.6	Experimental Setup - GutenStories	96
6.6.1	Evaluation Measures	98
6.6.2	Results and Discussion	98
6.7	Error Analysis	99
6.8	Conclusion	103
Chapter 7 Conclusions and Future Work		104
7.1	Main findings	104
7.1.1	Stepwise Illustration	104
7.1.2	News Image Caption Generation	105
7.2	Directions for future research	106
7.2.1	Stepwise Illustration	106
7.2.2	News Caption Generation	107
7.3	Summary	107
Appendix A Further Analysis and Results		109
A.1	Effects of introducing the dropout layer	109
A.2	Effects of varying the CNN architecture	109

List of Tables

2.1	Relevant Literature as per the Objectives	30
3.1	BBC News dataset statistics.	33
3.2	GutenStories dataset statistics.	38
3.3	mini-GutenStories dataset statistics.	38
3.4	(Primary) Stepwise Recipe dataset statistics.	39
3.5	(Augmented) Stepwise Recipe dataset statistics.	39
4.1	News image caption generation results in terms of BLEU scores.	55
4.2	News image caption generation results in terms of Meteor scores.	56
4.3	Comparison of BLEU scores over different image features	56
4.4	Comparison of Meteor scores over different image features	56
4.5	Human Evaluation results.	57
5.1	Text illustration performance using <i>Visual Saliency Recall@k</i> (<i>VSR@k</i>) on the Stepwise Recipe dataset. The best result in each column is highlighted in bold	75
5.2	Text illustration performance using <i>Recall@k</i> (<i>R@k</i>) and <i>Story Recall@k</i> (<i>StR@k</i>) on the Stepwise Recipe dataset. The best result in each column is highlighted in bold	75
5.3	Text illustration performance using <i>Textual Saliency Recall@k</i> (<i>TSR@k</i>) on the Stepwise Recipe dataset. The best result in each column is highlighted in bold	76
5.4	Human Evaluation results. The cell values indicate the number of images output by the corresponding model(s) that receive x number of votes ($x \in \{2, 3, 4, 5\}$) as majority.	77
5.5	Hubness analysis results. Text points as hubs in neighbourhoods of image points from Stepwise Recipe embeddings obtained using VRSS.	81
5.6	Hubness analysis results. Image points as hubs in neighbourhoods of text points from Stepwise Recipe embeddings obtained using VRSS.	81

5.7	Hubness analysis results. Text points as hubs in neighbourhoods of image points from Stepwise Recipe embeddings obtained using VSE++(R).	81
5.8	Hubness analysis results. Image points as hubs in neighbourhoods of text points from Stepwise Recipe embeddings obtained using VSE++(R).	82
6.1	Text illustration and Description Retrieval performance using <i>Recall@k</i> ($R@k$) and <i>Story Recall@k</i> ($StR@k$) and <i>Visual Saliency Recall@k</i> ($VSR@k$) on the mini-Gutenstories dataset. The best result in each column is highlighted in bold	97
6.2	Story Retrieval performance using <i>Story Recall@k</i> ($StR@k$) on the mini-Gutenstories dataset. The best result in each column is highlighted in bold	97
6.3	Text Illustration performance using <i>Visual Saliency Recall@k</i> ($VSR@k$) on the Gutenstories dataset. The best result in each column is highlighted in bold	99
6.4	Text Illustration performance using <i>Textual Saliency Recall@k</i> ($VSR@k$) on the Gutenstories dataset. The best result in each column is highlighted in bold	99

List of Figures

1.1	An instance from NY Times Journal of a news storyline	2
1.2	An instance from an online digital story 'The Tale of Peter Rabbit' authored by 'Beatrix Potter'	3
2.1	Encoder Decoder Architecture	13
2.2	Neural Image Captioning Architecture	14
2.3	Semantic Common Embedding Space Learning	17
2.4	Two Stage Architecture	20
2.5	Encoder Decoder Architecture	24
2.6	Encoder Decoder Architecture	25
2.7	Encoder Decoder Architecture	28
3.1	Example news articles and their accompanying images and image captions.	35
3.2	Web Crawler for automatic construction of the dataset.	36
4.1	Our proposed deep Neural Network (NN) architecture for news image caption generation.	47
4.2	Text and Image Representation Mechanism	48
4.3	A Deep NN Dual architecture for news image caption generation.	53
4.4	Sample Generated Captions providing comparison between LDA and NN methodology results	58
4.5	Error Analysis	60
5.1	Stepwise Recipe illustration example showing a few text recipe instruction steps alongside one full sequence of recipe images. Note that retrieval of an accurate illustration of Step 4, for example, depends on the model being able to use context information that was acquired in earlier steps.	62
5.2	Latent Variables	65
5.3	Variational Recurrent Sequence-to-Sequence (VRSS) model architecture.	66
5.4	Example images retrieved by the VRSS model.	78

5.5	Illustrative comparison of non-context (VSE++) and context models (VRSS) - VRSS result preferred by human evaluators.	79
5.6	Illustrative comparison of non-context and context models - VSE++(R) result preferred by human evaluators.	80
5.7	Illustrative comparison of non-context and context models - Neither VRSS nor VSE++(R) result preferred by human evaluators.	80
5.8	t-SNE plot for VRSS embeddings	82
5.9	t-SNE plot for VSE++(R) embeddings	83
6.1	Example automatic text illustration	85
6.2	Example text passage-image pairs in our dataset.	88
6.3	Deep NN Architecture	91
6.4	Illustrative comparison of correct and retrieved output of the Deep NN model	101
6.5	Illustrative comparison of correct and retrieved output of the Deep NN model	102
A.1	Visualisation of Top-1 retrieval scores of images and texts, over training iterations by varying dropout levels	110
A.2	Visualisation of Top-5 retrieval scores of images and texts, over training iterations by varying dropout levels	110
A.3	Visualisation of Top-10 retrieval scores of images and texts, over training iterations by varying dropout levels	112
A.4	Visualisation of Top-1, Top-5 and Top-10 retrieval scores of texts, over training iterations by varying image representation mechanisms	113
A.5	Visualisation of Top-1, Top-5 and Top-10 retrieval scores of images, over training iterations by varying image representation mechanisms	114

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Yulan He, for her valuable guidance and continuous support throughout my PhD. Yulan's deep insight to science and exceptional patience has encouraged me to attempt various interesting topics and her skillful supervision has helped me shape this thesis in the right direction. It has been a real honour and delightful experience to work with her during the past couple of years.

Many thanks to my second supervisor during my time at Aston University, George Vogiatzis, for his helpful and valuable advice. I really benefit from the thoughtful discussions with George, especially topics about deep learning and computer vision. I would also like to thank Hakan Ferhastosmanoglu and Tanaya Guha. I also want to thank my office mates and friends: Gabriele, Shengooshabad, Vasan, Shuang and Bowen. Thank you for making our office such a comfortable and easy-going place.

Finally, thank you to my family, especially my dear parents for whom I should do more. Your love, constant support and encouragement makes me what I am today.

I'd like to dedicate this thesis to my spiritual guide -
Her Holiness Satguru Mata Sudiksha Ji Maharaj.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Parts of this thesis have been published as full papers by the author in the following:

- **Vishwash Batra**, Yulan He, and George Vogiatzis. Neural caption generation for news images. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association(ELRA). URL <https://www.aclweb.org/anthology/L18-1273>
- **Vishwash Batra**, Aparajita Haldar, Yulan He, Hakan Ferhatosmanoglu, George Vogiatzis, and Tanaya Guha. Variational recurrent sequence-to-sequence retrieval for stepwise illustration. In Joemon M. Jose, Emine Yilmaz, Joao Magalhães, Pablo Castells, Nicola Ferro, M ario J. Silva, and Flavio Martins, editors, *Advances in Information Retrieval*, pages 50–64, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45439-5
- **Vishwash Batra** and Yulan He. Context-dependent text illustration and description retrieval. In preparation

Research was performed in collaboration during the development of this thesis, but does not form part of the thesis in the following:

- Shufeng Xiong, Ming Cheng, **Vishwash Batra**, Tao Qian, Bingkun Wang, and Ye Yangdong. Aspect terms grouping via fusing concepts and context information. *Information Fusion Journal*, 2020

Abstract

In this thesis, we investigate the task of sequence-to-sequence (seq2seq) retrieval: given a sequence (of text passages) as the query, retrieve a sequence (of images) that best describes and aligns with the query. This is a step beyond the traditional cross-modal retrieval which treats each image-text pair independently and ignores broader context. Since this is a difficult task, we break it into steps.

We start with caption generation for images in news articles. Different from traditional image captioning task where a text description is generated given an image, here, a caption is generated conditional on both image and the news articles where it appears. We propose a novel neural-networks based methodology to take into account both news article content and image semantics to generate a caption best describing the image and its surrounding text context. Our results outperform existing approaches to image captioning generation.

We then introduce two new novel datasets, GutenStories and Stepwise Recipe datasets for the task of story picturing and sequential text illustration. GutenStories consists of around 90k text paragraphs, each accompanied with an image, aligned in around 18k visual stories. It consists of a wide variety of images and story content styles. StepwiseRecipe is a similar dataset having sequenced image-text pairs, but having only domain-constrained images, namely food-related. It consists of 67k text paragraphs (cooking instructions), each accompanied by an image describing the step, aligned in 10k recipes. Both datasets are web-scrawled and systematically filtered and cleaned.

We propose a novel variational recurrent seq2seq (VRSS) retrieval model.

The model encodes two streams of information at every step: the contextual information from both text and images retrieved in previous steps, and the semantic meaning of the current input (text) as a latent vector. These together guide the retrieval of a relevant image from the repository to match the semantics of the given text. The model has been evaluated on both the Stepwise Recipe and GutenStories datasets. The results on several automatic evaluation measures show that our model outperforms several competitive and relevant baselines. We also qualitatively analyse the model both using human evaluation and by visualizing the representation space to judge the semantical meaningfulness. We further discuss the challenges faced on the more difficult GutenStories and outline possible solutions.

Chapter 1

Introduction

1.1 Motivation

Storytelling is central to human existence. Since time immemorial, humans have used narration as means for fostering ideas. Recent research, as in Botvin and Sutton-Smith [16], McKeough and Malcolm [96], Sun and Nippold [129], shows that storytelling has also been used as evaluation of development of language skills in children and adolescents. Therefore, it is considered to be an important task in machine learning methods for natural language processing. It is beyond simple perceptual tasks, like recognition and understanding of simple objects and concrete scenes, rather it requires a higher form of cognition. It requires understanding and interpretation of the underlying causal structure in narration. For machines, to be able to acquire storytelling, requires moving beyond static information. It requires to create an artificial intelligence (AI) that also needs to incorporate and model contextual information [54]. Furthermore, research in developmental psycholinguistics suggest the importance of visual context with textual narration in child language acquisition [106]. Therefore, the task of *Story Picturing*, or *Automatic Text Illustration* plays a huge role in automatic storytelling systems involving multiple modalities of data, for example images and texts [61].

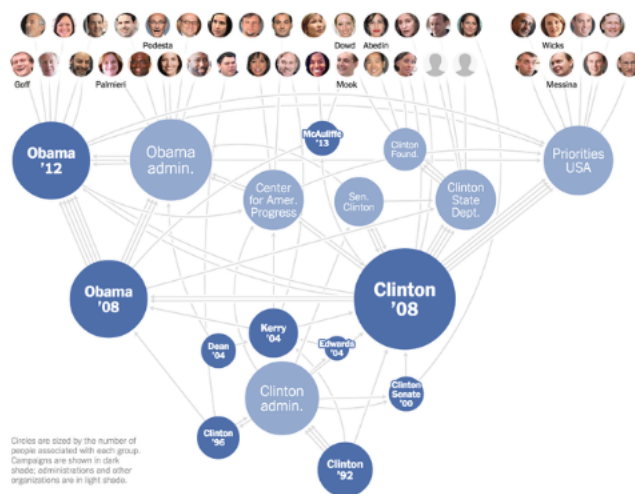
Also, due to creation of a huge amount of multimedia data on the Internet, present usually in the form of multimedia files containing images, videos and natural language texts. Such multimodal content and semantically related data is also often thematically collocated. The content is usually manifested as news articles, social media posts, personal or business blog posts. Many online news sites like CNN, Yahoo and BBC publish images with their stories and even provide photo feeds related to current events. Some news sites list all news articles related to a particular topic in a timeline of the events in which they occurred, creating a news storyline. One good news storyline telling example is from New York Times, is in Figure 1.1.



How Gun Traffickers Get Around State Gun Laws

NOVEMBER 13, 2015

The effect of state gun control laws is diluted by a thriving underground market for firearms brought from states with few restrictions.



Connecting the Dots Behind the 2016 Presidential Candidates

MAY 5, 2015

How the teams behind some likely and announced 2016 candidates are connected to previous campaigns, administrations and organizations.

Source: <https://www.nytimes.com/interactive/2015/us/year-in-interactive-storytelling.html>

Figure 1.1: An instance from NY Times Journal of a news storyline

Some of the recent popular news storylines include “Donald Trump winning US elections” or “Brexit- Britain leaving European Union”. Presence of such data in multimedia files has provided strong impetus to develop applications like automatic story summarisation or illustration systems [37]. Such applications have a huge impact and a lot of usage in personal digital assistants. Internet is also home to various digital books. For example, Project Gutenberg is an online library of over 60,000 free digital books. It contains several digitised versions of children’s stories with illustrations from several renowned authors like Beatrix Potter, Roald Dahl and many more. It has stories from different genres including Fairy Tales, Myths, Biographies, Science, History and other literature. One example is shown in Figure 1.2.



But Peter, who was very naughty,
ran straight away to Mr. McGregor's
garden, and squeezed under the
gate!

First he ate some lettuces and some
French beans; and then he ate
some radishes;



And then, feeling rather sick, he
went to look for some parsley.

Source: <https://www.gutenberg.org/files/14838/14838-h/14838-h.htm>

Figure 1.2: An instance from an online digital story 'The Tale of Peter Rabbit' authored by 'Beatrix Potter'

Figure 1.2 provides an excerpt from an online digital story “The Tale of Peter Rabbit” authored by “Beatrix Potter”. The backbone of the story is

the textual narration. The content creators usually place photos/illustrations carefully at appropriate places. These images typically represent specific highlights such as an event in the narration or may serve to depict feelings and the general emotion in an instance. It can be observed from this example that illustration of stories requires substantial human judgement and reasoning. A human would require thinking of an appropriate illustration. Even if the set of illustrations are provided in advance. A human would require paying attention to ensure the semantic coherence between the context and the corresponding illustration. This task is difficult and time-consuming for humans.

In this thesis, we aim to explore the ways in which Machine Learning and Natural Language processing can be employed to perform this task of automatic text illustration. Several neural models have been employed for Image Understanding [1] and Text Understanding [144], and also using joint models of images and texts. We focus on developing systems that can automatically illustrate a given sequence of text passages that best describe and align with a sequence of images. We study and employ various deep neural network architectures and compare its performance with several previously published methods. We refer to this problem as “Stepwise Text Illustration”.

1.2 Relevant prior work

There exists rich literature on multi-modal image-text representation, which can be broadly seen from two different angles, caption generation or retrieval for images, and natural-language based image retrieval or generation. An early story picturing system was developed by Joshi et al. [61]. It retrieved images suiting a very specific description. Their task is to select illustrative images from a large pool. However, the task is quite different from ours, as they are making a decision to select one picture at a time disregarding the context. Our task is rather, to illustrate a given story stepwise, considering prior context, making most of those approaches inadequate for the setting of this work.

- First, most of the prior work assumes that each image in the set already has supervised labels in the form of a set of tags, or an informative caption. We do not augment learning algorithms, and thus do not rely upon any predefined labels, tags or captions of the image. Therefore, we do not directly feed any supervised signals, although we use some pre-trained image classification models for semantic image feature representation, indirectly incorporating this information.
- Second, we aim to illustrate the given piece of text, by incorporating and modeling the prior context of the story.

- Third, we do not limit the length of the given text passages, to one-sentence descriptions, rather focus on longer pieces of text. While, most of the existing approaches focused on one-sentence description with an image.
- Fourth, we focus on retrieving images from a set of images expecting high semantic coherence with the given text passage.

The Visual Storytelling Dataset (VIST) dataset[56] was built with a motivation similar to our own, but for generating text descriptions of image sequences rather than the other way around. Relying on human annotators to generate captions, VIST contains sequential vision-to-language pairs with a focus on abstract visual concepts, temporal event relations, and storytelling. They highlighted the difference between a literal description of an image and the more figurative language used for an image in a wider story context. In our work, we focus on producing similar sequenced datasets in an automated manner by selecting sources such as cooking recipes, children stories or any form of sequential instructional illustrations. In Chapter 2, we provide with a comprehensive literature survey of the related work for this research and also provide a critical account of the some of the closest works to our problem.

1.3 Research Questions and Objectives

The broad research goal is to be able to develop intelligent information processing systems that can concisely summarize all the textual content with the retrieval or generation of visual content. More specifically, In this thesis, we investigate the new task of *sequence-to-sequence (seq2seq2) retrieval*: given a sequence (of text passages) as the query, retrieve a sequence (of images) that best describes and aligns with the query. This is a step beyond the traditional cross-modal retrieval which treats each image-text pair independently and ignores the broader context.

We have previously discussed that an automated system for *Automatic Stepwise Illustration* is highly desirable. However, the goal is not easy to achieve, as such a system would be very complex and multi-faceted. There are a lot of requirements it would need to satisfy. In this work, we only focus on a subset of research questions that lead to progress towards the ultimate goal. Here, we outline the research questions of this thesis with corresponding objectives required to answer them.

RQ1 How can we develop automatic text illustration systems that illustrate a

given narrative text passage with a sequence of illustrations, considering and incorporating prior context?

OBJ 1.1 Investigate existing literature in the field of automated text illustration to gain understanding of how joint-models of texts and images are utilised for the task of semantic and coherent image retrieval. Understand common approaches and identify gaps in the knowledge.

OBJ 1.2 Identify or build relevant datasets.

OBJ 1.3 Build stepwise illustration models using the identified datasets.

OBJ 1.4 Evaluate the models in a realistic scenario to test their ability to be employed as real-world applications.

RQ2 How can we fuse information from different modalities to summarise the given content for developing context-based models?

OBJ 2.1 Start with and focus on atomic text passage and image pairs for summarisation.

OBJ 2.2 Identify relevant, realistic datasets used for studying automatic joint image-text summarisation.

OBJ 2.3 Build joint-models using the identified datasets that incorporate semantic features from both the modalities (images and texts) and thus are able to fuse them as contextual information.

OBJ 2.4 Analyse the performance of different joint-models in order to identify good context-modeling approaches representing better semantic features.

RQ3 How can we study the automatic stepwise illustration systems in a domain-constrained setting, given narrative text passage in a limited domain with a sequence of illustrations, considering and incorporating prior context?

OBJ 3.1 Investigate existing literature in the field of automated text illustration in a limited domain-constrained setting to gain understanding of the joint-models of texts and images.

OBJ 3.2 Identify or build relevant datasets.

OBJ 3.3 Build stepwise illustration models using the identified datasets.

OBJ 3.4 Evaluate the models in a realistic scenario to test their ability to be employed as real-world applications.

1.4 Technical Challenges

TC1 How can we automatically create resources for the task of automatic stepwise illustration?

- Investigate existing literature to identify any existing joint corpora that can be utilised for the given task.
- Focus on developing an approach to create an unlabelled dataset of sequenced image-text pairs from any source.

1.5 Contributions

In this thesis, the following contributions are made:

1.5.1 Neural Caption Generation for News Images

We introduced a novel methodology for caption generation for images appearing in news articles. This task is different from traditional image captioning where a text description is generated given an image, in this case a caption is generated conditional on both image and the news article, where it appears. A novel neural-networks based methodology is proposed to take into account both news article content and image semantics to generate a caption that best describes the image and its surrounding text context. The results outperform existing approaches to image captioning generation. (see RQ2 and chapter 4)

1.5.2 GutenStories and Stepwise Recipe datasets

We constructed two datasets, GutenStories and Stepwise Recipe datasets, for the task of story picturing and sequential text illustration. GutenStories consists of around 90,000 text paragraphs, each accompanied with an image, aligned in around 18,000 visual stories. It consists of a wide variety of images and story content styles. Stepwise Recipe is a similar dataset having sequenced image-text pairs, but having only domain-constrained images, namely food-related. It consists of around 67,000 text paragraphs (cooking instructions), each accompanied by an image describing the step, aligned in around 10,000 recipes. Both datasets are web-scrawled and systematically filtered and cleaned. (see TC1 and chapters 3)

1.5.3 Variational Recurrent Sequence to Sequence Retrieval for Stepwise Illustration

We also propose a novel variational recurrent seq2seq (VRSS) retrieval model. We explore ways to guide the retrieval of a relevant image from the repository

to match the semantics of the given text. The model has been evaluated on both the Stepwise Recipe and GutenStories datasets. The results on several automatic evaluation measures show that our model outperforms several competitive and relevant baselines. We also qualitatively analyse the model both using human evaluation and by visualizing the representation space to judge the semantic meaningfulness. (see RQ3 and chapter 5)

1.5.4 Context-Dependent Text Illustration and Description Retrieval

We provide a study of existing as well as some new models for the task of context-dependent text illustration and description retrieval. We study several models incorporating different kinds of features to study the relevant importance of these features in modelling and retrieval of a relevant image from a repository to match the semantics of the given text. We study and compare performance of several previously published methodologies on GutenStories dataset. We also qualitatively analyse the models by visualizing the representation space to judge the semantic meaningfulness. (see RQ1 and chapter 6)

1.6 Thesis Outline

This PhD thesis follows a traditional outline, starting with background information, followed by four analysis chapters and closed with a conclusion. Chapters 1 and 2 provide the motivation and necessary background for the comprehension of this thesis. Chapter 3 describes the datasets used in the following analysis. Then, Chapters 4-6 describe the analysis performed on the task of stepwise illustration. Finally, we conclude in Chapter 7.

In Chapter 2, we provide the necessary background, followed by a comprehensive literature review of the existing related work done in the area. We finish the chapter by picking up literature that most closely relates to our work.

In Chapter 3, we discuss the datasets used in this thesis. We describe methods used for dataset creation and for systematic filtering and cleaning techniques.

In Chapter 4, we study existing systems which are employed for the task of automatic caption generation for news images. We particularly pick an existing system which uses Latent Dirichlet Allocation (LDA) based methodology to generate captions for news images but face several challenges. We reproduce results of their methodology and propose some improvement mechanisms. Sub-

sequently, we propose a novel deep neural-networks based summarisation model, which generates relevant captions for news images, and which is inspired by encoder-decoder architecture of neural translation systems.

In Chapter 5, we address and formalise the task of *sequence-to-sequence (seq2seq) cross-modal retrieval*. Given a sequence of text passages as query, the goal is to retrieve a sequence of images that best describes and aligns with the query. This new task extends the traditional cross-modal retrieval, where each image-text pair is treated independently ignoring broader context. Furthermore, we also propose a novel *variational recurrent seq2seq (VRSS) retrieval model* for this seq2seq task. We focus specifically on a domain-constrained dataset, namely Stepwise Recipe Dataset in this chapter.

In Chapter 6, we focus on the task of *sequence-to-sequence (seq2seq) cross-modal retrieval* on a broader domain, GutenStories dataset. We study existing methods in the literature for this task, and also propose some new methods.

In Chapter 7, we provide a conclusion and future research directions

Chapter 2

Background

Our work is related to a several different lines of research. There is very rich literature on multi-modal image-text representation and learning. We classify the work presented here broadly into four sections: Multimodal representation learning, Missing modality inference, Cross-modal retrieval and other related multimodal tasks. In the next few sections, we present research from all of the various related areas, specifically highlighting and emphasising upon those closest related to our work. However, we begin by providing a section about general introduction to the concepts in natural language processing. We have also provided some background about common deep neural-networks based architectures used in tasks in the sections ahead.

2.1 Prerequisites in Natural Language Processing

In this section, we discuss some of the prerequisites in natural language processing. Arthur Samuel coined the term Machine Learning in 1959 as “the ability for the computers to learn without explicitly being programmed” [119]. At the intersection of Machine Learning, Computational Linguistics and Data Mining lies Natural Language Processing. It studies the processing of data from natural languages (English, French etc) by computer programs. In this thesis we use Machine Learning methods to perform specific Natural Language Processing tasks related to the problem *Stepwise Illustration* and *Caption Generation*. A typical real-world Machine Learning task consists of a pipeline of steps or operations. It consists of steps like Data Preparation, Feature Engineering, Model Learning and Evaluation. Chapter 3 describes the Data Preparation and Evaluation stages. Data Preparation may include several steps like Data Collection, Data Preprocessing etc. Data Preprocessing is considered to be an important step in Machine Learning tasks. In the next section, we discuss the ways to represent natural languages data for many

tasks in Machine Learning. We also provide details of two widely accepted approaches to represent textual data, namely Bag-of-words and word2vec, which is based on distributed representations.

2.1.1 Representation

Machine Learning models work with numerical data, therefore it is essential to represent the textual data as numeric vectors to make them a suitable input. There are multiple ways of representing textual data in numerical format, for example bag-of-words. A word or a phrase is thus represented as a numeric vector, or a word embedding. Next, we discuss two most important ways to represent textual data.

Bag-of-words

Another method for text representations is bag-of-words. It is a special case of the n-gram model, where $n = 1$, and hence can be generalised to any n by creating the vocabulary out of phrases of length n . Although there are few drawbacks of using this representation. First of all, it ignores the word order in the sentence. Therefore, it loses this important contextual information. Secondly, it produces sparse representations, where high dimensional vectors have very few nonzero elements.

Distributed Representations

Bag-of-words are sparse representations. Distributional hypothesis states that “a word is characterised by the company it keeps”. There are many ways of obtaining representations based on the distributional hypothesis. Tomas Mikolov and Ilya Sutskever and Kai Chen and Greg Corrado and Jeffrey Dean [134] proposed a model called word2vec based on the distributional hypothesis. There have been many models proposed in recent literature that project sentences, paragraphs or even documents to vector space using principle of compositionality, that simpler constituent expressions’ meaning compose complex expressions’ meaning [80].

2.1.2 N-gram

An n-gram is a concept in the fields of computational linguistics and probability. It is a contiguous sequence of n items (phonemes, syllables, letters, words) from a given sequence of text or speech.

2.1.3 TF-IDF

TF-IDF or tf-idf is a technique used in text representation. tf-idf consists of Term Frequency and Inverse Document Frequency. Term frequency is the proportion of total number of times a given term t appears in the document against (per) the total number of all words in the document. The inverse document frequency provides a measure of how much information the word provides. IDF can be thought of showing how common or rare a given word is across all documents. As, some words appear more frequently in general, the value of tf-idf increases in proportion to the number of times a word appears in the document and is often offset by the frequency of the word in the corpus. Kim et al. [67] recently proposed many document representation algorithms for document classification.

2.2 Deep neural-networks based architectures

Deep Learning is a sub-field of Machine Learning that specifically studies artificial neural networks multiple layers deep. In this section, we provide literature related to some deep learning algorithms.

2.2.1 Convolutional Neural Networks (CNN)

A Convolutional Neural Network (CNN) is a common deep neural-networks based algorithm which takes in an image as the input and is able to assign importance (learnable weights and biases) to different aspects of objects in the image and be able to differentiate one from the other. Therefore, it has been commonly used for many classification problems. CNN has the ability of automatic feature extraction, as it can compute features from a raw image. Through the application of relevant filters, CNNs successfully capture the Spatial and Temporal dependencies in an image. Some of the commonly used CNN architectures are ResNet [50] and VGG [125].

2.2.2 Recurrent Neural Networks (RNN)

We also present research from recurrent neural networks (RNN) based algorithms, as we aim to develop models that can incorporate context, and the data is sequential in nature. Variational recurrent neural network (VRNN) [21], which introduces latent random variables into the hidden state of a recurrent neural network (RNN) by combining it with a variational autoencoder (VAE). They showed that through the use of high level latent random variables, VRNN can model the variability observed in structured sequential data such as natural

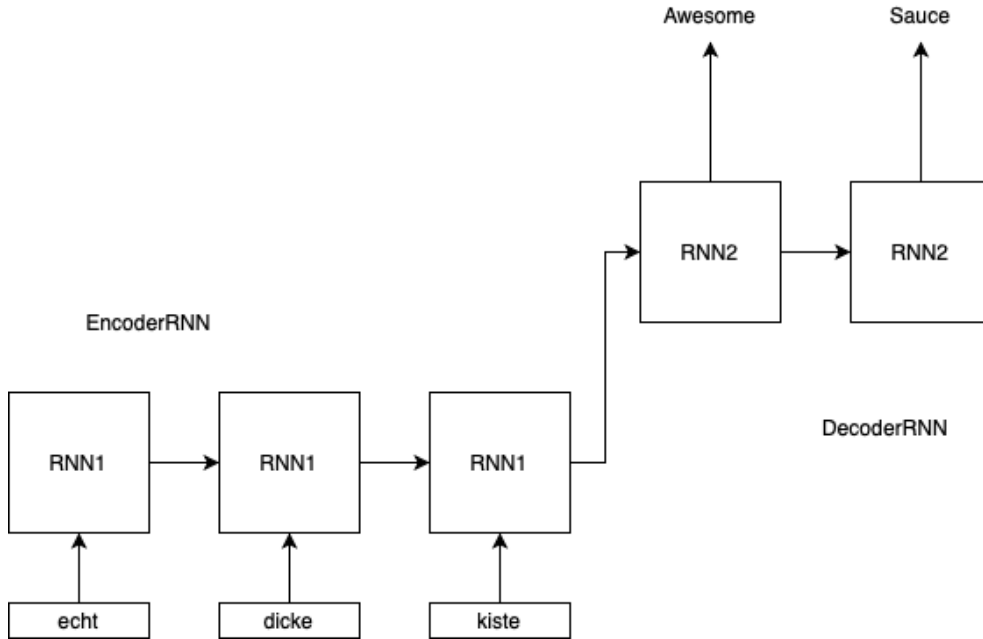


Figure 2.1: Encoder Decoder Architecture

speech and handwriting image data. VRNN has recently been extended to deal with other sequential modelling tasks such as machine translation [128].

Recently, an encoder-decoder architecture inspired from machine translation has been applied to image captioning and has achieved state-of-the-art performance. In the next subsection, we describe this architecture.

2.2.3 Encoder-Decoder Architecture

Recently, Deep learning solutions have been employed to address machine translation systems and have achieved state-of-the-art performance and are called neural translation systems. In neural translation systems, an encoder-decoder architecture is used. An encoder is used to read a sentence in the source language and is transformed into a rich fixed length embedding vector representation. This embedding vector is in turn fed to a decoder that generates the sentence in the target language.

Figure 2.7 shows the architecture, here each cell is a Recurrent Neural Network (RNN), and German words are encoded as sentence embeddings and are further fed to these RNN cells. h_i depicts the hidden vector that represents the hidden states of the RNN cell at timestep t . A variable length sequence “Echt”, “dicke” and “Kiste” is fed to the cells at different timesteps and system encodes all of this information into a fixed-length vector. This fixed length vector is fed to the decoder, represented with blue cells in this case. The decoder generates the vector embeddings of the outputs “Awesome”, “Sauce”

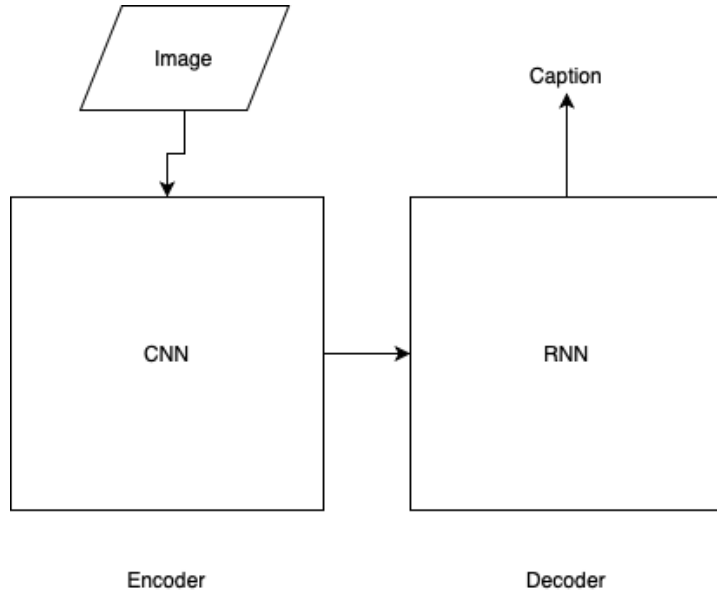


Figure 2.2: Neural Image Captioning Architecture

in target language that is English.

This architecture has been adopted in image captioning and a class of methods called ‘neural-image captioning’ methods have been developed. The idea here is to use a Convolutional Neural Network (CNN) as an encoder for the image and Recurrent Neural Network (RNN) as a decoder. They view captioning as a translation problem, where image is in the source language and target language is English. Figure 2.2 shows the neural image caption generator architecture. Here CNN is used to encode image to a fixed length embedding vector. The yellow circles represent RNN cells at different timesteps. x_t is the input vector at timestep t . h_t is the hidden state and y_t the output.

Over the last few years, it has been convincingly shown the CNNs can produce rich representation of the input image by embedding it to a fixed-length vector, such that this representation can be used by a variety of vision tasks [121]. Therefore, it is natural to use CNN as image encoder, by first pre-training it for classification task. This network is subsequently used as an off-the-shelf feature extractor, where the last hidden layer of the network is used as a feature vector. This hidden representation is fed to the decoder to generate descriptions for the image. [88] provide a model with similar architecture. [65] developed a deep neural network that infers the latent alignment between segments of sentences and region of image they describe. They use CNN for encoder and a bi-directional RNN over sentences.

Some of these models are end-to-end, that is they are fully trainable using stochastic gradient descent, sub-networks combine language and vision models.

[38] propose a model with similar architecture, they propose several methods in which the image information can be incorporated into the LSTM, they use

for language modelling. They use CNN for encoding images.

2.2.4 Neural language methods with attention

Rather than compress an entire image into a static representation, attention mechanisms have been introduced which allow salient features to dynamically come to forefront as needed. Using representations from the top layer of a Convolutional Network that distill information in an image down to the most salient objects is one effective solution. But it has a potential drawback of losing information present in the lower layers which could be useful for generating richer and more descriptive captions. Xu et al. [146] propose a soft and hard attention mechanism for image captioning tasks. They use a Convolutional Neural Network to encode the images and a Recurrent Neural Network with attention mechanism to generate a description. By visualising attention weights, they switch what the model is looking at while generating a word. You et al. [153] propose a Convolutional Neural Network with an attention mechanism that weights the image features and Recurrent Neural Network to generate captions to describe weighted image features.

The basic problem that the attention mechanism solves is that it allows the network to refer back to the input sentence, instead of forcing it to encode all information into one-fixed length vector.

2.3 Multimodal representation learning

In the world around us, humans perceive data originating from different modalities for example auditory, visual or tactile signals. Similarly, computers process information from various modalities for example, images, audio or texts. Although, information is spread out in different modalities, many times, the underlying semantic concept is the same. For example, when somebody mentions the word “Apple”, or comes up with an image of an apple, the underlying semantic concept that is being referred to, is the same. In order to bridge this semantic gap, various attempts have been made in order to learn joint models of data from different modalities. For example, Ngiam et al. [98] develops a speech recognition system where they jointly model audio and visual modality. They focus on learning representations of speech audio which are coupled with videos of the lips. Bordes et al. [14], Ma et al. [90], Socher et al. [126] only learn binary relations between objects of interest from different modalities.

Recently, many interesting applications have been developed by bridging language and vision modality together, for example automated image captioning systems. In human perception, visual information is the dominant

modality for acquiring knowledge, since a big part of brain is dedicated to visual processing. Whether or not, there are languages involved in the visual process is still an ongoing argument. However, to develop an intelligent system that tries to achieve AI, having languages provide a way for human-computer interaction. Multimedia applications that bridge language and vision mainly fall into either of two categories, visual description or visual retrieval [143].

Some of the applications in this domain include automatic image captioning systems, text-query based image retrieval systems. All of these applications have become possible due to two main reasons.

- emergence of machine learning algorithms, specifically some deep neural-networks based algorithms, which require a large amount of data
- the availability of this data due to recent growth in amount of digital information available on the Internet.

2.3.1 Image-Text correspondence

A fundamental problem, in joint-modelling research, is to associate images with some corresponding relevant, descriptive text. Such associations often rely on semantic understanding and go beyond traditional similarity search or image labelling tasks. One challenge is to provide a more human-like visual understanding, where the text is expected to reflect abstract ideas and events in an image, rather than simply identifying the objects in it. Related tasks within this domain include image caption generation [100, 148, 154], visual question answering [4, 145], and cross-modal image retrieval from text [139].

The aim of cross-modal retrieval is to return outputs of one modality from a data repository, while a different modality is used as the input query. The repository is therefore a multimodal one, usually consisting of paired objects from the two modalities, but may be labelled or unlabelled.

In the case of image retrieval from text queries, the applications are endless. Popular search engines today make a concerted effort to perform image search, where a text query is used to retrieve the relevant matching images. A variety of cross-modal retrieval methods have been explored over the years. Typically, data from disparate modalities are projected onto a common representation (feature) space to facilitate direct comparison in the search process. The algorithms for learning such common representation space may be categorised as supervised, rank-based, pairwise and unsupervised, as per a recent survey [139].

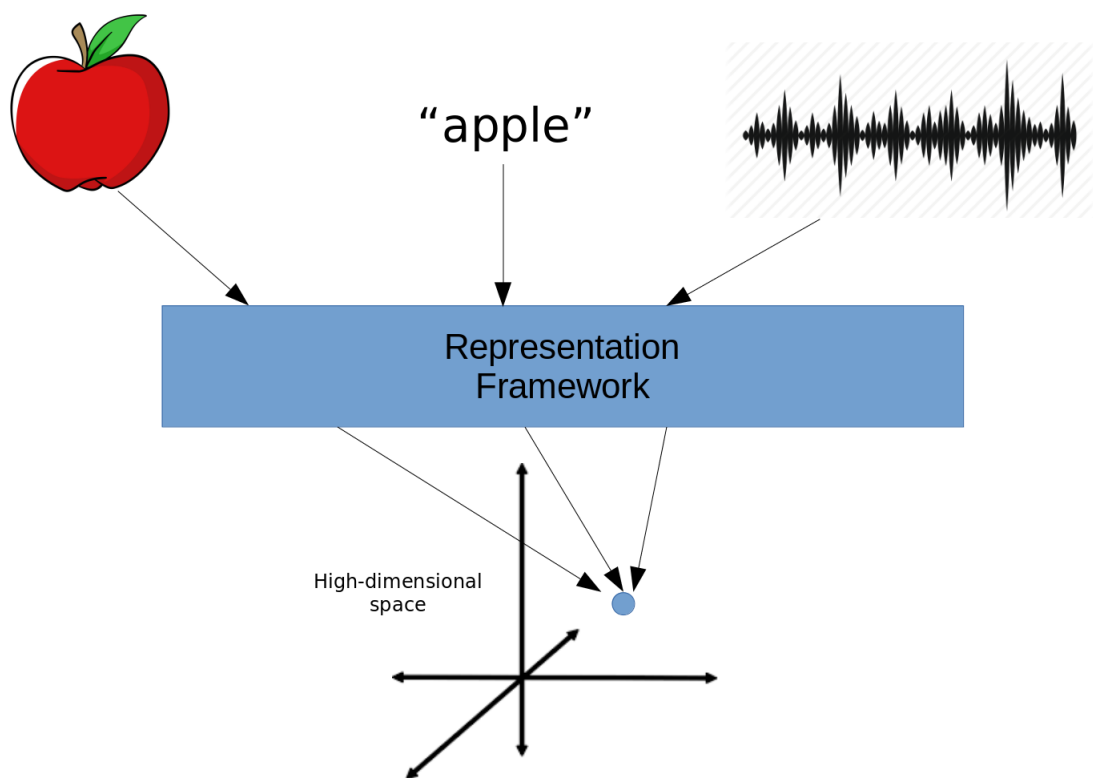


Figure 2.3: Semantic Common Embedding Space Learning

2.3.2 Hinge loss

There are many ways to learn semantic correspondence between data from two given different modalities, say images and texts. One possible approach involves using Canonical Correlation Analysis to learn the correspondence between the representations of both the modalities, thereby trying to preserve correlation between the text and images in the joint embedding, for example in Eisenschat and Wolf [29] and Klein et al. [74].

A common approach uses hinge loss function to reduce pairwise distances between corresponding data points. The hinge loss usually consists of the sum of two symmetric terms. The first one, is over all negative text representations, given the image query. The second one, is the sum over all negative images, given the text query. Each term is proportional to the expected loss over sets of negative samples. The hinge loss is zero, if a given image and text data point are closer to one another in the joint embedding space than to any other negative sample, by a fixed margin. In practice, for example in Karpathy and Fei-Fei [63], Kiros et al. [69] and Socher et al. [127], rather than summing over all negative samples in the training set, it is worth considering summing over the negatives in a mini-batch of stochastic gradient descent for computational efficiency.

The time complexity of computing these loss approximations is quadratic in the number of paired image-text data points in a mini-batch. There are many possible variants of this loss function. One is a pairwise hinge loss function in which elements of positive pairs are encouraged to lie within a hypersphere of a given radius in the joint embedding space, while negative pairs should be no closer than another fixed margin. This causes and puts too many constraints on the structure of the latent space, therefore is deemed to be problematic, and it entails the use of two hyper-parameters which can be very difficult to determine.

2.3.3 Order-embeddings

Mikolov et al. [97] showed a way to learn effective semantic embeddings for words, using distributional hypothesis. It is based on idea of quantifying and categorising semantic similarities between linguistic terms, on their distributional properties, in large collections of natural language data.

Most of the existing methods model the image-text relationships using distributional semantics. The general idea, as outlined before, is to map the objects under interest, texts and images into a very high-dimensional vector

space, such that semantically similar objects are mapped to nearby points in the common semantic space. The high-dimensional space is fully-ordered, therefore, distance-preserving [20, 127]. As the space is distance-preserving, common similarity functions, like Euclidean or cosine distance are used. Vendrov et al. [137] argued the visual-semantic hierarchy follows a partially ordered structure, than fully-ordered. They introduce order-embeddings, because their embeddings are not distance-preserving, but order-preserving and applied them to the tasks of hypernymy prediction and cross-modal image-text retrieval. Athiwaratkun and Wilson [5] introduced probability densities rather than point vectors in order embeddings.

2.4 Missing modality inference

The aim of missing modality inference is to generate or infer the outputs of one modality from a data repository, while a different modality is used as the given input. The repository is therefore a multimodal one, usually consisting of paired objects from the two modalities, but may be labelled or unlabelled.

2.4.1 Image Captioning

Automatic caption generation or description generation for visual data is one of the central tasks in computer vision and natural language processing research. Traditionally, there has been significant work in image classification, object detection and image annotation, but a relatively little focus on generating descriptions involving a full sentence. So, some of the obvious solutions consisted of combining the result of these methods with a stage, that arranges these keywords in the form of a sentence. We list all of these methods under two-stage architecture methods.

2.4.2 Two-stage architecture

As described above, these methods consist of a pipeline with two stages, namely content selection and surface realization. The former stage, content selection consists of an image annotation model that analyses the content of the image and identifies “what to say” of the image. The latter stage, surface realization consists of a language model, that analyses the keywords and identifies “how to say” of the image. Fig 2.4 depicts the overview of the two-stage architecture based methods in an illustrative diagram. We describe the content selection and surface realization components in detail ahead.

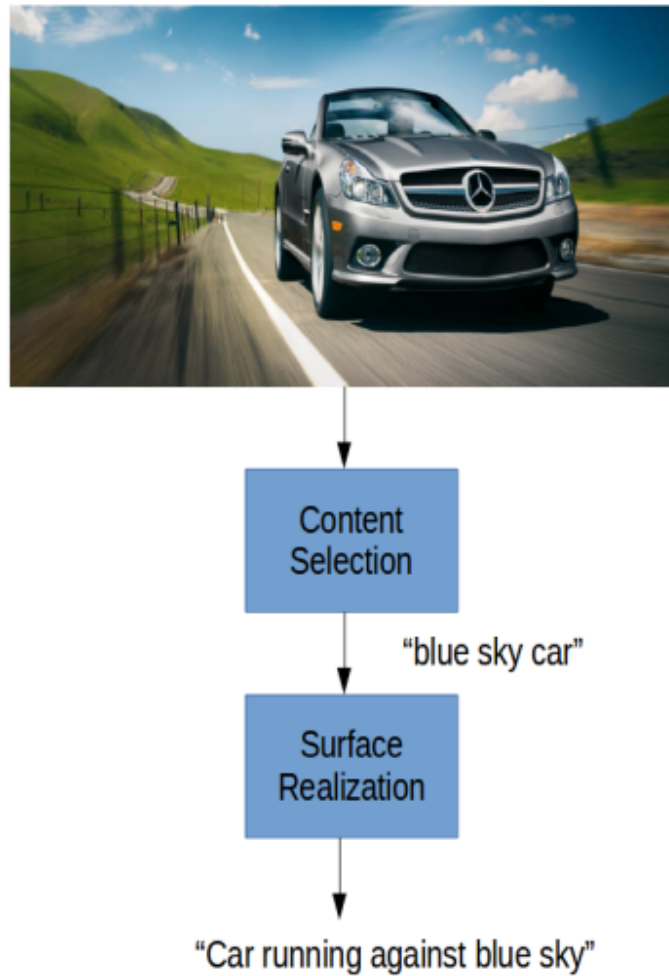


Figure 2.4: Two Stage Architecture

Content Selection

Much work within computer vision has focused on image annotation, a task which is very much related but distinct from image description generation. The goal in image annotation is to label an image with keywords relating to its content. When the keywords belong to a fixed set of categories, the problem is called image classification.

Supervised Image Annotation Methods

Here the problem is similar to image classification, as the keywords are fixed and pre-defined at training time. The fixed set of categories are identified usually in the form of classes of vocabulary words. Machine learning algorithms are applied to learn a one-to-one correspondence between an image and these classes. The core notion behind is to learn a mapping between visual feature vectors and semantics of the image. A detailed review of supervised methods for image annotation can be found in [58].

Unsupervised Image Annotation Methods

These methods do not have a fixed set of pre-defined classes. Instead, an attempt is made in order to learn the connections between visual features and words and automatically cluster them into classes of words, which will finally denote the semantics of the image. Typical solutions to this involve introducing latent variables. Standard latent semantic analysis (LSA) and its probabilistic variant (PLSA) have been applied to this task [101], [53]. [7] provide a more sophisticated model, they estimate the joint-distribution of words and regional image features while treating annotation as a problem of statistical inference in a graphical model. The final output is clusters of words, which appropriately describe the content of the image.

Surface Realization

The output of the previous stage is a set of keywords that appropriately describe the content of the image. The aim of this stage is to go from keywords to sentence.

Two methods are generally popular for this approach, namely extractive and abstractive methods:

Extractive methods

The main idea behind these methods is to use a database of sentences and retrieve a relevant sentence rather than constructing sentences using a language model. A sentence is retrieved to describe, or the description is generated by identifying and subsequently concatenating the most important sentences in the document. Various metrics could be used to calculate relevance of a sentence with a set of keywords for example, word-overlap based sentence selection score, vector-space based sentence selection score or topic-based sentence selection score. [60] provide a comprehensive overview.

Abstractive methods

Although extractive methods yield grammatically correct sentences and require relatively little linguistic analysis, there are few serious caveats to consider. Many a times, such is the case that there is no single sentence in the document that describes the image. These methods try to compose a sentence from the keywords based on language models learnt. These could be probabilistic generative models or neural-language based models.

In Farhadi et al. [32], images are parsed into $\langle \textit{object}, \textit{action}, \textit{scene} \rangle$ triplets. A more complex graph of detections beyond triplets is used by Kulkarni et al. [76]. State-of-art object recognition and language generation techniques are used in their model Babytalk. Feng and Lapata [37] provides a news article caption generator. They use an LDA-based methodology for image annotation and used a wide variety of surface realization techniques to generate the description/caption for the given news image.

All of these two-stage architecture methods have some serious limitations. A list of keywords is often ambiguous. A set of keywords “blue, sky, car” could depict “a blue sky” or a “a blue car”. Therefore, the models should be designed such that there is strong correlation between phrases and images or sentences and images that are semantically relevant. In other words, a direct leap should be taken from image to sentence and vice versa. Moreover, these approaches are heavily hand-designed and rigid when it comes to text generation. So, their applicability becomes limited, cannot be generalized for new domains. We refer to such qualities as “Sentential Integration” and “Generalized Applicability”. In our critical analysis, we check on these qualities for existing approaches.

It is important to note, recently many deep neural-networks based architectures have been employed for the task of automatic caption generation for images, called neural image caption generation. We further expand on these methods in the penultimate section.

2.4.3 Image Synthesis

The problem of image synthesis/generation from textual queries has also been studied in recent literature. Reed et al. [112] proposed usage of some recent conditional generative models, namely Generative Adversarial Networks (GAN) [43] for generating images from natural language text.

Zhang et al. [157] aimed to produce high quality photographic images conditioned on semantic text descriptions using their proposed model HDGAN. They introduced a hierarchical-nested adversarial objective inside the network hierarchy, with the intention of regularizing mid-level representations and assist generator training to capture the complex image statistics. They pushed generated images up to high resolutions of 500 pixels wide and 500 pixels deep using an extensible single-stream generator architecture to better adapt the jointed discriminators.

Zhang et al. [156] argued to decompose the hard problem of high-resolution image generation into more manageable sub-problems through a sketch-refinement process. They introduced a stacked GAN architecture, where the first-stage GAN sketches the primitive shape and colours of the object based on the given text description, yielding first-stage low-resolution images. The second-stage GAN takes first stage results and text descriptions as inputs and generates high-resolution images with photo-realistic details. It is also able to rectify any defects in results of the first stage and even add compelling details with the

refinement process. Zhang et al. [155] further improved upon the idea, while Xu et al. [149] incorporated attention.

2.5 Cross-Modal Retrieval

Methods in cross-modal retrieval can be broadly categorised into:

2.5.1 Atomic Cross-Modal Retrieval

In order to perform retrieval, the multiple modalities are typically represented in some shared common feature space to facilitate direct comparison. Classical approaches to compare data across modalities include canonical correlation analysis (CCA) [46], partial least squares (PLS) regression [113], and their numerous variants. But these are unsupervised and suffer from the problem of the “semantic gap” between modalities and their low-level features versus high-level semantic concepts. Further, Rasiwasia et al. [110] showed that class information and semantic matching could be leveraged to reduce the semantic gap.

More recently, various deep learning models have been developed to learn shared embedding spaces based on paired image-text data, either unsupervised, or supervised using image class labels. The deep models popularly used for cross-modal retrieval tasks include deep belief networks [99], correspondence autoencoders [34], deep metric learning [51], and convolutional neural networks for very large databases [140].

However, given the high dimensionality of the feature space, generalisability remains an issue and matching accuracy is also affected. Bokhari and Hasan [13] differentiate ways of combining information from separate modalities. Combining the individual classifier scores from different modalities is another approach, where some rule-based decisions are made.

With most of these models it is expected that by learning how to create embeddings from such pairwise aligned data, the common representation space will be capturing semantic similarities across the modalities.

2.5.2 Sequential Cross-Modal Retrieval

Most cross-modal retrieval and similarity search systems, however, do not consider sequences of related data, in the query and result. In traditional image

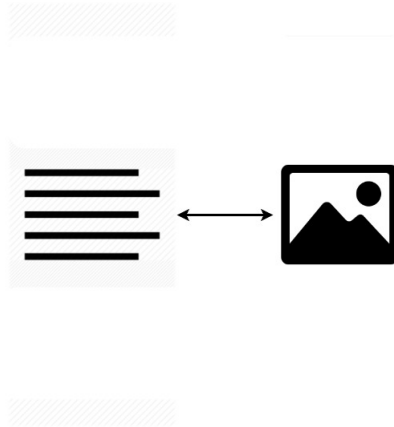


Figure 2.5: Encoder Decoder Architecture

retrieval using text queries, for example, each image-text pair are considered to be an “atomic” unit, as represented in 2.5. They have been considered in isolation and any broader ‘context’ has been ignored. The incorporation of such context is an important step towards semantic knowledge discovery. An image-from-text retrieval model that is aware of context must look beyond merely the associated paired text of an image in the repository to retrieve the appropriate result. It must also consider the sequential relationships between the entire set of texts and images.

Such context-aware cross-modal retrieval is possible using a model that takes into account *sequence-to-sequence (seq2seq) retrieval*, where contextual information and semantic meaning are both encoded and used to inform the sequence based retrieval from the data repository.

There are a variety of applications that can benefit from image-text sequence retrieval, such as *stepwise recipe illustration*, or a more generalised *story picturing*. In the context of recipe illustration, an effective retrieval system must produce a set of relevant images corresponding to each step of a given text sequence of recipe instructions, as represented in 2.6. Similarly, for the general task of automatic story picturing, a series of suitable images must be chosen that illustrate the events and abstract concepts found in a sequential input text taken from a story. Thus, the model must look beyond the single image-text pair and must also consider associations between the entire sequence of image-text pairs that together make up a recipe or story.

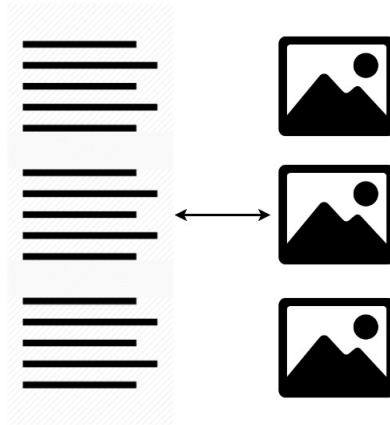


Figure 2.6: Encoder Decoder Architecture

Story Picturing

An early story picturing system [61] retrieved images that suit a very specific description. They used images from the Terragalleria dataset¹ and the AMICO dataset² to illustrate ten short stories based on key terms in the stories and image descriptions as well as a similarity linking of images. The idea was pursued further with a system [41] for helping people with limited literacy to read more easily. This system split a sentence into three categories and then retrieved a set of explanatory pictorial icons for each category.

An application [23] for making news articles more interesting by adding illustrations retrieved from the MIRFlickr-25000 dataset [57] used a sliding window over previous sentences to condition the retrieval. Another automatic illustration system [22] performed textual search and visual clustering over the repository in order to retrieve suitable illustrations. These systems are closer to text illustration systems than context-dependent story picturing systems.

To our knowledge, an application [68] that ranks and retrieves image sequences based on longer paragraphs as text queries was the first to suggest extending the pairwise image-text relationship to matching image sequences with longer paragraphs. They employed a structural ranking SVM with latent variables and used a custom-built Disneyland dataset, comprising of blog posts along with associated images, as the image-text parallel corpus from which to learn joint embeddings. Further, they augmented the same latent space with images from visitors' photo streams. We follow a similar approach for creating our image-text parallel corpus from cooking recipes rather than blog posts and design an entirely new seq2seq model to learn our joint embeddings.

¹<http://www.terrageria.com>

²<http://www.amico.org>

Coherence Neural Story Illustration (CNSI) was built on an encoder-decoder network, as described in Ravi et al. [111]. CNSI was used to first encode sentences using a hierarchical two-level sentence-story Gated Recurrent Unit (GRU), and then sequentially decode into a sequence of illustrative images corresponding to a passage of text. In their model, a previously proposed coherence model [103] was used to explicitly model co-references between sentences.

Visual storytelling

The main aim in visual storytelling is, given a sequence (of images) as the query, retrieve a sequence (of text descriptions) that best describes and aligns with the query and are coherent.

The Visual Storytelling Dataset (VIST) dataset ³ [56] was built with a motivation similar to our own, but for generating text descriptions of image sequences rather than the other way around. Relying on human annotators to generate captions, VIST contains sequential vision-to-language pairs with a focus on abstract visual concepts, temporal event relations, and storytelling. They highlighted the difference between a literal description of an image and the more figurative language used for an image in a wider story context. In our work, we focus on producing similar sequenced datasets in an automated manner by selecting sources such as cooking recipes.

In [87], a joint sequence-to-sequence model was formulated to learn a common image-text semantic space. After enforcing coherence of predicted sentences, they were able to generate paragraphs to describe photo streams. They performed experiments on both the above datasets for this text generation task. Recent work [111] has used the VIST dataset for the inverse problem of retrieving images when given text, similar to the illustration problem that we are interested in. They focus on sentences that are abstract and have a sequential aspect. An encoder-decoder network was used to illustrate a sequence of sentences that form a story. They use the VIST dataset despite this not being custom-built for such a task.

2.5.3 Common Embedding Space Learning

A number of pairwise-based methods over the years have attempted to address the cross-modal retrieval problem in different ways, such as metric learning [108] and deep neural networks [138] to learn a shared feature space. For

³<http://visionandlanguage.net/VIST/>

instance, [64] devised an alignment model that learns inter-modal correspondences between images and text using MS-COCO [85] and Flickr-30k [107] datasets. In [70], they proposed unifying joint image-text embedding models with multimodal neural language models, making use of an encoder-decoder pipeline. A later method [31] used hard negatives to improve their ranking loss function, which yielded significant gains in cross-modal retrieval performance. Such systems focus only on isolated image retrieval from the repository when given a text query, and do not address the seq2seq retrieval problem that we study in this work.

In a slight variation to the cross-modal retrieval problem, in [6] the goal was to retrieve an image-text multimodal unit when given a text query. For this, they proposed a gated neural architecture to create an embedding space from the query texts and query images along with the multimodal units that form the retrieval results set, and then performed semantic matching in this space. The architecture consisted of embedding layers and relevance matching layers, and the training minimized structured hinge loss, but there was no sequential nature to the data used.

Recently, a stacked cross-attention network was introduced by Lee et al. [82], to study the problem of discovering the full latent alignments using both image regions and words in a sentence as context and infer image-text similarity. It was thus employed for the cross-modal retrieval task, by learning a common embedding space.

Lu et al. [89] proposed a model for learning task-agnostic joint representations of image content and natural language. It is an extension over the popular BERT [27] architecture to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers.

2.6 Other related multimodal tasks

There are various interesting applications at the intersection of natural language processing and computer vision. In the next subsections, we categorize the literature based on the applications or the research problem that they tackle. Starting from image captioning systems, visual storytelling, visual question-answering to multi-modal summarisation.

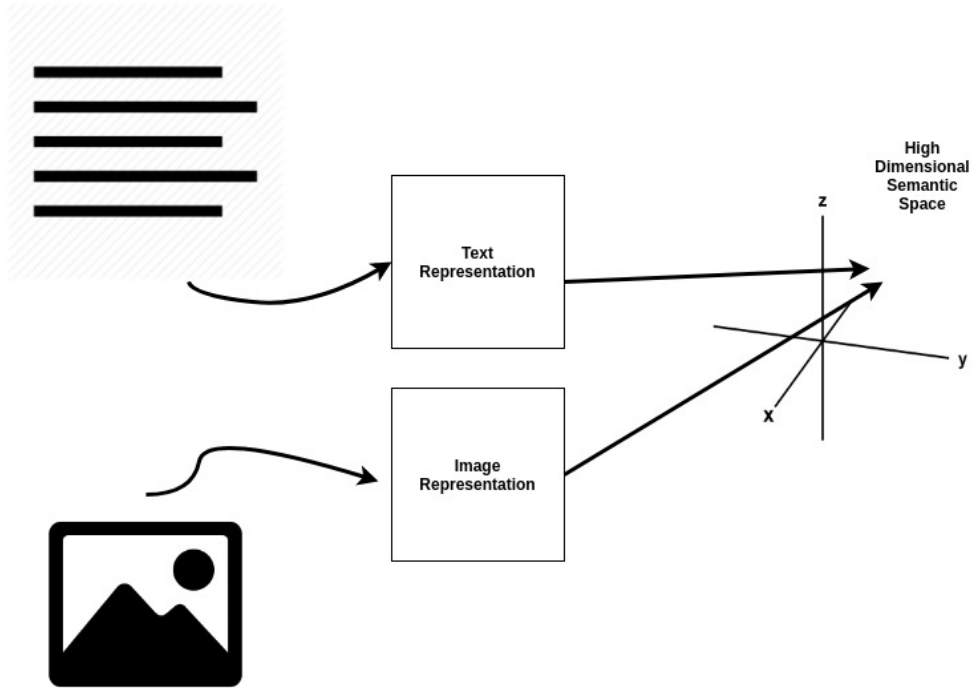


Figure 2.7: Encoder Decoder Architecture

2.6.1 Image-sentence Ranking

A large body of work has addressed the problem of ranking descriptions for a given image, for example in [42] improving image sentence embeddings using large weakly annotated photo collections. Such approaches are based on the idea of co-embedding of images and natural language texts.

2.6.2 Visual Question-Answering

This is form of a renewed experiment in AI, to develop understanding ability of machines for visual and textual data. Some of the works have focused on visual question answering interfaces, where given an image and a natural language question about the image, the task is to provide an accurate natural language answer [2].

2.6.3 Multimodal Summarisation

In this section, we provide a literature review of both text-only and multi-modal summarisation of documents. Traditional methods to date involve designed features such as sentence position and length. Features like words in the title, the presence of proper nouns, content features such as word frequency and sometimes event features such as action nouns [30].

Wang et al. [141] have proposed a low-rank approximation approach, where they use CNN to encode images and RNNs for sentences to learn joint-embeddings of news stories and images for timeline summarisation.

Recently, there has been a surge of interest in repurposing sequence transduction neural network architectures for Natural Language Processing [130]. Encoder-Decoder architecture modelled by a Recurrent Neural Networks has also been used.

Rush et al. [114] propose a neural attention model for abstractive sentence compression which is trained on pairs of headlines and first sentences in an article.

Cheng and Lapata [19] use a hierarchical document encoder and attention-based hierarchical content extractor, but is only for text-only documents. Their representation framework includes, they first use a single layer Convolutional Neural Network CNN with max-over pooling operation. The CNN operates at the word level. They use a Recurrent Neural Network that composes a sequence of sentence vectors into a document vector, which captures the sentence to sentence transitions. They propose two sentence extractors as well as word extractor models to generate summaries.

Recently many models to effectively represent semantics of a sentence were introduced [72]. Kiros et al. [72] introduced SkipThoughts model, it essentially attempts to take the skip-gram model [134] for learning representations at the sentence level. It is trained on the BookCorpus dataset consisting of 11,038 books and 74,004,228 sentences. The SkipThoughts model uses the framework of neural translation models which consists of an encoder and a decoder to learn sentence embeddings. That is, an encoder maps words to a sentence vector and a decoder is used to generate the surrounding sentences. Encoder is used to map an English sentence into a vector. The decoder then conditions on this vector to generate a translation for the source English sentence.

The SkipThoughts model uses RNN with Gated Recurrent Unit (GRU) activations as an encoder and RNN with a conditional GRU as a decoder. In Kiros et al. [71], two separate models were trained on the BookCorpus dataset. One is a unidirectional encoder with 2,400 dimensions, and the other is a bidirectional model with forward and backward encoders of 1,200 dimensions each which are subsequently concatenated to form a 2,400-dimensional vector. The SkipThoughts vectors used in our work are a concatenation of the vectors from both models, resulting in a 4,800-dimensional vector.

Table 2.1: Relevant Literature as per the Objectives

Model	Summary	General	Sentential	Features	Topic
Feng and Lapata [37]	yes	yes	no	no	no
Long et al. [88]	no	yes	yes	yes	no
Cheng and Lapata [19]	yes	no	yes	yes	no
Ravi et al. [111]	yes	yes	yes	yes	no

2.7 Critical Review

In this section, we list down some selected methodologies from various tasks/domains for the purpose of a critical review, as to what features they do or do not have. Therefore, the LDA methodology in Feng and Lapata [37], Neural Image Captioning in Long et al. [88], Neural Summariser in Cheng and Lapata [19], attention based model in Xu et al. [146] mentioned in the previous sections and Coherence Neural Story Illustration (CNSI) built on an encoder-decoder network, as described in Ravi et al. [111].

According to the objectives for the research question identified to be addressed in chapter 1. Following is the current state in the existing relevant literature.

Table 2.1 lists down some methodologies mentioned in the previous sections and provides a critical account on the basis of “Summary”, as in whether their model focuses on a bigger form of text or not, “General”, whether the model has generalised applicability to new domains or not, “Sentential”, as in if there is direct leap taken from the image to text or text to image domain or not, “Features” to represent if the methods compute the features automatically or not. The “Topic” represents if a latent topic vector is also generated in the methodology or not.

It can also be observed from the Table 2.1 that no methodology satisfies all of our listed objectives. We analyse their methodologies in detail in the next chapters and also compare results in the experimental set-up sections.

Chapter 3

Datasets and Evaluation

This chapter addresses TC1, that of “Can we automatically create resources for the task of automatic stepwise illustration?”. Chapter 2 provides an overall background and describes the previous works analysing the problem of stepwise text illustration and some related problems. Those works involve collecting and annotating datasets from those platforms with respect to the task for which they are used. In this chapter, we discuss some of the datasets that are relevant for our task. We also introduce two new novel datasets in the following sections. We describe the process of dataset construction and systematic cleaning and filtering carried out. In the next chapter, we utilise the publicly available BBC News corpus, to address some of our research questions. Additionally, we have utilised datasets from the cooking domain that are publicly available, described in the further sections. In the last section, we also provide with some commonly used evaluation measures for these datasets.

3.1 Cooking Recipe Datasets

In the cooking domain, the first attempt in exploring automatic classification of food images was the creation of the Food-101 dataset [15] which contained 101K images across 101 categories. Since then, the newer Recipe1M dataset [118] has gained wide attention. The Recipe1M dataset paired each recipe with several images, thereby building a large dataset of 13M food images for its 1M cooking recipes. Recent work [17] proposed a cross-modal retrieval model that aligns images and recipes in a shared representation space, using the Recipe1M dataset.

As this dataset does not offer any sequential data, it cannot be utilised for stepwise text illustration, this association is only between images of the final dish and its corresponding entire recipe text. We build and release a sequenced

recipe dataset, called Stepwise Recipe dataset. It provides, by comparison, an image for every step of the recipe instructions, resulting in a complete sequence of image-text pairs for each recipe in the repository.

RecipeQA [150] is another recent popular dataset in the cooking domain. It is used for multimodal comprehension and reasoning tasks on recipes. The dataset is primarily used for its 36K questions that pertain to the 20K recipes, but in addition it contains illustrative images for each step of the recipes. The RecipeQA dataset has been used in recent work [3] to analyse coherence relations in multimodal image-text contexts, thereby producing a human-annotated corpus that labels coherence relations between the image-text pairs. Different inferential relationships were characterised, and the annotators were asked questions that helped to identify the relationships in each image-text pair. These range from cases where the image directly depicts the action described in the text to cases where the image provides some information about the process or outcome but may either omit or add details. The RecipeQA dataset is comparable to our newly created Stepwise Recipe dataset and reveals similar associations between image-text pairs.

3.2 BBC News Dataset

BBC News [35] is a publicly available dataset consisting of articles trawled from the BBC News website. It satisfies the following criteria:

- First, it is a representative of real-world data, as it is created from BBC news web pages.
- Second, it includes images with annotations that will potentially help in providing visual-textual correspondences
- Third, it also consists of auxiliary information. This auxiliary information, which provides a context to the news article could allow the mining of related linguistic information in order to help us create human readable descriptions.
- Fourth, it also contains gold standard captions for the evaluation of the output produced by the proposed methodologies.

Previously, many image related datasets have been used in computer vision and image retrieval, but they are not directly suitable for caption generation, since they have been created and annotated for different purposes. Yao et al. [151] provides an example for image parsing. Many other examples include,

datasets published for the tasks of image annotation, segmentation, object recognition, scene analysis [94], [84], [44], [24], [8], [115], [120]. It is worth noting, that most of the existing datasets often contain at the maximum, one or two prominent objects against a relatively simple background. Also, in terms of annotation keywords, the range is between 1 and 220. Very few contain full captions. They are also limited in domain, as most of them belong to a specific object category or scene types, for example actions.

Table 3.1: BBC News dataset statistics.

Measure	Value
Number of documents	3,361
Image width	200
Image height	150
Average caption length	9.5 words
Average document length	421.5 words
Caption Vocabulary size	6,180 words
Document Vocabulary size	6,180 words

For the above following reasons, we decided to work on the publicly available BBC news dataset and find it the best fit for the tasks, addressing some of our objectives. It was created by downloading articles from the BBC News website. Table 3.1 provides detailed statistics of the BBC News dataset. The captions tend to use half as many words as the document sentences, and more than 50 percent of the time contain words that are not attested in the document (even though they may be attested in the collection).

The accompanying news article text, as can be seen in figure 3.1, can have denotative or connotative relations with the given image. Denotative refers to describing some of the objects the image depicts, while connotative refers to describing the sociological, political or economic attitudes reflected in the image. The dataset is weakly-labelled. Therefore, this presents several interesting challenges to the problem as well. It is also worth noting, the news images present in this dataset are often cluttered, they display several objects, not only a few prominent ones and consist of complex scenes, and are rendered in a low resolution, adding further challenges. The images, accompanied with collateral text, can be informative and make up for the noise. The presence of background information for a news article, which the corresponding image depicts, or supplements is a highly useful feature of this dataset. Furthermore, rich linguistic information present in the text can be exploited to address the caption generation process with methods related to text summarisation without

extensive knowledge engineering.

3.3 GutenStories

For the task of automatic stepwise text illustration, we present the *GutenStories* dataset and its mini version and show how they can be used to facilitate research in language and vision. Our final GutenStories dataset consists of around 18K visual stories. Each visual story consists of a sequence of text description and image pairs. We first give details about the construction of the dataset. The further subsections also provide some statistics of both versions of the dataset.

In the next section, we first demonstrate the approach to create this unlabelled dataset of sequenced image-text pairs from any source. We release the newly created data repository, *GutenStories*, consisting of 18K visual stories with a total of 90K associated images, respectively, where each segment of text is paired with its corresponding image.

3.3.1 Construction

The construction of this type of image-text parallel corpus has several challenges as highlighted in previous work [68]. The text in a recipe is often unstructured, and therefore we do not have information about the canonical association between image-text pairs. Each image in a recipe is semantically associated with some portion of the text in the same recipe. We assume that the images chosen by the author of the web content to augment the text segments of the post are semantically meaningful. Therefore, we must perform some text segmentation to divide the scraped text into segments that are each expected to be associated with a single image.

Figure 3.2 provides an illustrative diagram of the process of dataset construction. The above process is carried out in a systematic way as illustrated. We also release the source code of the crawler. It consists of the following components:

- **Web Crawler** Given a set of web pages from any web source that might be consisting of multimodal information thematically collocated. The crawler randomly extracts HTTP documents possible consisting of a multimodal story.



Caption: A report estimated more than 2,000 children a year are detained

Ministers are set to admit that they may have significantly under-estimated the number of failed asylum seekers living in Britain, the BBC has learnt. Last year the National Audit Office estimated that the figure could be as much as 283,000 - but at the time the Home Office insisted that was too high. But a trawl of files in the Immigration and Nationality department has produced between 400,000 and 450,000 case files. The news comes as Home Secretary John Reid is set to shake-up his department. Home Office sources say that because of poor record keeping, officials are unable to calculate the exact number of failed asylum seekers, but the figure is far higher than previous estimates. The revelation means that the backlog of asylum cases will therefore take longer to clear...



Caption: A report estimated more than 2,000 children a year are detained

The government is contravening legal guidelines by detaining children whose parents are seeking asylum, a report for a coalition of charities says. The report for the No Place for a Child campaign, co-written by a Labour peer and two opposition MPs, highlights concerns over the issue. Labour's Lord Dubs said there were "workable alternatives" to detention. The Home Office has said detention is used sparingly, especially when children are involved. A report earlier this year for the No Place for a Child campaign estimated that more than 2,000 children were locked up in UK immigration centres. In their report Lord Dubs, Conservative MP John Bercow and Liberal Democrat Evan Harris said the use of detention for children was contrary to standards set by the United Nations. Mr Bercow said detaining children was not a "proportionate response". "It's bad for children, their families, the taxpayer and this country's reputation," he said...

Figure 3.1: BBC News Corpus shows sample news articles containing text, image and caption in the bold.

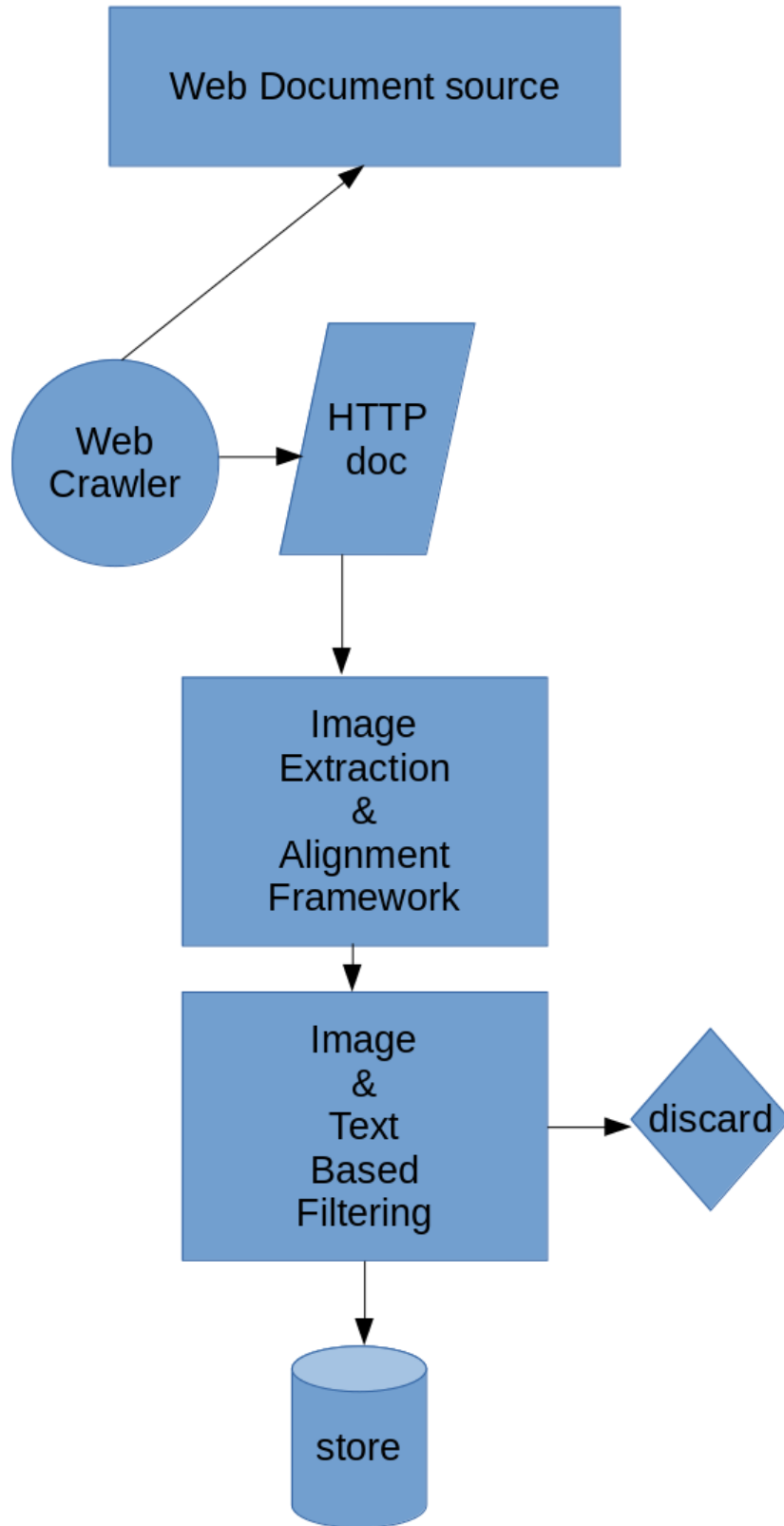


Figure 3.2: Web Crawler for automatic construction of the dataset.

- **Image Extraction and Alignment Framework** The image extraction and alignment framework, first of all segments text based on some existing methods. Then the given images, are allocated to the text segments based on relative image-text distance in the document.
- **Image and Text Based Filtering** We then filter out images based on the encoding format and by discarding images whose either dimension is less than 200 pixels. Finally, we perform text-based filtering by following [122] to ensure high text quality:
 - First, descriptions should have a high unique word ratio covering various Part-of-Speech (POS) tags therefore descriptions with high noun ratio are discarded.
 - Second, descriptions with high repetition of tokens are discarded.
 - Third, some predefined boiler-plate prefix-suffix sequences are also removed.

The GutenStories dataset has been programmatically created by crawling visual story web pages from Project Gutenberg website, following the process described in dataset construction section. Project Gutenberg, as mentioned before is an online catalogue containing more than 57,000 digital books. We also discarded non-English stories and discard stories which contain less than three images. We then performed image-based and text-based filtering, as described in the previous section.

To ensure a balance of cleanliness, informativeness, fluency and learnability of the resulting image, caption pairs, a set of quality control measures are followed ahead of image-based or text-based filtering. These quality measures are presented in the dataset construction section.

Originally, we crawled 600k image-text pairs, based on the above filtering steps, only around 15% of candidates were passed on to the later stages. After filtering, the GutenStories dataset consists of around 18k visual stories and 90k image/text pairs.

3.3.2 mini-GutenStories

We also introduce a mini-version of the GutenStories dataset, mini-GutenStories, containing super fine quality control checks and manual human intervention to remove errors. The mini version was created by hand-picking some key children storybook authors like Beatrix Potter, Roald Dahl, Rudyard Kipling and the

Table 3.2: GutenStories dataset statistics.

Measure	Value
Number of stories	5,774
Minimum number of words	165
Maximum number of words	2,325
Number of images	97,047

like. This version only consists of artificial images and stories particularly for an audience under the age of 9. The mini-GutenStories consist of 6K image, description pairs which can be aligned to 250 visual stories. Example text passage-image pairs are shown in Figure 6.2.

Table 3.3: mini-GutenStories dataset statistics.

Measure	Value
Number of stories	250
Number of authors	25
Minimum number of words	182
Maximum number of words	1,227
Number of images	6,427

Table 3.2 and Table 3.3 provides dataset statistics for the GutenStories and miniGutenStories dataset respectively.

3.4 StepwiseRecipeDataset

We construct the *Stepwise Recipe* dataset in a similar way as described before. It is composed of illustrated, step-by-step recipes from the following three websites: the Simply Recipes¹, Visual Recipes² and Olga’s Flavor Factory³. The primary dataset consists of about 2K recipes with 44K associated images (Table 3.4). After augmenting with RecipeQA [150] recipes, we obtain 10K recipes in total and 67K images (Table 3.5).

Figure 5.1 shows an example of the stepwise instructions and illustrations from a cooking recipe taken from our newly-built dataset. A few selected text recipe instruction steps are shown alongside the full sequence of recipe images. Note that retrieval of an accurate illustration of Step 4, for example, depends on the data model being able to encode the context from the previous steps of

¹<https://www.simplyrecipes.com>

²<http://www.visualrecipes.com>

³<http://www.olgasflavorfactory.com>

the recipe, as the current step adds to pre-existing information acquired from earlier steps.

Table 3.4: (Primary) Stepwise Recipe dataset statistics.

Measure	Value
Number of recipes	2,832
Minimum number of words	202
Maximum number of words	7,325
Mean number of words	1,040.5 \pm 673.2
Number of images	44,341

Table 3.5: (Augmented) Stepwise Recipe dataset statistics.

Measure	Value
Number of recipes	10,350
Minimum number of words	202
Maximum number of words	7,325
Mean number of words	1,180.5 \pm 541.6
Number of images	67,087

The following steps were taken during the dataset preparation:

- Recipes were automatically scraped from the websites and cleaned of HTML tags.
- The final recipes contain the recipe title, author, publication date (if available), and the description including ingredients list and comments uploaded by website users.
- In the recipe body text, images are referenced by image tag of the format "IMAGE IMGxxxxxx" where "xxxxxx" indicates the image ID.
- Each recipe is further supplied with metadata which contains the name of the author, the title, the topic (if available), the date and time of download, the date and time of publishing (if available), the URL, and a list of the recipe's image IDs.
- Each JPEG image is supplied with metadata containing its ID, the time and date of download, and its URL.

Furthermore, as mentioned before, we have augmented our Stepwise Recipe dataset with the RecipeQA dataset [150], which contains illustrative images

for each step of the recipes in addition to data for visual question answering. We follow the same filtering pipeline for images and texts on the RecipeQA dataset, and only augment data points that satisfy the criteria. After merging these two similar datasets, our final corpus contains 10K recipes and a total of 67K images.

3.5 Evaluation Measures

As also described in Chapter 2, Evaluation is an important stage in a typical Machine Learning pipeline. Evaluation of a model’s performance and its ability to generalise to unseen data can be done in various ways. This section discusses various model validation approaches and performance metrics for the tasks relevant in this thesis. Precision and Recall are two important measures in Information Retrieval. Precision is defined as the fraction of relevant data points out of retrieved data points. Recall is fraction of relevant data points amongst actually retrieved. Both are commonly used in Machine Learning and Information Retrieval.

Challenges

However, there are a few challenges involved in evaluating the correct retrieval of an image sequence with traditional metrics like Precision and Recall. As, there is no one visually correct sequence of images. There could be multiple image sequences in storytelling that describe a story without ambiguity. A fair evaluation metric must take into account not only the gold standard sequence, but also other visually correct sequences. Recently, in literature, such evaluation metrics were proposed [111]. These evaluation metrics involve checking for correct salient objects or scenes in the images, rather than checking correct image matches.

Following are some of evaluation measures commonly used. We also discuss some recently proposed evaluation measures and challenges that they overcome:

3.5.1 Precision

Precision is defined to be fraction of relevant items retrieved by a model to the total number of retrieved items. It is assumed that correct items exist as gold-standard items for unseen data. Here, *relevant* refers to those cases where retrieved items are exactly same as gold-standard items. *retrieved* is

the set of all retrieved items.

$$Precision = \frac{|\{relevant\} \cap \{retrieved\}|}{|retrieved|} \quad (3.1)$$

3.5.2 Recall

Recall is defined to be fraction of items that are relevant to the query that are retrieved.

$$Recall = \frac{|\{relevant\} \cap \{retrieved\}|}{|relevant|} \quad (3.2)$$

3.5.3 Recall@K or R@K

The above described retrieval measures can also be studied at a given cut-off rank, by considering only the topmost retrieved items by the system. Here, $topKretrieved$ are the set of top K most recommended items by the system.

$$Recall@K = \frac{|\{relevant\} \cap \{retrieved\}|}{|topKretrieved|} \quad (3.3)$$

3.5.4 Story Recall@K or StR@K

We also introduce a new metric called Story Recall @ K to ease down conditions in stepwise illustration task. One of a challenge associated with metrics like Recall@K are they only measure the degree of exact matches of the retrieved images with regards to the gold-standard images. This might not be appropriate for our text illustration task since a given text segment could be illustrated by multiple images expressing similar semantics. Therefore metrics like Visual Saliency Recall, Textual Saliency Recall which we define ahead were introduced.

The main concept underlying among saliency based recall measures is to redefine what constitutes relevant items among the ones that are retrieved. In Recall@K, only the gold standard entity is considered to be relevant. But in these measures, if both the images retrieved and gold-standard consist of same salient objects, they are considered to be relevant. We define Story Recall @ K measure by marking every image in the same gold-standard image sequence to be relevant. We refer to an image being *storyrelevant*, if it belongs to same sequence as the gold-standard image.

$$StR@K = \frac{|\{storyrelevant\} \cap \{retrieved\}|}{|topKretrieved|} \quad (3.4)$$

3.5.5 Visual Saliency Recall

Ravi et al. [111] introduces Visual Saliency Recall. They train a VGG19

[125] Network on ImageNet for 20,754 categories. They choose the top 10 most probable categories, as they are mostly interchangeable. Therefore, to define Visual Saliency Recall, *visualrelevant* is defined to be cases where the retrieved image is classified in the same category by the trained CNN network as gold-standard image.

Visual Saliency Recall @K or VSR@K is defined using the equation below:

$$VSR@K = \frac{|\{visualrelevant\} \cap \{retrieved\}|}{|topKretrieved|} \quad (3.5)$$

3.5.6 Textual Saliency Recall

Textual Saliency Recall is defined following similar concepts. However, instead of training a Convolutional Neural Network, the corresponding paired text for the retrieved image has entities that overlap with those entities found in the text query. It assumes a parallel image-text corpus. We use Latent Dirichlet Allocation [10] as topic modelling approach to extract topics or entities from the text. *textualrelevant* is defined if between the respective corresponding text segments of retrieved image and the gold-standard image have at least one common entity.

$$TSR@K = \frac{|\{textualrelevant\} \cap \{retrieved\}|}{|topKretrieved|} \quad (3.6)$$

3.5.7 Inception Score

Salimans et al. [117] introduced Inception Score. It is a metric for automatically evaluating the quality of image generative models. It was also shown to be correlating well with human scoring of the realism of generated images from the CIFAR-10 dataset [75].

3.5.8 BLEU

Bilingual Evaluation Understudy scores or also known as BLEU scores are basically the averaged percentage of n -gram matches, for each n -gram you calculate the percentage of matches. The BLEU [102] scores are typically used to evaluate machine translation models. They are calculated based on number of n -gram matches.

Meteor

The Meteor [78] score are computed in a similar way as BLEU scores, except they overcome the limitation of BLEU by also taking synonyms into consideration.

Chapter 4

Neural Caption Generation for News Images

This chapter addresses RQ2, “How can we fuse information from different modalities to summarise the given content for developing context-based models?”. Therefore, in this chapter, we focus on the first step, that of modelling context information of an atomic text passage and image pair. In this part, we focus on summarizing a pair of text passage and an image. More specifically, we focus on the problem of caption generation, in the domain of news images.

4.1 Introduction

There is rich information available on the Internet. Many online news sites like CNN, Yahoo, BBC etc. publish images with their stories and even provide photo feeds related to current events. These news sites are a good resource for multimedia files containing information in the form of videos, images and natural language texts.

News image caption generation, however, is different from the typical image captioning task. The input to news image caption generation is both a news article and its accompanying image, as opposed to the traditional image captioning task where the input is only an image. Hence, rather than enumerating objects in a given image and describing their properties or relationships to each other as in the traditional image captioning task, the output of news image caption generation is informative text not only describing the key semantics conveyed in the given image, but also summarising the content of its relating news article [9].

An example is shown in Figure 3.1. The figure shows two sample data points from the BBC News Corpus. On the right, the text passages are the given news articles, the caption is provided in the bold text under the image.

It can be seen that the captions of news images provide more information than what have been depicted in images only. For example, a reasonable caption for the second image would be “A building”. But its actual caption conveys much more information and it is evident that the text content of news articles would also need to be considered when generating good captions for news images.

News caption generation tools can assist journalists in creating descriptions for the images associated with their articles or in finding images that appropriately illustrate their text. It also helps in increased accessibility of web for visually impaired individuals (blind or people with partially impaired vision) users who cannot access the content of many sites in the same way sighted users can [39].

A wide variety of techniques exist for caption generation ranging from semantic space learning [63], where both supervised and unsupervised methods exist to learn associations between features extracted from image and words, to latent variable models [35]. There are models inspired by information retrieval and instantiations of noisy-channel model [79]. Semantic space learning models learn parameters to map an image to a caption, whereas latent variable models are probabilistic in nature. Recently, there has been a surge of interests in neural caption generation methods due to ground-breaking results produced by deep learning. Mainly, they all have a fundamental architecture in common which is inspired by encoder-decoder architecture from neural machine translation [152] [66] [18]. More details are provided in chapter 2.

In the encoder-decoder models, caption generation is seen as a translation problem where image is translated to a natural language. Convolutional Neural Networks (CNNs) are typically used as an image encoder, whereas Recurrent Neural Networks (RNNs) are used for decoding sentences, because of their sequence modeling capability. Although there are other variants proposed, for example, with attention mechanisms included, the encoder-decoder architecture is at the heart of these methodologies [147].

As also presented in Chapter 2, existing work in news image captioning generation is scarce. An early approach tackled the problem with a two-stage process, content selection and surface realization. The first stage consists of an image annotation model, where a given image is tagged with a set of keywords based on topics learned from both news article texts and images using a variant of Latent Dirichlet Allocation (LDA) [11]. The second stage uses extractive and abstractive summarisation techniques in forming a sentence from these set of keywords. Word-based models are highly specific in

nature and may result in ambiguous results. There is a need of sentential integration with the images, as a sentence describes an image without any ambiguity.

In this chapter, we propose a sequence-to-sequence deep neural-networks (NN) based model to address the news image caption generation problem. Specifically, we first encode each sentence of a given news article using an order-embedding vector and extract semantic features from the accompanying image using a pre-trained CNN Network, which are further projected to the same semantic space, such that both text and image vectors reside in a common semantic space [136]. We then feed the sentence vectors together with the image vector to a Long Short-Term Memory (LSTM) network [116] to generate a vector representation of the image caption. Finally, we use the generated vector to retrieve the most similar sentence from the original news article based on cosine similarity measurement as the caption of the given image. We also explore a number of variants of our proposed architecture and compare them with the previous work on the news image captioning task.

Our experimental results on the BBC News Corpus show that our proposed strategies outperform traditional methods according to automatic evaluation metrics like BLEU scores [102] and are comparable in terms of Meteor Scores [78]. Since automatic evaluation metrics are currently limited by their capability to measure the quality of caption generation models, a human evaluation experiment has also been conducted, where users were shown the news articles from our test dataset.

Our evaluation results show that captions generated by our proposed approach were more favoured than captions generated by an existing model based on LDA. In what follows, we first discuss related work and then describe our proposed methodology, followed by experiments and results, and finally conclude this chapter.

We make the following contributions:

- We provide a comprehensive analysis of different ways of modelling multimodal information.
- We show that incorporating a sentential structure into the modelling framework is beneficial as compared to previously published methods.
- We perform an exhaustive analysis of features relevant to news image caption generation or retrieval.
- We propose a novel deep neural-networks based methodology that uses

LSTM cells to process and caption a given pair of news article and its accompanying image.

- We compare the proposed methodology with several existing algorithms in terms of various evaluation measures.
- We also provide human judgement results.

4.2 Problem Formulation

Our problem is formulated as follows: given a news image I , and its associated article D , create a sentence description S that best describes the image given D . The training data thus consists of document-image-caption tuples like the ones shown in Figure 3.1. During testing, we are given a document and an associated image for which we need to generate a caption.

4.2.1 Dataset

As highlighted in Chapter 3, most of the existing datasets in this domain have been specifically designed for the problems at hand. For example, the dataset created by Farhadi et al. [32] and Hodosh [52] contain image descriptions but they are limited to specific object categories and scene types.

The dataset used in this chapter is the BBC News Corpus that contains news articles scraped from the BBC website. The details on data collection process, number of instances, class imbalance and evaluation procedure have already been provided in Chapter 3. Note, the BBC News Corpus is a weakly-labelled dataset, which treats the captions and associated news articles as image labels.

We use the BBC News dataset collected in Feng and Lapata [35], which contains 3,361 news documents in total. The dataset covers a wide range of topics. Each news article consists of a text article, an image which are normally 200 pixels wide and 150 pixels high, and a caption of the image which has an average length of 20.5 words. On an average each news article contains 421.5 words. The caption vocabulary is 6,180 words and the document vocabulary consist of 26,795 words. The vocabulary shared between captions and documents is 5,921 words. Some example news articles with their accompanying images and image captures are shown in Figure 3.1. The original dataset was split into a training set consisting of 3,115 news articles, and a test set consisting of 237 remaining news articles.

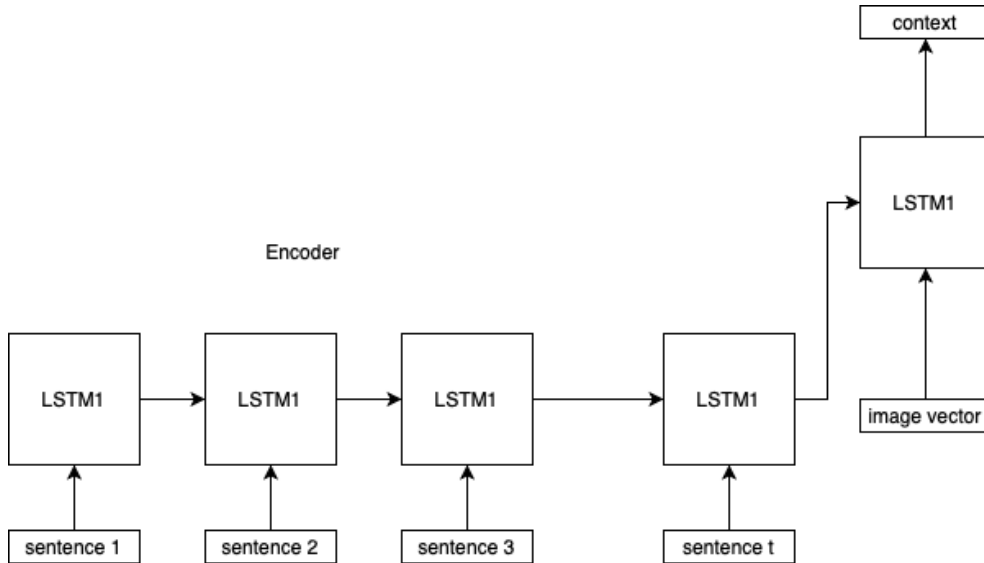


Figure 4.1: Our proposed deep Neural Network (NN) architecture for news image caption generation.

4.2.2 Existing Methods

Most of the relevant literature for this problem falls to image captioning domain. However, very scarce literature exists for the specific problem of news images. As discussed in chapter 2, an earlier methodology for this task proposed an LDA based methodology. Therefore, we consider it to be in our list of baselines for model benchmarking.

4.3 Proposed Methodology

In this section, we propose a novel deep Neural Networks (NN) based architecture to automatic caption generation of news images. Figure 4.1 provides a block diagram of the model architecture. We first convert sentences in a news article into a sequence of vectors using a pre-trained order-embedding model. For more details, refer to Vendrov et al. [136]. We then encode the accompanying image into an image embedding using the pre-trained Oxford VGG network [123] as an off-the-shelf feature extractor. The VGG features are further projected to the same order-embedding space. Both sentence and image vectors are represented in a 1,024-dimensional semantic space. The sequence of sentences from the news article thus convert to a sequence of vectors, followed by the encoded image vector.

The sentence vector sequence is then fed to a LSTM network, which is a specific type of Recurrent Neural Network (RNN). The output of the network is

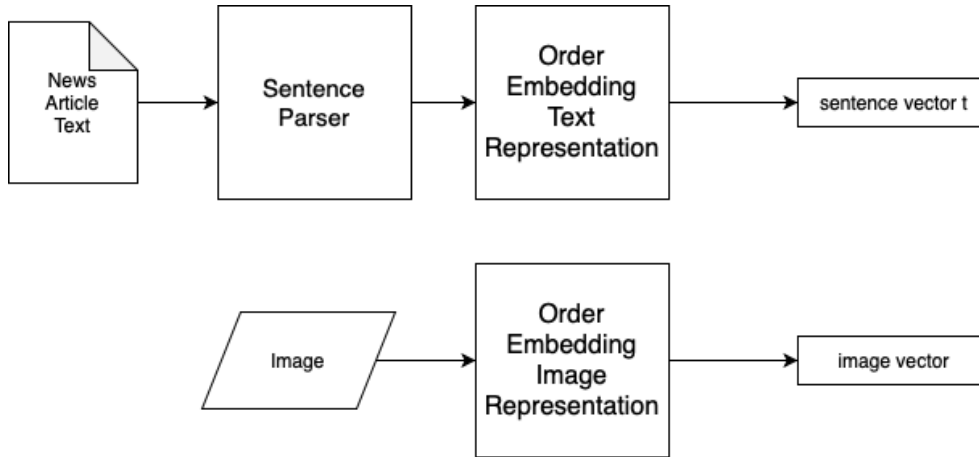


Figure 4.2: Text and Image Representation Mechanism

fed into another LSTM cell which also takes the image vector as an additional input. The final output is considered as a representation which captures the semantics conveyed in both text and image. The cross entropy between the output vector and caption order-embedding vector is used as an objective function to train the LSTM parameters.

4.3.1 Text and Image Representation

For encoding sentences, we use a pre-trained order-embedding model [136] to encode sentences using distributed representations. Order-embeddings exploit the partial order structure of the visual-semantic hierarchy by learning a mapping between sentences and semantic vector space. This projects each sentence into a 1024-dimensional embedding space.

For encoding images, we first use a pre-trained Convolutional Neural Network (CNN), which is an important class of learnable representations applicable, among others, to numerous computer vision problems. Deep CNNs, in particular, are composed of several layers of processing, each involving linear as well as non-linear operators. We use pre-trained Oxford VGG network as an off-the-shelf feature extractor. The whole network consists of 22 layers. We use the fc7 features, that is the output of the penultimate fully connected layer, as a representation for the image. The VGG features are projected to same order-embedding space, where sentence vectors reside. As such, both image and sentence vectors reside in a common semantic space which enables direct comparison between them. Figure 4.2 provides an illustrative diagram explaining the image and text framework.

4.3.2 LSTM Training

RNNs surely do a great job at modelling sequences. Unfortunately, the short-coming of such networks is that they are unable to carry forward information when the length of the chain grows beyond a measure. This is called vanishing gradient effect. To solve this problem, a forgetting mechanism has been proposed in LSTM. LSTMs have many variations. One cell consists of three gates i.e. input, output and forget. Gates typically use sigmoid activation, while input and cell state is often transformed with the hyperbolic tangent function, \tanh .

At timestep t , an LSTM has two inputs, x_t the input vector at that timestep and h_{t-1} , the hidden state vector of previous timestep. All the W are weight matrices and b are biases, which are learnable model parameters. In the forward pass, this is how updates are done in the input gate i_t , forget gate f_t , the output gate o_t , the input transform cin_t is taken and the state c_t and h_t is updated in this manner.

$$\begin{aligned}i_t &= g(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\f_t &= g(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\o_t &= g(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\cin_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_{cin}) \\c_t &= f_t c_{t-1} + i_t cin_t \\h_t &= o_t \cdot \tanh(c_t)\end{aligned}$$

In encoder-decoder based models, information is encoded to a context vector which is then fed to the decoder. For example, in Machine Translation, the encoder is Recurrent Neural Network (RNN), the sentence is the source language is encoded to a fixed-length context vector which is fed to another RNN decoder. Also, in image captioning, the image is encoded using CNN to a fixed-length vector, which is then fed to the decoder RNN. Our model architecture consists of an RNN decoder as well. For encoding, we use the representation framework as described above.

At training time, in the forward pass, both sentence vectors and an image vector are fed to a LSTM network to obtain a context vector, as shown in Figure 4.1. It is assumed that the context vector summarises the information conveyed in both textual and visual formats. The decoder uses this information

to generate a caption for the article. We also encode the image caption with the order-embedding model. We use the cross-entropy between the output of the LSTM network and the order-embedding vector of the image caption as the loss function to backpropagate and update model parameters. We set the learning rate to 0.6, momentum to 0.9 and train the model with 30 epochs using stochastic gradient descent.

During testing, given a news article and its accompanied image, we retrieve the most relevant sentence from the article based on the cosine similarity measurement between the output vector from the LSTM and the order-embedding vector of each sentence. We consider this as an extractive summarisation approach, although it could be extended to be an abstractive summarisation technique as well, in which a caption can be generated, details of which will be discussed later.

Algorithm 1 provides instructions for the main algorithm for the proposed LSTM-based methodology. As it could be observed in the algorithm, training data consists of tuples of the form (I, D, C) , where I is the raw image from the BBC News Corpus, D is the text document i.e. the accompanying news article in the corpus and C is the given caption in the dataset. Please note that C is a one-sentence caption, that may or may not be present in the article. The machine learning pipeline involves representing all the training and testing corpus to be present in numerical form, as also highlighted in Chapter 3. Therefore, we use a pretrained CNN architecture to convert all raw images in vector form. We use VGG19 [125] architecture, which is previously trained on image classification task. β_1 represents the semantic features obtained after passing it through all convolution and transformative layers of the CNN. We extract the features from the penultimate layer fc7. Therefore β_1 represent semantic features of the image.

Algorithm consists of instructions for converting the text in document to vector form. News article is first parsed using a sentence parser, we use Stanford NLP Toolkit [92]. These sentences are further projected to pretrained Order Embedding Space. Furthermore, the image features β_1 are also transformed to same order embedding space as that of sentences. β_3 represents the vector obtained by feeding the caption text to order embedding space using the same encoder, as that of sentences. All the sentence vectors are fed as inputs to the LSTM. At the last time step, image feature vector is also fed to the LSTM. Then the output of the LSTM at this time step is used to calculate the cross entropy loss with the β_3 . This loss function is used to backpropagate over LSTM parameters.

Algorithm 1 LSTM training for the proposed methodology

```
1: for batch in trainData do
2:
3:   for  $(I, D, C) \in \{(I_1, D_1, C_1), \dots, (I_N, D_N, C_N)\}$  do
4:     Generate  $\beta_1 = VGG(I)$ 
5:     Generate  $\beta_2 = OrderEmbeddingImageRepr(\beta_1)$ 
6:     SentencesList  $S = SentenceParser(D)$ 
7:
8:     for sentenceS in SentencesList do
9:       Generate  $\beta_s = OrderEmbeddingTextRepr(S)$ 
10:       $LSTM\_Input(\beta_s)$ 
11:    end for
12:     $LSTM\_Input(\beta_2)$ 
13:     $o = LSTM\_Output()$ 
14:    Generate  $\beta_3 = OrderEmbeddingTextRepr(C)$ 
15:     $LSTM\_backpropagate(CrossEntropyLoss(\beta_3, o))$ 
16:  end for
17: end for
```

Algorithm 2 provides instructions for the main testing procedure for the proposed methodology. Most of the instructions are similar to the training procedure and involve feeding forward all the inputs to their respecting components to conduct appropriate data transformations. After all required operations are performed, LSTM outputs the vector o . For extractive summarisation, we find the sentence nearest to the vector in the order embedding space and output this as the generated caption vector. This output caption can be used to compute several evaluation metric scores like BLEU, Meteor etc.

The methodology is inspired by encoder-decoder architecture of neural machine translation. The basic framework consists of RNN-RNN architecture. Recurrent Neural Networks are used, because of their great ability to model sequences. In natural language processing, An RNN has been found a great way to model sentence and to learn the grammatical rules of the language. This is because of the parameter sharing ability of these models across time index. A convolutional neural network across different of a sentence has also been used for modelling and is the basis for time-delay neural networks, introduced in Lang et al. [77]. This is because the convolution operation also allows the network to share parameters through time but is shallow. The output of the convolution is a sequence where each member of the output is a function of a small number of neighbouring members of the input. Therefore, we use RNNs, more specifically gated RNNs, as we describe ahead to model sentences of the news article. Before describing our main methodology, we describe the some of

Algorithm 2 main testing procedure for the proposed methodology

```
1: for batch in testData do
2:
3:   for  $(I, D, C) \in \{(I_1, D_1, C_1), \dots, (I_N, D_N, C_N)\}$  do
4:     Generate  $\beta_1 = VGG(I)$ 
5:     Generate  $\beta_2 = OrderEmbeddingImageRepr(\beta_1)$ 
6:     SentencesList  $S = SentenceParser(D)$ 
7:
8:     for sentenceS in SentencesList do
9:       Generate  $\beta_s = OrderEmbeddingTextRepr(S)$ 
10:       $LSTM\_Input(\beta_s)$ 
11:    end for
12:     $LSTM\_Input(\beta_2)$ 
13:     $o = LSTM\_Output()$ 
14:    Generate  $o_c = NearestNeighbour(o)$ 
15:     $ComputeBLEUScores(C, o_c)$ 
16:     $ComputeMeteorScores(C, o_c)$ 
17:  end for
18: end for
19:  $ComputeAverageBLEUScores()$ 
20:  $ComputeAverageMeteorScores()$ 
```

the fundamental principles of the encoder-decoder architecture, and how it is used in machine translation, of which we have drawn inspiration.

4.3.3 Variant Architecture

In this section, we discuss a variant architecture to our model. As, there are multiple ways, in which sequential information can be propagated through an LSTM network. Another variant of the proposed architecture is to feed the image vector at each timestep of the LSTM such that the input to each LSTM cell is a concatenation of a sentence vector and the image vector. Figure 4.3 shows a variant of our proposed architecture which is called the Deep NN Dual Architecture.

4.4 Experiments

Experiments are conducted to evaluate the performance of the proposed model and compare its performance with alternative approaches. We created a similar train-test split, as in Feng and Lapata [35], of image-text-caption triplets in the BBC news dataset.

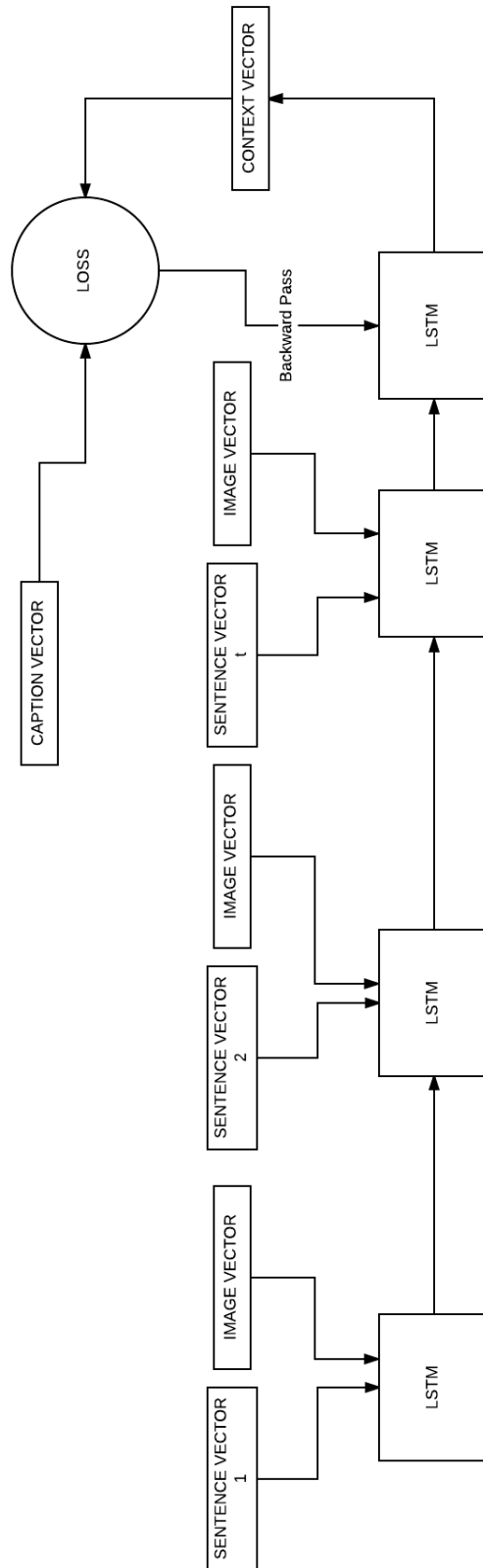


Figure 4.3: A Deep NN Dual architecture for news image caption generation.

4.4.1 Methods

LDA-based (KL)

We reproduced the results from Feng and Lapata [35]. For content selection, we first synthesized textual and visual dictionaries where a textual dictionary was created by assigning a unique token id to each word present in any of the articles and visual dictionary was made by clustering SIFT descriptors into 2,000 different visual words. We then trained an LDA model with 1,000 topics on the BBC news dataset containing both text and images. For surface realization, we only used extractive summarisation. It has been shown in Feng and Lapata [35] that retrieving sentences based on the Kullback-Leibler (KL) divergence between the topic distribution of a sentence and the topic distribution of a news article with its accompanying image gives the best results in terms of human evaluation. Therefore, we picked their best performing methodology for comparison.

LDA-based (word-based)

For extractive summarisation, word-based overlap strategy was also implemented, where the sentence with maximum overlap between extracted annotations and words in the sentence is picked.

Nearest Neighbour

We also implemented a Nearest Neighbour approach in the order-embedding space. Since both sentences and images are projected to the same semantic space, we can simply choose the sentence which is nearest to a given image as its caption. We use cosine similarity measurement to calculate the similarity score between a sentence vector and an image vector.

Ablation Experiments

We also conduct an ablation study of the proposed method by only feeding text only input and also conduct experiments with a variant architecture, which involves fusing information from different modalities in a different way.

Deep NN (text input only)

We explore a variant of our proposed architecture where the input is only text from news articles. This is similar to news headline generation based on text

input only except that what we generated here are image captions.

Deep NN (dual)

This is the variant of the architecture shown in Figure 4.3 where the input to an LSTM cell at each timestep is the concatenation of a sentence vector and the image vector.

4.4.2 Evaluation Metrics

We compare the generated image captions with the provided gold-standard captions. We compare the methodologies, both on the basis of automatic as well as human evaluation methods. For automatic evaluation, we compare using both BLEU and Meteor scores.

Human Evaluation

Apart from objective evaluation using BLEU and Meteor, we have also invited human participants to evaluate the generated results by various models. For human evaluation, we invited 16 human evaluators to choose between the caption generated by the baseline models and our approach for each pair of news article and image presented to them. If human evaluators found none of the captions generated can describe the image well, they can choose the option by selecting the “none” category.

4.4.3 Results

In this section, we first present the effects of incorporating sentential integration, by comparing against the proposed the neural-network based methodology and the given LDA based methodology.

Method	BLEU
LDA-based (KL)	0.3002
Nearest Neighbour	0.3237
Deep NN (text only)	0.3315
Deep NN (dual)	0.3303
Deep NN	0.3427

Table 4.1: News image caption generation results in terms of BLEU scores.

We compare our proposed NN approach with the baseline model based on LDA [37] using both objective evaluations including BLEU and Meteor. We also experimented with a variant of our model using only text content of news

Method	Meteor
LDA-based (KL)	0.0706
Nearest Neighbour	0.0672
Deep NN (text only)	0.0642
Deep NN (dual)	0.0609
Deep NN	0.0677

Table 4.2: News image caption generation results in terms of Meteor scores.

articles as input to our model.

The BLEU scores are shown in Table 4.1. It can be observed that the simple Nearest Neighbour approach already outperforms the LDA-based method in terms of the BLEU score. Deep NN with text input only improves Nearest Neighbour slightly on BLEU. Deep NN (dual) performs almost the same as Deep NN (dual). This shows that feeding an image vector at each time step somehow diffuse the semantic information captured in images. Our model (deep NN), where the image vector was only fed in the last timestep in the LSTM network, gives the best overall BLEU score of 0.3427, which outperforms the LDA approach by 4%.

The Meteor scores are shown in Table 4.2. In terms of Meteor scores, both Deep NN and Nearest Neighbour give similar results and they slightly outperform other variants of the deep NN model. Deep NN also performs on par with LDA since the difference of their Meteor scores is only 0.003.

The Tables 4.3 and 4.4 show results of the proposed methodoly Deep NN obtained by feeding, different image features. Here, we compare results based on image classification algorithms as described in He et al. [49].

Features	BLEU
VGG	0.3215
ResNet152	0.3310
Order-Embedding	0.3427

Table 4.3: Comparison of BLEU scores over different image features

Features	Meteor
VGG	0.0542
ResNet152	0.0613
Order-Embedding	0.0677

Table 4.4: Comparison of Meteor scores over different image features

For human evaluation, the study as described in the previous section was conducted by developing a web-based app, it can be observed that 38.3 percent of times, the caption generated by our approach was selected as the most appropriate image description by the users, whereas only 28.8 percent of times, the caption generated by the LDA-based model was preferred. We also notice that a staggering 32.91 percent of times, no caption was picked by the users, which could be due to the limited capability of extractive summarisation techniques. Figure 4.5 shows qualitative study of generated captions.

When only using text content of news articles as the input to our NN architecture, the original model reduces to one-sentence summarisation based purely on text content. As expected, without taking into account the image information, the model has a difficulty in producing appropriate description of a given image. As such, the results are worse than the full approach taking both text and image as input.

Method	Human Evaluation
LDA-based (KL)	28.8%
Deep NN	38.3%

Table 4.5: Human Evaluation results.

LDA Methodology gave an average BLEU score of 0.3508 and an average Meteor score of 0.06524, whereas an average BLEU score of 0.3358 and an average Meteor score of 0.077409 was observed for the proposed model. This high BLEU score in LDA methodology is evident because of the fact that sentence selection algorithm picks word with maximum overlap from the set of sentences. Although, a manual inspection of the outputs was done, and captions were of poor performance. A BLEU score of 0.2915 was observed for only sentence based neural captions.

4.5 Error Analysis

In this section, we present more results from the experiments conducted. Figure 4.5 shows three cases of results. The first case, shows the case, where majority of users picked “Deep NN” caption as a right caption for the given article. In this case, Deep NN methodology is clearly able to identify the subject “Chris Langham” in the picture. It is also able to capture background knowledge of the article. The second case, is where the majority of users picked “LDA” caption as a right caption for the given article. In this case, LDA method-



Ireland's state-owned airline Aer Lingus is to be partly privatized on September 29, it has been announced. The government will sell more than 50% of its shares to investors, with staff retaining their current 15% holding. The sale is expected to raise 400m euros allowing the firm to buy aircraft and expand its operations. The Irish government insists Aer Lingus needs to be partly sold to guarantee its future but union leaders say that it could result in job cuts. "This is the right decision for the company, its employees, customers and Ireland," Ireland's transport minister Martin Cullen said.

Caption - New planes will be bought with money raised from the float
LDA- The price of shares will be set shortly before the flotation, after which the government will retain a stake of at least 25% in the company.
Neural- Ireland's state-owned airline Aer Lingus is to be partly privatized on September 29, it has been announced.



The call is made by a man identified on the film as "Azzam the American", a convert also known as Adam Gadahn who is wanted for questioning by the FBI. He says ignorance of Islam leads Westerners to accept wars waged by their governments and Israel against Muslim countries. The video opens with an introduction by al-Qaeda number two Ayman al-Zawahiri. Separately, Palestinian militants who held two Fox News journalists hostage for almost two weeks last month vowed to target all non-Muslims who entered the Palestinian territories....

Caption - The video is introduced by al-Qaeda deputy leader Zawahiri
LDA- "Any infidel blood will have no sanctity," the Holy Jihad Brigades group said in a statement posted on the internet.
Neural- Al-Qaeda has urged non-Muslims - especially in the US - to convert to Islam, according to a new videotape.



A libel action brought in London by a UK TV presenter against Arnold Schwarzenegger and two of his aides has been settled, reports say. Anna Richardson was suing the California governor and two aides over comments they made about her claims that he groped her in December 2000. The parties' lawyers said in a joint statement that they were satisfied, the Associated Press reports. The Terminator star faces a re-election battle for governor in November. Details were not given of the settlement, which was confirmed for Reuters by one of the aides, former film publicist Sheryl Main. "Yes, it is true that it has been settled," she told the news agency by telephone. "I don't have any comment but it's settled and I think we're all very happy to put that behind us." No comment from Anna Richardson was immediately available.

Caption - Schwarzenegger faces a re-election battle in November
LDA- Anna Richardson was suing the California governor and two aides over comments they made about her claims that he groped her in December 2000.
Neural- The Terminator star faces a re-election battle for governor in November.

Figure 4.4: Sample Generated Captions providing comparison between LDA and NN methodology results

ology is able to identify the subject. However, the third case shows, where majority of users picked “No” caption as a suitable description for the given article. This is an example case, where both “LDA” as well as “Deep NN” methodologies have failed to capture the content of the articles. It is quite a challenging case. The gold standard caption is “Parts of Charlie and Chocolate factory were also filmed there.”, which is not clearly evident from the image.

38.3% of times, “Deep NN” caption has been picked as a right choice by the users. 32.9% of times, “No” caption has been picked as a suitable choice. 28.8% of times “LDA” caption has been picked as a right choice by majority of users.

We believe, this significant 32.926% of no caption is because of the limited applicability of extractive generation techniques. As many a times, there is no caption in the database, that best describes a given case.

4.6 Conclusion

In this chapter, we have proposed a novel deep NN-based architecture for the task of automatic caption generation for news images. The experimental evaluation on the BBC News corpus show that proposed methodology gives a better BLEU score than baseline models and performs similarly compared to the LDA approach on Meteor scores. Nevertheless, we notice that the captions generated by our approach were favoured over the captions generated by the LDA based model most of time by human evaluators. This chapter studied the problem on an atomic image-text passage, the next chapter provides a more formal introduction to problem of stepwise illustration.

Mr Langham, from Cranbrook, Kent, was previously charged in May with 15 separate counts of making indecent images of children. Mr Langham received the new charges when he answered bail at a police station in Kent. In a statement, the married actor said he was "determine to clear my name". In a further statement issued through the BBC, Mr Langham said he will withdraw from all BBC projects "until these matters are resolved". He has been bailed to appear at Sevenoaks Magistrates' Court on Thursday. He won a Bafta for best comedy performance in May this year. The award was for his portrayal of government minister Hugh Abbot in the BBC series 'The Thick of It'. Earlier on Wednesday, the corporation confirmed the actor would not be returning in a special Christmas edition of the programme. But a BBC spokeswoman denied he had been axed from the show, and said his absence was due to the show's focus...



Case 1:- Majority of users picked **Deep NN** Caption as the right caption

Gold Standard:

Chris Langham stars in BBC TV series The Thick of It

LDA: Mr Langham is a familiar face on BBC television shows, including his spoof documentary People Like Us, which transferred to the small screen from BBC Radio 4.

Deep NN: Comic actor Chris Langham has been charged with eight counts of indecent assault and one other sexual offence, police have said.

The Catholic Church has accused a BBC documentary of a "deeply prejudiced attack" on the Pope over claims of a systematic cover-up of child sex abuse. Panorama examined a document which allegedly encourages secrecy in dealing with cases of priests abusing children. It says this was enforced by Cardinal Joseph Ratzinger before he became Pope. The Most Reverend Vincent Nichols, Archbishop of Birmingham, said the claim was "entirely misleading" but the BBC said it stood by the programme. 'Misuse of the confessional'. The document called Crimen Sollicitationis was written in 1962 and apparently instructed bishops how to handle claims of child sex abuse. Programme makers asked Father Tom Doyle, a former church lawyer who was sacked from the Vatican for criticising its handling of child abuse, to interpret the document. He said it was an explicit written policy to cover up cases of child abuse, which stressed the Vatican's control and made no mention of the victims. The Catholic Church said the document was not directly concerned with child sex abuse, but with the misuse of the confessional...



Case 2:- Majority of users picked **LDA** Caption as the right caption

Gold Standard:

Archbishop Nichols said the BBC should be ashamed

LDA: The Catholic Church has accused a BBC documentary of a "deeply prejudiced attack" on the Pope over claims of a systematic cover-up of child sex abuse.

Deep NN: A BBC spokeswoman said the BBC has a well-defined complaints system and would reply to the letter once they receive it.

The James Bond stage destroyed by fire at the weekend "will need to be demolished and rebuilt", according to a statement from Pinewood Studios. The cause of the blaze at Iver Heath, Buckinghamshire, which left the celebrated stage completely gutted, has yet to be confirmed. However, Pinewood said the rest of its studios would be fully operational "by the end of today". The stage was housing sets built for Casino Royale, the next Bond movie. No filming was taking place at the time and there were no casualties. "The production had completed shooting and was in the process of removing its film sets," said Pinewood. It said the studio had "well established procedures" to deal with fires which had proved effective. "The Board has not been able to assess the full effects of this incident," the statement continued. Buckinghamshire Fire Brigade were alerted at 1118 BST on Sunday. At least eight fire engines tackled the blaze, the smoke from which was visible from up to 10 miles away...



Case 3:- Majority of users picked **No** Caption as the right caption

Gold Standard:

Parts of Charlie and the Chocolate Factory were also filmed there

LDA: It is the second time the stage, originally built for the 1977 Bond film The Spy Who Loved Me, has been destroyed by fire.

Deep NN: Buckinghamshire Fire Brigade were alerted at 1118 BST on Sunday. At least eight fire engines tackled the blaze, the smoke from which was visible from up to 10 miles away.

Figure 4.5: Error Analysis

Chapter 5

Variational Recurrent Sequence-to-Sequence Retrieval for Stepwise Illustration

This chapter addresses the RQ3, that “How can we study the automatic stepwise illustration systems in a domain-constrained setting, given narrative text passage in a limited domain with a sequence of illustrations, considering and incorporating prior context?”. Therefore, in this chapter, we propose the novel Variational Recurrent Sequence to Sequence Retrieval model and employ it on the Stepwise Recipe Dataset. We study and compare its performance with several relevant and competitive baselines.

5.1 Introduction

There is growing interest in cross-modal analytics and search in multimodal data repositories. In this chapter, we propose a variational recurrent learning model to enable seq2seq retrieval, called Variational Recurrent Sequence-to-Sequence (VRSS) model, which produces a joint representation of the image-text repository, where the semantic associations are grounded in context by making use of the sequential nature of the data. Stepwise query results are then generated by searching this representation space. More concretely, we incorporate the global context information encoded in the entire text sequence through the attention mechanism into a Variational Autoencoder (VAE) at each time step which converts the input text into an image representation in the image embedding space. In order to capture the semantics of the images retrieved so far (in

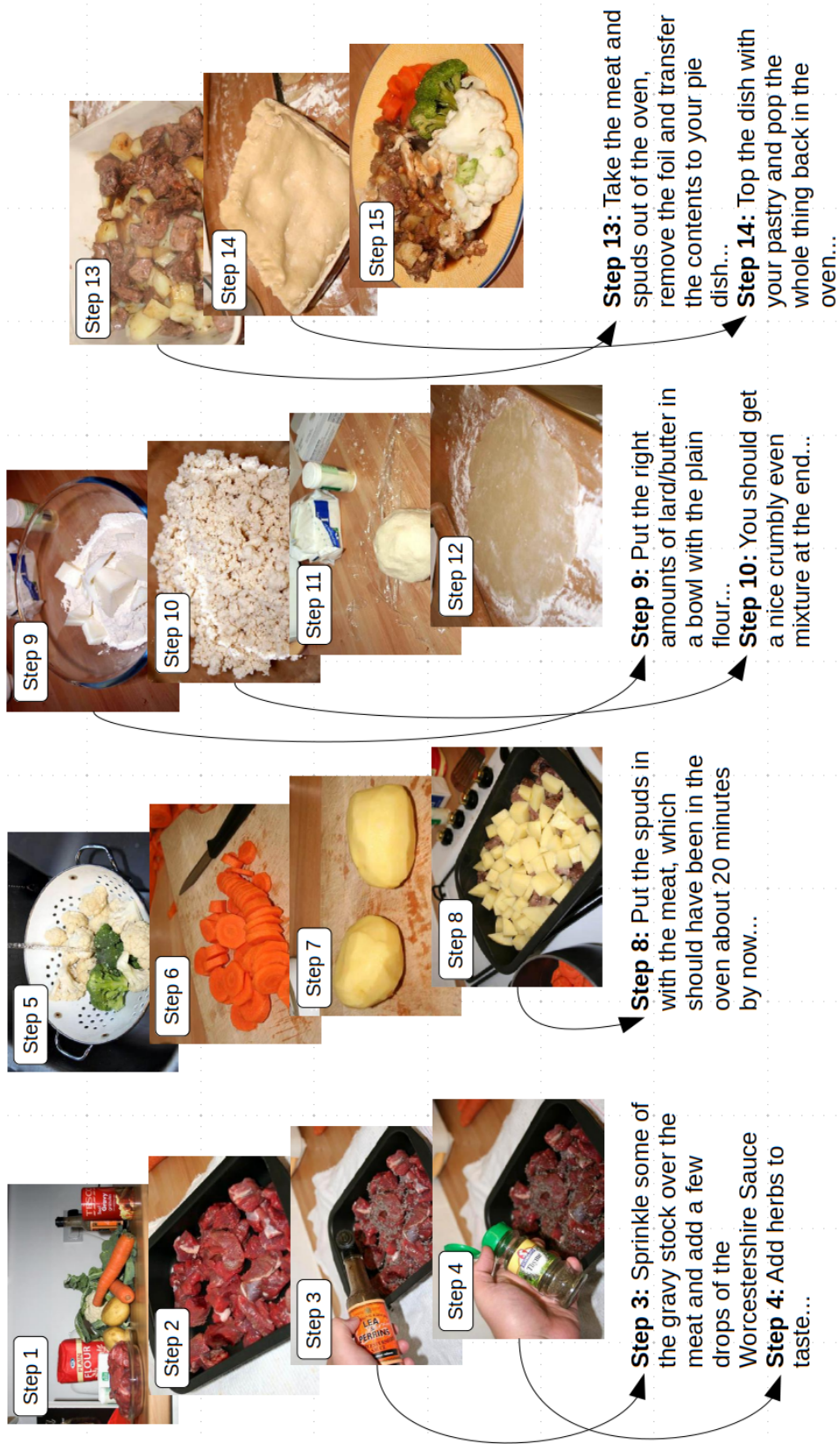


Figure 5.1: Stepwise Recipe illustration example showing a few text recipe instruction steps alongside one full sequence of recipe images. Note that retrieval of an accurate illustration of Step 4, for example, depends on the model being able to use context information that was acquired in earlier steps.

a story/recipe), we assume the prior of the distribution of the topic given the text input is no longer a standard Gaussian distribution, but follows the distribution conditional on the latent topic from the previous time step. By doing so, our model can naturally capture the sequential semantic structure of text/image.

The model is further used to search the data repository and generate step-wise query results in a sequential manner. A benefit of the proposed learned cross-modal embedding approach is that the repository may be easily extended without the need for repeatedly indexing the entire data. Once the embedding model is designed, any number of new inputs may be immediately inserted into the same existing space of data points. This is a desirable feature that may be useful in many related domains. For instance, the challenge of searching and indexing audio streams has been studied [142] where there is a need for real-time insertions of live audio streams into the index to include them in query results.

Our main contributions, in this chapter, can be summarised below:

- We formalise the task of *sequence-to-sequence (seq2seq) retrieval* for step-wise illustration of text.
- We propose a new *variational recurrent seq2seq (VRSS) retrieval model* for seq2seq retrieval, which employs temporally-dependent latent variables to capture the sequential semantic structure of text-image sequences.
- We study the effectiveness of the proposed methodology by comparing against several evaluation results on the stepwise recipe dataset.
- We also conduct a human evaluation study to test the effectiveness of the proposed methodology.
- We conduct an error analysis study of different cases to analyse possible reasons where the model may or may not work.
- We provide a hubness analysis of the embedding spaces learnt using the model and make appropriate inferences.

Our new VRSS model outperforms several cross-modal retrieval alternatives on this dataset, using a variety of performance metrics. Human evaluation of the retrieved illustrations also confirms that our model produces joint representations that are semantically meaningful. Furthermore, we qualitatively analyse the representations generated from VRSS and its competing model. Towards this aim, we analyse the quality of our embeddings and also show that our model outperforms several retrieval baselines.

5.2 Variational Recurrent Seq2seq (VRSS) Retrieval Model

In this section, we propose a novel variational recurrent sequence-to-sequence (VRSS) model. The proposed VRSS model also introduces temporally dependent latent variables to better capture the complex interplay among the text passages and images. Different from existing approaches, we have taken into account the global context information encoded in the whole query sequence and used VAE for cross-modal generation. The Figure 5.2 provides an illustrative diagram of latent variables generating the image space. We use the VAE model to convert the text into a representation in the image embedding space, instead of using it to reconstruct the text input. Finally, we used the max-margin hinge loss objective function to enforce that the converted text embedding must be close to its paired image embedding.

The Variational Recurrent Sequence-to-Sequence Retrieval (VRSS) model is based on the attentional neural encoder-decoder architecture [128], which is central to neural machine translation (NMT). Within this framework, semantic representations of the source and target sequences are learned in an implicit way. It is motivated by the recent successful applications of variational recurrent neural networks (VRNN). In the original paper describing such a VRNN model, Chung et al. [21] advocate this combination of variational autoencoder and recurrent network as a way to model both sequential dependencies as well as complex multimodal distributions. Figure 5.2 demonstrates the hypothesis behind, that of approximating the image semantic space with a hidden latent space.

5.2.1 Problem Formulation

The seq2seq retrieval task is formalised as follows: given a sequence of text passages, $x = \{x_1, x_2, \dots, x_T\}$, retrieve a sequence of images, where

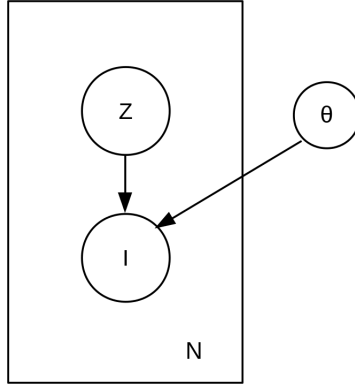


Figure 5.2: Latent Variables

each i_t is the image representation, $i = \{i_1, i_2, \dots, i_T\}$ (from a data repository) which best describes the semantic meanings of text passages, i.e., $p(i|x) = \prod_{t=1}^T p(i_t|x, i_{<t})$. For learning, we consider a training set (e.g., recipes or stories) $S = \{S^1, S^2, \dots, S^N\}$, where each S^n consists of a sequence of images and their associated text. Each such paired sequence is represented as $S^n = \{(x_1^n, i_1^n), (x_2^n, i_2^n), \dots, (x_{|S^n|}^n, i_{|S^n|}^n)\}$ where (x_1^n, i_1^n) is the first paired text and image, S^n , (x_2^n, i_2^n) is the second pair, and so on. That is, each text sequence $x^n = \{x_1^n, x_2^n, \dots, x_T^n\}$ and each image sequence $I^n = \{I_1^n, I_2^n, \dots, I_T^n\}$ has been paired element-wise.

We address the seq2seq retrieval problem by considering three aspects:

1. encoding the contextual information of text passages
2. capturing the semantics of the images retrieved (in a story/recipe)
3. learning the relatedness between each text passage and its corresponding image.

It is natural to use recurrent neural networks to encode a sequence of text passages. Here, we encode a text sequence using a bi-directional gated recurrent unit (bi-GRU). Given a text passage, we use the attention mechanism to capture the contextual information of the whole recipe. Because the text embeddings and image embeddings reside in different semantic spaces, we map the text embedding into a latent topic z_t by using a variational autoencoder (VAE). In order to capture the semantics of the images retrieved so far (in a story/recipe), we assume the prior of the distribution of the topic given the text input follows a distribution conditional on the latent topic z_{t-1} from the previous step. We decode the corresponding image vector i_t conditional on the latent topic, to learn the relatedness between text and image with a multi-layer perceptron (MLP) and obtain a synthetic image embedding point generated

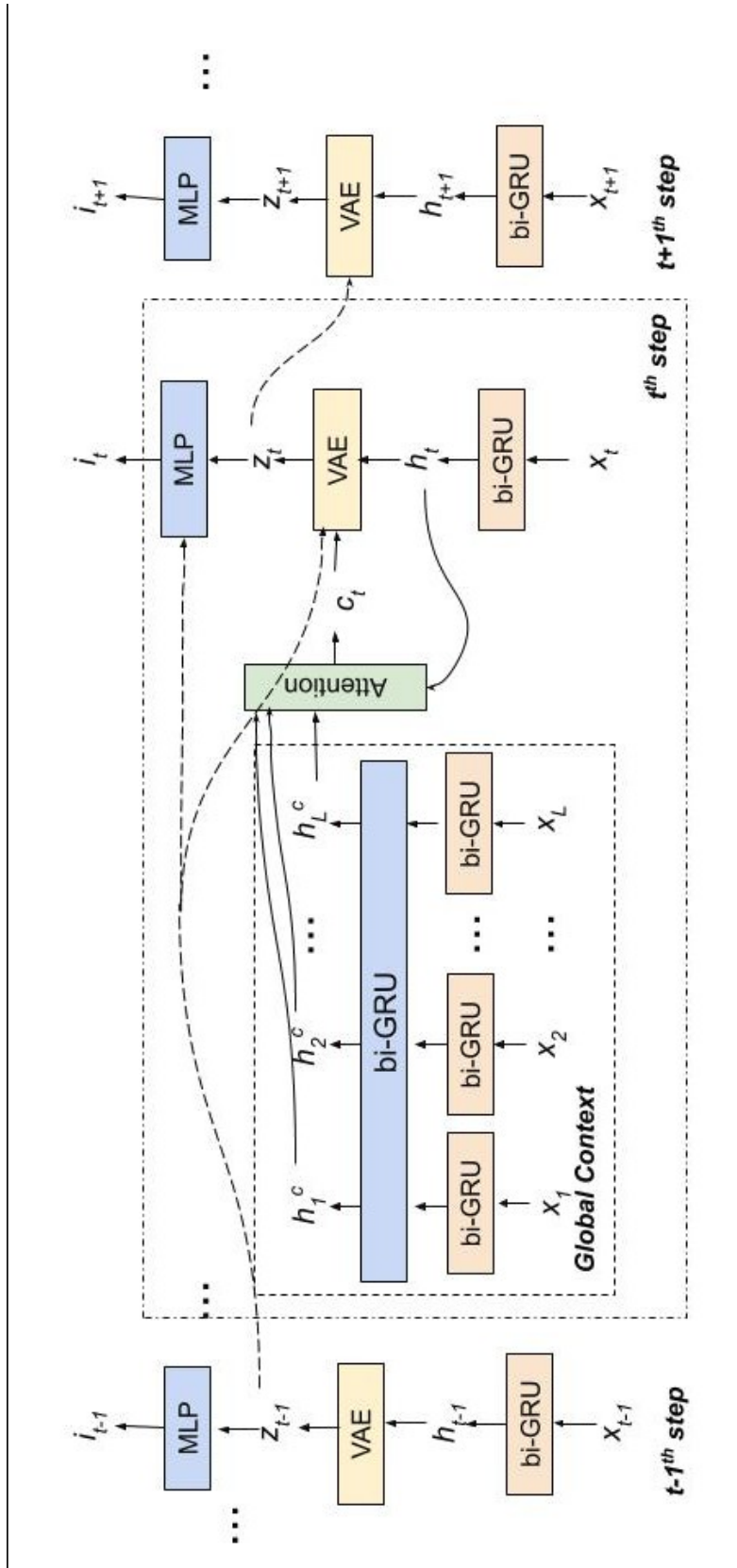


Figure 5.3: Variational Recurrent Sequence-to-Sequence (VRSS) model architecture.

from its associated text embedding point.

Our proposed *Variational Recurrent Seq2seq (VRSS) model* is illustrated in Figure 5.3. Below, we describe each of the main components of the VRSS model.

5.2.2 Text Encoder

We use a bi-GRU to learn the hidden representations of the text passage (e.g. one recipe instruction) in the forward and backward directions. The learned hidden states in the two directions are then concatenated to form the representation at the text segment level To encode a sequence of such text passages.

$$\{h_t = [\vec{r}_t, \overleftarrow{r}_t]\}$$

For example, for a given recipe, a hierarchical bi-GRU is used which first encodes each text segment and subsequently encodes the whole sequence to combine the all text segments. Where h_t encodes all contextual semantics of the t -th word with respect to all other surrounding recipe words.

5.2.3 Image Encoder

Let $\hat{i} = g(I)$ represent the features in the semantic space of the joint model in [93]. $g(I)$ serves as the vector representation for the raw image I from the repository. The model generates these feature representations $g(I)$, where To generate the vector representation of an image, we use the pre-trained modified ResNet50 deep convolutional model [93]. On experiments with alternative pre-trained deep convolutional architectures, it was found that most of the embeddings were clustered around some specific points. In experiments, this model produced a well distributed feature space when trained on the limited domain, namely food related images. This was verified using the PCA and t-SNE visualisations [91], which showed less clustering in the generated embedding space as compared to embeddings obtained from models pre-trained on ImageNet [26].

5.2.4 Incorporating Context

To capture the global context, we first encode each of the text passages using a bi-GRU and then feed the bi-GRU encodings into a top level bi-GRU. Assuming the hidden state output of each text passage x_l in the global context is h_l^c , we use an attention mechanism to capture its similarity with the hidden state

output of the t^{th} text passage h_t as $\alpha_l = \text{softmax}(h_t W h_l^c)$. The context vector is then encoded as the combination of L text passages weighted by the attentions as $c_t = \sum_{l=1}^L \alpha_l h_l^c$. This ensures that any given text passage (which could be a recipe instruction or a paragraph from a story) is influenced more by text passages that are semantically similar.

5.2.5 Latent Topic Modeling

At the t^{th} step text x_t of the text sequence (in a recipe/story) , we feed the corresponding text passage x_t to a the bi-GRU. The output h_t is combined with the context c_t and fed into a VAE to generate the latent topic z_t . We now define two prior networks f_{μ_θ} and f_{Σ_θ} to define the prior distribution of z_t , which is no longer a standard Gaussian distribution, but conditional on the previous z_{t-1} . We also define two inference networks f_{μ_ϕ} and f_{Σ_ϕ} which are functions of h_t , c_t , and z_{t-1} :

$$p_\theta(z_t|z_{<t}, x_{<t}) = \mathcal{N}(z_t|f_{\mu_\theta}(z_{t-1}), f_{\Sigma_\theta}(z_{t-1})) \quad (5.1)$$

$$q_\phi(z_t|z_{<t}, x_{\leq t}) = \mathcal{N}(z_t|f_{\mu_\phi}(z_{t-1}, h_t, c_t), f_{\Sigma_\phi}(z_{t-1}, h_t, c_t)) \quad (5.2)$$

Unlike the typical VAE setup where the text input x_t is reconstructed by generation networks, here we generate the corresponding image vector i_t . To generate the image vector conditional on z_t , the generation networks are defined which are also conditional on z_{t-1} :

$$p_\varphi(i_t|z_{\leq t}, x_{\leq t}) = \mathcal{N}(i_t|f_{\mu_\varphi}(z_{t-1}, z_t), f_{\Sigma_\varphi}(z_{t-1}, z_t)) \quad (5.3)$$

The generation loss for image i_t is then:

$$\begin{aligned} \mathcal{L}_{recons.}(i_t) = & \mathbb{E}_{q(z_{\leq T}|x_{\leq T})} \log p(i_t|z_{\leq t}, x_{<t}) \\ & - KL(q(z_t|x_{\leq t}, z_{<t})||p(z_t|x_{<t}, z_{<t})) \end{aligned} \quad (5.4)$$

5.2.6 Image Retrieval

We enable the search process by a timestep-wise hinge loss to model $p(i_t|x, z_t, i_{<t})$. The latent semantic variable z_t is used to predict the image at the given timestep t , with a hinge loss max-margin objective:

$$\mathcal{L}_{HL}(i_t) = \sum_j \max(0, \alpha - s(i_t, \hat{i}_t) + s(i_j, \hat{i}_t)) \quad (5.5)$$

Where α is the margin parameter, i_t is the image vector generated by the model, \hat{i}_t is the vector representation of the gold-standard image at time step t , i_j is the negative images, and $s(\cdot)$ denotes the similarity measurement function. In our experiments, we use the cosine distance function.

5.2.7 Overall Objective Function

The overall objective function is the total of the image reconstruction loss and the image retrieval hinge loss summing over all the time steps for the whole image sequence, $\{i_1, i_2, \dots, i_T\}$, where β is the weighting factor:

$$\mathcal{L}_{overall} = \sum_{t=1}^T \mathcal{L}_{recons.}(i_t) + \beta \mathcal{L}_{HL}(i_t) \quad (5.6)$$

5.2.8 Parameter Configuration

As the initial parameter setting of the VRSS architecture, we use bi-GRU with the hidden dimension of 500 and set the dimension of latent topics to 500. We also introduce a dropout layer in the RNNs with probability of 0.3. Each word in the text is represented in the 500 dimensional embedding space. The image encoder projects images to a 2,048 dimensional feature space. For training the objective function, we use AdaDelta optimisation function, with a learning rate of 1.0. The values of hyperparameters α and β were set to be 0.2 and 1.7 respectively.

Next, we present main algorithm for VRSS procedure. Algorithm 3 provides the instructions for the main training procedure in the VRSS methodology. As can be observed in the algorithm, for every batch in the training dataset, each story consisting of a sequence of image and text pairs are fed to the model. Please note, the lowercase i represents image features extracted by feeding the image to a pretrained CNN architecture, whereas the uppercase I represents the raw image. \hat{i}_t are the semantic features obtained after feeding through pretrained CNN and extraction from the penultimate layer. r_t represents the intermediate feature representations of the text article at time step t , these are

features obtained by concatenating the forward and backward hidden features of the BiGRU architecture. The Bi-GRU is not trained and is randomly initialised. Also, the attention vector α_l is computed at time step t as described in the algorithm. The final context vector c_t is computed as the weighted summation of hidden states in the attention component. The next instructions describes the inference process, where the latent topics z_t are inferred using the distributions as represented. Also, the latent topics are further used to generate image vectors i_t . The reconstruction loss of the image vector summed with the hinge loss with the precomputed gold standard image features \hat{i}_t are used to propagate the parameters of the VRSS model, after summing up over all time steps.

Algorithm 3 main training procedure for VRSS model

```

1: for  $batch$  in trainData do
2:
3:   for  $S^n \in \{S^1, S^2, \dots, S^{N_B}\}$  do
4:      $S^n = \{(x_1^n, i_1^n), (x_2^n, i_2^n), \dots, (x_{|S^n|}^n, i_{|S^n|}^n)\}$ 
5:     text sequence  $x^n = \{x_1^n, x_2^n, \dots, x_T^n\}$ 
6:     image sequence  $I^n = \{I_1^n, I_2^n, \dots, I_T^n\}$ 
7:
8:     for  $I_t \in \{I_1, I_2 \dots I_T\}$  do
9:        $\hat{i}_t = ResNet50(I_t)$ 
10:    end for
11:
12:    for  $(x_t^n, i_t^n) \in \{(x_1^n, i_1^n), (x_2^n, i_2^n), \dots, (x_{|S^n|}^n, i_{|S^n|}^n)\}$  do
13:       $[\vec{r}_t, \overleftarrow{r}_t] = BiGRU(x_t^n)$ 
14:       $h_t = [\vec{r}_t, \overleftarrow{r}_t]$ 
15:      Assume  $\alpha_l^c$  is the hidden state output of each text passage  $x_t$ 
16:      for for each text passage  $l$  do
17:         $\alpha_l = softmax(h_t W h_l^c)$ 
18:        Here  $\alpha_l$  represents the attention vector
19:      end for
20:       $c_t = \sum_{l=1}^L \alpha_l h_l^c$ 
21:      Infer  $z_t$  using Equations 5.1 and 5.2
22:      Generate  $i_t$  using Equation 5.3
23:      Compute Reconstruction Loss for image  $i_t$  using Equation 5.4
24:      Compute Hinge Loss using Equation 5.4
25:    end for
26:     $\mathcal{L}_{overall} = \sum_{t=1}^T \mathcal{L}_{recons.}(i_t) + \beta \mathcal{L}_{HL}(i_t)$ 
27:    backpropagate( $\mathcal{L}_{overall}$ )
28:  end for
29: end for

```

Algorithm 4 provides instructions for the main testing procedure for VRSS methodology. Here the function $VRSS$ represents all the forward instructions

as in the training procedure. o^n is the image sequence from the output of the VRSS procedure. Please note that o is obtained by searching for nearest neighbour images in the image embedding space. This output sequence is further fed to compute various forms of recall measures against the gold-standard features.

Algorithm 4 Testing procedure for VRSS methodology

```

for each batch in testData do

  for  $S^n \in \{S^1, S^2, \dots, S^{N_B}\}$  do
     $S^n = \{(x_1^n, i_1^n), (x_2^n, i_2^n), \dots, (x_{|S^n|}^n, i_{|S^n|}^n)\}$ 
    text sequence  $x^n = \{x_1^n, x_2^n, \dots, x_T^n\}$ 
    image features sequence  $\hat{i}^n = \{\hat{i}_1^n, \hat{i}_2^n, \dots, \hat{i}_T^n\}$ 
    output sequence  $o^n = VRSS(x^n)$ 
    computeVisualSaliencyRecall( $o^n, \hat{i}^n$ )
    computeTextualSaliencyRecall( $o^n, \hat{i}^n$ )
  end for
end for
computeAverageVisualSaliencyRecall( $o^n, \hat{i}^n$ )
computeAverageTextualSaliencyRecall( $o^n, \hat{i}^n$ )

```

5.3 Experimental Setup

Experiments are conducted to evaluate the performance of the VRSS model and compare its performance with alternative approaches. We create a train-test split of 60K/6K image-text pairs and 9K/1K recipes in the Stepwise Recipe dataset. The split is done author-wise to ensure style consistency but having overlapping authors in train and test splits.

We measure and compare the performances of these models using various evaluation measures like Recall@k, Story Recall@k, TextualSaliency@k, VisualSaliency@k and also conduct a human evaluation study. The details of these evaluation metrics have been described in later sections.

5.3.1 Models for Comparison

We compare VRSS with the following models: LDA, VSE++ and RNN. Both LDA and VSE++ are *non-context models* as they treat each text passage in isolation without taking into account the preceding or succeeding text passages. RNN and VRSS are *context models* since they capture the contextual information for the retrieval of relevant images.

LDA

We re-implement the topic modelling based approach [36] to jointly generate words in text and visual words in image assuming each image-text pair share the same set of topics. For text illustration, we first synthesise textual and visual dictionaries where a textual dictionary is created by assigning a unique token id to each word presented in any of the documents and a visual dictionary is constructed by clustering feature descriptors extracted from images into 750 different visual words. We then train a latent Dirichlet allocation (LDA) model with 100 topics on the dataset containing both text and images. Retrieval of images is based on computing the probabilities of visual terms marginalised over document topics [36]. The retrieved image is the one with maximum overlap of visual terms.

Visual Semantic Embeddings (VSE++)

Following [31], we implement a deep neural network approach which maps the text representations and image vectors into the same semantic embedding space. Since both texts and images are projected to the same semantic space, we apply the traditional similarity search and choose the image which is nearest to a given text representation. Cosine similarity measurement is used to calculate the similarity score between a text vector and an image vector. A hierarchical bidirectional LSTM (bi-LSTM) with max pooling is employed to represent sentences in the vector space [132]. A pre-trained ResNet-152 architecture is used as the image encoder [48]. The image features are extracted from the penultimate fully-connected layer. The text and image vectors are then projected using a fully-connected linear layer to the semantic space. Triplet Ranking Loss function using the hard negatives [31] are used to align the image and text vectors in the semantic space. End-to-end network training is done for 50 epochs. The parameters of the pre-trained CNN network are kept fixed during training. We also introduce a dropout layer in the linear layers with a dropout value of 0.2.

VSE++(R)

Triplet loss function following the same mechanics of VSE++ but the image representations are obtained in the same way as the VRSS model.

Coherence Neural Story Illustration (CNSI)

We use the encoder-decoder CNSI model proposed in [111], with coherence capturing the co-reference relations among sentences, to retrieve a sequence of images illustrating a passage of text.

VRSS-VAE

This follows the same encoder-decoder architecture of our VRSS model, using two bi-GRU architectures as encoders and decoders with the same learning objective, but without latent variables. Therefore, it is treated as an ablation study of our VRSS model without the VAE module.

VRSS-globalCon

This is a variant of our VRSS model without the incorporation of the global context.

In all the neural models evaluated here, the image representation is extracted using the ResNet50 model [93] pre-trained on food-related images.

For training VSE++, we use both ResNet152 and VGG19 as our image encoders. Features are extracted from the penultimate fully-connected layer. The dimensionality of image embeddings is set to 2,048 for ResNet and 4,096 for VGG. We use a Bi-LSTM with max-pooling for representing text, following the architecture in [132]. we project both text and image vectors onto a 1024-dimensional common semantic space using two MLPs. End-to-end network training was done using triplet-ranking loss function for 300 epochs. The parameters of the pre-trained CNN network were kept fixed during training. We also introduced a dropout layer in the linear layers with a dropout value of 0.2.

5.3.2 Evaluation methods

The evaluation metric commonly used in similarity search and information retrieval tasks is *Recall@k*, which is the ratio of the number of correct images retrieved to the total number of retrieval queries. *Recall@k* indicates that the retrieved image was among the top k best matches out of the set of candidate images. We also define *Story Recall@k*, which considers the retrieved image as

correct if it is from the same data sequence.

Visual Saliency Recall

Further, we provide *Visual Saliency Recall@k* values. We implement *Visual Saliency Recall* following [111] and train a VGG-19 network and classify the images of the story test set using this network. The visual features from [93] for initialization. *VisualSaliencyRecall@k* results of the retrieved images when compared with gold-standard images. We follow the same mechanics as suggested in [111] to define a ‘visual saliency’ metric that compares the retrieved images to the gold-standard images using an image classifier.

Textual Saliency Recall

We also provide *Textual Saliency Recall@k* values. This demonstrates whether the corresponding paired text for the retrieved image has entities that overlap with those entities found in the text query.

Visual Feature Similarity

In addition, we report *Visual Feature Similarity* using the average cosine similarity between gold-standard image and retrieved image, considering image features generated using [93].

Human Evaluation

Previous work [111] highlights that existing quantitative retrieval metrics may be too harsh for a task of this description. Therefore, it is imperative that we use human evaluators to judge how appropriate and coherent the retrieved illustration sequences are. For our human evaluation, we pick a random sample (164 recipes, 1564 image-text pairs) from the test set (1K recipes, 6K image-text pairs). We present each evaluator with a sequence of recipe instruction steps that make up one complete recipe. Alongside each text segment, they are given three possible illustrations that depict that step, which are randomly shuffled images of the gold-standard, the non-context model, and the proposed VRSS model. The evaluator is asked to select all image options that may be appropriate illustrations for the corresponding text segment. Finally, the evaluator must also indicate which of the overall illustration sequences are coherent and flow well together. A total of 5.1K ratings are obtained from 12

evaluators, ensuring that every sample has received at least 2 ratings.

5.4 Results and Discussion

5.4.1 Automatic Evaluation

Table 5.1: Text illustration performance using *Visual Saliency Recall@k* ($VSR@k$) on the Stepwise Recipe dataset. The best result in each column is highlighted in **bold**.

Models	VisualSaliencyRecall@k		
	VSR@1	VSR@5	VSR@10
<i>Non-Context Models</i>			
LDA	3.2	6.7	12.5
VSE++	7.8	21.5	23.5
VSE++(R)	8.1	23.1	26.6
<i>Context Models</i>			
CNSI	16.6	31.8	39.8
VRSS-VAE	11.3	29.2	33.2
VRSS-gc	15.1	28.9	32.7
VRSS	18.4	33.4	45.1

Table 5.2: Text illustration performance using *Recall@k* ($R@k$) and *Story Recall@k* ($StR@k$) on the Stepwise Recipe dataset. The best result in each column is highlighted in **bold**.

Models	Recall@k			Story Recall@k
	R@1	R@5	R@10	StR@1
<i>Non-Context Models</i>				
LDA	1.4	3.4	8.9	4.1
VSE++	5.2	15.1	19.5	18.1
VSE++(R)	7.7	18.6	24.6	21.3
<i>Context Models</i>				
CNSI	3.6	8.9	13.7	18.4
VRSS-VAE	6.4	19.7	23.1	
VRSS-gc	5.2	19.9	26.5	21.1
VRSS	8.2	21.3	29.8	24.4

Table 5.2 reports the retrieval performance of different methods using *Re-*

call@k and *Story Recall@k* metrics. It can be observed that LDA gives the worst results, which shows that using a generative model for capturing the semantic topics from both text and image does not work well in the seq2seq retrieval task. By mapping both text and image into the same embedding space, Using the image encoder pre-trained on recipe images, VSE++ outperforms LDA. Our VRSS model without the VAE component (VRSS-VAE) gives similar performance compared to the non-context model VSE++ despite considering the contextual information. VRSS without the incorporation of global context (VRSS-GlobalCon.) performs similarly as VRSS-VAE. CNSI gives worse results compared to both VRSS variants in *Recall@k* and *Story Recall@1*. Our new VRSS model, which maps each hidden state of the RNN into a latent topic and also further incorporates global context information, gives the best results across all metrics. This indicates the importance of representing semantics encoded in both text and images in a more abstract manner and the benefit of incorporating global context.

Table 5.3: Text illustration performance using *Textual Saliency Recall@k* (*TSR@k*) on the Stepwise Recipe dataset. The best result in each column is highlighted in **bold**.

Evaluation	TSR@1	TSR@5	TSR@10
<i>Non-Context Models</i>			
LDA	6.5	18.7	26.8
VSE++	15.6	42.3	56.2
VSE++(R)	19.7	48.9	60.0
<i>Context Models</i>			
RNN	20.2	41.1	54.2
VRSS	25.5	56.7	68.9

Recall@k and *Story Recall@k* metrics only measure the degree of exact matches of the retrieved images with regards to the gold-standard images. This might not be appropriate for our text illustration task since a given text segment could be illustrated by multiple images expressing similar semantics. Example image retrieval results are shown in Figure 5.5 where both the gold-standard and the VRSS retrieved images are displayed for some recipe instructions. It can be observed that although VRSS failed to retrieve the gold-standard images in these examples, its output images are still appropriate illustrations of the corresponding texts. For this reason, we also report the evaluation results using more semantics-based and feature-based metric, *Visual Saliency Recall@k*.

It can be observed from Table 5.1 that VRSS performs significantly better

Table 5.4: Human Evaluation results. The cell values indicate the number of images output by the corresponding model(s) that receive x number of votes ($x \in \{2, 3, 4, 5\}$) as majority.

# Votes	2	3	4	5
Gold-standard only	0	442	171	47
Gold-standard and VRSS	255	41	0	0
Gold-standard and VSE++	88	9	0	0
Gold-standard, VRSS and VSE++	75	0	0	0

than baselines on *Visual Saliency Recall@k*. These recall scores indicate that VRSS is able to retrieve images that are described by text segments that are semantically related to the query text, even if the images themselves do not match the gold-standard image.

We also calculate the *Visual Feature Similarity* which measures the average cosine similarity between the gold-standard image and the retrieved image in the feature space. For VRSS, this is 0.51 and for VSE++ it is 0.37, and for CNSI it is 0.45 Hence, VRSS retrieves illustrations that are visually similar to the gold-standard image.

5.4.2 Human Evaluation

For the human evaluation, we count the number of votes received for the gold-standard images, the VRSS model output images, and the VSE++ (non-context based) model output images. We only count a vote if there is majority consensus among the evaluators. Hence, in Table 5.4, the ‘# Votes’ column indicates the number that constitutes a majority among voters. We also highlight the fraction of recipes for which VRSS received equal or higher preference compared to gold standard.

In Table 5.4, we see the preference results obtained from human evaluation of the retrieved recipe illustrations. Considering majority agreement as 2 votes, gold-standard was never preferred in isolation. Rather, in 61% of the cases, both the gold-standard image and the image retrieved using VRSS were deemed to be appropriate illustrations for the given text query. In 18% of the cases, gold standard as well as the retrieved images from both models were considered appropriate. In the remaining 21% of the cases, the VRSS output was not judged as being appropriate.

Taking 3 votes as the majority, gold-standard alone was picked in 88% of the

cases and picked in combination with the VRSS output in 8% of the remaining cases, with a negligible number of cases for the other combinations. Where the majority consensus is above 4 votes, evaluators chose gold-standard alone in every case. Therefore, VRSS outperforms other models particularly in ambiguous cases where the text is likely to contain an indirect description of the image.

The VRSS output is about 3 times more likely to be selected compared to the VSE++ output. Over 60% of the time, at least 2 human evaluators believe that the VRSS output is as appropriate as the gold-standard image. These results indicate that the context based VRSS model significantly outperforms the non-context-based model.







TEXT	EXPECTED OUTPUT	VRSS OUTPUT
<p>In a medium bowl over simmering water, melt the unsweetened chocolate and butter. When melted, set aside to cool slightly for a few minutes.</p>		
<p>In a large bowl over simmering water, slowly melt your 6oz white chocolate with ¼ cup cream, stirring occasionally.</p>		
<p>Broil pork for 20-30 minutes. Turn and baste at 8 minute intervals...some of the edges will and should get a little crispy...</p>		

Figure 5.4: Example images retrieved by the VRSS model.

In Figure 5.4, we manually inspect some result images obtained using the VRSS model.

5.5 Error Analysis

In this section, we discuss different cases where the VRSS model performs successfully or not and analyse possible reasons.

Figure 5.5 shows examples where the VRSS output was preferred by human evaluators, and Figure 5.6 shows examples where VRSS did not retrieve an appropriate image but the non-context model retrieved a more appropriate image. In Figure 5.7, neither model was deemed to have produced an appropriate output. Further, Figure 5.5 highlights some cases where metrics other than recall are beneficial. The first and second rows depict semantically related entities in both the gold-standard and the VRSS output, and images that have similar feature representations. The third row depicts a result retrieved due to the context-aware nature of the model, corresponding to an adjacent text segment in the sequence, which is counted favourably when using the *Story Recall* metric.










TEXT	EXPECTED OUTPUT	CONTEXT MODEL OUTPUT	NON-CONTEXT MODEL OUTPUT
Chop half your mushrooms coarsely. And the rest finely. If you're using dried mushrooms, soak for 20 minutes in enough hot water to cover.			
Boil your pasta in salted water. Drain, but reserve about quarter cup of water. Toss in the edamame and ham, and drizzle with about 2 tablespoons of olive oil.			
Getting darker.			

Figure 5.5: Illustrative comparison of non-context (VSE++) and context models (VRSS) - VRSS result preferred by human evaluators.

Further, it highlights some cases where metrics other than recall such as *Visual Feature Similarity*, and *Story Recall* are beneficial. The first and second rows depict semantically related entities in both the gold-standard and the VRSS output, and images that have similar feature representations. The third row depicts a result retrieved due to the context-aware nature of the model, corresponding to an adjacent text segment in the sequence, which is counted favourably when using the *Story Recall* metric.

5.6 Embedding Analysis

In order to analyse the limitations and inform future work in similar retrieval applications, we perform a qualitative analysis of our generated embedding space. We compare the embeddings produced using the VRSS model with those







TEXT	EXPECTED OUTPUT	CONTEXT MODEL OUTPUT	NON-CONTEXT MODEL OUTPUT
Grate some cheese finely (about a cup). Cut up your ham. I used intentionally rough cuts to give it some texture.			
While mac is baking, heavily salt the pork belly pieces, and cook until crispy in the same skillet you used earlier.			

Figure 5.6: Illustrative comparison of non-context and context models - VSE++(R) result preferred by human evaluators.







TEXT	EXPECTED OUTPUT	CONTEXT MODEL OUTPUT	NON-CONTEXT MODEL OUTPUT
Get a wok or saute pan hot and put in the oil. Add onions, lower heat to low, and cook until onions start to turn brown. Add pork and raise heat to medium-high.			
Remove from heat and put in bowl. Let cool to room temp then refrigerate for 4 hours or up to overnight.			

Figure 5.7: Illustrative comparison of non-context and context models - Neither VRSS nor VSE++(R) result preferred by human evaluators.

from the VSE++(R) model, to reveal some characteristics of these embeddings that might influence the performance on this task.

We first assess the distribution of “popularity” of points in the image-text shared embedding space which is used for performing retrieval. Consider a set of points D , with $N_k(x)$ as the nearest neighbourhood of a point x in D . The number of times a point x occurs among the nearest neighbourhoods $N_k(x)$ of all other points in D can be thought of as a measure of its *hubness*, reflecting its popularity in the space. It has been studied [109] that the distribution of this $N_k(x)$ becomes considerably skewed when considering higher dimensionality of points in D , revealing the emergence of “hubs”. Conversely, “anti-hubs” are points that occur in very few or none of the nearest neighbourhood lists.

We use the Hub Toolbox implementation [33], after modifying it to handle multimodal data. We compare hubness results for VRSS and VSE++(R), using the embeddings generated for the test split of the data. Table 5.5 and

Table 5.6 show the calculated hubness when considering the k -nearest text neighbours of all image points and vice versa respectively ($k = 5, 10$), in the embedding space obtained using VRSS.

Table 5.5: Hubness analysis results. Text points as hubs in neighbourhoods of image points from Stepwise Recipe embeddings obtained using VRSS.

Hubness ($S^k = 5$)	0.90
Anti-hubs at k= 5	2.27%
Hubness ($S^k = 10$)	0.48
Anti-hubs at k=10	0.12%

Table 5.6: Hubness analysis results. Image points as hubs in neighbourhoods of text points from Stepwise Recipe embeddings obtained using VRSS.

Hubness ($S^k = 5$)	6.90
Anti-hubs at k= 5	34.79%
Hubness ($S^k = 10$)	6.23
Anti-hubs at k=10	28.34%

The values in Table 5.6 suggests high hubness. This can be more intuitively understood by looking at the high fraction of points that behave as anti-hubs. We see clear indications of hubness in one direction, that is, many image points are hubs and anti-hubs when we are given the neighbouring image point lists for all text points. This is likely to have an impact on retrieval accuracy.

This is contrasted against Table 5.7 and Table 5.8 respectively, which report the calculated hubness in the embedding space obtained using VSE++ in a similar fashion.

Table 5.7: Hubness analysis results. Text points as hubs in neighbourhoods of image points from Stepwise Recipe embeddings obtained using VSE++(R).

Hubness ($S^k = 5$)	1.90
Anti-hubs at k= 5	14.46%
Hubness ($S^k = 10$)	1.50
Anti-hubs at k=10	6.22%

Hubness in the VSE++(R) embeddings, by comparison, is not perceived to be as significant. Therefore, hubness analysis suggests that the retrieval

Table 5.8: Hubness analysis results. Image points as hubs in neighbourhoods of text points from Stepwise Recipe embeddings obtained using VSE++(R).

Hubness ($S^k = 5$)	2.29
Anti-hubs at k= 5	5.71%
Hubness ($S^k = 10$)	1.81
Anti-hubs at k=10	1.13%

performance in VSE++(R) is less adversely affected compared to that of our VRSS model and cannot solely account for the difference in the model performances.

In order to determine why VRSS significantly outperforms VSE++(R) according to both human evaluators and common retrieval accuracy metrics, we then turn to simple visualisations of the embedding spaces.

The t-SNE technique [91] is popular for efficiently visualising high-dimensional data in two or three dimensions, retaining local structure while revealing some global structure. By visualising 2-dimensional maps using t-SNE, we are able to study clustering and separability across the two modalities in the embedding space.

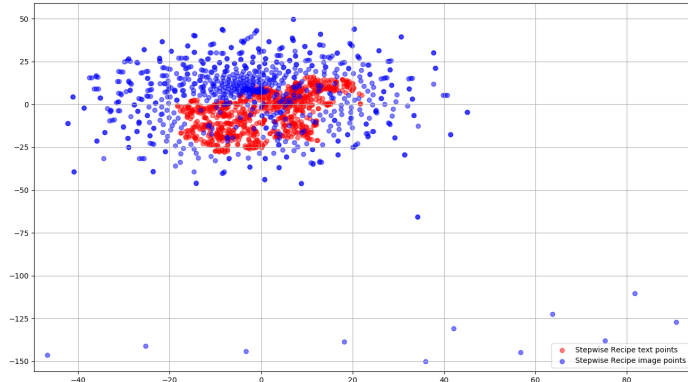


Figure 5.8: t-SNE plot for VRSS embeddings.

We examine a random sample of 850 points from each image-text shared embedding space, as generated by VRSS and VSE++(R) for the test split of the data. The visualisation in Figure 5.9 reveal that VSE++(R) embeddings suffer from disparate features in the two modalities, leading to a separation of the modalities in the shared embedding space. The t-SNE plot in Figure 5.8, however, demonstrates that VRSS is able to handle this challenge of multimodal data. The embeddings are better distributed in the shared space, owing to

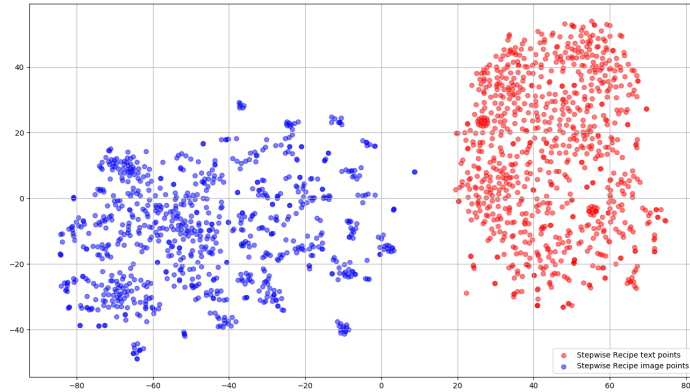


Figure 5.9: t-SNE plot for VSE++(R) embeddings.

the reconstruction of image embeddings from the input text during our latent topic modelling step. These reconstructed points are more comparable to the image encodings themselves.

5.7 Conclusion

We presented VRSS model that given a sequence of text passages, retrieves a sequence of images best describing the semantic content of text and introduced the Stepwise Recipe dataset which will facilitate further research on this problem. Our results on the Stepwise Recipe dataset show that VRSS significantly outperforms competitive baselines in terms of both automatic and human evaluations. This chapter provided a formal introduction to the problem of stepwise illustration but studied it on a limited domain. The next chapter studies it on a more general domain.

Chapter 6

Context-Dependent Text Illustration and Description Retrieval

In this chapter we address the RQ1 of “How can we develop automatic text illustration systems that illustrate a given narrative text passage with a sequence of illustrations, considering and incorporating prior context?”. Therefore, in this chapter, we focus on the task of *Stepwise Illustration* on a more generalised domain, in this part, we study existing methods, as well as propose new methods, and employ them on the Gutenstories dataset, which contains images from variety of domains.

6.1 Introduction

Automatic text illustration, as described by Hartmann and Strothotte [47], is the task of automatically illustrating a piece of text with relevant images. Joshi et al. [61], described the task of *automatic story picturing* slightly differently from automatic text illustration. The main objective in *text illustration* is to provide an illustration to a piece of text, whereas in *story picturing*, it is to depict or explain, the events and ideas conveyed by a text in the form of a few representative pictures.

For example, for a given piece of text, “*But Peter Rabbit, who was very naughty, ran straight away to Mr. McGregor’s garden, and squeezed under the gate!*”, for *text illustration*, any picture of Peter Rabbit, would be considered a relevant illustration. However, for *story picturing*, the objective is to retrieve a specific image of a rabbit trying to squeeze under a gate as shown in Figure 6.1. That is, *story picturing* should illustrate the text in the context of the story and should be able to convey the idea depicted in the text. Therefore, specificity is



....flopsy, mopsy, and
cotton-tail, who were
good little bunnies, went
down the lane to gather
blackberries:but peter,
who was very naughty,
ran straight away to mr.
mcgregor's garden, and
squeezed under the
gate!....

Figure 6.1: Example automatic text illustration

the key difference between the two tasks.

Story picturing defined in Joshi et al. [61] aims to illustrate a rather short piece of text by one or more images. We are instead interested in automatically retrieving images for specific passages from a longer piece of text in a context-dependent way. In a long text, ideas are usually unfolded sequentially. Therefore, the retrieval of an image in the sequence should condition not only on the given text passages, but also on the preceding text passages or images retrieved in the story sequence. This is also more natural to how human understanding evolves. There could be lots of redundant information. Models should have the ability to select relevant information.

Although there are various image resources available on the Web, such as Flickr5k, Flickr30k [133], ImageNet [25] and COCO [85], which can be used for training cross-modal retrieval algorithms, they only contain isolated image-text pairs and cannot be used to develop approaches for story picturing. More recently, the Visual Storytelling (VIST) dataset [55] was introduced for the task of generating text descriptions given a sequence of four to five photos collected from Flickr. In VIST, each photo is only accompanied with one sentence as its literal caption and one sentence as a description in context. It does not contain rich contextual information as can be seen in children story books.

To the extent of our knowledge, a suitable dataset does not exist for the tasks described above. Existing datasets do not have strong links between text passages and explaining images. Storytelling models can be trained that actually analyse content of the images and can capture the key semantics of

the image. Besides search and retrieval, these models can help in clustering visual stories into various categories. They can also be used in automatic text illustration.

In this chapter, we focus on the two versions of a dataset, *GutenStories* introduced in Chapter 3, for automatic story picturing and description retrieval for children stories. *GutenStories*, which has been programmatically constructed by processing web pages retrieved from Project Gutenberg, an online book catalogue containing more than 57,000 digital stories. Details of the processing pipeline that extracted, filtered and transformed raw data into a set of stories, each with a sequence of text passages with their accompanied images, can be found in the datasets chapter. We also provide evaluation results of various approaches for cross-modal retrieval on our dataset and highlight the challenges faced.

Our main contributions are as follows:

- We introduced two versions of a dataset *GutenStories* in Chapter 3, which has been programmatically constructed by processing web pages retrieved from Project Gutenberg, an online book catalogue containing more than 57,000 digital stories. We provided details of the processing pipeline that extracted, filtered and transformed raw data into a set of stories, each with a sequence of text passages with their accompanied images.
- We formalise the task of story picturing and description retrieval, and outline challenges faced with these tasks.
- We provide preliminary evaluation results of various approaches for cross-modal retrieval on our dataset. We compare the performance of some traditional probabilistic methods with some neural-networks based approaches for the tasks presented.

The rest of the chapter is organised as follows. In the subsequent sections, we provide a problem formulation and list down some key challenges involved. The sections following provide details of the experimental setup and discuss the results obtained.

6.2 Challenges

One of the main challenges of the dataset is the amount of variety of texts and images, that are present in the dataset. There are stories from dozens of

genres. Stories presented can be categorised to History, Science, Technology, Folklore, Children Fairy Tales. Dataset consists of mostly artificial, but also natural images. As most of the books are digitised from textbooks. Some of the images are also book covers.

Some of the key challenges involved in this task are as follows:

- GutenStories mostly consist of artificial images, most of the existing datasets in the computer vision community are a representative of real-world data. Therefore, the existing pre-trained models that are trained on a different domain of images, need to transfer the knowledge from other domains, to be utilised.
- There could be indirect semantic relationships among segmented image regions and given text. To infer, the visual-text correspondence, models would need to look and identify the most important aspects, both within text and image.
- Rich linguistic information is naturally embedded in the text, as both the images and texts come from a variety of story and content styles.
- Images, as can also be seen in figure 6.2 are usually cluttered depicting more complex scenes, contain more and less prominent objects, and are often rendered in varying resolutions.
- Text passages are also admittedly noisy since these are not written in order to exhaustively list every object in the image but to describe the main event or related aspects thereof in the story.
- A concern here is that detecting and recognizing all objects from images under such noisy conditions is still beyond the capability of current computer vision and natural language processing research.

It is worth noting, our aim is to employ this dataset to observe the correlation between the visual and textual modalities without explicitly performing object recognition.

6.3 Problem Formulation

GutenStories dataset consists of a wide variety of images. mini-GutenStories only consists of children stories. Each story is a sequence of image-text pairs. We assume, due to the nature in which dataset is constructed, that the corresponding texts describe the content of the image either directly or non-directly.



Peter said he hoped that it would rain. At this point old Mrs. Rabbit's voice was heard inside the rabbit hole, calling: "Cotton-tail! Cotton-tail! fetch some more camomile!" Peter said he thought he might feel better if he went for a walk.



They went away hand in hand, and got upon the flat top of the wall at the bottom of the wood. From here they looked down into Mr. McGregor's garden. Peter's coat and shoes were plainly to be seen upon the scarecrow, topped with an old tam-o'-shanter of Mr. McGregor's.

Figure 6.2: Example text passage-image pairs in our dataset.

Given the constraints defined above, our goal is threefold: Automatic story picturing, description retrieval and story retrieval.

Automatic text illustration

Automatic text illustration or story picturing is the task of retrieving the image I , that corresponds to text article x , in a story sequence S .

Description Retrieval

Description retrieval is the task of retrieving the text article x , given the image I , in the story sequence S .

The task of story picturing requires the understanding of text and context information in preceding text passages and images retrieved so far. Since it is a difficult task, we start with a relatively simpler task of retrieving an image given a text passage. Note that our text passage here may contain multiple sentences. As is common in information retrieval, we measure the performance by $Recall@K$, which is the fraction of the number of correct images retrieved at the top K position out of a total of number of text queries. We also define another metric, $Story@K$, which considers the retrieved image as correct if it is from the same story sequence.

Story Retrieval

We formulate the story retrieval task as follows, given a text article x , and its accompanying image with image representation i , which belongs to the story sequence S . The objective is to retrieve any image from story sequence S , given a text article x . For the task of story retrieval, we measure the performance by $StR@K$ i.e the fraction of the queries for which the correct story is retrieved, given K closest points in the embedding space.

6.4 Methods

In this section, we describe several models incorporating different kinds of features to study the relevant importance of these features in modelling and retrieval of a relevant image from a repository to match the semantics of the given text or the task of description retrieval and story picturing. We study and compare performance of these several previously published methodologies on GutenStories dataset and its mini version. We also propose some new methodologies for this task.

6.4.1 LDA-based

As proposed in Feng and Lapata [36], a mixed-LDA model can be trained on documents consisting of images and texts. They proposed several extractive and abstractive summarisation techniques to generate summaries. They also proposed ways to retrieve images. We re-implemented the topic modelling based approach in Feng and Lapata [36] to jointly generate words in text and visual words in image assuming each text-image pair share the same set of topics. For text illustration, we first synthesised textual and visual dictionaries where a textual dictionary was created by assigning a unique token id to each word presented in any of the documents and visual dictionary was constructed by clustering SIFT descriptors extracted from images into 750 different visual words. We then trained a LDA model with 100 topics on the dataset containing both text and images. It has been shown in Feng and Lapata [36] that retrieving images based on the computed probabilities of visual terms marginalising over document topics. The retrieved image is the one with maximum overlap of visual terms.

word-overlap based

For extractive summarisation, word-based overlap strategy was also implemented, where the sentence with maximum overlap between extracted annotations and words in the sentence is picked.

6.4.2 Joint embedding learning

We also implemented a deep neural-networks based approach which maps the text representations and image representations into the joint common semantic embedding space. Since both texts and images are projected to the same semantic space, we can simply choose the image which is nearest to a given text representation. Cosine similarity measurement is used to calculate the similarity score between a text vector and an image vector. A hierarchical bidirectional LSTM (Bi-LSTM) with max-pooling is employed to represent sentences in the vector space, following Talman et al. [132]. A pre-trained ResNet-152 architecture is used as our image encoder [48]. The image features are extracted from the penultimate fully connected layer. The text and image vectors are then projected using a fully connected linear layer to the semantic space. Triplet Ranking Loss function using the hard negatives, as described in Faghri et al. [31], are used to align the image and text vectors in the semantic

space. We refer to this model as Deep-NN-res0.2 in the result tables.

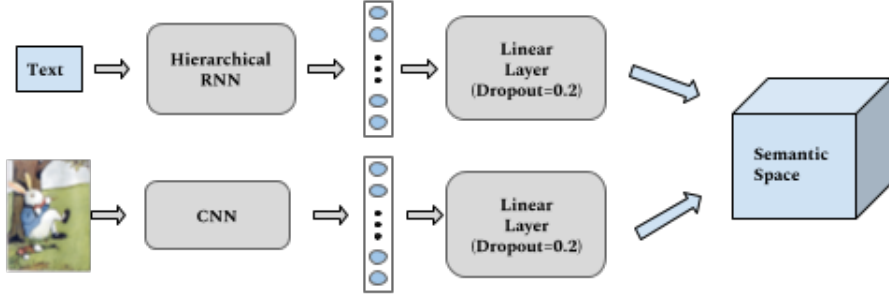


Figure 6.3: Deep NN Architecture.

In this section, we also present instructions in the main training procedure for joint embedding learning model in Algorithm 5. As it can be observed in the algorithm, the procedure involves transforming passages of text and images to a space, where they could be directly compared. For each batch in the training dataset, the text sequences and image sequences are transformed accordingly. For each image and text pair in a story, the image I_T^n is fed to the pretrained CNN model to extract semantic features. The image features are represented by F^i . Similarly, for the text passages, a RNN is used, more specifically, in this case a hierarchical RNN is used to represent the text in vector form. F^x represents features in the text space. Both F^x and F^i are further transformed using two linear layers into a common semantic space. Please note, we also use dropout in this layer to account as regularisation measure. We dropout with a probability of 0.2. Triplet ranking loss is computed in the semantic space and model parameters are back propagating over this loss function.

The testing procedure is also fairly similar to training procedure, as presented in Algorithm 6. The key difference is that the respective computed features are used to retrieve their key images/texts. Recall measures are used to evaluate the model.

Pre-training with SkipThoughts

We explored a number of variants of the Deep NN architecture presented above. For text encoder, instead of using the hierarchical Bi-LSTM, we use the pre-trained SkipThoughts model [73] to extract the representation for each sentence in the text passage. The sequence of sentence embeddings is then fed to a Recurrent Neural Network (RNN). The last hidden layer of the RNN is used as the text passage embedding.

Algorithm 5 Training procedure for Joint Embedding Learning methodology

```
for each batch in trainData do

  for  $S^n \in \{S^1, S^2, \dots, S^{N_B}\}$  do
     $S^n = \{(x_1^n, I_1^n), (x_2^n, I_2^n), \dots, (x_{|S^n|}^n, I_{|S^n|}^n)\}$ 
    text sequence  $x^n = \{x_1^n, x_2^n, \dots, x_T^n\}$ 
    image sequence  $I^n = \{I_1^n, I_2^n, \dots, I_T^n\}$ 

    for  $(x_t^n, I_t^n) \in \{(x_1^n, I_1^n), (x_2^n, I_2^n), \dots, (x_{|S^n|}^n, I_{|S^n|}^n)\}$  do
       $F^x = RNN(x_t^n)$ 
       $F^i = CNN(I_t^n)$ 
       $S^x = LinearLayer(F^x)$ 
       $S^i = LinearLayer(F^i)$ 
       $backpropagate(TripletRankingLoss(S^x, S^i))$ 
    end for
  end for
end for
```

Algorithm 6 Testing procedure for Joint Embedding Learning methodology

```
for each batch in testData do

  for  $S^n \in \{S^1, S^2, \dots, S^{N_B}\}$  do
     $S^n = \{(x_1^n, I_1^n), (x_2^n, I_2^n), \dots, (x_{|S^n|}^n, I_{|S^n|}^n)\}$ 
    text sequence  $x^n = \{x_1^n, x_2^n, \dots, x_T^n\}$ 
    image sequence  $I^n = \{I_1^n, I_2^n, \dots, I_T^n\}$ 

    for  $(x_t^n, I_t^n) \in \{(x_1^n, I_1^n), (x_2^n, I_2^n), \dots, (x_{|S^n|}^n, I_{|S^n|}^n)\}$  do
       $F^x = RNN(x_t^n)$ 
       $F^i = CNN(I_t^n)$ 
       $S^x = LinearLayer(F^x)$ 
       $S^i = LinearLayer(F^i)$ 
       $O^i = NearestNeighbour(S^i)$ 
       $O^x = NearestNeighbour(S^x)$ 
       $computeVisualSaliencyRecall(O^i, O^x, I_t^n, x_t^n)$ 
       $computeTextualSaliencyRecall(O^i, O^x, I_t^n, x_t^n)$ 
       $computeTextRecall(O^i, O^x, I_t^n, x_t^n)$ 
       $computeImageRecall(O^i, O^x, I_t^n, x_t^n)$ 
    end for
  end for
end for

ComputeAverageTextRecall()
ComputeAverageImageRecall()
ComputeAverageVisualSaliencyRecall()
ComputeAverageTextualSaliencyRecall()
```

Variant Image Representation Mechanisms

For image encoding, we explored different Convolutional Neural Networks (CNN) based architectures, the use of VGG [124] instead of ResNet-152 for the generation of image embeddings.

Dropout-based regularisation

We also study the effects of introducing a dropout layer to tackle the noisy nature of the dataset, as highlighted in the challenges section. So, we explore another variant, by introducing a dropout in the baseline architecture. A dropout layer is introduced at the intermediate linear layer. Several dropout values were experimented with, and best results were obtained using $p=0.2$. Here p is the dropout rate.

6.4.3 Event Representation Embeddings

As there is lot of event related data in the textual stories, we explored another text representation variant. We implemented a Named-Entity-Recognition (NER) based strategy. As GutenStories dataset is often cluttered with lots of objects and scenes, both in the text and images. We aim to extract entities from different categories, namely person, location, organisation and keywords from the text. Stanford NLP Toolkit [40] was used for extraction. We represent the extracted entities using the fastText library [12] (also in [62]).

We concatenated summed representations of all entities, from each category, to form the text representation vector. Subsequently, we fed these extracted features into the same Deep NN architecture as described above and project the data into common semantic space, to reduce corresponding distances, with a triplet ranking based loss function.

Several other event representation methods have also been explored in recent research [28, 95].

6.4.4 Stacked Cross Attention Network (SCAN)

To infer the latent semantic alignment between objects in an image or other salient stuff for example, snow or sky and the corresponding words in a piece of text, for the problem of image-text matching, Lee et al. [81] presented a model based on stacked cross-attention network. They argue to look at the similarity of all possible pairs of regions and words by attending accordingly

to more or less important words. Therefore, they incorporate attention over all sub regions of a given image and words in a text. They use a multi-step attentional process to capture the possible number of semantic alignments to discover the full latent alignments using both image regions and words in a sentence as a context to infer the image-text similarity. We use the model presented in Lee et al. [81].

6.4.5 Context-based Models

In Chapter 5, we saw the importance of incorporating context for the task of text illustration. Therefore, we also experiment with some context-based models on the GutenStories dataset.

VRSS

We employ the model presented in Chapter 5 VRSS on this task.

VRSS(order)

We experimented with the presented VRSS model in the same way as described previously. However, the embeddings pre-trained on the task of image-text retrieval are used in this case. We used pre-trained embeddings from [137], both for text and image representation.

6.5 Experimental Setup – mini-GutenStories

In this section the experimental design for evaluating the performance of the models presented above is discussed. Note, in this section, we only present results on the mini version of the dataset, as described before. We provide details of the training procedure and parameter estimation.

Experiments were conducted on the mini-GutenStories dataset. We created an author-wise train-test split for image-text pairs. Train set consists of 5k pairs. Test set consists of 1k pairs. We use both ResNet152 [48] and VGG19 [125] as our image encoders. Features are extracted from the penultimate fully connected layer. The dimensionality of image embeddings is set to 2048 for ResNet and 4096 for VGG. We use a hierarchical Bi-LSTM with max-pooling for representing text, following the architecture in [132]. Features were further projected using two linear layers for images and texts from the feature space to 1024-dimensional semantic space. End-to-end network training was done using

triplet-ranking loss function for 300 epochs. The parameters of the pre-trained CNN network were kept fixed during training. We also introduced a dropout layer in the linear layers with a dropout value of 0.2.

For the text illustration experiments, the proposed dataset was evaluated with three baselines. The first one is essentially LDA methodology for retrieving images. We select the image based on the topic-mixture framework. The second one is a straightforward implementation of the vector space model where documents and images are represented by vectors whose components. We followed common practice in weighting terms by their tf-idf values and used the cosine similarity measure to find the image whose vector representation is most similar to the test document vector.

6.5.1 mixed-LDA training

The mixed-LDA model training was done by first extracting SIFT features from all the images in the dataset. Subsequently, visual words were formed by clustering all the extracted features into 750 clusters, using standard clustering algorithm, here we used KMeans. About 12 Million key points were extracted from around 7,000 images.

The mixed-LDA model was trained using scikit-learn library [105] in PyTorch [104], and the model was trained for 1000 topics. KL divergence was computed among the topic distributions using the SciPy library [59].

6.5.2 Results and Discussion

Table 6.1 report the retrieval performance of different methods on mini-GutenStories using $Recall@K$ and $Story@K$ metrics, respectively. We compare the performance of the probabilistic LDA model with some deep neural-networks based architectures. Table 6.2 provides results of the models on $Story@K$ evaluation measure.

It can be observed that hierarchical Bi-LSTM gives better results compared to SkipThoughts in text encoding. Also, using ResNet consistently outperforms the model with VGG for image feature extraction. We also notice that most of the baseline methods achieve quite low performance on this dataset, partly due to the nature of the data used that unlike existing image datasets that each image is usually aligned with a single sentence (or multiple alternative captions) all directly describe what is depicted in the image, in our dataset,

each image is paired with a text passage consisting of multiple sentences. Text passage provides description of the events in the context of the story and may not directly describe the image. This can be seen from examples in Figure 6.2. It depicts the challenging nature of the problem. It shows examples of two image-text pairs, where Ideally, the model should learn the latent alignment of the key entities/relations described in text with the objects and layouts depicted in images.

We have experimented with alternating ways of encoding text and image, for example using LDA topics, in the hope that more abstract representations from text and image could help with the cross-modal retrieval. However, our results show that it performs even worse compared to deep NN approaches.

Also, GutenStories mainly consists of artificial images, whereas image encoders VGG19 and Resnet152 were trained on natural images. Future work could consider training image encoders from large-scale artificial images in order to extract better image representations.

The above highlights some of the challenges faced with the dataset and there is a need of developing models that could capture the key semantics of the text and the contextual information present in a story sequence, and also better image representations in order to learn associations between the text-image pairs. We hope our constructed datasets could encourage more work in tackling the challenging problem of story picturing or context-dependent text illustration from children’s stories.

6.6 Experimental Setup - GutenStories

In this section, the experimental design for evaluating the performance of the models presented above is discussed. Note, in this section, we present results on the full version of the dataset, as described before. We provide details of the training procedure and parameter estimation.

Experiments were conducted on the GutenStories dataset. We created an author-wise train-test split for image-text pairs. Train set consists of 62k pairs. Test set consists of 1k pairs. We use both ResNet152 [48] and VGG19 as our image encoders. Features are extracted from the penultimate fully connected layer. The dimensionality of image embeddings is set to 2048 for ResNet and 4096 for VGG. We use a hierarchical Bi-LSTM with max-pooling for representing text, following the architecture in [132]. Features were further

Table 6.1: Text illustration and Description Retrieval performance using $Recall@k$ ($R@k$) and $Story Recall@k$ ($StR@k$) and $Visual Saliency Recall@k$ ($VSR@k$) on the mini-Gutenstories dataset. The best result in each column is highlighted in **bold**.

Models	ImageRecall@k			TextRecall@k		
	R@1	R@5	R@10	R@1	R@5	R@10
<i>LDA-based</i>						
LDA-based	0.4	1.2	2.7	1.2	1.8	2.0
LDA(word)	0.2	1.1	2.1	1.1	1.6	1.8
<i>Joint embedding learning</i>						
Event Repre	0.5	1.1	2.1	1.1	1.8	1.9
Deep-NN-skip	0.9	2.3	4.8	1.4	3.4	5.8
Deep-NN-vgg	0.8	3.2	5.4	0.5	4.2	8.3
Deep-NN-vgg0.2	0.8	3.2	5.4	0.8	4.7	7.2
Deep-NN-res	1.3	3.3	6.2	1.5	3.5	8.5
Deep-NN-res0.2	1.2	3.3	6.4	2.5	6.2	10.5

Table 6.2: Story Retrieval performance using $Story Recall@k$ ($StR@k$) on the mini-Gutenstories dataset. The best result in each column is highlighted in **bold**.

Models	StoryRecall@k		
	StR@1	StR@5	StR@10
<i>LDA-based</i>			
LDA-based	2.3	3.4	5.4
LDA(word)	1.4	4.6	10.1
<i>Joint embedding learning</i>			
Deep-NN-skip	1.4	4.6	10.1
Deep-NN-vgg0.2	1.8	5.9	9.1
Deep-NN-res0.2	5.5	14.1	20.8

projected using two linear layers for images and texts from the feature space to 1024-dimensional semantic space. End-to-end network training was done using triplet-ranking loss function for 300 epochs. The parameters of the pre-trained CNN network were kept fixed during training. We also introduced a dropout layer in the linear layers with a dropout value of 0.2.

For the text illustration experiments, the proposed dataset was evaluated

with three baselines. The first one is essentially LDA methodology for retrieving images. We select the image based on the topic-mixture framework. The second one is a straightforward implementation of the vector space model where documents and images are represented by vectors whose components. We used the cosine similarity measure to find the image whose vector representation is most similar to the test document vector.

6.6.1 Evaluation Measures

Visual Saliency Recall

We implement *Visual Saliency Recall* by following [111] and use ResNet152 pretrained on ImageNet dataset and classify the images of the story test set using this network [131].

Textual Saliency Recall

We also provide *Textual Saliency Recall@k* values. This demonstrates whether the corresponding paired text for the retrieved image has entities that overlap with those entities found in the text query.

6.6.2 Results and Discussion

It can be observed in tables 6.3 and 6.4, SCAN methodology outperforms all other methods. We believe, it is due to its nature of selectively attending to sub-constituents of both images and texts. It is able to learn the latent semantic alignment based on possible subunit matches. It even outperforms the context-based models that incorporate the preceding and succeeding information in a given story as well.

Both in terms of textual saliency recall and visual saliency recall, SCAN outperforms. However, it still reaches the maximum top-1 recall of 4.5%, which clearly denotes there is a margin to improve. With a manual inspection of the results obtained, we found sometimes the Visual Saliency Recall and Textual Saliency Recall measures could also be inefficient at evaluating this task, which outlines a need for a better automatic evaluation measurement.

Table 6.3: Text Illustration performance using *Visual Saliency Recall@k* ($VSR@k$) on the Gutenstories dataset. The best result in each column is highlighted in **bold**.

Models	VisualSaliencyRecall@k		
	VSR@1	VSR@5	VSR@10
<i>Non-Context</i>			
LDA-based	1.1	2.0	3.3
Deep-NN-Res0.2	2.7	3.4	5.2
SCAN	4.5	6.3	12.1
EventRepr	2.1	2.9	4.0
<i>Context</i>			
VRSS-VAE	2.5	3.7	5.2
VRSS-gc	2.3	3.3	4.9
VRSS	2.5	3.9	5.6
VRSS(order)	4.3	6.1	9.3

Table 6.4: Text Illustration performance using *Textual Saliency Recall@k* ($VSR@k$) on the Gutenstories dataset. The best result in each column is highlighted in **bold**.

Models	TextualSaliencyRecall@k		
	TSR@1	TSR@5	TSR@10
<i>Non-Context</i>			
LDA-based	1.3	1.9	3.4
Deep-NN-Res0.2	2.8	3.2	6.1
SCAN	3.3	5.1	6.9
EventRepr	1.1	2.1	2.9
<i>Context</i>			
VRSS-VAE	2.1	3.2	4.4
VRSS-gc	1.9	2.3	4.7
VRSS	2.4	3.3	5.0
VRSS(order)	3.1	4.9	6.3

6.7 Error Analysis

In this section, we discuss different cases where the different models perform successfully or not successfully and analyse possible reasons.

Figure 6.4 shows a manual inspection of the some of the results obtained

using the best performing Deep-NN methodology on mini-GutenStories. It can be clearly seen the challenging nature of the task. It also depicts the challenge in automatic evaluation measures of this task. The figure presents the cases, where the correct image was not retrieved using the best performing model. Further, Figure 6.4 highlights some cases where metrics other than recall such as *Textual Saliency Recall*, *Visual Feature Similarity*, and *Story Recall* are beneficial. An image is retrieved corresponding to an adjacent text segment from the same sequence, which is counted favourably when using the *Story Recall* metric.

Figure 6.5 provides another error analysis study of the Deep NN Model. The figure presents the cases, where the correct image was not retrieved using the best performing model.

The cases presented in both Figure 6.4 and Figure 6.5 demonstrate the challenges associated with the problem. We present some points below listing these challenges and also propose future directions to overcome these.

- Firstly, images in the dataset are artificial images and not natural images. Most of the methodologies presented here indirectly use semantic knowledge from these images by using pretrained CNN models. These models are pretrained on sets of natural images. One possible alternative is to use a pretrained CNN classifiers specifically trained on the domain of artificial images or related images to extract semantic features.
- Secondly, we also previously highlighted challenges associated with the representation of the text. Some of the models depend on the semantic representation of the text. In recent literature, transformers-based representation mechanisms [27] have been proposed and achieved excellent results on various benchmarks. The inclusion of such mechanisms can help boost the performance of the some of the methodologies we have presented in the sections above.
- A methodology presented before showed how attending to different sub-constituents of images and texts can help improve the performance previously. There are several possible alternatives in this direction by experimenting with many attention variants to attend to different parts. As, it can also be observed from the images, the complexity of scenes can sometime act like hindrance in obtaining appropriate semantic representations of these images. This challenge can be overcome by attention mechanisms.

TEXT	EXPECTED OUPUT	RETRIEVED IMAGE
<p>Says Simple Simon to the pieman, "Let me taste your ware." Says the pieman to Simple Simon, "Do you mean to pay?" Says Simon, "Yes, of course I do!" And then he ran away.</p>		
<p>THERE WAS A CROOKED MAN There was a crooked man, and he went a crooked mile, He found a crooked sixpence against a crooked stile</p>		
<p>Then up she took her little crook, Determined for to find them; She found them indeed, but it made her heart bleed, For they'd left all their tails behind 'em.</p>		

Figure 6.4: Illustrative comparison of correct and retrieved output of the Deep NN model

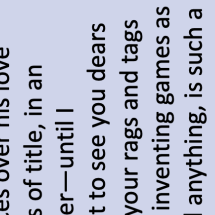
Text	Expected Output	Retrieved Output
<p>Lord Rupert glances over his love letters from ladies of title, in an aristocratic manner—until I could <i>scream</i>. Just to see you dears dancing about in your rags and tags and laughing and inventing games as if you didn't mind anything, is such a relief."</p>	 <p>"Lord Rupert glances over his love letters."</p>	
<p>every broken window and chair leg and table and ragged blanket— and the tears ran down their faces for the first time in their lives. About six o'clock in the morning Peter Piper made a last effort.</p>	 <p>"They were all over the house together."</p>	
<p>fall into fits of giggles and they could only stop them by all joining hands together in a ring and dancing round and round and round and kicking up their heels and laughing until they tumbled down in a heap.</p>	 <p>"They were all over the house together."</p>	

Figure 6.5: Illustrative comparison of correct and retrieved output of the Deep NN model

- Knowledge-bases constructed based on datasets from the real world can also be beneficial in this problem context. A formal way to represent this information can be explored and can provide boost to the various performance metrics discussed above.

6.8 Conclusion

We present two versions of novel GutenStories dataset and also investigated the new task of automatic stepwise text illustration. We provide some evaluation results on text illustration using both using approaches based on topic modelling and deep NN architectures, and highlight the challenges faced with the constructed dataset.

Chapter 7

Conclusions and Future Work

7.1 Main findings

In the current section we summarise our main findings with respect to each of the research questions set up in chapter 1. We have grouped the research questions introduced there according to the two main tasks we have considered in this thesis, stepwise illustration and caption generation.

7.1.1 Stepwise Illustration

In the first part of this thesis, we address the research questions concerned with the task of *Stepwise Illustration*. We formally introduced the problem in Chapter 5.

Research Question 3 How can we automate the process of creating resources for the task of automatic stepwise illustration?

To address this research question. We demonstrated an approach to create an unlabelled dataset of sequenced image-text pairs from any source. We also released two new data repositories, *Stepwise Recipe* and *GutenStories*, consisting of 10K recipes with a total of 67K associated images and 18k visual stories with a total of 90k associated images, respectively, where each segment of text is paired with its corresponding image. Figure 5.1 shows an example of the stepwise instructions and illustrations from a cooking recipe taken from our newly-built dataset. A few selected text recipe instruction steps are shown alongside the full sequence of recipe images. Note that retrieval of an accurate illustration of Step 4, for example, depends on the data model being able to encode the context from the previous steps of the recipe, as the current step adds to preexisting information acquired from earlier steps. Both datasets are web-crawled and systematically filtered and cleaned.

We also highlight the challenges associated with these datasets. We conduct several experimental studies involving traditional as well as some recently published methodologies for evaluating these datasets.

Research Question 4 How can we study the automatic stepwise illustration systems in a domain-constrained setting, given narrative text passage in a limited domain with a sequence of illustrations, considering and incorporating prior context?

We presented VRSS model that given a sequence of text passages, retrieves a sequence of images best describing the semantic content of text and introduced the Stepwise Recipe dataset which will facilitate further research on this problem. Our results on the Stepwise Recipe dataset show that VRSS significantly outperforms competitive baselines in terms of both automatic and human evaluations.

Research Question 1 How can we develop automatic text illustration systems that illustrate a given narrative text passage with a sequence of illustrations, considering and incorporating prior context?

We have studied several approaches to automated text illustration systems for text passages. We classified the approaches into four main categories: atomic image retrieval for a given text passage, atomic text retrieval given an image, sequential image retrieval given a sequence of text passages and sequential text retrieval given a sequence of images. Following a thorough investigation of the existing literature in the field of automated text illustration to gain understanding of how joint-models of texts and images are utilised for the task of semantic and coherent image retrieval. We highlighted some common methods used and proposed new methodologies. We also evaluated the models in a realistic scenario to test their ability to be employed as real-world applications.

7.1.2 News Image Caption Generation

In this second part, we address the research questions concerned with the task of caption generation for images appearing in news articles. We formally introduced the problem in Chapter 4.

Research Question 2 How can we fuse information from different modalities to summarise the given content for developing context-based models?

To address this question, we focused on the atomic image-text passages. We studied the problem of news image caption generation. We conducted a thorough investigative literature survey of the existing approaches employed in

this domain. We identified several datasets that could be utilised and found BBC News Data to be most suitable for this task. An earlier methodology used a mixed-LDA to do cross-modal retrieval for this task. They proposed several abstractive and extractive summarisation based techniques for caption generation. We treated their best performing methodology, which is an extractive summarisation technique, that retrieves the caption having closest topic distribution, computed using KL divergence, as a baseline. We investigated several ways to incorporate semantic features from both the modalities (images and texts) and thus were able to fuse them as contextual information. Furthermore, we proposed a novel deep neural-networks based architecture for the task of automatic caption generation for news images. The experimental evaluation on the BBC News corpus shows that proposed methodology gives a better BLEU score than baseline models and performs similarly compared to the LDA approach on Meteor scores. Nevertheless, we noticed that the captions generated by our approach were favoured over the captions generated by the LDA based model most of time by human evaluators. In future, this model might be extended to a full-fledged encoder-decoder architecture, where the context vector from the LSTM cell used in our model can be passed to another LSTM cell, which acts as a decoder for word sequence generation.

7.2 Directions for future research

There are a few directions in which future work can focus. In this final section, we outline some of the major directions, based on the tasks that were tackled in this thesis.

7.2.1 Stepwise Illustration

Stepwise illustration is a challenging task that can be applied and extended to many domains. While automatic story illustration is a task deemed quite useful for educational and teaching purposes for children. Our proposed illustration models achieved reasonable performance, there is a lot of scope for improvement through addressing the following aspects of the problem.

In our experiments we have highlighted the use of evaluation measures like Visual Saliency Recall and Textual Saliency Recall, which can help in realistically evaluating these models. The models can also be extended using knowledge graphs [45]. In recent literature, knowledge graphs have been incorporated into existing text understanding and image understanding systems.

Also, in recent literature, many innovative methods to extract and represent events have been discussed. Our models can be extended with better text and image representation mechanisms to improve performance. Several examples can be found in Ding et al. [28], Martin et al. [95].

The presented VRSS model can also be extended to perform image synthesis. The synthesised image embedding point associated with every text embedding point can be employed for either image generation or image retrieval as desired. For example, Li et al. [83] propose a sequential image generation model to visualise a sequence of sentences. Turkoglu et al. [135] discuss a general framework for sequential image generation.

7.2.2 News Caption Generation

The proposed novel deep NN-based architecture for the task of automatic caption generation for news images can easily be extended in several ways. The experimental evaluation on the BBC News corpus show that proposed methodology gives a better BLEU score than baseline models and performs similarly compared to the LDA approach on Meteor scores. Nevertheless, we notice that the captions generated by our approach were favoured over the captions generated by the LDA based model most of time by human evaluators. In future, this model can be extended to a full-fledged encoder-decoder architecture, where the context vector from the LSTM cell used in our model can be passed to another LSTM cell, which acts as a decoder for word sequence generation.

7.3 Summary

In this thesis, we addressed and formalised the task of *sequence-to-sequence (seq2seq) cross-modal retrieval*. Given a sequence of text passages as query, the goal is to retrieve a sequence of images that best describes and aligns with the query. This new task extends the traditional cross-modal retrieval, where each image-text pair is treated independently ignoring broader context. We proposed a novel *variational recurrent seq2seq (VRSS) retrieval model* for this seq2seq task. Unlike most cross-modal methods, we generate an image vector corresponding to the latent topic obtained from combining the text semantics and context. This synthetic image embedding point associated with every text embedding point can then be employed for either image generation or image retrieval as desired. We evaluate the model for the application of *stepwise*

illustration of recipes, where a sequence of relevant images are retrieved to best match the steps described in the text. To this end, we built and released two new multimodal data repositories *Stepwise Recipe* dataset and *GutenStories* dataset. To our knowledge, these are the first publicly available dataset to offer rich semantic descriptions in a sequenced manner. We also provide qualitative analysis of how semantically meaningful the results produced by our model are through human evaluation and comparison with relevant existing methods. We also proposed new models and studied existing models for the task of context-dependent text illustration and description retrieval. We studied several models incorporating different kinds of features to study the importance of these features in modeling and retrieval of a relevant image from a repository to match the semantics of the given text. We studied and compared the performance of several previously published and newly proposed methodologies on GutenStories dataset.

Appendix A

Further Analysis and Results

In the section ahead, we present results by varying dropout levels from the joint embedding learning methodology in Chapter 6.

A.1 Effects of introducing the dropout layer

In this section, we study the effects of introducing the dropout layer in some of the discussed models. As most of the data we are dealing with, contains some amount of noise, as also discussed in the challenges section. We introduce a dropout layer in our deep NN architecture.

We also conducted experiments over MS-COCO dataset [86] and observed an improved performance throughout all training iterations by introducing dropout. Figures A.1, A.2 and A.3 show visualisation of top-1, top5 and top-10 retrieval scores of images and texts respectively, over training iterations by varying dropout levels. Here the top graph in each figure represents text retrieval and bottom one represents image retrieval. The orange, blue, red and sky blue represents dropout levels of 0, 0.2, 0.4 and 0.8 respectively.

A.2 Effects of varying the CNN architecture

In this section, we study the effects of varying the CNN architecture in the deep-NN methodology presented in Chapter 6.

Here, the experiments are conducted over mini-GutenStories dataset and observed an improved performance throughout all training iterations by using ResNet architecture. Figures A.4 and A.5 show visualisation of top-1, top5 and top-10 retrieval scores of texts and images respectively, over training iterations by varying CNN architectures. Here the top graph in each figure represents

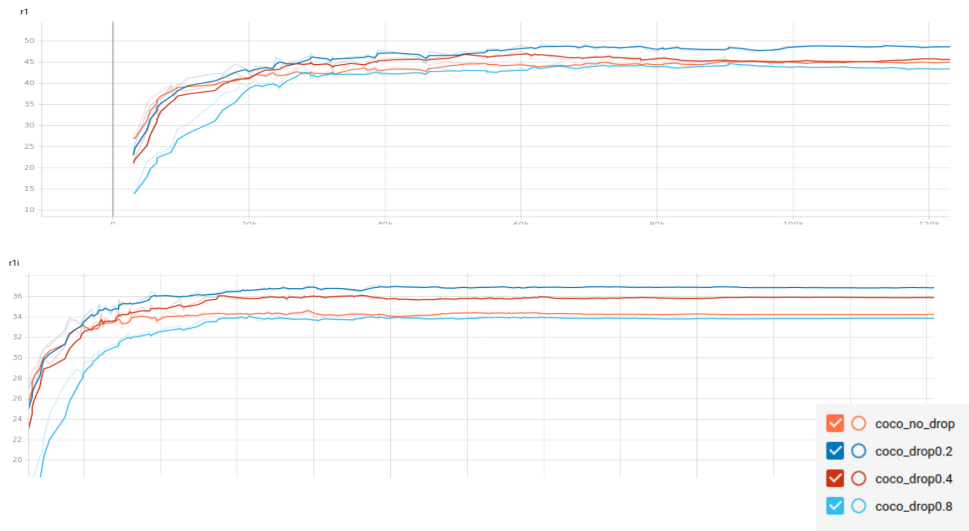


Figure A.1: Visualisation of Top-1 retrieval scores of images and texts, over training iterations by varying dropout levels

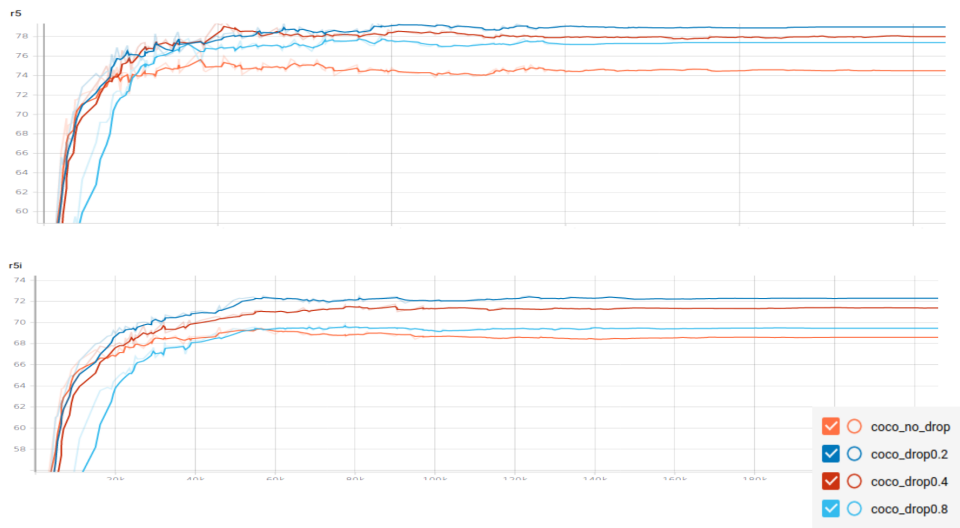


Figure A.2: Visualisation of Top-5 retrieval scores of images and texts, over training iterations by varying dropout levels

top-1, middle one top-5 and bottom one top-10 retrieval scores. Orange is the best performing with ResNet152 and dropout of 0.2.

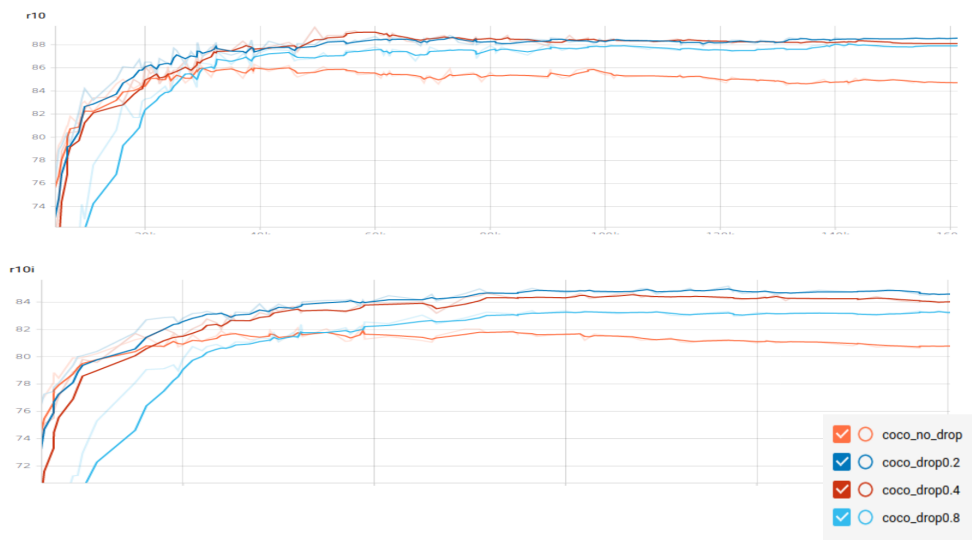


Figure A.3: Visualisation of Top-10 retrieval scores of images and texts, over training iterations by varying dropout levels

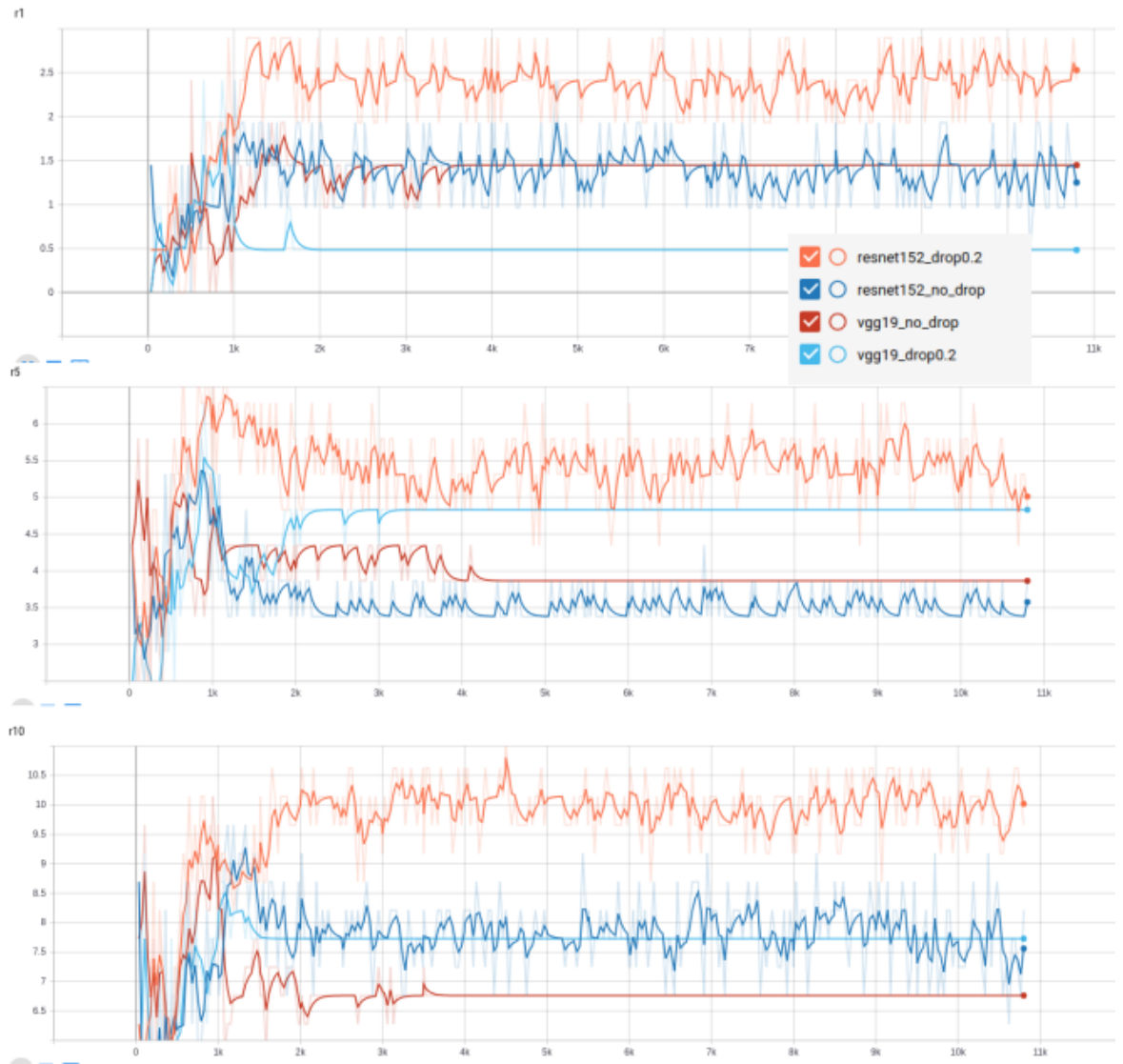


Figure A.4: Visualisation of Top-1, Top-5 and Top-10 retrieval scores of texts, over training iterations by varying image representation mechanisms

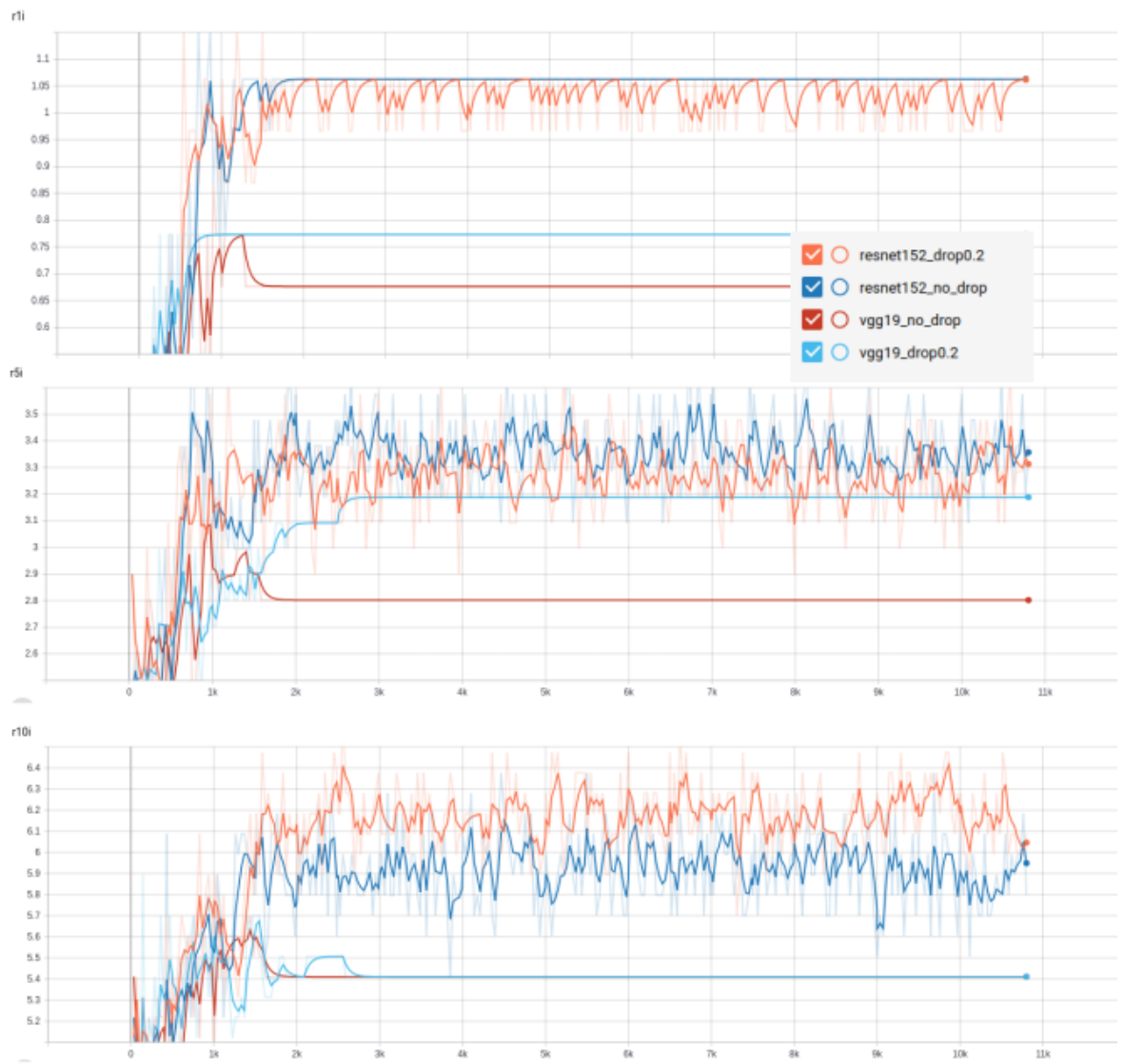


Figure A.5: Visualisation of Top-1, Top-5 and Top-10 retrieval scores of images, over training iterations by varying image representation mechanisms

Bibliography

- [1] Somak Aditya, Yezhou Yang, and Chitta Baral. Integrating knowledge and reasoning in image understanding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6252–6259. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/873. URL <https://doi.org/10.24963/ijcai.2019/873>.
- [2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual Question Answering. pages 1–23, 2015. URL <http://arxiv.org/abs/1505.00468>.
- [3] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. Cite: A corpus of image–text discourse relations. *arXiv preprint arXiv:1904.06286*, 2019.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [5] Ben Athiwaratkun and Andrew Gordon Wilson. Hierarchical density order embeddings. *CoRR*, abs/1804.09843, 2018. URL <http://arxiv.org/abs/1804.09843>.
- [6] Saeid Balaneshin-kordan and Alexander Kotov. Deep neural architecture for multi-modal retrieval based on joint embedding space for text and images. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 28–36. ACM, 2018.
- [7] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003. ISSN 1532-4435. doi: 10.1162/153244303322533214.

- [8] Kobus Barnard, Quanfu Fan, Ranjini Swaminathan, Anthony Hoogs, Roderic Collins, Pascale Rondot, and John Kaufhold. Evaluation of localized semantics: Data, methodology, and experiments. *International Journal of Computer Vision*, 77:199–217, 05 2008. doi: 10.1007/s11263-007-0068-6.
- [9] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–848–II–854 Vol.2, June 2004. doi: 10.1109/CVPR.2004.1315253.
- [10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003. ISSN 1532-4435.
- [11] David Meir Blei. *Probabilistic Models of Text and Images*. PhD thesis, Berkeley, CA, USA, 2004. AAI3183785.
- [12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016. URL <http://arxiv.org/abs/1607.04606>.
- [13] Mohammad Ubaidullah Bokhari and Faraz Hasan. Multimodal information retrieval: Challenges and future trends. *International Journal of Computer Applications*, 74(14), 2013.
- [14] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI’11*, pages 301–306. AAAI Press, 2011. URL <http://dl.acm.org/citation.cfm?id=2900423.2900470>.
- [15] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014.
- [16] Gilbert J. Botvin and Brian Sutton-Smith. The development of structural complexity in children’s fantasy narratives. 1977.
- [17] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44. ACM, 2018.

- [18] X. Chen and C. L. Zitnick. Mind’s eye: A recurrent visual representation for image caption generation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2422–2431, June 2015. doi: 10.1109/CVPR.2015.7298856.
- [19] Jianpeng Cheng and Mirella Lapata. Neural Summarization by Extracting Sentences and Words. *Arxiv*, pages 484–494, 2016. URL <http://arxiv.org/abs/1603.07252>.
- [20] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 1, pages 539–546 vol. 1, June 2005. doi: 10.1109/CVPR.2005.202.
- [21] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.
- [22] F. Coelho and C. Ribeiro. Automatic illustration with cross-media retrieval in large-scale collections. In *2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 25–30, June 2011. doi: 10.1109/CBMI.2011.5972515.
- [23] Diogo Delgado, Joao Magalhaes, and Nuno Correia. Assisted news reading with automated illustration. In *Proceedings of the 18th ACM International Conference on Multimedia, MM ’10*, pages 1647–1650, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874311. URL <http://doi.acm.org/10.1145/1873951.1874311>.
- [24] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

- [27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>.
- [28] Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, and Junwen Duan. Event representation learning enhanced with external commonsense knowledge, 2019.
- [29] Aviv Eisenschat and Lior Wolf. Capturing deep correlations with 2-way nets. *CoRR*, abs/1608.07973, 2016. URL <http://arxiv.org/abs/1608.07973>.
- [30] Erkan G. and D Radev. LexRank: Graph-based Lexical Centrality As Saliency in Text Summarization. *Journal of Artificial Intelligence*, 22(1): 457–479, 2004. ISSN 1076-9757. doi: 10.1613/jair.1523.
- [31] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.
- [32] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6314 LNCS(PART 4):15–29, 2010. ISSN 03029743. doi: 10.1007/978-3-642-15561-1_2.
- [33] Roman Feldbauer, Maximilian Leodolter, Claudia Plant, and Arthur Flexer. Fast approximate hubness reduction for large high-dimensional data. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, pages 358–367. IEEE, 2018.
- [34] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014.
- [35] Y. Feng and M. Lapata. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4): 797–812, April 2013. ISSN 1939-3539. doi: 10.1109/TPAMI.2012.118.
- [36] Yansong Feng and Mirella Lapata. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, Stroudsburg, PA,

- USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL <http://dl.acm.org/citation.cfm?id=1857999.1858124>.
- [37] Yansong Feng and Mirella Lapata. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2012.118.
- [38] Basura Fernando. Guiding Long-Short Term Memory for Image Caption Generation.
- [39] Leo Ferres, Avi Parush, Shelley Roberts, and Gitte Lindgaard. Helping people with visual impairments gain access to graphical information through natural language: The igrph system. In *Proceedings of the 10th International Conference on Computers Helping People with Special Needs, ICCHP’06*, pages 1122–1130, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-36020-4, 978-3-540-36020-9. doi: 10.1007/11788713_163. URL http://dx.doi.org/10.1007/11788713_163.
- [40] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL ’05*, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL <https://doi.org/10.3115/1219840.1219885>.
- [41] Andrew B. Goldberg, Xiaojin Zhu, Charles R. Dyer, Mohamed Eldawy, and Lijie Heng. Easy as abc?: Facilitating pictorial communication via semantically enhanced layout. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL ’08*, pages 119–126, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-48-4. URL <http://dl.acm.org/citation.cfm?id=1596324.1596345>.
- [42] Yunchao Gong, Liwei Wang, Micah Hodosh, and Julia Hockenmaier. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. pages 1–16.
- [43] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>.

- [44] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL <http://authors.library.caltech.edu/7694>.
- [45] Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2026–2037. Curran Associates, Inc., 2018. URL <http://papers.nips.cc/paper/7473-embedding-logical-queries-on-knowledge-graphs.pdf>.
- [46] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
- [47] K. Hartmann and Th. Strothotte. A spreading activation approach to text illustration. In *Proceedings of the 2Nd International Symposium on Smart Graphics*, SMARTGRAPH '02, pages 39–46, New York, NY, USA, 2002. ACM. ISBN 1-58113-555-6. doi: 10.1145/569005.569012. URL <http://doi.acm.org/10.1145/569005.569012>.
- [48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL <http://arxiv.org/abs/1512.03385>.
- [49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL <http://arxiv.org/abs/1512.03385>.
- [51] Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7):1363–1377, 2016.
- [52] Micah Hodosh. Cross-caption coreference resolution for automatic image understanding. (July):162–171, 2010.
- [53] Thomas Hofmann. Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196, 2001. ISSN 08856125. doi: 10.1023/A:1007617005950.
- [54] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet

- Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016.
- [55] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. June 2016. URL <https://www.microsoft.com/en-us/research/publication/visual-storytelling/>.
- [56] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016.
- [57] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08*, pages 39–43, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-312-9. doi: 10.1145/1460096.1460104. URL <http://doi.acm.org/10.1145/1460096.1460104>.
- [58] Chihli Hung and Chih-Fong Tsai. Automatically Annotating Images with Keywords: A Review of Image Annotation Systems. *Recent Patents on Computer Science*, 1:55–68, 2008. ISSN 22132759. doi: 10.2174/2213275910801010055.
- [59] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL <http://www.scipy.org/>.
- [60] K.S. Jones and Others. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12, 1999. doi: 10.1145/375551.375604. URL <http://books.google.com/books?hl=en&lr=&id=YtUZQaKDmzEC&oi=fnd&pg=PA1&dq=Automatic+summarising++factors+and+directions&ots=ZnsusoI{ }9D&sig=a497DHP{ }NwK5jA-qnw-yCVvCc>.
- [61] Dhiraaj Joshi, James Z. Wang, and Jia Li. The story picturing engine—a system for automatic text illustration. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):68–89, February 2006. ISSN 1551-6857. doi: 10.1145/1126004.1126008. URL <http://doi.acm.org/10.1145/1126004.1126008>.

- [62] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759, 2016. URL <http://arxiv.org/abs/1607.01759>.
- [63] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, April 2017. ISSN 1939-3539. doi: 10.1109/TPAMI.2016.2598339.
- [64] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, April 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2598339. URL <https://doi.org/10.1109/TPAMI.2016.2598339>.
- [65] Andrej Karpathy and Fei Fei Li. Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June: 3128–3137, 2015. ISSN 10636919. doi: 10.1109/CVPR.2015.7298932.
- [66] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, pages 1889–1897, Cambridge, MA, USA, 2014. MIT Press. URL <http://dl.acm.org/citation.cfm?id=2969033.2969038>.
- [67] DongHwa Kim, Deokseong Seo, Suhyoun Cho, and Pilsung Kang. Multi-co-training for document classification using various document representations: Tf-idf, lda, and doc2vec. *Inf. Sci.*, 477:15–29, 2019.
- [68] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. Ranking and retrieval of image sequences from multiple paragraph queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1993–2001, 2015.
- [69] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. URL <http://arxiv.org/abs/1411.2539>.
- [70] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. URL <http://arxiv.org/abs/1411.2539>.
- [71] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-Thought Vectors. (786):1–9, 2015. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004. URL <http://arxiv.org/abs/1506.06726>.

- [72] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015.
- [73] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015. URL <http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf>.
- [74] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4437–4446, June 2015. doi: 10.1109/CVPR.2015.7299073.
- [75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL <https://doi.org/10.1145/3065386>.
- [76] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2012.162.
- [77] Kevin J. Lang, Alex H. Waibel, and Geoffrey E. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural Netw.*, 3(1):23–43, January 1990. ISSN 0893-6080. doi: 10.1016/0893-6080(90)90044-L. URL [http://dx.doi.org/10.1016/0893-6080\(90\)90044-L](http://dx.doi.org/10.1016/0893-6080(90)90044-L).
- [78] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07*, pages 228–231, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
- [79] Victor Lavrenko, R. Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 553–560. MIT Press, 2004. URL <http://papers.nips.cc/paper/2474-a-model-for-learning-the-semantics-of-pictures.pdf>.

- [80] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, page II–1188–II–1196. JMLR.org, 2014.
- [81] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *CoRR*, abs/1803.08024, 2018. URL <http://arxiv.org/abs/1803.08024>.
- [82] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*, 2018.
- [83] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David E. Carlson, and Jianfeng Gao. Storygan: A sequential conditional GAN for story visualization. *CoRR*, abs/1812.02784, 2018. URL <http://arxiv.org/abs/1812.02784>.
- [84] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, June 2004. doi: 10.1109/CVPR.2004.383.
- [85] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL <http://arxiv.org/abs/1405.0312>.
- [86] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL <http://arxiv.org/abs/1405.0312>.
- [87] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [88] Jonathan Long, Evan Shelhamer, Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Karel Lenc, Andrea Vedaldi, Emily Denton, Soumith Chintala, Arthur Szlam, Rob Fergus, Philipp Fischer, H Philip, Caner Hazırbas, Patrick Van Der Smagt, Daniel Cremers, Thomas Brox,

- Fandong Meng, Zhengdong Lu, Zhaopeng Tu, Hang Li, Qun Liu, Vijay Mahadevan, and Student Member. Show and Tell: A Neural Image Caption Generator. *arXiv*, 32(1):1–10, 2014. ISSN 9781467369640. doi: 10.1109/CVPR.2015.7298935. URL <http://arxiv.org/abs/1411.5908v1>.
- [89] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13–23. Curran Associates, Inc., 2019.
- [90] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. *CoRR*, abs/1504.06063, 2015. URL <http://arxiv.org/abs/1504.06063>.
- [91] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [92] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- [93] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m: A dataset for learning cross-modal embeddings for cooking recipes and food images. *arXiv preprint arXiv:1810.06553*, 2018.
- [94] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, July 2001. doi: 10.1109/ICCV.2001.937655.
- [95] Lara J. Martin, Prithviraj Ammanabrolu, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. Event representations for automated story generation with deep neural nets. *CoRR*, abs/1706.01331, 2017. URL <http://arxiv.org/abs/1706.01331>.
- [96] Anne McKeough and Jennifer Malcolm. Stories of family, stories of self: Developmental pathways to interpretive thought during adolescence. *New Directions for Child and Adolescent Development*, 2011(131):59–71, 2011.

doi: 10.1002/cd.289. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/cd.289>.

- [97] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
- [98] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal Deep Learning. *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pages 689–696, 2011. doi: 10.1145/2647868.2654931.
- [99] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.
- [100] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. *Nips*, pages 1143–1151, 2011. ISSN 9781618395993.
- [101] JY Pan, Hyungjeong Yang, and Christos Faloutsos. MMSS: Multi-modal story-oriented video summarization. *Data Mining, 2004. ICDM'04. . . .*, (22):1–4, 2004. doi: 10.1109/ICDM.2004.10033. URL http://ieeexplore.ieee.org/xpls/abs/_all.jsp?arnumber=1410343.
- [102] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://doi.org/10.3115/1073083.1073135>.
- [103] Cesc C. Park and Gunhee Kim. Expressing an image stream with a sequence of natural sentences. In *NIPS*, 2015.
- [104] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [105] F. Pedregosa, 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.

- Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [106] S Pinker. *Language Learnability and Language Development (1984/1996)*. Cambridge, MA: Harvard University Press, 1996.
- [107] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan Carlos Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.
- [108] Novi Quadrianto and Christoph Lampert. Learning multi-view neighborhood preserving projections. In *Proceedings of the 28th International Conference on Machine Learning; Washington, USA; 28 June-2 July 2011*, pages 425–432. Association for Computing Machinery, 2011.
- [109] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010.
- [110] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.
- [111] Hareesh Ravi, Lezi Wang, Carlos Muniz, Leonid Sigal, Dimitris Metaxas, and Mubbasir Kapadia. Show me a story: Towards coherent neural story illustration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7613–7621, 2018.
- [112] Scott E. Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *CoRR*, abs/1605.05396, 2016. URL <http://arxiv.org/abs/1605.05396>.
- [113] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop” Subspace, Latent Structure and Feature Selection”*, pages 34–51. Springer, 2005.
- [114] Alexander M Rush, Sumit Chopra, and Jason Weston. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (September):379–389, 2015. ISSN 19909772. doi: 10.1162/153244303322533223. URL <http://arxiv.org/abs/1509.00685>.

- [115] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1):157–173, May 2008. ISSN 1573-1405. doi: 10.1007/s11263-007-0090-8. URL <https://doi.org/10.1007/s11263-007-0090-8>.
- [116] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, abs/1402.1128, 2014. URL <http://arxiv.org/abs/1402.1128>.
- [117] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. URL <http://arxiv.org/abs/1606.03498>.
- [118] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028, 2017.
- [119] Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3:210–229, 1959.
- [120] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):754–766, April 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.133. URL <https://doi.org/10.1109/TPAMI.2010.133>.
- [121] Pierre Sermanet and David Eigen. OverFeat : Integrated Recognition , Localization and Detection using Convolutional Networks arXiv : 1312 . 6229v4 [cs . CV] 24 Feb 2014.
- [122] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/P18-1238>.
- [123] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL <http://arxiv.org/abs/1409.1556>.

- [124] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL <http://arxiv.org/abs/1409.1556>.
- [125] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.
- [126] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 926–934. Curran Associates, Inc., 2013.
- [127] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. doi: 10.1162/tacl.a-00177. URL <https://www.aclweb.org/anthology/Q14-1017>.
- [128] Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. Variational recurrent neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [129] Lei Sun and Marilyn A. Nippold. Narrative writing in children and adolescents: Examining the literate lexicon. *Language, Speech, and Hearing Services in Schools*, 43(1):2–13, 2012. doi: 10.1044/0161-1461(2011/10-0099).
- [130] Ilya Sutskever. Sequence to Sequence Learning with Neural Networks. pages 1–9.
- [131] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17*, pages 4278–4284. AAAI Press, 2017. URL <http://dl.acm.org/citation.cfm?id=3298023.3298188>.
- [132] Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. Natural language inference with hierarchical bilstm max pooling architecture. *CoRR*, abs/1808.08762, 2018.
- [133] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015. URL <http://arxiv.org/abs/1503.01817>.

- [134] Tomas Mikolov and Ilya Sutskever and Kai Chen and Greg Corrado and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *CoRR*, abs/1310.4546, 2013. URL <http://arxiv.org/abs/1310.4546>.
- [135] Mehmet Ozgur Turkoglu, William Thong, Luuk J. Spreeuwers, and Berkay Kicanaoglu. A layer-based sequential framework for scene generation with gans. *CoRR*, abs/1902.00671, 2019. URL <http://arxiv.org/abs/1902.00671>.
- [136] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *CoRR*, abs/1511.06361, 2015. URL <http://arxiv.org/abs/1511.06361>.
- [137] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *CoRR*, abs/1511.06361, 2015.
- [138] Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. Image-text cross-modal retrieval via modality-specific feature learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 347–354. ACM, 2015.
- [139] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.
- [140] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. Effective deep learning-based multi-modal retrieval. *The VLDB Journal - The International Journal on Very Large Data Bases*, 25(1): 79–101, 2016.
- [141] William Yang Wang, Yashar Mehdad, Dragomir R Radev, and Amanda Stent. A Low-Rank Approximation Approach to Learning Joint Embeddings of News Stories and Images for Timeline Summarization. *Naacl2016*, pages 58–68, 2016.
- [142] Zeyi Wen, Xingyang Liu, Hongjian Cao, and Bingsheng He. Rtsi: An index structure for multi-modal real-time search on live audio streaming services. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1495–1506. IEEE, 2018.
- [143] Peratham Wiriyathamabhun, Douglas Summers-Stay, Cornelia Fermüller, and Yiannis Aloimonos. Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Comput. Surv.*, 49(4):71:1–71:44, December 2016. ISSN 0360-0300. doi: 10.1145/3009906. URL <http://doi.acm.org/10.1145/3009906>.

- [144] Chenyan Xiong, Zhengzhong Liu, Jamie Callan, and Tie-Yan Liu. Towards better text understanding and retrieval through kernel entity salience modeling. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 575–584, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3209982. URL <http://doi.acm.org/10.1145/3209978.3209982>.
- [145] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.
- [146] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, and Ruslan Salakhutdinov. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. ISSN 19410093. doi: 10.1109/72.279181.
- [147] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL <http://proceedings.mlr.press/v37/xuc15.html>.
- [148] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [149] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017. URL <http://arxiv.org/abs/1711.10485>.
- [150] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*, 2018.
- [151] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, Aug 2010. ISSN 1558-2256. doi: 10.1109/JPROC.2010.2050411.

- [152] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *CoRR*, abs/1603.03925, 2016. URL <http://arxiv.org/abs/1603.03925>.
- [153] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image Captioning with Semantic Attention. *Cvpr*, (1):10, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.503. URL <http://arxiv.org/abs/1603.03925>.
- [154] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.
- [155] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, Aug 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2856256.
- [156] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1612.03242, 2016. URL <http://arxiv.org/abs/1612.03242>.
- [157] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. *CoRR*, abs/1802.09178, 2018. URL <http://arxiv.org/abs/1802.09178>.