Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Cross Media Relations across news domains

Anoushka Mittal

August 19, 2022

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MSc (Computer Science - Data Science)

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed: ______________________ Date: ______________________

Abstract

Recent years have seen a tremendous up flux of data. In NLP, the massive amount of data growth became the motivation for the development of text processing systems which can effectively process data. However, the data on the internet tends to be more versatile where multiple modalities like images, audio, video and text are intertwined. Hence there is a need to develop systems that can process the different modalities together.

Visual entailment is a task involving multiple modalities. It is inspired by the textual entailment task in linguistics which aims to classify the relation between a text premise and text hypothesis. The Visual Entailment task aims to classify the relation between an image premise and a text hypothesis. This study aims to analyse the use of the VE task to observe how the cross-media relations of image-caption pairs in news articles differ across categories of news. This is a challenging task as it involves both computer vision and natural language processing. Image-caption pairs along with article text are scraped using links to news articles provided in the Kaggle News Category data set. The One-For-All framework is utilised to caption images reducing the VE task to textual entailment. CLIP is applied to rank the actual and generated captions with the intent to draw conclusions depending based upon which type of caption ranks higher. Clustering and statistical analyses are performed on the CLIP generated ranks. BERT is used to find the semantic similarity between actual and generated captions.

It is found that captioning systems like OFA cannot be employed for reducing VE to TE for news articles. While the cross-media relations do not differ across the categories of news, there is significant interaction between the CLIP ranks for the actual and generated image captions and the entailment relations. A survey of different existing datasets is provided followed by a new dataset of 40K news articles containing image-caption pairs from the HuffPost website. It is hoped that this research work can act as a reference for future work in this area.

Acknowledgements

I first and foremost thank my supervisor, Dr.Carl Vogel for his constant support and continuous guidance during this project, and utmost understanding and empathy during some difficult times over the course of the past year. I thank Trinity College Dublin and the School of Computer Science and Statistics for providing me the opportunity to work on this project.

I thank Dr.Fuxiao Liu, Researcher at the University of Maryland for providing access to the VisualNews Dataset.

I thank my family for supporting me through my education and especially my mother for always motivating me, without which I would not have been able to present this work. I also thank my friends for making my past year in Dublin a memorable one and for their words of encouragement through the course of the project.

Contents

List of Figures

List of Tables

Nomenclature

BERT	Bidirectional Encoder Representations from Transformers	
CLIP	Contrastive Language Image Pre-training	
COSMOROE	CrOSs-Media inteRactiOn rElations	
CST	Cross-Document Structure Theory	
NLI	Natural Language Inference	
NLP	Natural Language Processing	
OFA	One-For-All Framework	
RST	Rhetorical Structure Theory	
RTE	Recognising Textual Entailment	
SNLI	Stanford Natural Language Inference	
SOTA	State-Of-The-Art VE	Visual Entailment

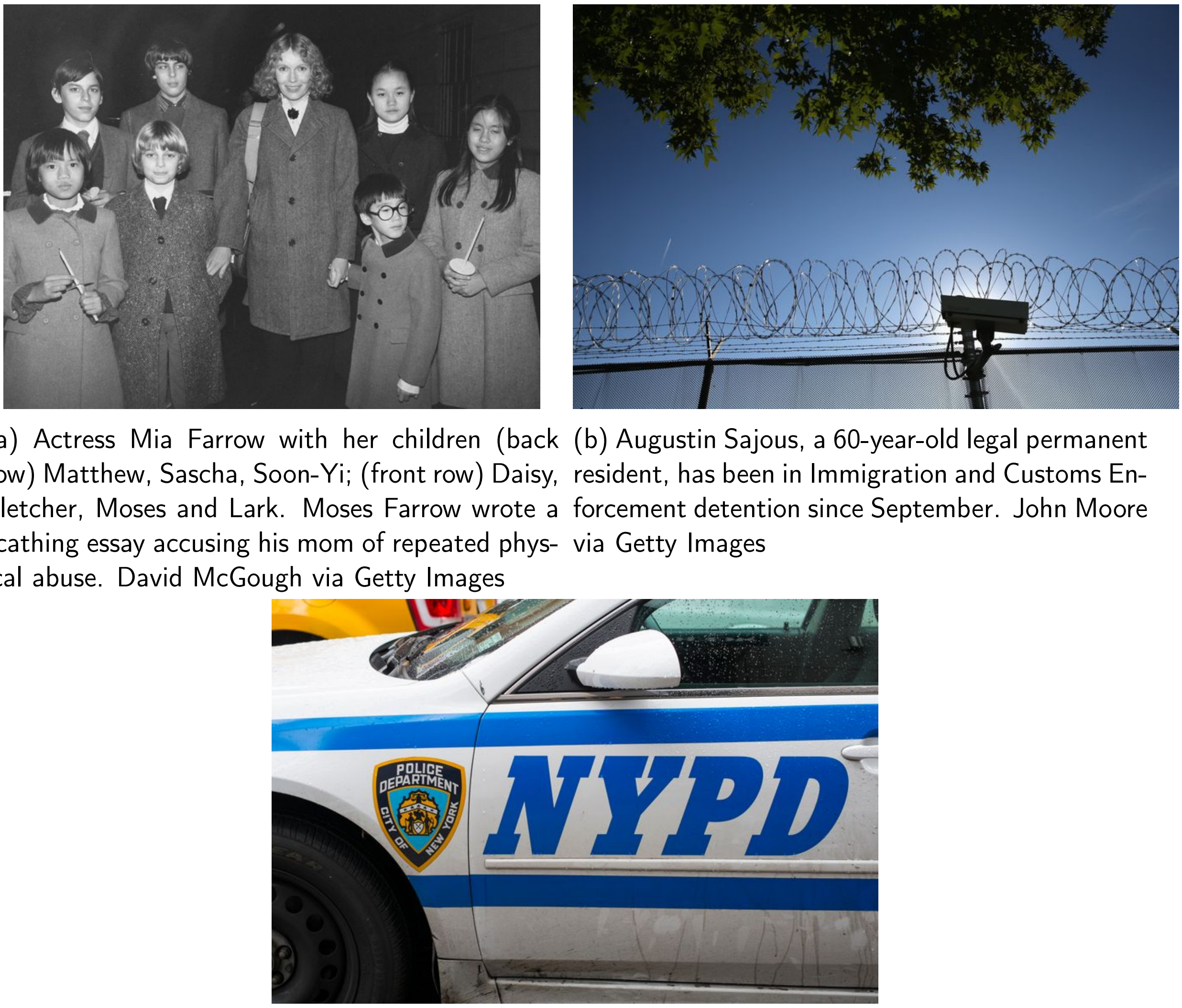
1 Introduction

Ever since the pandemic, there has been a tremendous increase in the amount of data involving multiple modalities like video, audio, images, and text on account of work and academics shifting online. And even prior to the pandemic, multimodal data was present in the form of videos and their captions on YouTube, images and their captions in news or social media. The field of multimodal research has seen much interest recently and it acts as a bridge between the computer vision and the Natural Language Processing communities. The goal of multimodal research is to make computers understand semantics and learn to correlate and connect the information being presented to them in various ways. This is required in order to create systems which can comprehend and process such data effectively similar to the motivation behind the development of text processing systems to deal with the massive amounts of textual data (Binti Zahri et al. (2012)). Recognising Textual Entailment (RTE) is a text processing task in which the relation between a text hypothesis and a text premise has to be classified. Table 1.1 presents examples of the RTE task.

The Visual Entailment task is an extension of the RTE task and requires classification of the relation between a text hypothesis and an image premise. This study focuses on how the Visual entailment relations between image-caption pairs in articles vary across categories of news. Figure 1.1 explains the problem in further detail. More examples of the Visual Entailment task

A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C	The man is sleeping
An older and a younger man smiling	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C	A man is driving down a lonely road
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella

Table 1.1: Examples of the textual entailment task, taken from Bowman et al. (2015). Here, the premise and hypothesis are both from the textual medium



(a) Actress Mia Farrow with her children (back row) Matthew, Sascha, Soon-Yi; (front row) Daisy, Fletcher, Moses and Lark. Moses Farrow wrote a scathing essay accusing his mom of repeated physical abuse. David McGough via Getty Images

(b) Augustin Sajous, a 60-year-old legal permanent resident, has been in Immigration and Customs Enforcement detention since September. John Moore via Getty Images

(c) Jamilla Clark and Arwa Aziz say the New York Police Department's policy requiring women to remove religious head coverings for booking photos violates their civil rights.icenando via Getty Images

Figure 1.1: RTE relations in a cross media context in the VE task. Figure 1.1a(article accessed 19.08.2022) is an entailment relation because most named entities are present in both image and caption. Figure 1.1b(article accessed 19.08.2022) is a contradiction as the image does not represent any of the entities mentioned in the caption. Figure 1.1c(article accessed 19.08.2022) is neutral because while a form of an entity mentioned in the caption is present in the picture however, the other entities are not. Since this cannot be called entailment or contradiction, it is classified as a neutral relation. In the RTE Task, both the premise and hypothesis are from the same medium whereas in the Visual Entailment task, the premise is an image and the hypothesis is a text.

are presented in Figure 2.9 found in Section 2.2.9. There are 3 main relations: Entailment, Contradiction and Neutral. The VE relations have been borrowed from the RTE task.

The nature of this study is an exploratory one surveying different existing tools, techniques and data sets to try and form a methodology for addressing the problem of VE in news articles. The previous existing data sets are presented by comparing the number of data points, attributes, structure, features and some issues. A data set of approximately 40K news articles with image-caption pairs is collected from the Huff Post website which was not covered in previous data sets. The collection of the data set does not require using an API and so can be scraped as required. This document hopes to act as a resource collating the many different data sets and relevant terms encountered giving a brief overview.

The first chapter introduces the topic of the dissertation and while giving a brief overview of the same with regards to the motivation for pursuing this topic, the various challenges faced and the possible challenges for future researchers pursuing work in this field as well as a brief summary of the structure and content of the rest of the dissertation.

1.1 Motivation

Images are omnipresent and are often followed by captions or text and such image-text pairs can be found everywhere on the Internet ranging from social media sites like Instagram, Twitter, Facebook to news articles and blogs. Images facilitate better understanding of convoluted concepts and also increase retention. They are also preferred when spatial information is involved. Images are often followed by caption to provide the context in which the image is being used. It is also often said of talks/papers/textbooks that they are improved by including images additional to the text. Work in this area will help in answering questions like: will any image/text relation suffice, or is there an optimal distribution of cross-media relations that can be said to suffice, and if so, knowing the standard distributions of cross media relations provides an "image/text style" checker. In social media, image captions can be very versatile. Similarly, in news articles, images are often used to depict varying concepts, scenes and entities. For example, if a request has to be issued to the public about a wanted criminal, a textual medium would be inefficient at getting the information across. However, if only the image is included, one might need to read the entire article to be able to understand what the image is about and captions help in giving the context in which the image is being used. Captions are used for more purposes than just providing context. They may be used to present analogies or satires (Refer Figure 1.2). Sometimes the captions may not exactly relate to the image or could even be contradictory. Hence it is a very challenging and interesting problem to try and attempt to recognise the relations between the images and captions in news articles. And it is also interesting to know how the relations differ across the different categories of news. Research in the Visual Entailment field is also required to advance identification of multimedia interaction relationships for intelligent image-language multimedia systems such as Embodied Conversational Agents (ECA), robots and intelligent systems. For example, if a video of a human talking is presented then can the agent correlate what is being said to the expression of the human. While this project does not deal with such a high level application,

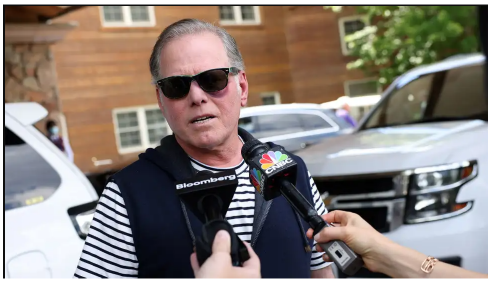


Figure 1.2: Article text from The Onion(accessed 19.08.2022): "BURBANK, CA—Visibly shaken with fresh bruising on his face, Warner Bros. Discovery CEO David Zaslav announced Friday that the studio had no plans to scrap upcoming DC Comics film The Flash at this time."

the current existing solutions for Visual Entailment are based on data sets like MS-COCO in which the images describe very general everyday life scenes and the associated captions are also very generic and descriptive in nature (Liu et al. (2020)). Generally in news articles, the captions involve named entities and places and are also linguistically richer than being plain simple descriptive captions. Hence, attempting to apply the Visual Entailment task in the news domain has potential to take this task further into real world scenarios. In the field of fake news detection, some works have started to consider making use of multimodal analysis of the images involved for making better predictions and similarly, this has potential applications or can be further extended to how the cross-media relations vary or differ across the fake news and real news. Apart from this, the analysis of cross-media relations in news articles can be utilised to see how they vary across sites which are generally referred to as tabloids and sites which are referred to as reputed. Or how the cross-media relations vary within the same category of news. For example in politics, whether or not there is a difference in the cross-media relations between articles which can be regarded as propaganda and articles which are not.

1.2 Objectives

The previous section highlights many interesting applications and extensions of the proposed research question. Formally our research question is to find out the distribution of cross media relations in the different news domains. In order to answer our question, the following tasks require analysis:

1. Identifying the relevant data sets to the research project or its sub-tasks
2. Understand the currently available solutions in the field of Visual Entailment
3. Identify the distribution of the relations exhibited across the data
4. Develop automated means to support the observational study

1.3 Contributions

Some of the contributions of this work are as follows:

1. Presents a survey of the different data sets existing in the Visual Entailment space and other data sets which are not directly related but could be beneficial for applying the VE task to the news domain
2. Contributes a data set with 40K points containing the images and corresponding captions along with article text (similar to the N24News data set and the VisualNews Data set) for the HuffingtonPost website
3. Performs an evaluation of techniques which could be used to solve the problem of multimodal entailment on a manually annotated data set of 100 articles annotated with the labels entailment, contradiction, neutral and also with relations considering the COSMOROE framework : equivalence, independence, complementarity
4. Compares some existing Image Captioning systems
5. Provides a comprehensive review of the concepts involved
6. Proposes use of a different cross-media relations framework rather than making use of relations borrowed from the Recognising Textual Entailment task.

1.4 Challenges

This topic is particularly challenging because

1. It lies at the intersection of the 2 very challenging fields of Computer Vision and Natural Language Processing. Apart from this, as it involves the news domain, the computer vision capabilities also need to be able to recognise named entities such as particular faces and places.
2. Visual Entailment relations have never been explored in news articles. Visual Entailment solutions exist for scenes where everyday life common scenes are presented in the picture but news domains present scenarios never encountered earlier.
3. There exist data sets which allow for the captioning of images in news articles and there also exist data sets for Visual Entailment however, there do not exist data sets which tell us about the entailment relations between images and captions in news articles. Also because in normal Visual Entailment tasks, there is no other content whereas in news articles there is the article content which can affect the entailment relations between the image and the caption

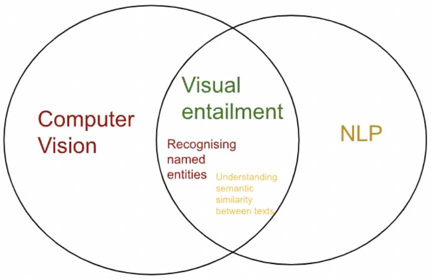
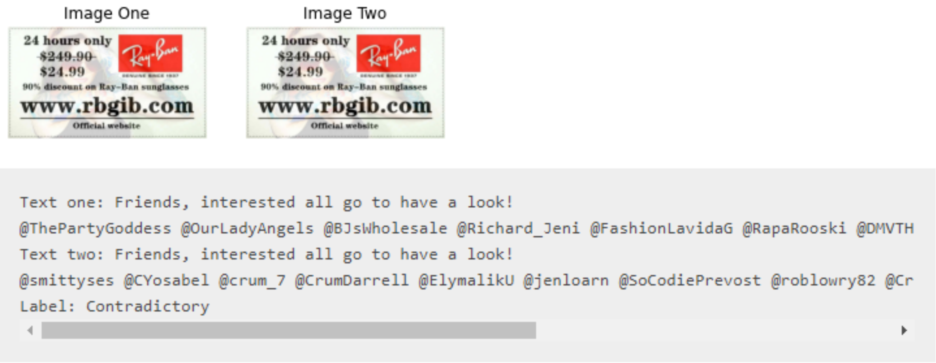

Computer Vision
Visual entailment
NLP
Recognising named entities
Understanding semantic similarity between texts

Figure 1.3: The project involves the fields of Computer Vision and Natural Language Processing. It also requires the ability to be able to recognise named entities like people, places. It also involves semantic similarity which in itself is an open problem

Apart from these challenges, some hardware challenges were also faced. On the Mac M1, the implementations requiring PyTorch could not be run. Even in general, the complexity of the models being used in the existing solutions either requires training which takes very long amounts of time. For example, the Semantic Similarity with BERT implementation took about 10 hours to train for 2 epochs on 250K samples on the Mac M1 which shows the level of time investment required for this project. In order to tackle this problem, ready-to-use models were made use of. The data sets involved in projects like these are also humongous. For example, the VisualNews Data set origin zip file has a size of 91 GB which makes it infeasible to run it on the hardware I had access to. This field is also very subjective in nature as different people may have different notions about the relations and there are issues of indeterminacy involved due to entity and event co-reference. One has to declare their assumptions about entity and event co-reference which can vary across different works making the task even more challenging. Another example is the assumptions made or the quality of the labelled data set. In one of the works, an already existing data set containing the image-caption pairs is used to create out of context 'falsified' image-caption pairs by using some complex procedures for which they assume that the pairs in the original data set are pristine as they are from reputed sources. While it is a valid assumption, there is always some doubt regarding such claims (please note that I am not calling their assumption or technique invalid or that I present something which is much better. These are presentations of some doubts I had while considering their methodology). The issue is that even if the image-caption pairs are not falsified, it is not necessary that they entail each other or if the pairs are falsified then they necessarily contradict each other. Figure 1.4 represents an example from the Google's Recognizing multimodal entailment data set where it can be seen how labelled data sets are prone to errors (which is also acknowledged, there are no claims that the data set is perfect, these are examples of the different types of issues which are encountered)

The choices of assigning the relations are also very subjective and humans themselves get



```
Text one: Friends, interested all go to have a look!
@ThePartyGoddess @OurLadyAngels @BJsWholesale @Richard_Jeni @FashionLavidaG @RapaRooski @DMVTH
Text two: Friends, interested all go to have a look!
@smittyses @CYosabel @crum_7 @CrumDarrell @ElymalikU @jenloarn @SoCodiePrevost @roblowry82 @Cr
Label: Contradictory
```

Figure 1.4: The images and the captions are both exactly the same yet it is labelled as a contradiction. It was confirmed from the blog author that it could be a mislabelling. Screenshot taken from Multi-modal Entailment Keras Blog(accessed 19.08.2022)

confused. In the textual entailment task there is a text premise and a text hypothesis, in Visual Entailment there is an Image Premise and a Text hypothesis. However, when it comes to Visual Entailment in news articles, there is not only the image and caption, but also the article text which could potentially affect the decision of which relation has to be assigned to the image caption pair; similar to the variation of the Visual Entailment task as presented in Huang et al. (2020) where an image and text are the premise along with a text hypothesis. However, the current project focuses on the Visual entailment task where the premise is only an image and the hypothesis is a text. It was also challenging to put the vast amount of literature involved with this project in a way which is coherent enough for someone else to understand. The author hopes that it is not too convoluted to be understood.

1.5 Dissertation Overview

The first chapter presents the motivation of pursuing this work along with the objectives, contributions and the challenges encountered. The second chapter contains the literature review which presents the related work in this field with descriptions of the different data sets which could be relevant to future researchers pursuing work in the field of visual entailment. It is hoped that the second chapter is comprehensive enough without requiring the need to refer to outside sources and that by the end of the second chapter the reader will have accumulated a considerable amount of knowledge with regards to the different data sets and methods applied in the field of Visual Entailment. The third chapter describes the methodologies used for the data set collection and the attempted labelling heuristic and how the original research question is answered. The fourth chapter presents the evaluations of the different Image Captioning systems and the results of the proposed pipeline. This is followed by the conclusion section which brings together the major and minor results obtained, the limitations of the project along with the scope and possibilities of future work along with the lessons learnt.

2 Background

2.1 Related Work

Research in Visual Entailment has seen much attention in the recent years with a plethora of data sets, methodologies being tried out in order to solve this problem. However, to the best of our knowledge, visual entailment in news articles has not been attempted yet.

This section briefly describes some of the related work in the field of Visual Entailment and provides descriptions about the various available data sets meant to serve as a ready reference for knowledge about the different data sets and also some of the main terms generally encountered when working in this field.

2.1.1 Visual Entailment

Our discussion of previous research starts off with the work of Xie et al. (2019) which is regarded as a standard visual entailment benchmark by Thomas et al. (2022). Xie et al. (2019) introduces the Visual Entailment (VE) task as a new inference task which aims to predict whether an image semantically entails the given text. The VE task is heavily drawn upon the Textual Entailment task where the main aim is to classify the semantic relation between a text premise and a Hypothesis into either Entailment, Contradiction and Neutral or Unknown relations (Murakami et al. (2010)). The main distinction between the VE and the TE task is that in TE, both the premise and the hypothesis are texts whereas in VE the premise is an image and the hypothesis is a text.The TE task was actually a challenge which was an attempt to form an encompassing application-independent framework in order to develop and evaluate the generic semantic approaches. The main view underlying the RTE task is that different NLP applications need to address the problem of variability of language and to recognise that a particular target meaning can be inferred from different text variants. RTE abstracts this inference need, suggesting that many NLP applications can benefit from the generic models born out of solving the problem of textual entailment (Dagan et al. (2005)).

There are several techniques which have been employed in order to solve the textual entailment

Type	Mononuclear		Multinuclear
Definition	Segments play different roles and one segment is more important than the other, playing the role of nucleus and satelite respectively.		Both segments are nuclei and have equal status
	Presentational Relations They affect and act upon the hearer	Subject Matter Relations They inform the hearer	
Number of subcategories	10	15	7
Examples of subcategories	Concession, Evidence, Background	Elaboration, Interpretation	Sequence, List, Conjunction

Table 2.1: The taxonomy of RST textual discourse relations

task. For example, methods based on the measurement of the degree of lexical overlap between the bag-of-words. Other methods include aligning syntactic or semantic dependency graphs, reference rule generation and statistical classifiers leveraging a wide range of features. Murakami et al. (2010) note that while the approaches show great promise in solving the task, more robust models for recognizing the logical relations are desirable. Xie et al. (2019) define the 3 relations for the VE task as follows:

1. Classify as an entailment relation if the premise presents enough evidence to conclude that the hypothesis is true

2. Classify as a contradiction if the premise presents enough evidence to conclude that the hypothesis is false

3. Classify as neutral or unknown if the premise present insufficient evidence to draw any conclusions about the hypothesis

These relations are taken from the RTE task and there have been discussions about the possibility of use of other relations in textual processing as well as for multimodal analysis which is now discussed.

2.1.2 Discourse relations

In order to retrieve information from texts for processing, it is crucial to understand the interpretation of how the different textual elements relate to one another. Discourse relations facilitate such interpretation (Hou et al. (2020)). Different Discourse structure theories (DSTs) have been presented over the previous years such as the Rhetorical Structure Theory (RST), Cross-document Structure Theory (CST), RST Treebank, Discourse GraphBank, Lexicalised Tree-Adjoining Grammar based discourse. Apart from the Discourse Graph, the other theories present the discourse relations between textual units as a hierarchical structure whereas the Discourse GraphBank represents them using a graphical structure. Hou et al. (2020) found that discourse relations the connection between 2 sentences, but also the amount of similarity in terms of their contents. Table 2.1 presents the different textual discourse relation in Rhetorical Structure theory.

However, as noted by Murakami et al. (2010), information on the web requires a broader set of semantic relations than the ones typically defined for the Recognising Textual Entailment (RTE) task, Cross-Document Structure Theory (CST). CST is another task which recognises the semantic relations between different texts and is an expanded rhetorical structure analysis based on the Rhetorical Structure Theory (RST). Murakami et al. (2010) present a prototype semantic relation identification system building upon the RTE and CST tasks to create relations which are simple enough to facilitate the automatic recognition of semantic relations between the statements found in internet text. However, the relations they have proposed are more suited to facts and opinions on the web rather than News articles and are targeted at scenarios where the entities involved are textual. Pastra (2008) presents CrOSs-Media inteRactiOn rElations (COSMOROE), a broader corpus based relations framework for describing the semantic interrelations between images, languages and body movements. They utilise the framework to annotate a corpus of TV travel programmes and appears to be better suited for the Visual Entailment task in News articles. To the best of our knowledge, the current work is the first to propose use of a different relations framework to be utilised for the Visual Entailment task rather than making use of the TE task entailment relations.

The characteristics of a descriptive framework for cross-media relations as outlined by Pastra (2008) are briefly discussed below:

1. Descriptive power: the framework should not be medium specific and should be general enough to describe the relation between any media pair. It should also be applicable across different domains and genres in which the interaction could manifest.

2. Computational applicability: the framework should be able to model than just describe the media-interaction; it should be able to guide the computational modelling of the multimedia dialectics similar to the way humans work with multimodal data

The paper highlights that the above mentioned criteria are rooted within the need for intelligent multimedia systems and that describing the cross media relations should have both wide scope deepening one's understanding of the multimedia dialectics and facilitating the computational modelling and wide coverage scaling beyond specific media pairs.

It also discusses the Rhetorical Structure Theory and its shortcomings in using it to describe cross media relations. It underscores that the RST was formulated to describe the rhetorical relations in the textual discourse rather than the multimedia discourse.

We now briefly discuss the core COSMOROE relations with a broad overview of the same provided by Figure 2.1

1. Equivalence: information expressed by the different media (the image and the caption) is semantically equivalent and refers to the same entity which could be an object, a state, an event or a property.

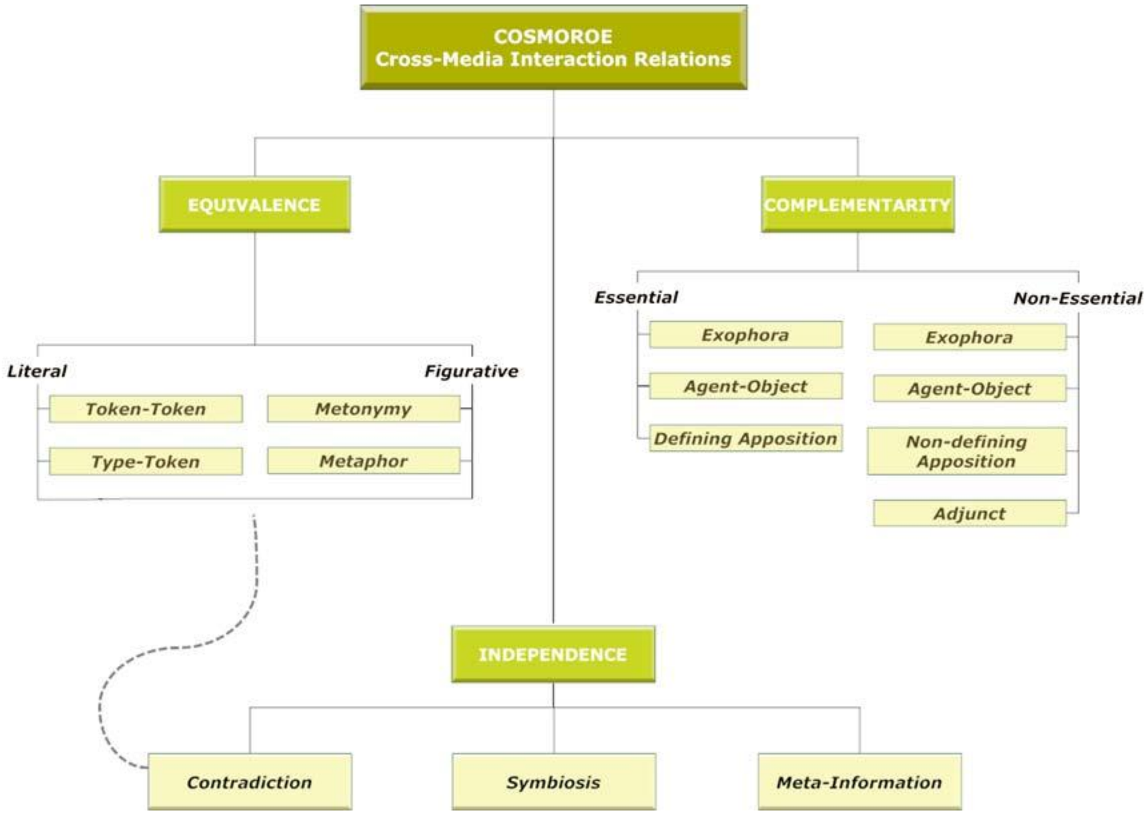


Figure 2.1: Summary of the cross-media relations proposed in the COSMOROE framework (Image taken from Pastra (2008).)

2. Complementarity: information expressed in one of the mediums is a complement of the information presented in the other medium. It must be noted that the information may be essential like the text indicating to an image or part of the image or non-essential like an image showing the means used by the speaker to reach a destination which is mentioned in the text caption.

3. Independence: the mediums carry independent messages which could be coherent or incoherent with the document topic (here for example, a news article which can be considered for the cases of advertisements with images and captions which may or may not be related to the article).

As can be seen in Figure 2.1, there are finer relations defined beyond the broad three, details of which while omitted here, can be found in the paper Pastra (2008). It is also worth noting that it would be prudent to first create systems which are able to classify the image-text pairs accurately in the 3 broad categories before moving on to the finer classifications. It should also be noted that the author of this dissertation believes that the finer relations can help in dealing with certain gray areas associated with the classification of the image-text semantic relations. Pastra (2008) also serves as a reference for annotators to follow when creating data sets should they choose the COSMOROE framework for their VE task.

The challenge to develop algorithms to identify the COSMOROE relations is far from trivial as the author of the paper also notes. Previous works have focused on using the relations from the Textual Entailment tasks for the Visual Entailment tasks. One may correlate the different relations as follows:

1. the entailment relation is closely related to the equivalence relation of the COSMOROE framework
2. the neutral relation is closely related to the complementarity relation of the COSMOROE framework
3. the contradiction relation and the independence relation

2.1.3 Image Captioning

We now revert to our discussion of the related works in the Visual Entailment field by continuing the discussion of Xie et al. (2019). They conduct various experiments involving Image Captioning, Relational Network, Attention top-down and bottom-up and their own solution Explainable Visual Entailment (EVE) with its variations EVE-Image and EVE-ROI. In the Image Captioning experiment, they first generate a caption for the image premise which essentially converts it into a text premise and then use a Textual Entailment Classifier. By making use of Image Captioning, the Visual Entailment task gets reduced to a Textual Entailment task. For the Image Caption Generator, they use a PyTorch Tutorial Implementation in which the image encoder is a pre-trained ResNet152 whereas the decoder for the caption is a Long Short Term Memory (LSTM) Network. The Text Entailment classifier is composed of 2 text processing components which extract the text features from the premise and the hypothesis (The hypothesis was the original image caption and the premise is the generated Image Caption). These extracted text features are fused and sent through 2 Fully Connected layers which have input and output dimensions of $[600, 300]$ and $[3003, 3]$ for the final prediction.

This approach to the Visual Entailment problem inspired the search for different Image-Captioning approaches which form the major part of the discussion following next.

The main components to the Image Captioning architecture from Sharma et al. (2018) are:

1. A deep Convolutional Neural Network which takes in a pre-processed image and gives a vector of image embeddings
2. An encoder which transforms the image embeddings into a tensor
3. A decoder which generates output at each time step which is conditioned on the tensor which was the output of the encoder module and the decoder inputs

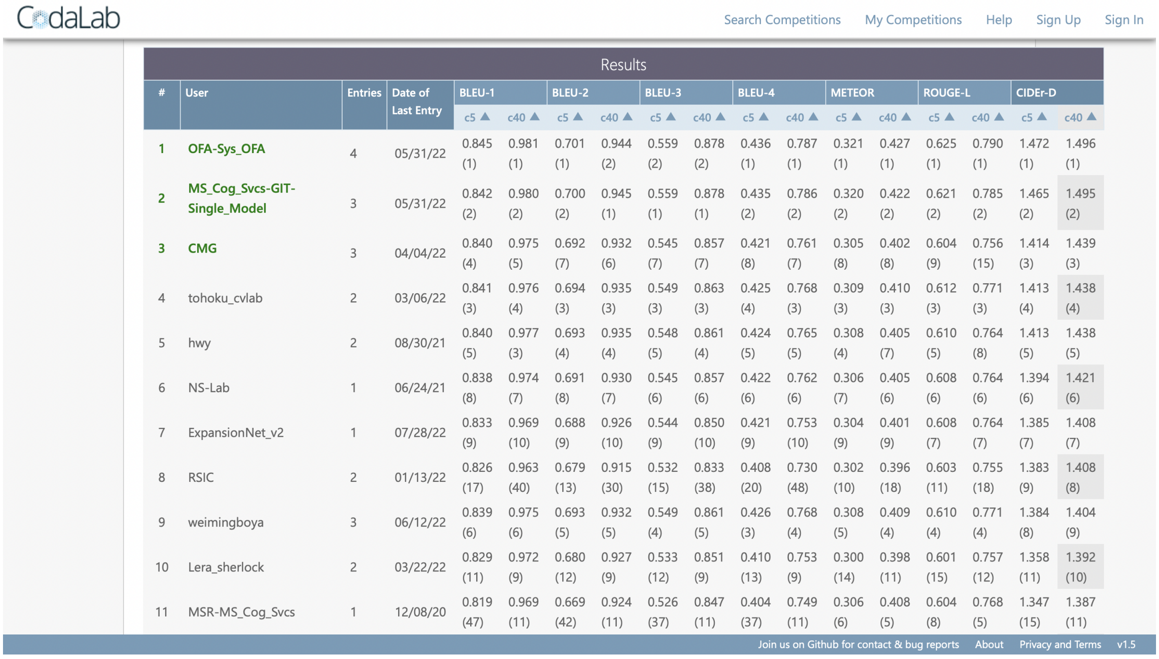
Generally, the main architecture remains the same with variations mainly occurring in the implementation of the encoder and decoder functions of the encoder and decoder modules. The key goals in Image Captioning are: Automatic Caption Generation, Automatic Caption Evaluation, which in itself are both challenging problems. For the automatic Caption Evaluation, there exist commonly used metrics CIDEr, METEOR, BLEU, ROUGE which are n-gram based metrics and their while they are not able to capture the semantics of the text and cannot correlate well to human judgements (Cui et al. (2018)), they are still very popular metrics. SPICE is another metric which was proposed to be able to correlate well to human judgements however, it then is not able to capture the syntactic structure of the sentence. Research still continues to create better metrics making this problem even harder. Many image captioning systems make use of the metrics mentioned above like Liu et al. (2020).

Image Captioning is basically performing the inverse task of Dall-E, which generates images from text by using a 2-stage model consisting of a prior which given the text input, generates the CLIP image embeddings and then employing a decoder that can generate an image which is conditioned on the image embedding (Ramesh, Dhariwal, Nichol, Chu, and Chen (Ramesh et al.)). In order to perform the inverse of this task, Tewel et al. (2021) employ Contrastive Language-Image Pretraining to perform zero-shot Image Captioning. Here, the input is an image which is fed into a model employing CLIP along with the GPT-2 Language model in order to generate the caption for the input image. Examples of the captions generated by this approach have been presented in Section 4.2 along with some discussion about the obtained captions. There exist many previous works and implementations addressing Image Captioning and in the following discussion, the current State-Of-The-Art on the COCO captions data set (Figure 2.3, Figure 2.2), One-For-All (OFA), a unified Sequence-to-Sequence framework is discussed. OFA supports Task Comprehensiveness while being both task and modality agnostic.

ONE-For-All framework

OFA satisfies the following 3 properties (as noted by Wang et al. (2022)) that omnipotent models should have in order to support better generalization to open-ended problems while also maintaining the ease of use as well as performance.

1. Task-Agnosticness (TA): support different types of tasks using a unified task representation and model should also be agnostic to pretraining or fine tuning steps
2. Modality-Agnosticness (MA): make use of unified input and output representation which can be shared among all the tasks in order to be able to handle the different modalities
3. Task-Comprehensiveness (TC): there should be enough task variety to robustly accumulate the generalisation ability



Results

#	User	Entries	Date of Last Entry	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
				c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲
1	OFA-Sys_OFA	4	05/31/22	0.845 (1)	0.981 (1)	0.701 (1)	0.944 (2)	0.559 (2)	0.878 (2)	0.436 (1)	0.787 (1)	0.321 (1)	0.427 (1)	0.625 (1)	0.790 (1)	1.472 (1)	1.496 (1)
2	MS_Cog_Svcs-GIT-Single_Model	3	05/31/22	0.842 (2)	0.980 (2)	0.700 (2)	0.945 (1)	0.559 (1)	0.878 (1)	0.435 (2)	0.786 (2)	0.320 (2)	0.422 (2)	0.621 (2)	0.785 (2)	1.465 (2)	1.495 (2)
3	CMG	3	04/04/22	0.840 (4)	0.975 (5)	0.692 (7)	0.932 (6)	0.545 (7)	0.857 (7)	0.421 (8)	0.761 (7)	0.305 (8)	0.402 (8)	0.604 (9)	0.756 (15)	1.414 (3)	1.439 (3)
4	tohoku_cvlab	2	03/06/22	0.841 (3)	0.976 (4)	0.694 (3)	0.935 (3)	0.549 (3)	0.863 (3)	0.425 (4)	0.768 (3)	0.309 (3)	0.410 (3)	0.612 (3)	0.771 (3)	1.413 (4)	1.438 (4)
5	hwy	2	08/30/21	0.840 (5)	0.977 (3)	0.693 (4)	0.935 (4)	0.548 (5)	0.861 (4)	0.424 (5)	0.765 (5)	0.308 (4)	0.405 (7)	0.610 (5)	0.764 (8)	1.413 (5)	1.438 (5)
6	NS-Lab	1	06/24/21	0.838 (8)	0.974 (7)	0.691 (8)	0.930 (7)	0.545 (6)	0.857 (6)	0.422 (6)	0.762 (6)	0.306 (7)	0.405 (6)	0.608 (6)	0.764 (6)	1.394 (6)	1.421 (6)
7	ExpansionNet_v2	1	07/28/22	0.833 (9)	0.969 (10)	0.688 (9)	0.926 (10)	0.544 (9)	0.850 (10)	0.421 (9)	0.753 (10)	0.304 (9)	0.401 (9)	0.608 (7)	0.764 (7)	1.385 (7)	1.408 (7)
8	RSIC	2	01/13/22	0.826 (17)	0.963 (40)	0.679 (13)	0.915 (30)	0.532 (15)	0.833 (38)	0.408 (20)	0.730 (48)	0.302 (10)	0.396 (18)	0.603 (11)	0.755 (18)	1.383 (9)	1.408 (8)
9	weimingboya	3	06/12/22	0.839 (6)	0.975 (6)	0.693 (5)	0.932 (5)	0.549 (4)	0.861 (5)	0.426 (3)	0.768 (4)	0.308 (5)	0.409 (4)	0.610 (4)	0.771 (4)	1.384 (8)	1.404 (9)
10	Lera_sherlock	2	03/22/22	0.829 (11)	0.972 (9)	0.680 (12)	0.927 (9)	0.533 (12)	0.851 (9)	0.410 (13)	0.753 (9)	0.300 (14)	0.398 (11)	0.601 (15)	0.757 (12)	1.358 (11)	1.392 (10)
11	MSR-MS_Cog_Svcs	1	12/08/20	0.819 (47)	0.969 (11)	0.669 (42)	0.924 (11)	0.526 (37)	0.847 (11)	0.404 (37)	0.749 (11)	0.306 (6)	0.408 (5)	0.604 (8)	0.768 (5)	1.347 (15)	1.387 (11)

Figure 2.2: CodaLab Microsoft COCO Image Captioning Challenge Results. Screenshot taken from the CodaLab website(accessed 19.08.2022)

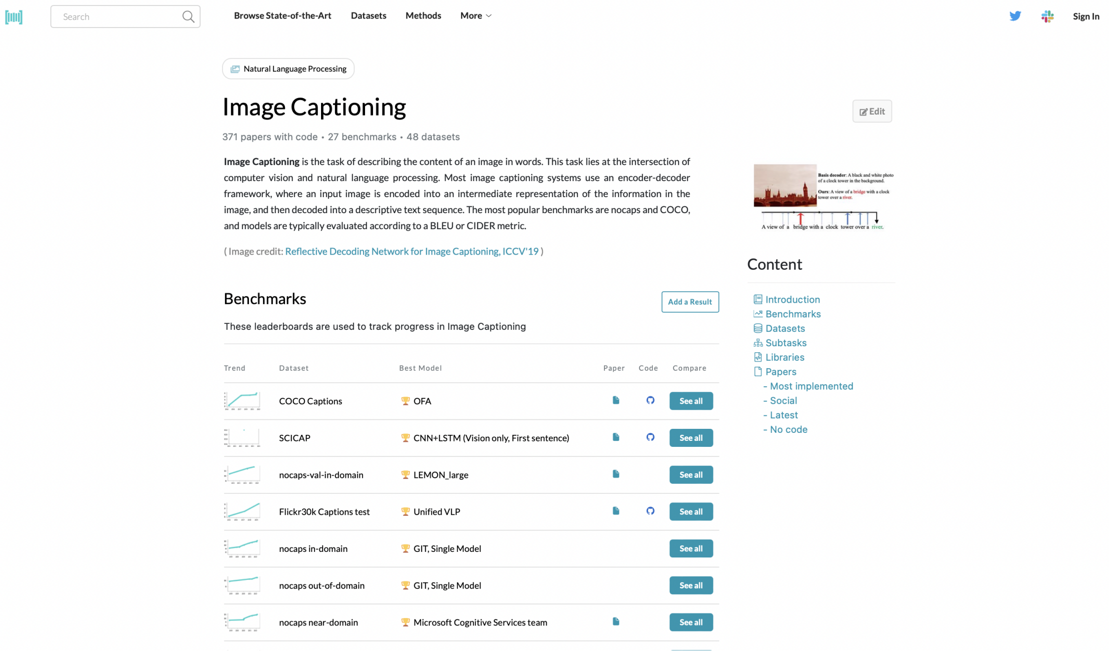

Search
Browse State-of-the-Art
Datasets
Methods
More
Sign In
Natural Language Processing
Edit
Basis decoder: A black and white photo of a clock tower in the background.
Ours: A view of a bridge with a clock tower over a river.
A view of a bridge with a clock tower over a river.
Add a Result

Image Captioning

371 papers with code • 27 benchmarks • 48 datasets

Image Captioning is the task of describing the content of an image in words. This task lies at the intersection of computer vision and natural language processing. Most image captioning systems use an encoder-decoder framework, where an input image is encoded into an intermediate representation of the information in the image, and then decoded into a descriptive text sequence. The most popular benchmarks are nocaps and COCO, and models are typically evaluated according to a BLEU or CIDER metric.

(Image credit: Reflective Decoding Network for Image Captioning, ICCV'19)

Benchmarks

These leaderboards are used to track progress in Image Captioning

Trend	Dataset	Best Model	Paper	Code	Compare
	COCO Captions	🏆 OFA	▇	○	See all
	SCICAP	🏆 CNN+LSTM (Vision only, First sentence)	▇	○	See all
	nocaps-val-in-domain	🏆 LEMON_large	▇		See all
	Flickr30k Captions test	🏆 Unified VLP	▇	○	See all
	nocaps in-domain	🏆 GIT, Single Model			See all
	nocaps out-of-domain	🏆 GIT, Single Model			See all
	nocaps near-domain	🏆 Microsoft Cognitive Services team	▇		See all

Content

- Introduction
- Benchmarks
- Datasets
- Subtasks
- Libraries
- Papers
  - Most implemented
  - Social
  - Latest
  - No code

Figure 2.3: Image Captioning SOTA for different data sets along with relevant papers. Screenshot taken from the Paperswithcode website(accessed 19.08.2022)

They discuss the architecture in which they use ResNet modules for convolving $x_v \in R^{H \times W \times C}$ to P patch features of hidden size. For the text modality, they apply Byte-Pair Encoding (BPE) in order to transform a given text sequence into a sub-word sequence and then embedding them to features.

In order to process the different modalities while retaining task-agnostic behaviour, there is a need to represent the different modalities in a unified space. They utilize text-to-image synthesis strategies for their target-side image representations. Dong et al. (2018) also present a strategy where they perform image-caption retrieval by relying on a visual space rather than a joint subspace. Their deep neural network architecture Word2VisualVec learns to predict from the textual input, a visual feature representation. However, this may not be able to satisfy the properties for an omnipotent model.

OFA uses Transformers for the backbone architecture, adopting encoder-decoder framework as unified architecture for the different tasks like pre-training, fine tuning and the zero-shot tasks. The encoder and decoder are stacks of Transformer layers. Vaswani et al. (2017) note that most of the competitive neural sequence transduction models are composes of an encoder-decoder structure. An encoder consists of a self-attention layer and a Feed Forward Network (FFN) which maps an input sequence containing symbol representations to a sequence containing continuous representations, call it z. The decoder consists of a self attention layer, an FFN along with a cross attention layer which generates an output sequence of symbols using z as the input, one element at a time. Vaswani et al. (2017) describe an attention function as mapping a vector query and a set of vector key - vector value pairs to a vector output. The vector output is a weighted sum of the vector values, where the weights are computed using a compatibility function of the query and the corresponding key. Figure 2.4 presents the Transformer model architecture.

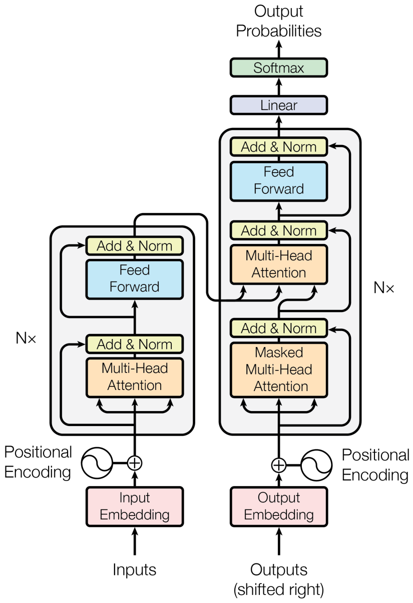


Figure 2.4: Transformer Model Architecture. Figure taken from Vaswani et al. (2017)

Wang et al. (2022) design 5 tasks for the cross modal representation learning: Grounded Captioning (GC), Visual Question Answering (VQA), Visual Grounding (VG), Image Captioning (IC), Image-Text Matching (ITM). It should be noted that for ITM, they use the original image-text pairs as the positive samples and in order to create the negative samples, they pair the image with a randomly substituted caption. In order to create the negative pairs in a systematic way, Luo et al. (2021) present a strategy where they use the image and corresponding caption along with other images and captions to create falsified pairs. More details on this strategy can be found in Section 2.2.13. Our methodology makes use of OFA's Image Captioning Cross Model Representation learning task.

As previously mentioned, many implementations have been rolled out for the Image Captioning task however, news article images are challenging because of the following reasons (Tran et al. (2020)):

1. highly dependent on real-world knowledge, especially when concerning named entities such as particular places, people or events
2. captions are linguistically richer including uncommon words. Previous Image Captioning models generate captions which tend to use much simpler language perhaps at the level of a high school student rather than a professional journalist with extensive vocabulary.

Entity aware Image Captioning

Tran et al. (2020) note that the linguistic simplicity of the generated captions could be amounted to the use of Long Short Term Memory networks. They make use of the Transformer architecture. However, this may raise a question about the advantage of this captioning method over Wang et al. (2022) which also uses the transformer architecture. The Transform and Tell captioner has 2 additional specialised modules in their model. One of the modules aims to detect faces whereas the other focuses on detecting the objects. This helps in improving the accuracy of generated entity names especially when concerning people. It is also worth noting that a framework like OFA and an entity aware Image Captioner like Transform and tell cannot be directly compared as OFA focuses on task and modality agnosticness whereas the Image captioner works with a particular task and set of modalities. Tran et al. (2020) also contribute the NYTimes800k data set, more details of which are provided in Section 2.2.12.

Liu et al. (2020) presents another entity aware Image Captioner called Visual News Captioner which is also built upon the Transformer architecture equipped with novel multi-modal feature fusion techniques as well as attention mechanisms. The mechanisms aim to generate the named entities with better accuracy and slightly better prediction results while also utilising lesser parameters. They also contribute the VisualNews data set, further detailed in Section 2.2.4.

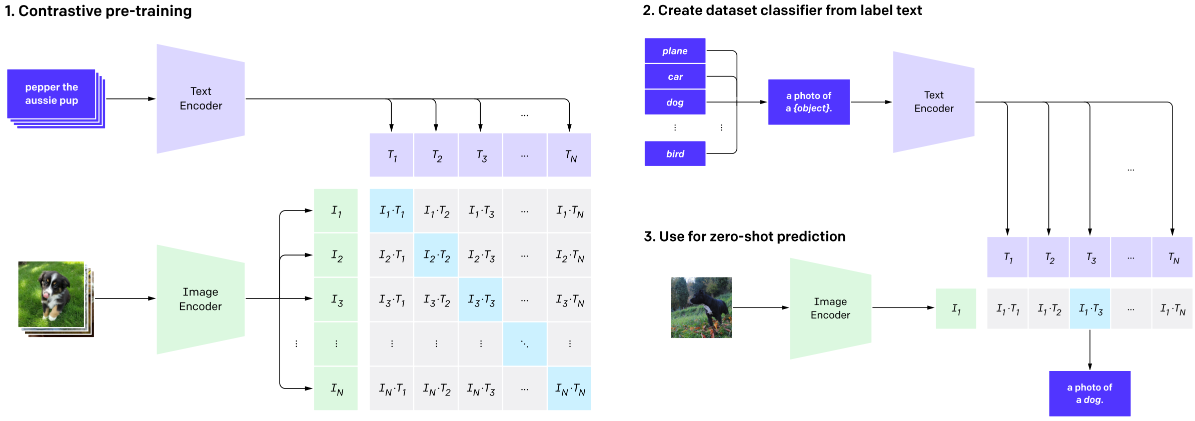

1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
T_1 T_2 T_3 ... T_N
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
a photo of a dog.

Figure 2.5: Summary of the CLIP approach. Image taken from OPENAI CLIP blog

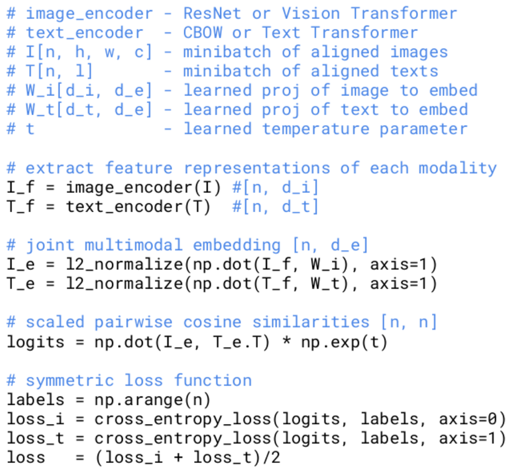
2.1.4 Image-Caption Ranking

Apart from Image Captioning, another multi-modal task is Image Caption Retrieval. In this task, an image is specified, and the most suitable captions for the image are returned. Basically, if an image is provided along with a certain number of captions, the system assigns ranks to the captions in terms of similarity with the image. There have been many works and implementations for this particular task as well. We now discuss one of the very popular implementations, Contrastive Image Language Pre-training or CLIP for short.

CLIP

Radford et al. (2021) demonstrate how the Image Caption retrieval as a pre-training task can be a scalable and efficient way to learn the State-Of-The-Art image representations from scratch. Generally image models train image feature extractor together with a linear classifier in order to predict labels for the image, the Contrastive Language Image Pre-training trains image encoder together with text encoder in order to predict the correct pairings in a batch of image-text training pairs. The main highlight of their work is that they try to use Natural Language Supervision for image representation learning. They define CLIP as a simpler Con-VIRT which is trained from scratch as an efficient method to learn from Natural Language Supervision. Figure 2.5 presents the summary of the CLIP approach and Figure 2.6 presents the CLIP Numpy-like pseudocode. Two separate image encoder architectures were considered for CLIP. CLIP uses the ViT-L/14 Vision Transformer as the image encoder which they pre-train at a pixel resolution of 336 pixels hence referring to this model as ViT-L/14@336px. For the text encoder they use a Transformer which operates on a lower-cased BPE encoding. The transformer they use also has some architecture modifications which are described in Radford et al. (2019).

Dong et al. (2018) try to work on Image-Caption retrieval by transferring the caption to the visual space rather than using a joint subspace. Here, the sentences which describe the same



```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 2.6: CLIP implementation pseudocode. Screenshot taken from Radford et al. (2021)

image tend to lie closer and sentences which describe different images tend to lie far from each other. They claim that this textual embedding captures not just visual but also semantic similarities.

2.1.5 Extensions/Variations of the Visual Entailment task

Apart from the benchmark Xie et al. (2019) and Wang et al. (2022), works like Huang et al. (2020), Thomas et al. (2022) present extensions/slight variations of the original Visual Entailment task. Our work also proposes a slight variation of the VE task where the difference is the entailment relations being utilised instead of the original 3 relations.

The Recognising Cross-media Entailment (RCE) task is an extension of the RTE task to recognising entailment where the premises can be different media types. They propose Heterogeneous Interactive Learning (HIL) to recognize the entailment relationships from image-text premises to text hypotheses. It must be noted however, that Huang et al. (2020) focus on Recognizing Cross Media Entailment (RCE) task which is slightly different from the Visual Entailment task. The VE task has an image premise and text hypothesis, whereas the RCE task has image-text premises and a text hypothesis. This can be applied in the news articles domain by considering the article text or the headline as the premise along with the image with the image caption as the text hypothesis. This can help decide the image-caption relation in the context of the article. They make use of the Stanford Natural Language Inference (SNLI) data set (Section 2.2.8) and the Flick30K data set (Section 2.2.3) in order to establish SNLI-RCE. The HIL approach has 2 main parts as follows:

1. Cross-media hybrid embedding: for performing the cross embedding of different media being used (premise and hypothesis) to create their fine-grained representations via cross-media alignment of inference cues.

2. Heterogeneous Joint Inference: for constructing a heterogeneous interaction tensor space and and also extract the semantic features via modelling cue interaction between hypothesis and premise.

The goal of the task presented in Thomas et al. (2022) is to predict the logical relationships between knowledge elements in text to image. In the VE task, the relationship is predicted of the entire text hypothesis to the image whereas in the 'Fine-grained Visual Entailment task' the relationship between each knowledge element in the text hypothesis has to be established with the image. They define knowledge elements as claims which collectively make up the entire hypothesis. They represent the hypothesis as an Abstract Meaning Representation (AMR) graph as AMR graphs can capture the semantic meaning of the text regardless of syntax. They make note that for training their methods, they are lacking fine-grained labels for which they propose a multi-instance approach which uses sample-level supervision in order to learn fine-grained labels. However, in order to evaluate their approach, they use a data set with manually annotated knowledge elements. This work is extremely recent and their GitHub was populated with their code and data around mid-July 2022. This extension of the VE task can be very useful when it comes to news articles.

There have also been some works surveying the different approaches in multimodal analysis. Guo et al. (2019) describe and present a survey on deep multimodal representation learning. Multimodal representation learning aims to narrow the heterogeneity gap amongst the different modalities. They categorise the various methods into 3 frameworks: coordinated representation, joint representation, encoder-decoder. The summary of the different frameworks as presented in the paper is contained in Table 2.2.

Considering the amount of research in this field and the need for data sets in order to create solutions, many data sets have been created which have been summarised in Section 2.2. The aim of Section 2.2 is to present an overview of existing data sets in this field for future researchers to make a quick decision on which data set is suitable for their study as the search for data sets and making the right decision can sometimes end up taking a lot of time.

2.2 Data sets

2.2.1 MS-COCO

The Microsoft Common Objects in COntext data set was originally created around 2015 to advance the SOTA in object recognition (Lin et al. (2014)). It contains 91 classes or object types with a total of 328K images in which the individual object instances have been segmented. This data set addresses core problems in scene understanding like detecting objects in non-iconic views, localising objects precisely in 2 dimensions and contextual reasoning between objects. Non-iconic views mean scenarios where the object may not be in the focus

Frameworks and models	**Key issues**	**Advantages**	**Disadvantages**
Joint Representation	obtaining modality-invariance fusing complementary semantics	fuses several modalities	cannot infer individual modalities
Coordinated Representation (CR) Framework	maximising cross modal correlation	infers each modality individually	hard to coordinate more than 2 modalities
Cross Modal Similarity (CR Framework)	preserving inter and intra modality similarity	measures cross-modal similarity	
DCCA (CR Framework)	maximising cross modal correlation	*unsupervised learning*	
Encoder-decoder	capturing shared semantics	generates novel samples	encodes only one of the modalities
PGM	maximising the joint distribution	generates the missing modality *unsupervised learning*	high computational cost
Multimodal autoencoders	minimzing the reconstruction loss	preserves modality specific characteristics *unsupervised learning*	designed to work in general purpose settings
Generative Adversarial Networks	narrowing the distribution difference	generates high quality novel samples *unsuupervised learning*	suffers from training instability
Attention mechanism	evaluating imporance of features selecting complementary features	selects the salient localized features filters out noise	no obvious drawbacks reported

Table 2.2: Summary of the different frameworks presented. The table was created using the Tables generator website(accessed 19.08.2022). Cross Modal Similarity and DCCA belong to the Coordinated Representation (CR) framework. Table taken from Guo et al. (2019)

of the picture and maybe be somewhere in the background or even occluded. Iconic views are much cleaner versions where the object of interest is more prominent in the picture. Contextual reasoning also forms a part of basic human visual understanding. For example, an object in the air is more likely to be a bird than a pig. Context comes in handy when there are doubts about the identity of an object perhaps because it is too small or blurry. 2D localization is about defining an object's spatial location.

The MS-COCO data set is used for Image Captioning (Sharma et al. (2018), Wang et al. (2022)), Image-Text Matching (Diao et al. (2021)) among many other studies. It is the data set which is most frequently used to compare performance of new models and new data sets for various computer vision tasks.

2.2.2 Flickr8K

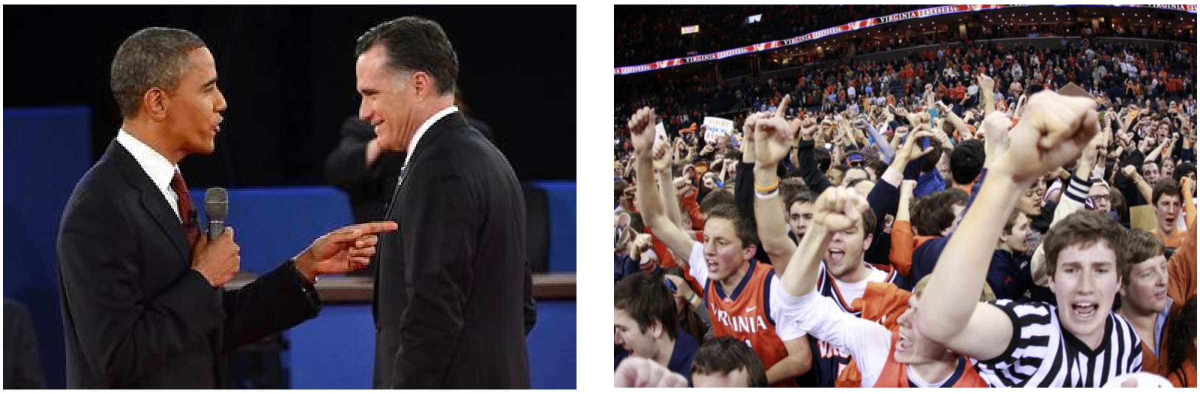
The Flickr8K data set has 8,108 images where each image is associated with five different captions assigned using Crowd-sourcing using the popular Amazon Mechanical Turk (Hodosh et al. (2010)). The captions describe the events and the entities in the images which are "action" images with scenes featuring people and animals. Just like the MS-COCO data set, the Flickr8K data set is used for Image-Captioning (Radford et al. (2021))

2.2.3 Flickr30K

Flickr30K data set as the name suggests is similar to the Flickr8K data set and also contains annotations for images which contain non-iconic views reflecting the composition of everyday real life scenes just like the MS-COCO data set (Sharma et al. (2018)). The Flickr30K data set contains 31,783 pictures of everyday real life scenes and 158,915 captions which were assigned using crowd sourcing (just like the Flickr8K data set as this is basically an extension of that work). The data set is widely used for image captioning.

2.2.4 VisualNews data set

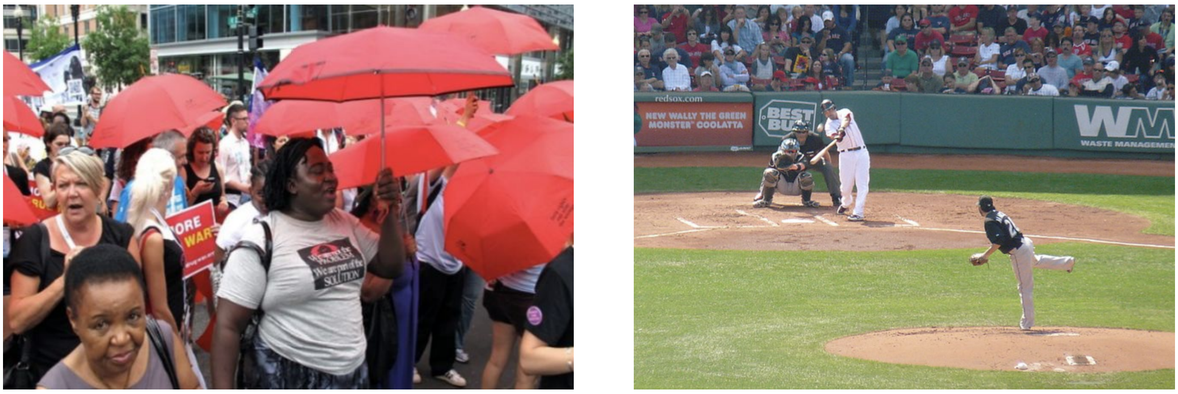
A benchmark and data set consisting of more than 1 million news images along with their captions and their article text. The images were scraped from publicly available data and from varying sources and topics in English language. Some of the news sources include TheGuardian, USA TODAY, BBC, and the Washington Post). The data set is created with the goal of News Image Captioning. Image captioning models generally make use of the MS-COCO data set in which the images present certain everyday objects. The captions in the MS-COCO data set describe the images in a descriptive way rather than in an interpretative way. For example, if the image shows people standing with umbrellas, the caption would be something similar like: "a bunch of people holding blue umbrellas" whereas a news article may have something like "The Men Rights Activists stand to show solidarity with the new



(a) President Obama and Mitt Romney debate in Hempstead NY on Tuesday.

(b) Virginia Cavaliers fans celebrate on the court after the Cavaliers game against the Duke Blue Devils at John Paul Jones Arena.

Figure 2.7: Examples from the VisualNews data set. Note how the captions involve particular named entities, places and events



(a) A bunch of people who are holding red umbrellas.

(b) A baseball player hitting the ball during the game.

Figure 2.8: Examples from the COCO data set. Note how the captions are descriptive rather than interpretative. The captions fail to address the higher level situation which may be occurring in the picture. For example, in Figure 2.8a, the caption is not wrong but it does not give information on "why the people are holding the red umbrellas?"

movement" and hence the captions in the COCO data set do not fail to describe the image but they cannot capture the language generally found in news articles. This data set has to be requested from the authors (Liu et al. (2020)).

2.2.5 LAION-400M

The LAION-400M data set was built and released in order to address the issue of large image-text data sets existing but not being made publicly available (Schuhmann et al. (2021)). The data set consists of 400 million image text pairs which were filtered using CLIP, along with their CLIP Embeddings and the k Nearest Neighbours indices. More specifically, they provide 400M pairs of the image URL and corresponding metadata. They provide parquet

Table 2.3: Image Size Distribution LAION400M

Height	OR	Width	≥ 1024	26M
Height	AND	Width	≥ 1024	9.6M
Height	OR	Width	≥ 512	112M
Height	AND	Width	≥ 512	67M
Height	OR	Width	≥ 256	268M
Height	AND	Width	≥ 256	211M

files consisting of the following attributes for every pair: the sample ID, URL, Not-Safe-For-Work tag (detected using CLIP), height, width of image, cosine similarity score between text and image embedding as well as the type of Creative Commons License if applicable. For the data collection, they use Common Crawl's WAT files to parse out HTML Image tags with an alt-text attribute. From these, they drop image-text pairs where the caption is less than 5 characters long and images are less than 5KB in size. They more duplicates using a bloom filter and they compute the image and alt-text embeddings using CLIP. They find the similarity of the image and text using the cosine similarity between the image and text embeddings and drop pairs with a cosine similarity less than 0.3 (threshold selected based upon human inspections). They also use the image and text embeddings to filter out the illegal content. Table 2.3 presents the Image Size distribution of the LAION400M data set. All the image-text pairs in this data set are in English language.

2.2.6 LAION-5B

Following the LAION-400M data set, a data set 14 times bigger with 5.85 Billion CLIP-filtered image-text pairs LAION-5B has come into existence. It seems to be pretty recent and the research paper Schuhmann, Beaumont, Gordon, Wightman, Coombes, Katta, Mullis, Schramowski, Kundurthy, Crowson, et al. (Schuhmann et al.) still seems to be under review however, it might soon be published and would be very relevant to people pursuing research in the field of Visual Entailment. The main difference between LAION-400M and LAION-5B is not only the difference in data set size but also the face that LAION-400M contains image-text pairs only in the English language whereas LAION-5B contains 2.3B pairs in English, 2.2B pairs from 100+ non-English languages and 1B pairs cannot be assigned a particular language because they are say, named entities. The attributes of the data set are as follows: the image URL which covers millions of domains, the captions, the width and height of the image, the language (since this data set contains more than 1 language), the cosine similarity of the image and text using their ViT-B/32 embeddings with CLIP for English and mCLIP for others, the probability of being a watermarked image and the probability of being an unsafe image. More details can be found at their website LAION-5B. The image height and width distribution is provided in Table 2.4.

Table 2.4: Image Size Distribution LAION-5B

English	Height	AND	Width	$\geq$ 1024	76M
English	Height	AND	Width	$\geq$ 512	488M
English	Height	AND	Width	$\geq$ 256	1324M
Not English	Height	AND	Width	$\geq$ 1024	57M
Not English	Height	AND	Width	$\geq$ 512	480M
Not English	Height	AND	Width	$\geq$ 256	1299M

Table 2.5: Distribution of 24 categories in the N24News data set

Category	Count	Category	Count
Health	3000	Books	3000
Science	3000	Art and Design	3000
Television	3000	Style	2681
Travel	3000	Media	3000
Movies	3000	Food	3000
Dance	3000	Wellness	681
Real Estate	3000	Fashion	3000
Economy	1761	Technology	3000
Sports	3000	Your Money	1263
THeater	3000	Education	825
Opinion	3000	Automobiles	1825
Music	3000	GLobal Business	3000

2.2.7 N24news

Wang et al. (2021) present the data set N24News created using the New York Times news website. Unlike other data sets associated to news article category classification which mostly contain the article text, N24News contains both the image and text information of the news along with the category of the article. The data set consists of 60K data points and 24 categories and the distribution of the same has been provided in Table 2.5. They collected the data using the New York times API (similar to Tran et al. (2020)) and obtained the published links between 2010 and 2020. They only retain the articles in text form. They choose to include 1 image from each article and they preprocess the data by excluding any articles which do not have any images in them. They have noted that they decided not to merge the categories which are similar like science and technology. The upper limit of articles from each category was 3000. The data points consist of attributes like the category, headline, abstract, article text, image and its corresponding caption. The data set is split randomly using a 8:1:1 ratio for the training, validation and testing sets respectively. The average lengths of the Headline, the caption, abstract and the body progressively increase from headline towards the body; the average lengths are 52.33, 115.27, 129.42 and 4701.08 respectively.

2.2.8 SNLI

The Stanford Natural Language Inference Corpus is a collection of sentence pairs labeled with their entailment relation of entailment, contradiction or neutral. The data set has 570K sentence pairs and is used for learning Natural Language Inference (Bowman et al. (2015)). Natural Language Inference (NLI) is also known as Recognizing Textual Entailment (RTE) which is the task of determining the inference relation between 2 sentences, generally a hypothesis and a premises. They present indeterminacies of event and entity co-reference as challenges which their data set aims to address by grounding the examples in specific scenarios along with a constraint that the premise and hypothesis in the examples describe the scenario from the same perspective. As with most of the data sets, Amazon Mechanical Turk was used for the data collection where each individual worker was provided premise scene descriptions in which the captions from the Flickr30K corpus were used; then the workers were asked to supply the hypotheses for the 3 different entailment labels of entailment, contradiction and neutral thereby also making the data balanced among the 3 classes. An important thing to note is that only the captions from the Flickr30K data set were used without using the captions. Although this data set does not pertain to the VE task, the description of this data set was included because this data set is utilised in another study to extend and create a data set for such tasks 2.2.9.

2.2.9 SNLI-VE

SNLI-VE (Xie et al. (2019)) is a Visual Entailment data set which was built atop of the SNLI and Flickr30K data sets. The data set contains an image which is considered as the premise and each image has 3 natural language sentences associated with it which are considered as the hypotheses accompanied by the appropriate Visual entailment like entailment, neutral or contradiction. The labels are assigned on the basis of the relationship between the image (premise) and the natural language sentence (hypothesis). They connect the images from the Flickr30K data set to the SNLI data set through the annotations allowing them to construct a structured Visual Entailment data set. They use the captions to connect the 2 data sets and in the final data set, they replace the premise (caption) with the image so that the image becomes the premise. Figure 2.9 presents some examples from the SNLI-VE data set.

For the sake of completeness, the four criteria outlined by Xie et al. (2019) for developing an effective data set are included here:

1. The premise should be based on real-world scenes (be non-iconic) and one premise (image) can be paired up with different hypotheses for the different labels of entailment, contradiction and neutral
2. Fine-grained reasoning about subtle changes in the hypotheses needs to be enforced as

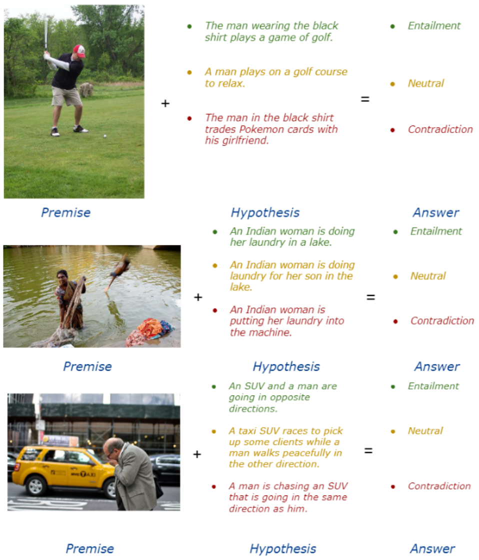

The man wearing the black shirt plays a game of golf.
A man plays on a golf course to relax.
The man in the black shirt trades Pokemon cards with his girlfriend.
+
=
Entailment
Neutral
Contradiction
Premise
Hypothesis
Answer
An Indian woman is doing her laundry in a lake.
An Indian woman is doing laundry for her son in the lake.
An Indian woman is putting her laundry into the machine.
+
=
Entailment
Neutral
Contradiction
Premise
Hypothesis
Answer
An SUV and a man are going in opposite directions.
A taxi SUV races to pick up some clients while a man walks peacefully in the other direction.
A man is chasing an SUV that is going in the same direction as him.
+
=
Entailment
Neutral
Contradiction
Premise
Hypothesis
Answer

Figure 2.9: Examples from the SNLI-VE data set. Image taken from Xie et al. (2019)

small changes in the hypotheses can lead to a different label

3. An image may only exist in a single partition

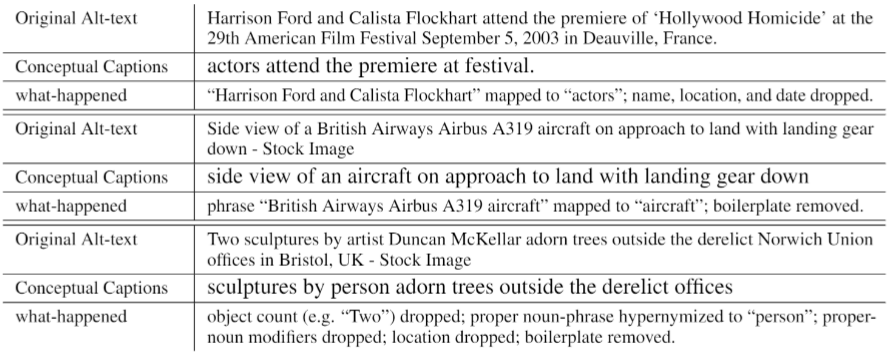
4. A major concern about the image captioning data sets is because if the captions have been assigned by humans, there will naturally be an element of bias present which can cascade into the models trained using that data. For example, GPT-3 generated stories contain gender and representation bias even though it was improved over GPT-2 which suffered from biasness (Lucy and Bamman (2021)). Since biases seem to be hard to avoid, the data set bias has to be measured as well as baselines provided so that evaluations can use them as the lower bound for performance

Xie et al. (2019) note that SNLI-VE meets criterion 1 and 2, some adjustments were made to meet criterion 3 and with regards to criterion 4 that the SNLI data set contains a hypothesis-conditioned bias and since SNLI-VE is built upon SNLI, it inherits that bias for which a hypothesis-only baseline was provided by the creators.

2.2.10 Conceptual Captions

Conceptual Captions is a Google research data set consisting of around 3.3 Million pairs of images and their descriptions. Unlike COCO where the images were the captions were assigned by humans in an Amazon Mechanical Turk assignment, here the images and their descriptions were sources from the web allowing for a wider variety of styles. They use the alt-text attribute of the images for their description which is also what was used in the dissertation for the image captions. The images and their descriptions are collected and then filtered out using the Flume pipeline which can process billions of pages parallelly. This is done so that the there is a balance of fluency, learnability, informativeness, cleanliness of the captions contained in the data set thereby leading to better models generating better captions. They describe their filtering process in much detail in their paper Sharma et al. (2018) which is included here (for the sake of completeness)

They only keep images with height and width greater than 400 pixels and ratio of either to the other is not more than 2. The pipeline also excludes pornographic images or one which trigger profanity detectors which leads to only 35% of the candidates being retained (65% are discarded). For the descriptions they used Part-Of-Speech, pornography/profanity, sentiment polarity annotations from the Google Cloud Natural Language APIs. Apart from these annotations they also had some heuristics like discarding candidates with high token repetition, no determiners or nouns or prepositions and high noun ratio. They choose to retain candidates which have high unique word ratio covering many POS(Part-Of-Speech) tags. They also use a vocabulary consisting of 1 Billion token types. These tokens need to have appeared in the English Wikipedia at least 5 times. Candidates with tokens not in this vocabulary are also discarded. This however, means that the pipeline would discard the captions which may have names of regular civilians which would rid the data of many candidate captions especially in domains like crime, politics, entertainment, technology. So before making use of the pipeline, this step will have to be modified. However, it must be noted that for an initial testing of systems it is a fair solution. They also remove captions which trigger the pornography/profanity detectors or captions with very high or very low polarity annotation scores. They then use the Google Cloud Vision API to filter out images where none of the text tokens could be mapped to the image contents. This particular step can be used to classify such image-caption pairs as contradicting each other. They have considered around 5 billion images from around 1 Billion web pages in the English language. Due to their filtering criteria, only 0.2% of the image,caption pairs could get through. Figure 2.10 presents examples of the original text and the hypernymed texts. As it can be seen that the hypernymed text are similar in form to the captions generated by image captioning systems, comparing ranks of hypernymed captions and generated captions can yield better results than comparing the actual and generated captions.



Original Alt-text	Harrison Ford and Calista Flockhart attend the premiere of 'Hollywood Homicide' at the 29th American Film Festival September 5, 2003 in Deauville, France.
Conceptual Captions	actors attend the premiere at festival.
what-happened	"Harrison Ford and Calista Flockhart" mapped to "actors"; name, location, and date dropped.
Original Alt-text	Side view of a British Airways Airbus A319 aircraft on approach to land with landing gear down - Stock Image
Conceptual Captions	side view of an aircraft on approach to land with landing gear down
what-happened	phrase "British Airways Airbus A319 aircraft" mapped to "aircraft"; boilerplate removed.
Original Alt-text	Two sculptures by artist Duncan McKellar adorn trees outside the derelict Norwich Union offices in Bristol, UK - Stock Image
Conceptual Captions	sculptures by person adorn trees outside the derelict offices
what-happened	object count (e.g. "Two") dropped; proper noun-phrase hypernymized to "person"; proper-noun modifiers dropped; location dropped; boilerplate removed.

Figure 2.10: Examples of the hypernymisation in Sharma et al. (2018)

On the remaining candidates, text transformation using hypernymisation was performed using the Google Cloud Natural Language APIs and Google Knowledge Graph API and from the transformed text, very short or inconsistent sentence samples were discarded amounting to 20% of the candidates. They then cluster the resolved entities and consider only the candidates for which the detected types have a count of more than 100 which amounted to about 55% of the candidates. The hypernymisation can be used on the actual image captions in order to compare them with the descriptive type captions generated by current image captioners. Concluding, the data set contains 3.3 Million images in the training set, 28K in validation set and 22.5K in the testing set.

2.2.11 GoodNews

The GoodNews data set contains an article, image and its corresponding caption. The training-validation-test split provided by Biten et al. (2019) is 421K, 18K, 23K for training, validation and testing respectively. The data set contains about 466K image URLS of which Tran et al. (2020) were able to download 99.2% of the original data set which amounts to about 463K images. This means that the data set contained 0.8% broken image links. Tran et al. (2020) also note some other issues with the GoodNews data set like incomplete or non-English articles included or irrelevant sidebar images included. Nevertheless, till the work by Tran et al. (2020), the GoodNews data set was the largest available data set for News Image Captioning.

2.2.12 NYTimes800K

This data set is the contribution of the work Tran et al. (2020) which aims to work upon the issues of the GoodNews data set. This data set contains about 793K images. It should be noted that this is the number of images but the number of articles covered is lesser than the number of images which means that more than one image was extracted from the articles (the same is observed with the GoodNews data set i.e. the the number of images is greater than number of articles hence articles contribute more than 1 image). Tran et al. (2020) compare NYTimes800K and GoodNews using attributes like number of articles, images; average article, caption length; % of nouns, pronouns, proper nouns and so on. The only main difference observed between the 2 data sets is the average data set length and the average article length which for NYTimes800K is 974 whereas for GoodNews is almost half which is 451. Just like the GoodNews data set, NYTimes800K was also created by making use of The New York Times public API. NYTimes800K also makes use of a custom parser to ensure that non-English articles, sidebar images and captions are not included. They also collected information on where the image is located in the article. However, since articles are contributing more than one image and apart from the key image which can be found at the top of the article, there can be other images lying somewhere in the middle of the article and they underscore that the text surrounding the image and its placement are important when it comes to image captioning because intuitively, text more relevant to the image will tend to lie closer to the image.

2.2.13 NewsCLIPpings data set

NewsCLIPpings (Luo et al. (2021)) is a large-scale automatically constructed data set with out-of-context as well as real news based on the VisualNews data set (Section 2.2.4). The automatic generation procedure makes use of only the image and caption and does not make use of the article text. They posit that all the pairs in the VisualNews data set are pristine as they have been collected from reputable news sources. Basically they are automatically creating the out-of-context news but following a different approach as opposed to previous works which would assign some random caption to an image to call it falsified. Here, they consider 4 different scenarios where they design the matches as follows:

1. caption-image similarity: They use the SOTA CLIP Text-Image Similarity to retrieve the samples which have the highest similarity between the image and the caption
2. caption-caption similarity: They use CLIP Text-Text similarity to match captions which have the highest text-text similarity and retrieve the image corresponding to those. In this case, the captions are semantically similar but have different named entities. Suppose there is an image and a caption. The caption is designated as source and is matched to a set of candidate captions. The caption with the highest similarity to the source is designated as target and the image of target is assigned the source caption.
3. Person match: They retrieve an image which has the named entity from the source caption but in a different scene or context. As one can tell, this in itself is a challenging task which they handle as follows:
   (a) The candidate captions must contain "PERSON" entities and candidate images must contain a person related Faster-RCNN bounding box.
   (b) There can also be a scenario where a person may be mentioned in the caption but the image does not contain the person. In order to avoid such cases, they use spaCy's dependency parser to determine if the person is in the possessive form or is the object of a sentence
   (c) In order to ensure that the image being retrieved has a distinct context, they place a constraint that the Places365 ResNet similarity must be lesser than 0.9.
4. Scene Match: An image is retrieved which contains the same scene but in a different context i.e. the intent here is to mislabel the event. They do this as follows:
   (a) Candidate captions should not have any "PERSON" named entities. This aims to get rid of head shots or images without much scene information

(b) They match samples which have highest Places365 ResNet image similarity. This is determined by a dot product of their ResNet embeddings. (Basically, they embed the source image and the target image using ResNet and they take a dot product of those embeddings in order to get a measure of the similarity of the images, similar to cosine similarity)

They also make use of Adversarial CLIP Filtering in order to deal with a distributional shift in CLIP Text-Image Scores between a pair of source image and source caption (pristine pair) and another target image and source caption. The problem here is that the similarity score of those 2 pairs could provide enough 'signal' to be able to classify an image-caption pair as falsified or pristine. Concluding, NewsCLIPpings data set is built upon the Visual News data set. Individual captions appear twice in the NewsCLIPpings data set once in pristine image-caption pairs and once in falsified with a total of 988K pairings with half of them pristine and half falsified. Basically, this data set provides identifiers for the VisualNews data sets. Luo et al. (2021) also discuss previous data sets such as MAIM, MEIR, TamperedNews, COSMOS however they are omitted here due to a low degree of relevance. One striking aspect about their paper is their Ethical Considerations section. They to some extent have tried to discuss some of the criteria which were referred to in the paper Xie et al. (2019) (Section 2.2.9) for a good data set and apart from those, they have also included a section on the Carbon Emissions of the research. This is something which should be considered by each research work and it is hoped that the same can be followed in our future works.

3 Methodology

This chapter describes the methodologies and implementations which were followed for the data set collection and the attempt at the labelling heuristic. The following discussion gives a brief context on how the research on this project evolved followed by the different methodologies. The research was commenced with a search for relevant data sets. Different sites for multimodal data were considered including social media sites such as Instagram, Twitter or Facebook. Generally, media on the internet is multimodal in nature including audio-visual data, text data or images. In order to limit the scope of the research, a constructive simplification to focus only at still images was made. News websites were also included in the potential candidates. Initially the research question was formulated differently. The search for data sets was begun and methods to collect our own data were made and many different methods were considered for the same. After some exploratory analysis, news websites were chosen as our domain. The initial plan was to collect the data set using pygooglenews, further details and explanation of why it was not used can be found in Section 3.2.1. For some time the focus was on finding Image classification data sets and some data sets along with ready-to-use image classifiers were also found. At the very outset of this project, the initial goal was to classify the objects in the image on a broad level and if the objects were contained in the caption then that would have been a sufficient condition for classifying the relation as an entailment. However, systems to make finer classifications were researched. Hence initially some image classification systems from the ModelZoo website were analysed. The ModelZoo website presents the different available image classification and object detection implementations in one place. Then the question was how to correlate the classification to the actual caption which was much more than just a list of objects. Some image-caption pairs were collected using pygooglenews and then hypothesised that if the image is a stock image, then it means that the article does not have an actual image of the event and hence the caption may not be closely related to the image. We explored APIs of some stock image websites to find whether or not the image is a stock picture. Onyshchak (2020) sent the research into a better direction as it introduced us to Dong et al. (2018) thereby inspiring the search in the field of Image-Caption retrieval. Then slowly the research got directed towards Image Captioning with systems like GPT-3 being explored. This led us to Contrastive Language Image Pre-training or CLIP. RTE was encountered during the literature review which led to the Visual entailment task.

3.1 Experiments

The major experiments which the research is split up into are as follows:

1. Collection of articles from different categories from a news website without using APIs
2. Process the images and utilise existing systems to caption the images
3. Use CLIP ranking to rank the actual image caption in comparison to other candidate captions.
4. Calculate the semantic similarity between the actual image caption and other candidate captions
5. Perform analysis and evaluations using unsupervised and statistical methods

Experiment 1 is performed in order to create a data set of news articles from a website which to the best of our knowledge was not covered in previous data sets. Experiment 2 is performed so that once the image has been captioned, our Visual Entailment task is reduced to a Textual Entailment task (the baseline method used in the benchmark for the visual entailment task (Xie et al. (2019))). Experiment 3 is performed so as to create additional features to support our exploratory study. Since this is an unsupervised problem (there are no data sets available which tell us about the entailment relations between images and captions in news articles) so we have to try and extract as much information as possible using the images and the already existing systems. Experiments 2 and 4 can be encapsulated as the main Visual Entailment methodology. The evaluation of these experiments forms the content of the next chapter.

3.2 Experiment 1: Data Collection

In this section the different data set collection methods are discussed.

3.2.1 Data Extraction using pygooglenews

The Pygooglenews library was used initially to collect the news articles. It has functions which allow users to retrieve the top stories for particular country and languages, users can also retrieve stories by topic/news category. It is a very powerful library however, it allows retrieval of only 100 articles at a time. While there are ways to circumvent it, alternative ways had to be considered as the techniques which allow bypassing the restriction would take up a lot of collection hours which may make it infeasible to collect data which could be of any significance. It should also be noted that when the data collection was attempted using this library, it was working fine on Colab and a Windows machine, but it would not work on a Mac book M1 machine.

3.2.2 Data set Gathering using Web Scraping

The data set is gathered by parsing articles whose URLS were found on the online Kaggle News Category data set (Misra (2018)). The online Kaggle data set is a JSON file which consists of the category of the article, the date, the author, the article title, short description. This JSON is parsed and by making use of the BeautifulSoup Python library and requests library, three images from each article are obtained and its corresponding caption using the Alt-Text HTML attribute.

Since the Kaggle data set is available online, anyone who wishes to gather this data set can download it from the Kaggle website and run it through the scraping code. The URLS are links to articles from the HuffingtonPost website due to which the articles are not from varied sources unlike the VisualNews Data set. According to the description provided for the VisualNews Dataset in the work Liu et al. (2020), it seemed that categories of the articles were not included, however upon looking at the demo for accessing the data set, it was found that the categories are included as well. The currently created data set contains 40K data points, and the main difference is the source of the news. The current data set is similar to the N24News data set which contains the images and corresponding captions along with the article text and category. However the difference is that the articles are from a different website. N24News gets articles from the New York Times Website using an API whereas the current data set uses the data set from Kaggle for the links to the Huffington Post articles. Using these links and web scraping with BeautifulSoup the image links and the captions are collected initially. However, unlike the Visual News Data set, another version of the data set contains the links to the news articles, the category, headlines, authors, date of article, short description of article, image link, image captions, and article text. For the VisualNews format version of the data set, the images were downloaded and placed into the images folder with the appropriate ID information contained in the JSON file along with the other attributes outlined above. Similarly the articles were placed in text files in the articles folder. The purpose of this activity was to make the dataset conformant to the format of the VisualNews Dataset. The access to the data set was requested and the author very kindly obliged, however, the zip folder containing the articles is 4.2GB large and the zip folder containing the data.json file along with the images and the articles has a size of 91GB.

Initially, using the links from the Kaggle data set, 3 images and corresponding captions are collected from each article. If the caption length is greater than 50 characters, the article text is retrieved and the article goes into the data set for which the images are retrieved later on. This is done so that useless images do not have to be downloaded. This can also help with reducing the number of image links from which the images are downloaded thus slightly reducing the risk of downloading possibly dangerous images. If the caption length is shorter than 50 then the article is not included. This rids the data set of images of authors (which have image captions of very less length as the caption is the author name), and it also rids the data set of any images without captions or images with irrelevant captions which only have the name of the stock photography website the image is from. This is done in a phased manner such that 100 articles are accessed at a time, followed by the data being converted to a CSV and then the next 100 are collected. This was done due to unexpected kernel crashes leading to loss of data collected till then hence collecting in a phased manner allows salvaging of the already collected data in the face of unexpected scenarios like kernel crashes, network failures or system failures. Once this preprocessing is performed, the 4 categories with the highest number of articles are chosen. However, since the choice of categories can be very subjective, the data at the different stages can be provided or scraped by the user with the help of instructions provided in the GitHub Repository. The image IDs are stored in a JSON file along with the other attributes like the category of the data, the image caption, the article text. The reason why the images should be downloaded is because the links may break or change any time thereby undermining the amount of data which can be obtained. However, it is useful to also have a version of the data set where the image links are contained along with the other attributes for easier distribution and the images could be downloaded by the user on their side. (It must be noted that due to legal restrictions, one needs to use the Kaggle data set and use the scraping code to reproduce the data) Since our data set also contains the date of the article, the data set can also be used to perform temporal analyses such as the evolution of cross-media relations over time. Figure 3.1a gives an overview of the collection procedure. Section 4.1 presents additional information about the data sets.

3.3 Experiment 2: Image Captioning

This section describes the Image Captioning systems used for captioning the images in the news articles. However, this section describes the same briefly as the systems have been explained in more detail in Section 2.1.3.

3.3.1 OFA Image Captioning

The One-For-All task and modality agnostic framework was employed to generate the captions for the images. Some examples of its caption generation capabilities have been provided in Section 4.2. They provide a very detailed tutorial on how to run their implementation on their GitHub page along with demos for the various tasks they implement: Image Captioning, text to image generation, Visual Grounding and Visual Question Answering. They also provide the Colab notebooks for Image Captioning, Visual Question answering, Referring Expression Comprehension along with a generic interface which gives the ability to perform various different tasks using only one model by using different instructions. It must be noted that the OFA captioning system could not be run on the Mac M1 and was used on Colab itself. Future researchers ought to download the checkpoint and save it to their Google Drive as it would save them a lot of time especially if they themselves have a Mac with a silicon chip.

3.3.2 Zero-shot Image-to-text generation

The other Image Captioning system which was considered and was successfully run was Tewel et al. (2021)'s Zero Shot Image to text generation. They employ CLIP to perform zero-shot Image captioning where the input is an image which is fed into the model consisting of CLIP and GPT-2 Language model to generate the caption. Their main aim of using the image to text generation for Visual-semantic arithmetic is also another very interesting task. While the visual semantic arithmetic examples presented may not really benefit our task too much directly, it has the potential to be employed for its solution. The image captions were generated using their Python implementation on their Github Page and also used their Image captioning demo provided on the same page to compare the results. They results and some discussions are presented in Section 4.2.

It must be noted that in the final pipeline, the OFA captioning system was used as it is regarded as the State-Of-The-Art System for Image Captioning on the MS-COCO data set 2.3.

The author of this dissertation acknowledges the fact that more time could have been invested in employing the Entity-aware Image Captioners which could have improved the results however, the analysis of general Image Captioning systems is also beneficial.

3.4 Image-Caption Ranking

This section describes the method used for the Image Caption ranking. Here, the system is provided with an image and a set of captions and the system returns a ranking for the images.

There are innumerable implementations for this as well, however, OpenAI's official CLIP implementation is used. Refer to Section 2.1.4 for more details on how this system works. The CLIP system is provided with the image from the news article along with a set of 4 captions:

1. Headline
2. Short description of the article

3. The actual image caption
4. The generated caption for the image from the OFA system

Initially, the cosine similarity between the image and the actual caption was to be calculated by making use of the image-text embeddings which CLIP returns but then it was decided to use the above set of 4 candidate captions. The motivation behind this is as follows: If the actual image caption ranks highest among all of them, one can conclude that the relation between the image and the caption is that of entailment (equivalence). On the other hand, if the OFA generated caption which basically describes the image in a literal sense is ranking higher than the actual caption then while it cannot be concluded whether the relation is contradiction (independence) or neutral (complementarity) but one can conclude that the relation most likely is not entailment (equivalence). Hence the headline and short description were also included and the motivation for that is if the headline ranks highest, say, then it can be said that the image may not be related to the actual caption provided but it does correlate to the article material and hence the relation can be said to be neutral. However, if the OFA generated caption ranks the highest, then that means the caption which is descriptive in nature correlates better to the image than the actual caption, headline or short description which means that the image is most likely out of context and hence a contradiction relation can be concluded. It must be noted that in this particular field a lot of assumptions have to be made which may sometimes not even seem feasible but in order to inch closer to the solution, one can approach it after assumptions which can make the problem easier to solve and then build from there. Section 4.3 presents some examples of the ranks produced by this system.

3.5 Experiment 4: Semantic Similarity

This experiment is not dependent on Experiment 3 but has high dependence on Experiment 2. As was the baseline method used in the paper Xie et al. (2019), after performing the Image Captioning which reduces our task to textual entailment, our task is to find the semantic similarity between the generated and the actual caption. Bidirectional Encoder Representations from Transformers or BERT is used with some fine tuning in order to perform the semantic similarity task. The Semantic Similarity with BERT blog on the Keras Website was followed. It uses the Stanford Natural Language Inference Data set. Fortunately, since this does not require Pytorch, it could be run on the Mac M1 but was very time consuming. Training for 2 epochs on 250K samples took 10 hours on the Mac M1 with results not being that satisfactory. Some results have been presented in Section 4.4. It must also be noted that they also provide a ready to use Demo, however, for practical purposes, one will want to run the code on their own system. While making a script to feed the sentences into the demo and retrieving the results can save training time, and when considering smaller data sets it may be not the worst idea to do so, for bigger data sets it is not a very tangible solution.

They also provide a pre-trained model which can be used with a few changes to the code provided in their Google Colab notebook. Using the pre-trained model it takes just 10 minutes to give results and the provided results are as good as their demo.

3.6 Experiment 5: Unsupervised methods and Statistical Analysis

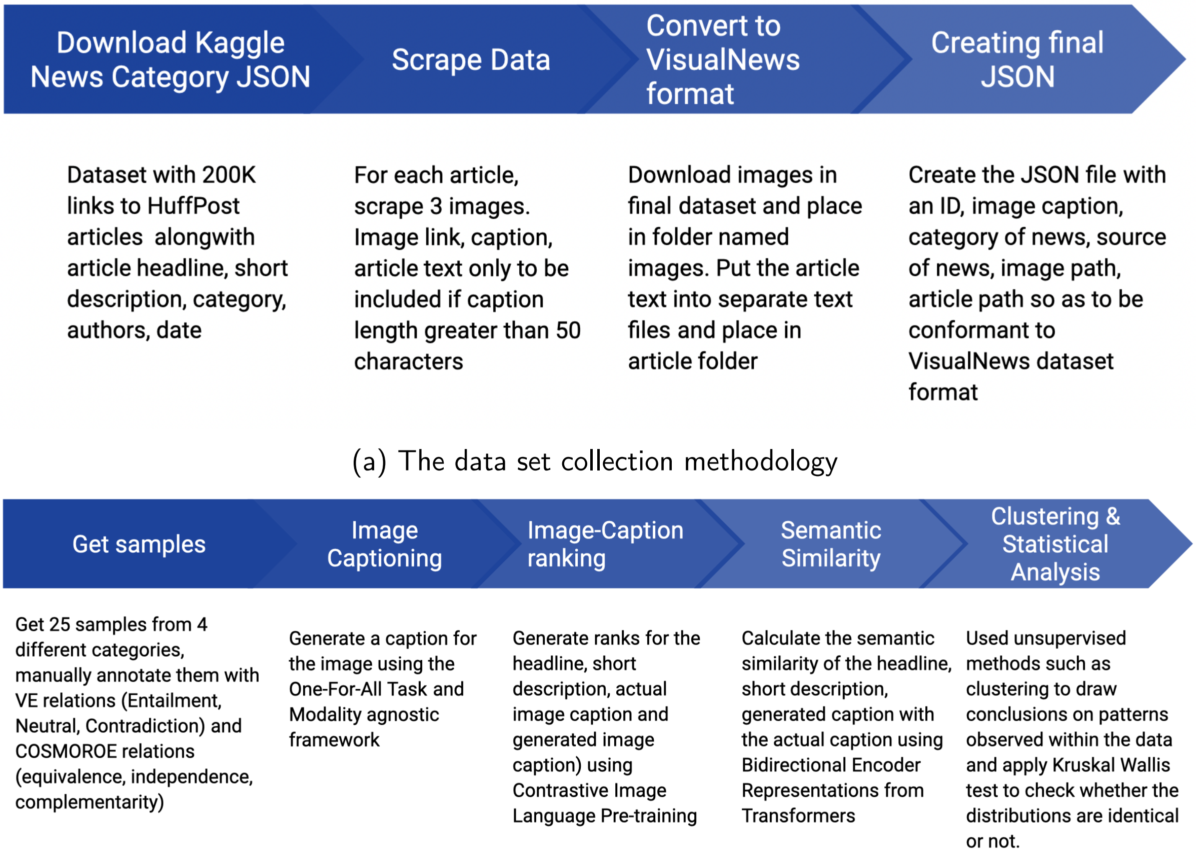
Since there are no data sets which contain labelled image-caption pairs from news articles, in order to evaluate our methods,a data set of 125 data points had to be manually annotated. The data set contains 25 data points from 5 different categories. The data set was annotated with the original textual entailment relations as well as the CrossMedia relations proposed by Pastra (2008). It was found that the relations can indeed be mapped in the following way:

1. Entailment: equivalence
2. Neutral: Complementarity
3. Contradiction: Independence

This is actually good news because this means that the previous data sets which were annotated using crowd sourcing efforts do not need to be annotated again and the research for the finer cross media relations as presented in Figure 2.1 can be carried out in the future. K-Means clustering is applied along with the different combinations of the generated features. Just to summarise, the generated features used are CLIP ranks with the headline, short description, image caption and actual caption. Apart from this statistical tests like the Chi-Square test and the Kruskal Wallis Test are considered to analyse the distributions. However, the Chi Square test was not used on the current data set as it requires the frequencies of the items to be at least 5 whereas in our data set there were frequencies which fell below that threshold hence the Chi-Square test was not used. The results are presented in Section 4.5.

As one can tell, the scope of this project is really vast with individual experiments being challenging problems in themselves. While it seems that almost everything already has ready implementations and models to be used, the journey from researching the solution to the problem, to finding an appropriate solution, and successfully making use of it is an extremely long one. However, this also proves to be a great learning experience and while the journey can sometimes be very frustrating, at the end the knowledge gained is invaluable.

Figure 3.1 presents the methodology in a more succinct way.


Download Kaggle News Category JSON
Scrape Data
Convert to VisualNews format
Creating final JSON

Dataset with 200K links to HuffPost articles alongwith article headline, short description, category, authors, date

For each article, scrape 3 images. Image link, caption, article text only to be included if caption length greater than 50 characters

Download images in final dataset and place in folder named images. Put the article text into separate text files and place in article folder

Create the JSON file with an ID, image caption, category of news, source of news, image path, article path so as to be conformant to VisualNews dataset format

(a) The data set collection methodology

Get 25 samples from 4 different categories, manually annotate them with VE relations (Entailment, Neutral, Contradiction) and COSMOROE relations (equivalence, independence, complementarity)

Generate a caption for the image using the One-For-All Task and Modality agnostic framework

Generate ranks for the headline, short description, actual image caption and generated image caption) using Contrastive Image Language Pre-training

Calculate the semantic similarity of the headline, short description, generated caption with the actual caption using Bidirectional Encoder Representations from Transformers

Used unsupervised methods such as clustering to draw conclusions on patterns observed within the data and apply Kruskal Wallis test to check whether the distributions are identical or not.

(b) The heuristic methodology

Figure 3.1: The methodologies presented in a pictorial form

4 Evaluation

This chapter presents the results and discussions of the various experiments presented in the previous chapter. They shall be addressed by each experiment.

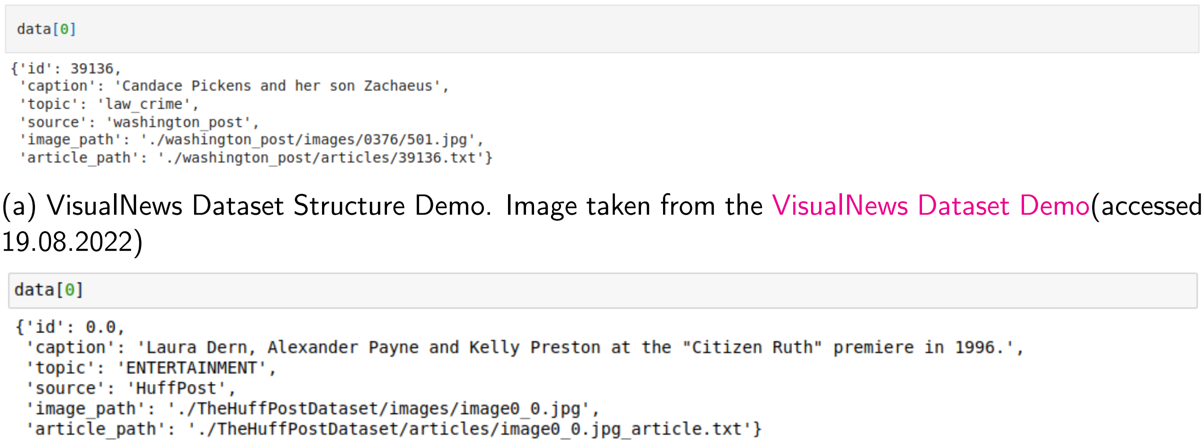
4.1 Experiment 1: Data Collection

This section provides some insights into the collected data set. Figure 4.1 presents the similarities between the attributes of the collected data set and the VisualNews data set. Figure 4.2 presents the similarity and ease of access of the different attributes present in the data set for both the collected and the Visual News data set.

It should be noted that creating datasets with similar structures can encourage better collaboration and allow creation of very rich datasets containing data from many diverse sources. In order to get access to the VisualNews Dataset, one must follow the instructions on their GitHub Page.

4.2 Experiment 2: Image Captioning

In this section the results from the different captioning systems are presented which could assist the decision of future researchers regarding which systems to use or whether to even use the systems at all. The manually annotated dataset has 125 data points which have been annotated with the textual entailment relations: entailment, contradiction, neutral and also with the relations from the COSMOROE framework Pastra (2008): equivalence, independence, complementarity. The dataset has 25 data points from the following categories: CRIME, SPORTS, ENTERTAINMENT, WOMEN, POLITICS. The images are captioned using the OFA captioning system; the output of which is then considered as another candidate to compare against the ranks of the actual image caption. Once the image captioning has been performed, after that the following are considered as the candidates to be ranked using the CLIP system: article headline, article short description, actual image caption, generated image caption. The hypothesis is that if the generated image caption ranks higher than the actual image caption then the relation between the image-caption pair is likely contradiction



```
data[0]
```

```
{'id': 39136,
 'caption': 'Candace Pickens and her son Zachaeus',
 'topic': 'law_crime',
 'source': 'washington_post',
 'image_path': './washington_post/images/0376/501.jpg',
 'article_path': './washington_post/articles/39136.txt'}
```

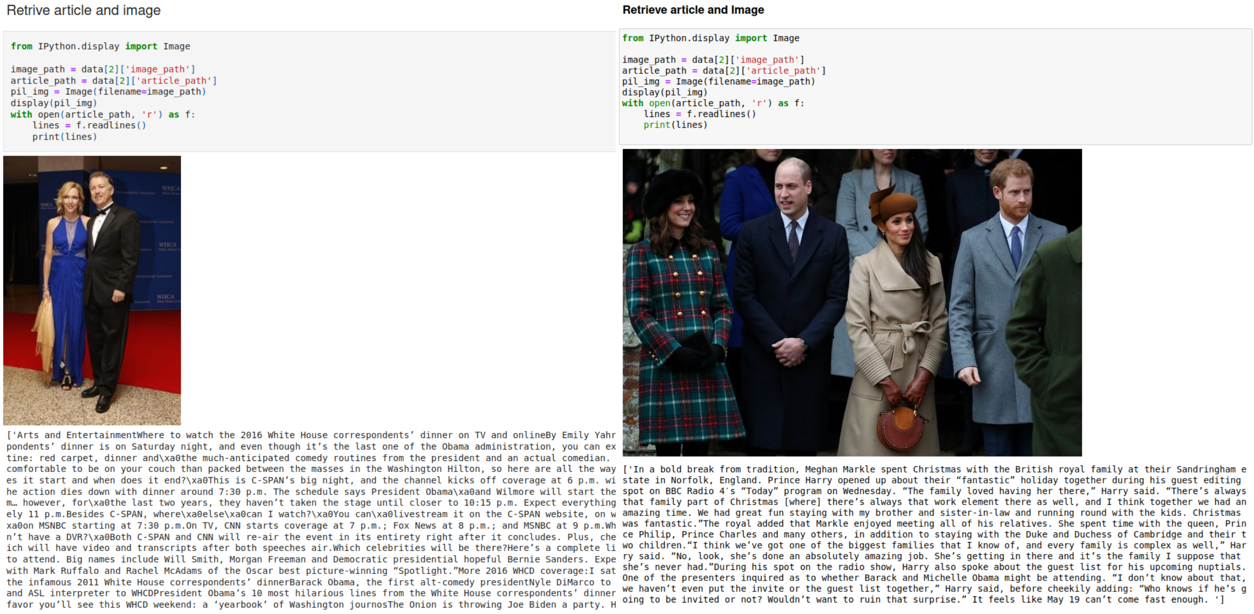
(a) VisualNews Dataset Structure Demo. Image taken from the VisualNews Dataset Demo(accessed 19.08.2022)

```
data[0]
```

```
{'id': 0.0,
 'caption': 'Laura Dern, Alexander Payne and Kelly Preston at the "Citizen Ruth" premiere in 1996.',
 'topic': 'ENTERTAINMENT',
 'source': 'HuffPost',
 'image_path': './TheHuffPostDataset/images/image0_0.jpg',
 'article_path': './TheHuffPostDataset/articles/image0_0.jpg_article.txt'}
```

(b) TheHuffPost Dataset Structure Demo

Figure 4.1: Presenting the format of the 2 datasets. The main aim of the same is to present the structure similarity of the 2 datasets.



Retrive article and image

```
from IPython.display import Image

image_path = data[2]['image_path']
article_path = data[2]['article_path']
pil_img = Image(filename=image_path)
display(pil_img)
with open(article_path, 'r') as f:
    lines = f.readlines()
    print(lines)
```

```
['Arts and EntertainmentWhere to watch the 2016 White House correspondents’ dinner on TV and onlineBy Emily Yahr
pondents’ dinner is on Saturday night, and even though it’s the last one of the Obama administration, you can ex
tine: red carpet, dinner and\xa0the much-anticipated comedy routines from the president and an actual comedian.
comfortable to be on your couch than packed between the masses in the Washington Hilton, so here are all the way
es it start and when does it end?\xa0This is C-SPAN’s big night, and the channel kicks off coverage at 6 p.m. wi
he action dies down with dinner around 7:30 p.m. The schedule says President Obama\xa0and Wilmore will start the
m… however, for\xa0the last two years, they haven’t taken the stage until closer to 10:15 p.m. Expect everything
ely 11 p.m.Besides C-SPAN, where\xa0else\xa0can I watch?\xa0You can\xa0livestream it on the C-SPAN website, on w
xa0on MSNBC starting at 7:30 p.m.On TV, CNN starts coverage at 7 p.m.; Fox News at 8 p.m.; and MSNBC at 9 p.m.Wh
n’t have a DVR?\xa0Both C-SPAN and CNN will re-air the event in its entirety right after it concludes. Plus, che
ich will have video and transcripts after both speeches air.Which celebrities will be there?Here’s a complete li
to attend. Big names include Will Smith, Morgan Freeman and Democratic presidential hopeful Bernie Sanders. Expe
with Mark Ruffalo and Rachel McAdams of the Oscar best picture-winning “Spotlight.”More 2016 WHCD coverage:I sat
the infamous 2011 White House correspondents’ dinnerBarack Obama, the first alt-comedy presidentNyle DiMarco to
and ASL interpreter to WHCDPresident Obama’s 10 most hilarious lines from the White House correspondents’ dinner
favor you’ll see this WHCD weekend: a ‘yearbook’ of Washington journosThe Onion is throwing Joe Biden a party. H
```

Retrieve article and Image

```
from IPython.display import Image

image_path = data[2]['image_path']
article_path = data[2]['article_path']
pil_img = Image(filename=image_path)
display(pil_img)
with open(article_path, 'r') as f:
    lines = f.readlines()
    print(lines)
```

```
['In a bold break from tradition, Meghan Markle spent Christmas with the British royal family at their Sandringham e
state in Norfolk, England. Prince Harry opened up about their “fantastic” holiday together during his guest editing
spot on BBC Radio 4′s “Today” program on Wednesday. “The family loved having her there,” Harry said. “There’s always
that family part of Christmas [where] there’s always that work element there as well, and I think together we had an
amazing time. We had great fun staying with my brother and sister-in-law and running round with the kids. Christmas
was fantastic.”The royal added that Markle enjoyed meeting all of his relatives. She spent time with the queen, Prin
ce Philip, Prince Charles and many others, in addition to staying with the Duke and Duchess of Cambridge and their t
wo children.“I think we’ve got one of the biggest families that I know of, and every family is complex as well,” Har
ry said. “No, look, she’s done an absolutely amazing job. She’s getting in there and it’s the family I suppose that
she’s never had.”During his spot on the radio show, Harry also spoke about the guest list for his upcoming nuptials.
One of the presenters inquired as to whether Barack and Michelle Obama might be attending. “I don’t know about that,
we haven’t even put the invite or the guest list together,” Harry said, before cheekily adding: “Who knows if he’s g
oing to be invited or not? Wouldn’t want to ruin that surprise.” It feels like May 19 can’t come fast enough. ']
```

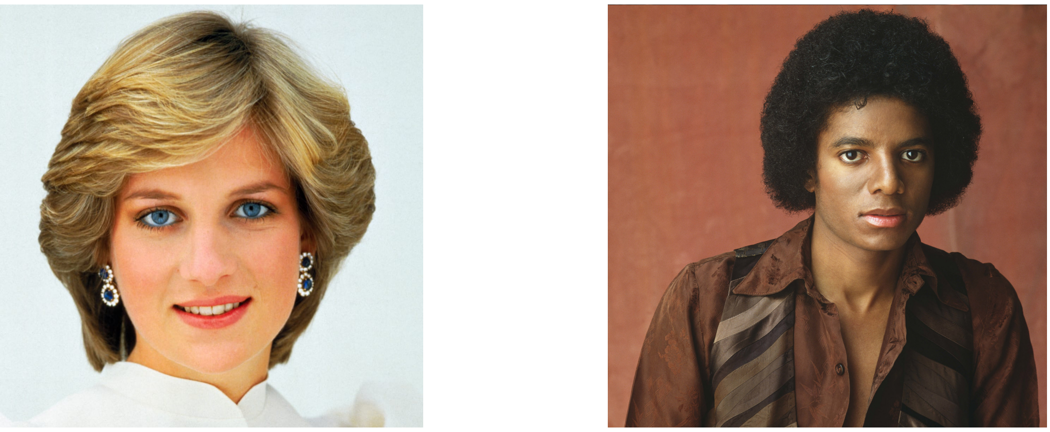
(a) VisualNews Dataset reading Image, Article Demo. Image taken from the VisualNews Dataset Demo(accessed 19.08.2022)

(b) TheHuffPost Dataset reading Image, Article Demo

Figure 4.2: Accessing the Image, Article from the 2 datasets. The main aim of the same is to present access similarity

or neutral. Please note, the caption and the image may be in accordance to the article but here, only the relations between the image and the caption have to be considered.

However, it must be noted that the annotations may not be precise as they were not annotated by an expert in this subject matter of entailment relations as it may involve/require a high level of knowledge in literature and English.



(a) YoadTew Python: Image of a British British Prime Minster Margaret Thatcherism in the 1980s.
YoadTew Demo: Image of Princess Diana in 1989.
OFA Python: a portrait of a woman with short blonde hair
OFA Demo: a portrait of a woman with short blonde hair
Image taken from link(accessed 19.08.2022)

(b) YoadTew Python: Image of a black man man and the colour of a shirt with a black.
YoadTew Demo: Image of Michael via PBS Newshour via YouTube.
OFA Python: a portrait of a woman wearing a brown jacket and a striped shirt
OFA demo: a portrait of a woman wearing a brown jacket and a striped shirt
Image taken from link(accessed19.08.2022)

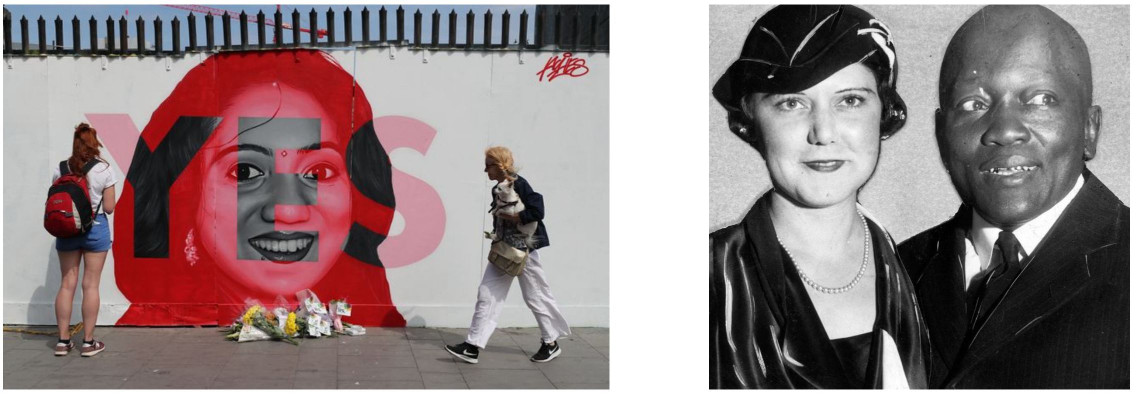
Figure 4.3: The captions generated by the 2 systems.

COSMOROE helps assign meaning to the pairs which were just assigned as neutral and then they can also become informative as being assigned neutral does not really do a lot. Also, in news articles, the image cannot always exactly represent the caption. For example, Image from HussPost article(accessed 16.08.2022) has caption 'Rep. Mo Brooks (R-Ala.) is a plaintiff in a suit seeking to bar the Census Bureau from counting undocumented residents.BLOOMBERG / GETTY IMAGES' so the image cannot literally show whatever is mentioned in the caption. Hence using normal entailment relations would just assign these as neutral perhaps which tends to be overlooked whereas the COSMOROE relations can provide more information.

While OFA is regarded as SOTA on the MS-COCO dataset, the results are not very good. However, it must be noted that these results are for the OFA base model which could be run feasibly on hardware I had access to. While the YoadTew Demo results are more fine grained, they also did take much longer to generate. While it remains a worthy candidate to be analysed in the future, Entity aware Image Captioners like Liu et al. (2020), Tran et al. (2020) would be better candidates.

4.3 Experiment 3: Image Caption Ranking

This section presents a few examples of the ranks assigned by the CLIP system to actual examples taken from the HuffPost website. As can be seen in Table 4.1, for Figure 4.4a the headline ranks highest and if the actual caption and the image are considered, it can be seen



(a) The landmark referendum heralds a new era for women's rights in a government that for centuries operated as a theocracy.
Image taken from HuffPost Article(accessed 19.08.2022)

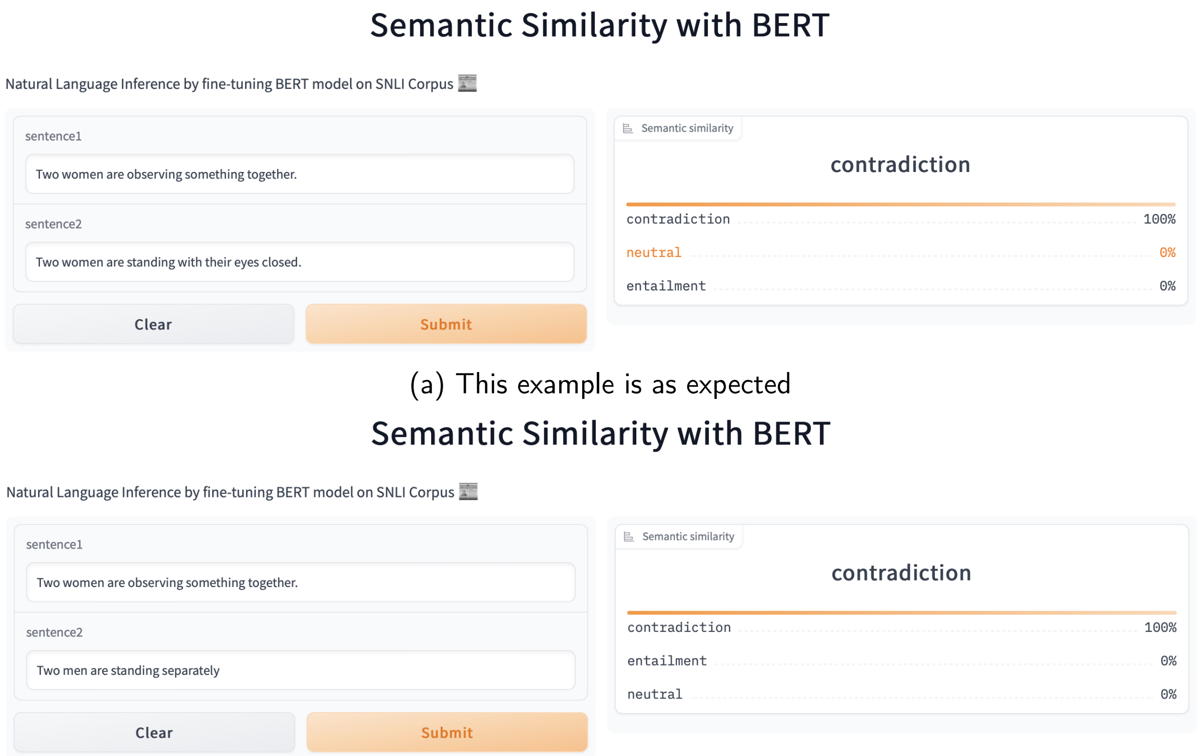
(b) Jack Johnson in 1920 with his second wife, Lucille Cameron. He was often criticized for his romantic relationships with white women. HuffPost article(accessed 19.08.2022)

Figure 4.4: The images and their captions whose ranks assigned by the CLIP system are presented in Table 4.1

	Actual Caption	Generated Caption	Headline	Short description
Figure 4.4a	The landmark referendum heralds a new era for women's rights in a government that for centuries operated as a theocracy.	two people walking in front of a mural on a wall	Ireland Votes To Repeal Abortion Amendment In Landslide Referendum	Irish women will no longer have to travel to the United Kingdom to end their pregnancies.
	0.2996749	0.007292813	0.68584234	0.007190041
Figure 4.4b	Jack Johnson in 1920 with his second wife, Lucille Cameron. He was often criticized for his romantic relationships with white women.	a black and white photo of a man and a woman	Jack Johnson Was Pardoned, But Taboo Sex Is Still Being Criminalized	A new law to fight sex trafficking targets some of the people it ostensibly aims to protect.
	0.23145692	0.7367733	0.031766154	3.59E-06

Table 4.1: The CLIP ranks for the headline, Generated Caption, Headline, Short Description for Figures 4.4a, Figure 4.4b

that there is a neutral relation between the two as they neither entail nor contradict each other. However, when one considers Figure 4.4b, it is seen that the image and caption do entail each other yet the OFA caption is ranked the highest. This is not a fault of either of the systems. They are performing in the way they are supposed to. Hence, this gives us a hint that there needs to be either a different image captioner used which can provide captions more similar to the ones seen in news articles or to use techniques like hypernymisation (Sharma et al. (2018)) on the actual captions so that the named entities in the caption can be converted into tags which can be better compared against the captions generated by OFA.



Semantic Similarity with BERT

Natural Language Inference by fine-tuning BERT model on SNLI Corpus

sentence1: Two women are observing something together.

sentence2: Two women are standing with their eyes closed.

Semantic similarity: contradiction

```
contradiction ........ 100%
neutral .............. 0%
entailment ........... 0%
```

(a) This example is as expected

Semantic Similarity with BERT

Natural Language Inference by fine-tuning BERT model on SNLI Corpus

sentence1: Two women are observing something together.

sentence2: Two men are standing separately

Semantic similarity: contradiction

```
contradiction ........ 100%
entailment ........... 0%
neutral .............. 0%
```

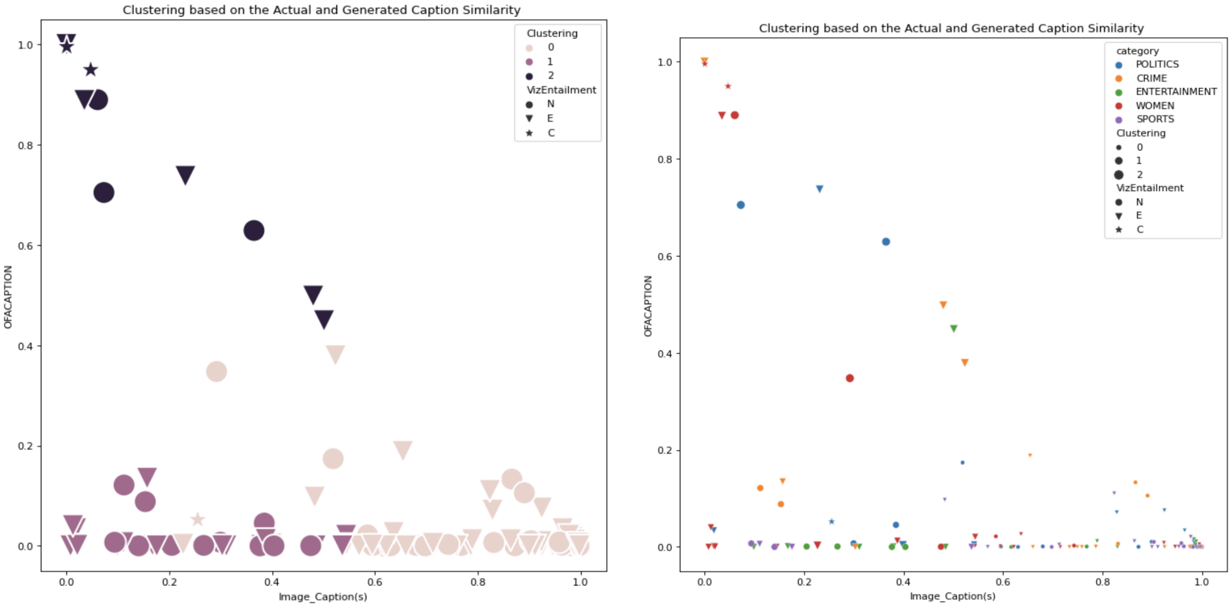
(b) This example is confusing as it talks about 2 men standing separately which does not have anything to do with the women. They can be said to not be related or to have a neutral relation. But it would be better to say that they are 'independent' (thereby an example of how the COSMOROE relations make more sense)

Figure 4.5: An example presenting how the Semantic Similarity system works. Taken from the Semantic Similarity Demo(accessed 19.08.2022)

4.4 Experiment 4: Semantic Similarity

Here the Semantic Similarity with BERT implementation demo is presented by making use of some examples. Figure 4.5 presents 2 different examples discussing the subjectivity issues in this sort of task and presenting how COSMOROE relations are better suited to label cross-media relations than RTE relations.

However, when the semantic similarity implementation was run on python, it was noticed that it was tagging the relation between almost all of the actual captions and the OFA generated captions as contradictions. This gives us another hint that a different captioner system ought to be used as OFA is being too descriptive for our use case.



(a) The colours represent the different clusters, shapes represent the manually annotated relations

(b) The colours represent the different categories, the size represents the cluster number assigned and the shape represents the manually annotated relations

Figure 4.6: Clustering using CLIP ranks for actual and generated image captions. It can be observed that the all the relations are lying in the different clusters and so there are no set patterns for relations based on the CLIP ranking score. This was also observed in the other combinations which were used for the clustering and one of the examples is presented. Figure 4.6b presents how there are no patterns even when it comes to categories of news articles.

4.5 Experiment 5: Unsupervised methods and Statistical Analysis

Upon inspection of the manually annotated data set, it was found that the entailment relation is the most frequent across different categories with 78 of the image-caption pairs entailing each other, followed by the neutral relation with 41 of the image-caption pairs followed by only 3 pairs contradicting each other. This also indirectly gives us another result. Luo et al. (2021)'s work makes use of the VisualNews data set to generate out-of-context image caption pairs. It is posited that the image-caption pairs in the VisualNews dataset are pristine as they are from reputed sources. Initially, this assumption seemed a little inappropriate and the author of the paper was mailed in order to clarify this. They noted that since the image-caption pairs in VisualNews are from reputed sources like the Guardian, BBC hence it could be assumed that they are not falsified. And now considering the results of the manually annotated data set, their assumption seems justified.

We apply the Kruskal Wallis test to the data in order to find out the interactions between the different entailment relations, categories to the features like the CLIP rank assigned to the actual caption, generated caption, headline, short description. Table 4.2 presents a summary of the obtained results. From the p-values it is observed that the rank assigned by the CLIP system for the Actual Image Caption and the Visual Entailment category are not independent. The same applies for the generated captions as well but not for the headline or short description (which makes sense as the relations were assigned based on the captions and not the headline or short description). It is seen that the Actual Image Caption rank is not independent of the category implying that the image caption similarities show some variance across the different categories which also applies to the generated captions.

The fact that there is a significant interaction between the Visual Entailment relation and the CLIP ranks for Image Captions shows that the existing systems which were not specifically designed for this task showed results where there is significant interaction between the manually assigned classes and the similarity assigned by the CLIP ranker. Similar results are obtained when considering the CLIP ranks for the Visual Entailment relation and the OFA caption CLIP ranks. This shows that even though the OFA Captioner is providing very descriptive captions, there still is a significant interaction between the relations and the similarity. Considering the above, it can be safe to assume that using Entity aware captioners will have a better interaction with the entailment relation thereby progressing towards the solution for automatic classification of the entailment relations.

This motivates that if Image Captioner systems such as Entity aware captioners which are specifically designed to caption images in news articles are used, they have potential to have a high interaction with the entailment relations and thus towards automatic classification using systems for which datasets already exist.

When we consider the mean values for Entailment and contradiction in Table 4.3 it is seen that for Image Caption the mean for Entailment is higher than Contradiction whereas for OFA Caption the mean for contradiction is higher than for entailment which is in line with our initial hypothesis that if the OFACAPTION is ranking higher than the Actual Image Caption then it is most likely a contradiction. In order to generalise better, an analysis with a larger sample is required. From Table 4.4 it can be seen that the actual image caption mean value for contradiction relation is lower for the articles in the 'Women' category. On the contrary the mean value of the contradiction relation for OFA captioner is highest in the 'Women' category. These 2 observations are suggestive of the fact that pictures in the news articles in the 'Women' category contain more named entities than other categories because of which the CLIP ranker gives a higher rank to the captions generated by OFA as the CLIP ranker would not be able to correlate the name of a celebrity to the person in the picture (similar to the case in Figure 4.4b.

Table 4.2: Kruskal Wallis test results

Measure	Categorical Atrribute	Kruskal-Wallis p-value
Image Caption	Visual Entailment	0.00457
OFA Caption	Visual Entailment	0.01894
Headline	Visual Entailment	0.4148
Short Description	Visual Entailment	0.2592
Image Caption	Category	0.0106
OFA Caption	Category	0.003773
Headline	Category	0.4189
Short Description	Category	0.00256

Table 4.3: Mean Scores based on Visual Entailment relation

Measure	Contradiction value	Entailment value	Neutral value
Image Caption	0.1011983	0.6583530	0.6340100
OFA Caption	0.66568493	0.06338676	0.07956814
Headline	0.01155467	0.23020050	0.25785666
Short Description	0.22156219	0.04805976	0.02856520

Table 4.4: Mean Scores based on Category of News

Measure	CRIME	ENTERTAINMENT	POLITICS	SPORTS	WOMEN
Image Caption	0.6884417	0.6833690	0.6554918	0.7388412	0.4222895
OFA Caption	0.11136564	0.02004280	0.10435541	0.01164739	0.16883265
Headline	0.1890416	0.2931468	0.1706899	0.1917225	0.3234410
Short Description	0.01115103	0.00344144	0.06946287	0.05778885	0.08543688

5 Conclusion and Future Work

This section outlines the main conclusions, the limitations, the future work and the lessons learnt.

5.1 Contributions

This research project was of an exploratory nature where the objectives were to identify relevant data sets, understand the current solutions in the field of Visual Entailment and how it can be applied in the news domain. The other objectives were to identify the distribution of the relations exhibited across the data and to attempt the development of automated means to support the observational study.

The project reached a certain level of success in identifying the data sets relevant to this study and presents brief overviews of the different important data sets found which can give new researchers quick insight about the data available in this field and for experienced researchers to use as pointers. Inspired by the previous work, by applying web scraping on the links provided by an already existing corpus, a data set of 40K news articles from the HuffPost news website is contributed. There are 2 versions of the data set. One contains the category, headline, authors, link, short description, date, image caption, image link and article text. This version of the data set is also useful and might be better for distribution purposes as this data set is just 170MB (as one can retrieve the images on their end for the other version of the data set). However, the data set is not made public because of some legal restrictions. Apart from this the other version of the data set in JSON format has an ID, caption, topic (or category) of news, source of news (which in our case is just HuffPost), image path (path to the downloaded image), article path (path to the text file containing the article). This was done so that the format is similar to the VisualNews data set (Figure 4.2). Apart from this, the COSMOROE relations as presented as the more suitable relations for the visual entailment task in the news domain and also illustrate the reasons. The cross media relations do not differ a lot across different categories of news, however, considering the size of the manually annotated data set, this needs to be investigated further. It was also found that there is significant interaction between the entailment relations and the rankings assigned by the CLIP ranking system. And it is also concluded that for the particular task of visual entailment in news article domain, the image-captioning task of the One-For-All framework is not very suitable. A heuristic to label the image-caption pairs from news articles automatically was also attempted however, it did not yield any results which could support the automatic annotation. The most important aspect of this research project however, are the learning outcomes. The author believes a vast amount of information was explored as the scope of the project is also very wide and it consists of a lot of challenging tasks. Considering the initial research plan, a lot more ground was covered than was expected however, the author also acknowledges that many aspects of the research could have been improved.

5.2 Limitations

Given the ambitious nature of this research project, there are many limitations. A few of the main ones shall now be discussed. Since there were no labelled data sets for image-caption pairs, a manual annotation was performed for 125 articles. However, this number is not large enough for supervised learning methods to yield good results. The unsupervised methods could have been more complex than were used. The evaluation methodology for the Image-Captioning system could have been improved by making use of different metrics which exist to evaluate such systems (however it should be noted that most popular metrics used are N-gram based and therefore do not perform a good job when it comes to matching the semantics of the texts which is necessary in the news domain as captions tend to be interpretative than pure descriptions). Finally, more Image Captioning systems could have been considered to be compared.

5.3 Future Work

This field is so diverse that there is a lot of scope for future work as the work of visual entailment relations in the field of news articles is considerably new and considering the time scales in academia, the field of multimodal research is still relatively new. However, after the research methodology was performed it was concluded that Image Captioning systems like OFA cannot be employed for the Visual Entailment task in the news article domain and that different systems like Entity aware image captioners should be used or to perform preprocessing upon the image captions before applying ranking or semantic similarity to use procedures like the ones explained in Sharma et al. (2018).

Bibliography

Binti Zahri, N. A. H., F. Fukumoto, and S. Matsuyoshi (2012, 12). Exploiting discourse relations between sentences for text clustering. pp. 17–32.

Biten, A. F., L. Gómez, M. Rusiñol, and D. Karatzas (2019). Good news, everyone! context driven entity-aware captioning for news images. *CoRR abs/1904.01475*.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Cui, Y., G. Yang, A. Veit, X. Huang, and S. Belongie (2018). Learning to evaluate image captioning.

Dagan, I., O. Glickman, and B. Magnini (2005). The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, Berlin, Heidelberg, pp. 177–190. Springer-Verlag.

Diao, H., Y. Zhang, L. Ma, and H. Lu (2021). Similarity reasoning and filtration for image-text matching. *ArXiv abs/2101.01368*.

Dong, J., X. Li, and C. G. M. Snoek (2018). Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia 20*(12), 3377–3388.

Guo, W., J. Wang, and S. Wang (2019). Deep multimodal representation learning: A survey. *IEEE Access 7*, 63373–63394.

Hodosh, M., P. Young, C. Rashtchian, and J. Hockenmaier (2010, July). Cross-caption coreference resolution for automatic image understanding. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Uppsala, Sweden, pp. 162–171. Association for Computational Linguistics.

Hou, S., S. Zhang, and C. Fei (2020). Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications 157*, 113421.

Huang, X., Y. Peng, and Z. Wen (2020, feb). Rce-hil: Recognizing cross-media entailment with heterogeneous interactive learning. *ACM Trans. Multimedia Comput. Commun. Appl. 16*(1).

Lin, T.-Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár (2014). Microsoft coco: Common objects in context.

Liu, F., Y. Wang, T. Wang, and V. Ordonez (2020). Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

Lucy, L. and D. Bamman (2021, June). Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Virtual, pp. 48–55. Association for Computational Linguistics.

Luo, G., T. Darrell, and A. Rohrbach (2021, November). NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, pp. 6801–6817. Association for Computational Linguistics.

Misra, R. (2018, 06). News category dataset.

Murakami, K., E. Nichols, J. Mizuno, Y. Watanabe, H. Goto, M. Ohki, S. Matsuyoshi, K. Inui, and Y. Matsumoto (2010, August). Automatic classification of semantic relations between facts and opinions. In *Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, Beijing, China, pp. 21–30. Coling 2010 Organizing Committee.

Onyshchak, O. (2020). Image recommendation for wikipedia articles.

Pastra, K. (2008, 11). Cosmoroe: A cross-media relations framework for modelling multimedia dialectics. *Multimedia Syst. 14*, 299–323.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). Learning transferable visual models from natural language supervision.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners.

Ramesh, A., P. Dhariwal, A. Nichol, C. Chu, and M. Chen.

Schuhmann, C., R. Beaumont, C. W. Gordon, R. Wightman, T. Coombes, A. Katta, C. Mullis, P. Schramowski, S. R. Kundurthy, K. Crowson, et al. Laion-5b: An open large-scale dataset for training next generation image-text models.

Schuhmann, C., R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Sharma, P., N. Ding, S. Goodman, and R. Soricut (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Tewel, Y., Y. Shalev, I. Schwartz, and L. Wolf (2021). Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic.

Thomas, C., Y. Zhang, and S.-F. Chang (2022). Fine-grained visual entailment.

Tran, A., A. Mathews, and L. Xie (2020). Transform and tell: Entity-aware news image captioning.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Wang, P., A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang (2022). Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

Wang, Z., X. Shan, X. Zhang, and J. Yang (2021). N24news: A new dataset for multimodal news classification.

Xie, N., F. Lai, D. Doran, and A. Kadav (2019). Visual entailment: A novel task for fine-grained image understanding. *ArXiv abs/1901.06706*.