

memeBot: Automatic Image Meme Generation for Online Social Interaction

by

Aadhavan Sadasivam

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2020 by the
Graduate Supervisory Committee:

Yezhou Yang, Chair
Hasan Davulcu
Chitta Baral

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Internet memes have become a widespread tool used by people for interacting and exchanging ideas over social media, blogs, and open messengers. Internet memes most commonly take the form of an image which is a combination of image, text, and humor, making them a powerful tool to deliver information. Image memes are used in viral marketing and mass advertising to propagate any ideas ranging from simple commercials to those that can cause changes and development in the social structures like countering hate speech.

This work proposes to treat automatic image meme generation as a translation process, and further present an end to end neural and probabilistic approach to generate an image-based meme for any given sentence using an encoder-decoder architecture. For a given input sentence, a meme is generated by combining a meme template image and a text caption where the meme template image is selected from a set of popular candidates using a selection module and the meme caption is generated by an encoder-decoder model. An encoder is used to map the selected meme template and the input sentence into a meme embedding space and then a decoder is used to decode the meme caption from the meme embedding space. The generated natural language caption is conditioned on the input sentence and the selected meme template.

The model learns the dependencies between the meme captions and the meme template images and generates new memes using the learned dependencies. The quality of the generated captions and the generated memes is evaluated through both automated metrics and human evaluation. An experiment is designed to score how well the generated memes can represent popular tweets from Twitter conversations. Experiments on Twitter data show the efficacy of the model in generating memes capable of representing a sentence in online social interaction.

DEDICATION

This is dedicated to my family, Sadasivam Krishnasamy, Amsavalli Karuppana Gounder and Madhuranthagi Sadasivam, who support and inspire me through every journey of my life.

ACKNOWLEDGEMENTS

I express my sincere gratefulness to Dr. Yezhou Yang for providing me the opportunity to work under his guidance. I admire Dr. Yang, he has been the inspiration for me since I joined ASU. His enthusiasm, guidance and motivation has made a big impact in shaping me as an individual and a student researcher. I'm confident that his mentoring would help me through my future endeavors. I extend my special thanks to Dr. Hasan Davulcu for his invaluable guidance and providing me the opportunity to work on the most exciting projects. I am honored to work with him, a person with such humbleness and wisdom.

Thanks to Dr. Chitta Baral for agreeing to be a presiding member of the thesis defense committee and for offering Topics in Natural Language Processing (CSE 576) course which played a crucial role in my research. I am grateful to Dr. Sadagpoan Narasimhan from Indian Institute of Information Technology Design and Manufacturing, Kancheepuram, India without whom none of this would have possible. I thank him for carefully guiding me through my initial days of research and showing me the path of research.

I am grateful to Mr. Karthik Manoharan for providing me the internship opportunity at PayPal and guiding me throughout my internship where my learning process was refined and helped me to move further with new approaches whenever I was stuck with my research. I thank Mr. Mike Nakamura for mentoring me in the best Software Engineering practices which enabled me to build the software models I have used throughout my research.

I feel blessed to be a part of both Active Perception Group (APG) and Cognitive Information Processing Systems (CIPS) Labs. A special thanks to Mert Ozer, Varun Chandra Jammula, Zhiyuan Fang, Tejas Gokhale, Shibin Zheng, Kausic Gunasekar, Mohammad Farhadi, Zhe Wang, Maryam Mousavi and Sanchit Pruthi for being wonderful colleagues.

I am indebted to my friends, Arun Prakash Sivakumar, Ashvant Ram Selvam, Goutham Maniarasu, Jitendra Prasad Arunachalam, Manikandan Pitchai, Mohana Kumar Poonga-

vanam, Raghavan Vellore Muneeswaran, Sai Shibi Ragupathy, Senthil Kumar Pandian, Ajay Gunasekaran and Dhruvi Desai who have been the emotional support and the reason to move forward every time I felt demoralized. I thank them all for their love and support.

TABLE OF CONTENTS

	Page
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER	
1 INTRODUCTION	1
1.1 Overview	1
1.2 Motivation	4
1.3 Challenges	5
1.4 Contribution	6
1.5 Outline	7
2 BACKGROUND	8
2.1 Meme Generation	8
2.2 Text Classification	9
2.3 Neural Machine Translation	10
2.4 Image Caption Generation	11
2.5 Controlled Text Generation	12
2.6 Transformer	12
3 DATASET	14
3.1 Meme Caption Dataset	14
3.2 Twitter Data	17
4 OUR APPROACH	18
4.1 Transformer Architecture	19
4.2 Meme Template Selection Module	21
4.3 Meme Caption Generator	23
4.3.1 Meme Template to Meme Caption	23

CHAPTER	Page
4.3.2 Sentence to Meme Caption	25
5 EXPERIMENTS AND RESULTS	29
5.1 Training Details	29
5.1.1 Meme Template Selection Module	29
5.1.2 Meme Caption Generator	30
5.2 Evaluation Metrics	32
5.2.1 Automated Metrics	32
5.2.2 Human Evaluation Metrics	32
5.3 Results and Analysis.....	33
5.3.1 Automated Evaluation	33
5.3.2 Human Evaluation Task Setup.....	34
5.3.3 Human Evaluation	34
5.4 Inter Rater Reliability.....	37
5.5 Controlling Meme Generation	38
6 TWITTER BOT	41
7 CONCLUSION AND FUTURE WORK.....	42
REFERENCES	44
APPENDIX	
A MEME BOT DEMO.....	47
B AMAZON MECHANICAL TURK QUESTIONNAIRE	49

LIST OF TABLES

Table	Page
3.1 Sample Examples (Template name, Captions and Meme Image) from the Meme Caption Dataset.	16
5.1 Dataset Statistics.	29
5.2 Meme Template Selection Performance on Meme Caption Test Dataset. Bold Font Highlights the Best Scores Obtained.....	30
5.3 Hyper-parameters Used in the Transformer Model.....	30
5.4 BLEU Scores for the Transformer Variants. Bold Font Highlights the Best Scores Obtained. Higher the Score Is Better.....	32
5.5 Human Evaluation Scores on Twitter Data. The Scores in the above Table Represent the Mean Scores of All Tweets. Relevance And Coherence are Scored on a Range of 1-4. User Likes Score Represents the Percentage of Total Raters Who Liked the Meme.	36

LIST OF FIGURES

Figure	Page
1.1 Memes Used by the Online Deep Learning Community on Social Media to Ridicule the State of the Art Pre-training Models.	3
1.2 An Illustrative Figure of memeBot, an End to End Neural and Probabilistic Architecture for Automatic Meme Generation. It Generates an Image Meme for a given Input Sentence by Combining the Selected Meme Template Image and the Generated Meme Caption.	5
3.1 Distribution of Meme Caption Count for the Scraped Meme Templates.	15
4.1 The Transformer (Vaswani <i>et al.</i> , 2017) - Model Architecture.	18
4.2 Attention Mechanisms in the Transformer (Vaswani <i>et al.</i> , 2017). (left) Scaled Dot-Product Attention. (right) Multi-Head Attention.	20
4.3 Meme Template Selection Module.	22
4.4 Meme Template to Meme Caption Generator - Model Architecture.	24
4.5 (left)Multi-Head attention. (right) Scaled Dot-Product Attention.	25
4.6 An Illustrative Figure of the Sentence to Meme Caption Generator Architecture.	26
4.7 Sentence to Meme Caption Generator - Model Architecture.	28
5.1 Memes Generated by Transformer Model Variants for the Given Input Sentence.	31
5.2 Memes Generated by the Transformer Variants for the Input Tweet - "Please save the world from Corona".	33
5.3 Human Evaluation Scores	35
5.4 User Likes Score Distribution	36
5.5 Qualitative Figure of Memes Grouped by Coherence Score.	37
5.6 Qualitative Figure of Memes Grouped by Relevance Score.	38
5.7 Inter Rater Reliability Scores for the Human Evaluation Metrics.	39

Figure	Page
5.8 Controlled Meme Generation. Memes Generated Using Different Meme Templates for a Given Input Sentence.	40
6.1 Twitter Bot - Illustrative Figure.	41
B.1 Sample AMT Questionnaire - Coherence Metric.	50
B.2 Sample AMT Questionnaire - Relevance and User Like Metric.	50
B.3 Sample AMT Questionnaire - Disclaimer.	51

Chapter 1

INTRODUCTION

1.1 Overview

A meme is defined as an element of a culture or system of behavior that may be considered to be passed from one individual to another by non genetic means, especially imitation. The memes are composed of ideas and beliefs. Memes have played an important role in the evolution of human life and are the successful behaviors that were and are being used to pass on the essential and important behavioral aspects of life. Understanding what causes a meme to replicate and how it replicates is an important aspect of this study.

An essential or successful meme is the one that has stayed for a long time and passed on or used by the majority of the population. Examples of such successful memes would be language, religion, culture, etc., The evaluation of successful meme or a useful meme depends upon the population evaluating it. An idea that is considered as essential by a set of people may not be considered important by another set of people. This leads to including the process of replication across population it replicates on as an integral part of the meme study.

The process of replication of a meme is an interesting phenomenon which depends upon the process of observation and learning. Since memes are composed of ideas, concepts and beliefs, they are very subjective and their meaning and understanding will vary based on the interpretation. A few memes are replicated in their original form and few memes are subject to variation during the process of replication I.e., the memes mutate during replication based on the interpretation. The memes mutate based on the needs and other diverse set of factors associated with the people spreads across.

Another important aspect of memes is their rate of transmission. The rate at which a meme replicates upon factors like its importance, how easy it is to replicate, people it spreads across but the most important factor is the medium it replicates on. The modern era has given birth to a new and powerful medium which is the internet. The internet facilitates the transmission of ideas and concepts instantaneously across the globe and has enabled us to share an idea and opinion by virtue of platforms like social media, blogs and open or closed messengers.

Over the internet, any idea is shared easily and is accessible to anyone. The internet not only has facilitated the spread of the traditional memes faster and easy, it has also given birth to a new type of memes called the internet memes. “An Internet meme, commonly known as just a meme is an activity, concept, catchphrase, or piece of media that spreads, often as mimicry or for humorous purposes, from person to person via the Internet” - Wikipedia¹. The connected and social nature of the internet allows the internet memes to propagate more rapidly.

An Internet meme could be anything from an image to video or a GIF. However it commonly takes the form of an image and is composed by combining a meme template (image) and a caption (a natural language sentence). The image typically comes from a set of popular image candidates. The caption conveys the intended idea or message through natural language, and most commonly in English.

Over the internet, information exists in the form of text, images, video, or a combination of these. Yet the information existing as a combination of image or video and short text often gets viral. Image memes are a combination of image, text, and humor, making them a powerful tool to deliver information and are used in viral marketing and mass advertising to propagate any ideas ranging from simple commercials to those that can cause changes and development in the social structures like countering hate speech.

¹https://en.wikipedia.org/wiki/Internet_meme



Figure 1.1: Memes Used by the Online Deep Learning Community on Social Media to Ridicule the State of the Art Pre-training Models.

The image memes are also important and popular because they portray the culture and social choices embraced by the internet community and they have a strong influence on the cultural norms of how a specific demographics of people operate. For example, in Figure 1.1, we present the memes used by an online deep learning community to ridicule how the new pre-training methods are outperforming the previous state of the art models.

The importance of image memes can be attributed to the fact that visual information is easier to process and understand when compared to reading large blocks of text, and this fact is evident in Figure 1.1. An organic and socially meaningful image meme has the potential to become viral through its powerful combination of visual and textual information, which has a strong influence in online social interaction to bring up topics, express ridicules, and sometimes create smearing effects, owing to the medium it strews upon.

The potential of memes to deliver any range of ideas and the key role played by them in shaping the popular culture of the internet community makes automatic meme generation an interesting research topic to delve into.

1.2 Motivation

By taking a deep look into the process of meme generation, we propose to co-relate meme generation to Natural Language Translation. To translate a sentence from a source to target language, one has to decode the meaning of the sentence in its entirety, analyze its meaning and then encode that meaning of the source sentence into the target sentence. Similarly, a sentence can be translated into a meme by encoding the meaning of the sentence into a pair of image and caption capable of conveying the same meaning or emotion as that of the sentence. Motivated by this intuition, we adopt an encoder-decoder model for our meme generation task.

Davison (2012) separates a meme into three components - **Manifestation, Behavior, and Ideal**. In an image meme, the **Ideal** is the idea that needs to be conveyed. The **Behavior** is to select a suitable meme template and caption to convey that idea and, the **Manifestation** is the final meme image with a caption conveying the idea. Wang and Wen (2015) and Peirson *et al.* (2018) focus on the behavior and manifestation of a meme, not much importance is given to the ideal of a meme. Oliveira *et al.* (2016) developed a pipeline to automatically generate image memes for news headlines. Their approach of meme generation is limited to a set of rules for selecting a meme template and using the words from the news to fill a predefined meme caption rule.

In this work, We intend to automatically generate a meme that can contextually represent a given input sentence (Ideal) as illustrated in Figure 1. The Behavior will be to select a meme template and generate a text caption to convey the Ideal and the Manifestation will be the generated meme that can represent the input sentence in an online social interaction, e.g. a twitter post. This is a challenging NLP task with immediate practical applications for online social interaction.

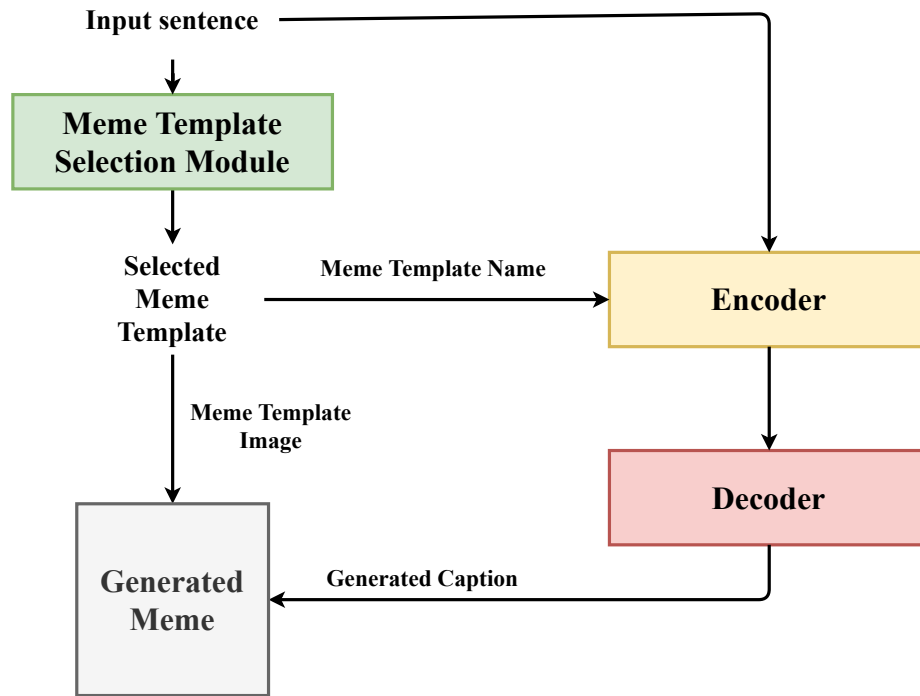


Figure 1.2: An Illustrative Figure of memeBot, an End to End Neural and Probabilistic Architecture for Automatic Meme Generation. It Generates an Image Meme for a given Input Sentence by Combining the Selected Meme Template Image and the Generated Meme Caption.

1.3 Challenges

The images used in the image memes are usually associated with ideas or concepts from some popular pop culture, TV shows or movies references. Internet users when they come across a similar concept, they use the image and add their own concept through the caption. By this means, the concept behind the meme template (meme image) mutates. The newly created meme replicates when introduced on the internet and shared among the internet users through social media or open messengers.

Meme generation requires the ability to semantically and contextually understand the input sentence along with the contextual knowledge of the image memes. Even with the

understanding of the input sentence and meme images, one has to possess a good fluency in natural language to generate a meme caption that is compatible with the meme image. The generated meme should also be relevant to the input sentence.

Meme generation is a complex task for human beings and it demands exposure to the multiple domains associated with meme and, good semantical and contextual exposure. In our work, we aim to learn the ideas or concepts associated with the meme templates by just using their language features.

1.4 Contribution

Following the insights from biological meme generation pipeline, intuition behind natural language translation and the recent success in various encoder-decoder architectures motivated us to use model an end to end neural and probabilistic approach to generate an image-based meme for any given sentence using an encoder-decoder architecture. The proposed meme generation pipeline can be scaled to any input sentence and can be used in real world settings. We summarize our contributions as follows:

- We analyze and present the understanding of **Ideal**, **Behaviour** and **Manifestation** of a meme in creating a successful meme generation pipeline and compiled the first large-scale Meme Caption dataset.
- We present an end to end encoder-decoder architecture to generate a meme for any given sentence. We select a compatible meme template for the given sentence and generate a meme caption that can represent the input sentence through the selected template.
- We design experiments based on human evaluation and provide a thorough analysis of the experimental results on using the generated memes for online social interaction.

1.5 Outline

In **Chapter 2**, we give a concise historical review on prior research on automatic meme generation and works related to image meme generation.

In **Chapter 3**, we present the first large scale meme caption dataset.

In **Chapter 4**, we present the detailed architecture of memeBot, our end-end probabilistic and neural model for automatic image meme generation.

In **Chapter 5**, We design experiments based on human evaluation and provide a thorough analysis of the experimental results on using the generated memes for online social interaction.

In **Chapter 6**, We present the Twitter bot architecture.

In **Chapter 7**, we end this work with conclusion and give a short discussion about the future work.

Chapter 2

BACKGROUND

2.1 Meme Generation

There are only a few studies on automatic meme generation and the existing approaches treat meme generation as a caption selection or caption generation problem.

Wang and Wen (2015) combined an image and its text description to select a meme caption from a corpus using a ranking algorithm. In their work, they use an empirical cumulative density function to map both image and textual features to a compatible embedding space to estimate the pair-wise correlation between the text and image features. They train a nonparanormal model for ranking meme descriptions for an image and during inference, they use a meme generation pipeline to select a text description for a given image using the trained model. The meme generation pipeline of Wang and Wen (2015) is limited to selecting a text description from a corpus and they do not generate a meme caption for an image or an input sentence.

Peirson *et al.* (2018) extends Natural Language Description Generation to generating a meme caption using an encoder-decoder model with an attention (Luong *et al.*, 2015b) mechanism. The authors of this work use a Long Short Term Memory network, a variant of the Recurrent Neural Network in their meme generation pipeline. In their meme generation pipeline, the authors encode the visual and textual features of a meme template into a latent space using an LSTM network with attention mechanism and use an LSTM decoder to generate meme captions from the encoded meme template representation.

The meme generation pipeline from Peirson *et al.* (2018) generates a meme with a **Manifestation** and **Behaviour** as defined out by Davison (2012) but the scope of their

meme generation pipeline is limited in the aspect of generating memes inspired from an **ideal** because of the fixed number of meme template images and descriptions.

Although there is not much work on automatic meme generation, the task of meme generation can be closely aligned with tasks like Sentiment Analysis, Linguistic Acceptability Neural Machine Translation, Image Caption Generation and Controlled Natural Language Generation.

2.2 Text Classification

In the domain of Natural Language Understanding (NLU), researchers have explored classifying a sentence based on their sentiment (Socher *et al.*, 2013) and their linguistic acceptability (Warstadt *et al.*, 2018).

Socher *et al.* (2013) aims at understanding and analyzing the compositionality of textual features in determining the sentiment of a sentence. They explore the capabilities of neural networks to analyze the intricacies of sentiment and to capture complex linguistic phenomena using a Recurrent Neural Network. They encode a given sentence into a fixed size representation and classify the sentence into five sentiment classes.

(Warstadt *et al.*, 2018) aims at investigating the ability of neural networks to judge the grammatical acceptability of a sentence with the goal of testing their linguistic acceptance. Acceptability judgements are the primary behavioral measure and still a hard task for human beings. The authors of the paper use a bidirectional LSTM to encode the input sentence into a latent space and use a classifier to classify the grammatical acceptability of the sentence. The authors put out the capabilities of the neural networks in encoding the textual features of the sentence to classify its grammatical acceptance.

Given the success of encoding the textual features of a sentence into a latent space to perform complex Natural Language Understanding tasks, we aim to explore the capability of neural networks in predicting the correct meme template for a given sentence by leveraging

the fact that in a image meme, the meme caption sits on the meme template because the meme template is capable of sharing the same idea or concept as that of the sentence.

2.3 Neural Machine Translation

The idea of creating an encoded representation and decoding it into a desired target is well establish in Neural Machine Translation. Sutskever *et al.* (2014), Bahdanau *et al.* (2014a) use an encoder-decoder model to encode and decode a sentence from a source to a targeted language.

Sutskever *et al.* (2014) presents a general end to end approach using a multi layered Long Short Term Memory (LSTM) network to encode the input sentence into a vector of fixed dimensionality and used a decoded LSTM to decode the latent vector to the target sentence. They experiment this architecture on English to French translation task. In this work, the authors claim that their model has learned the sensible phrases and sentence representations that are sensitive to word order and is relatively invariant to the active and passive voice. This work establishes the success of using a straightforward end to end encoder-decoder models on complex Natural Language Understanding tasks.

Bahdanau *et al.* (2014a) use a similar end to end encoder-decoder approach using multi layered LSTM. But this work exposes the limitation of encoding the entire sentence into a single vector of fixed size representation in the latent space. Compressing the long sentences into a fixed size representation is difficult and they introduce attention mechanism to align and translate the input sentence to the target sentence using a straight forward encoder-decoder model.

The important factor of the attention approach is that to creates a sequence of vectors during encoding and chooses a subset of these vectors during decoding thus freeing the model to compress all the information into a single vector of fixed size. This enables the model to handle long sentences and capture the intricacies of the sentence during the

translation task. Our proposed model of meme generation shares similar spirit with the above mentioned problems where we encode the given input sentence into a latent space followed by decoding it into a meme caption that can be combined with the meme image to convey the same meaning as that of the input sentence.

2.4 Image Caption Generation

In the domain of computer vision and natural language, Vinyals *et al.* (2015), Xu *et al.* (2015) and Karpathy and Fei-Fei (2015) address the image captioning task by encoding the visual features of an image and decoding the natural language description of the image.

Vinyals *et al.* (2015) uses a Convolution Neural Network to encode the visual features of the image into a fixed size vector and generate the natural language description from the encoded embedding using an LSTM decoder. They use a straight forward end to end neural and probabilistic model to generate the description of the image by maximizing the likelihood of generating the caption given the image. Xu *et al.* (2015) introduces two attention-based image caption generators for image captioning. A soft deterministic attention mechanism using standard back-propagation and a hard stochastic attention mechanism which can be trained by maximizing an approximate variational lower bound.

Karpathy and Fei-Fei (2015) uses a R-CNN to extract the visual features to encode the visual information and a bidirectional LSTM with attention to decode the natural language description. Their work captures the inter-modal correspondences between language and visual data. Given the success of encoding visual and textual information into a fixed or a set of variable size vectors and decoding them them into a natural language, we aim to encode the input sentence and the selected meme template into a meme embedding and then decode the meme embedding into the meme caption. However, the generated meme caption should represent the input sentence through the selected meme template, making it a conditioned or controlled text generation task.

2.5 Controlled Text Generation

Meme generation inspired from an idea or concept involves constraint on the generated sentences. The constraints could be a hard constraint such as inclusion of keywords and soft constraints like the generated caption should be compatible with the selected meme template, follow the semantics of the meme template, represent the context of the input sentence. Su *et al.* (2018) and Miao *et al.* (2019) generate a sentence with desired emotions or keywords using sampling techniques.

Su *et al.* (2018) and Miao *et al.* (2019) use Gibbs and Metropolis-Hastings sampling techniques respectively to iteratively generate a sentence with needed constraints from a given set of keywords by sampling new words from the vocabulary at every iteration. The candidate sentences are revised and updated iteratively, where the sampled new words replace the old ones. They use proposal, replacement, insertion and deletion strategies to add, remove or replace a word at every iteration.

Controlled Natural Language Generation with desired emotions, semantics and keywords have been studied previously. Huang *et al.* (2018) generate text with desired emotions by embedding emotion representations into Seq2Seq models. Hu *et al.* (2017) concatenate a control vector to the latent space of their model to generate text with designated semantics. Keskar *et al.* (2019) uses a control sequence to generate text which satisfies the control property.

2.6 Transformer

Recurrent neural networks, long short-term memory (LSTM) Hochreiter and Schmidhuber (1997) and gated recurrent neural networks (GRU) Chung *et al.* (2014) were the popular encoder decoder architectures used for the prominent natural language tasks like sequence modelling, language modelling and machine translation. The models built using recurrent

networks for an encoder-decoder architecture have to traverse sequentially over the input and output sequences and are not capable of capturing the long-term dependencies among the sequences. Attention mechanisms (Luong *et al.* (2015a), Bahdanau *et al.* (2014b)) are used along with sequence models to capture the dependencies among the sequences through a context vector. The recurrent networks with attention have produced promising results but it impossible to parallelize a sequential model and processing long sequences requires huge memory and computational power. To overcome this problem, the authors of Vaswani *et al.* (2017) propose a new architecture called the Transformer which relies entirely upon the attention mechanism which can capture the global dependencies between the input and output through an encoder-decoder architecture and this architecture enables parallelization of computation in the network.

Chapter 3

DATASET

3.1 Meme Caption Dataset

The main contribution of our work is that we have compiled the first large scale meme caption dataset. Analysing and understanding the **Ideal**, **Beviour** and **Manifestation** of a meme as defined by Davison (2012) is essential to create a successful meme generation pipeline. In an image meme, the ideal is the idea or concept that needs to be conveyed through the meme. The behavior is to choose the right meme template that is capable of conveying the idea or concept in the intended manner and coming up with a text caption that can convey the intended idea through the selected meme template. The manifestation is the final meme created by combining the text caption and the image that conveys the intended idea.

To study these aspects of a meme and analyze how the image memes are being used by the internet community, we compile a meme corpus dataset which contains the meme templates, meme images and the meme captions as shown in Table 3.1. To create this meme corpus, we adopt the open online resource [imgflip](https://imgflip.com)¹ which is one of the most commonly used meme generators to add text captions to the meme templates that are well established on the internet. To automatically crawl the data, we developed a web crawler and scraped 808 meme templates with total of 748,571 meme captions. In Figure 3.1, we present the distribution of the meme captions across the scrapped templates to show how the overall set of collected meme templates are used by the meme generators.

¹<https://imgflip.com>

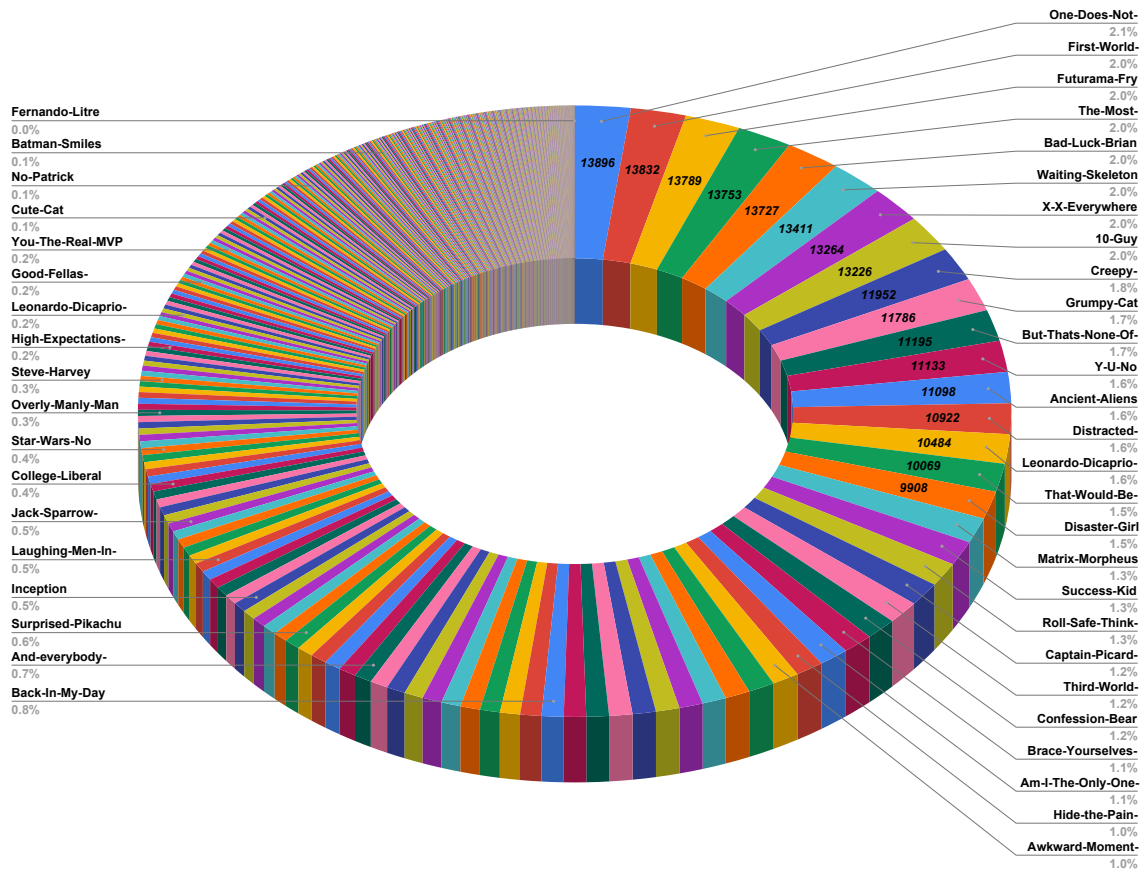


Figure 3.1: Distribution of Meme Caption Count for the Scraped Meme Templates.

From Figure 3.1, it can be seen that only a small set of templates are repeatedly used. We investigate this skewed distribution along with factors that can make a meme popular or viral. Replication of a meme depends on the mental processes of observation and learning of the group of people across which it is being shared Davison (2012). Popular meme templates are those that make a content shareable and viral over the internet. The skewed meme caption distribution illustrates that the meme template images with high caption counts are replicated often because of their capability to a make content viral or shareable over the internet. To this end, we experiment on automatic meme generation using the popular meme templates. We preprocess the text and remove captions with less than 5 words. We then

filter only the templates with $\#meme_captions > 5000$ and finally select 177,942 meme captions from 20 templates selected based on the highest caption count per template. Our dataset consists of meme template (image & template name) and meme caption pairs. A sample from the dataset is illustrated in Table 3.1.


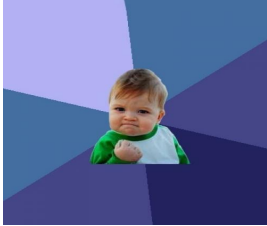

Template Name	Captions	Template Image
Leonardo Dicaprio Cheers	<ul style="list-style-type: none"> ● to those who have been fortunate enough to have known true love ● cheers to us making the future bright 	
Success Kid	<ul style="list-style-type: none"> ● carries the laundry didn't drop a single sock ● when you win your first fortnite game 	
Bad Luck Brian	<ul style="list-style-type: none"> ● Escapes the burning building gets hit by the firetruck ● Spends all night sleeps through exam 	

Table 3.1: Sample Examples (Template name, Captions and Meme Image) from the Meme Caption Dataset.

3.2 Twitter Data

We use the most retweeted tweets from Twitter to evaluate the efficacy of our model. Tweets with high retweet count are highly shared and are potential candidates to become viral. We randomly sampled 1000 tweets using a threshold of > 5000 retweets. The goal is to prompt our model to generate a meme by inputting a tweet and evaluate if the generated meme is relevant to the tweet.

Chapter 4

OUR APPROACH

In this section, we describe our approach: an end-to-end neural and probabilistic architecture for meme generation. Our model has two components. First, a meme template selection module to identify a compatible meme template (image) for the input sentence. Second, a meme caption generator inspired from the popular encoder-decoder transformer model of (Vaswani *et al.*, 2017) which is presented in Figure 4.1. We combine the selected meme image and the generated caption to create a meme.

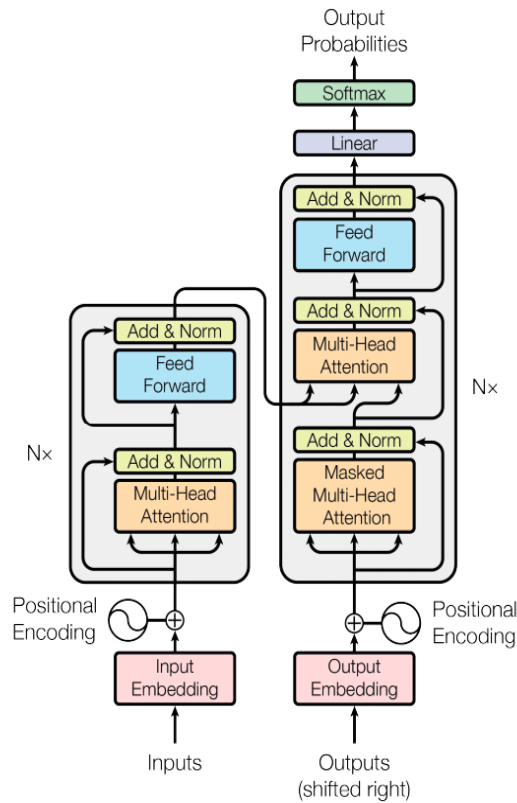


Figure 4.1: The Transformer (Vaswani *et al.*, 2017) - Model Architecture.

4.1 Transformer Architecture

We provide a brief overview of the transformer (Vaswani *et al.*, 2017) architecture since our meme generation pipeline has model architectures inspired from the original work. A transformer is composed of four essential components which are named as Scaled Dot-Product Attention, Multi-Head Attention, Position-wise Feed-Forward Networks and positional embedding and, are briefly explained below as a summary from the original work.

Scaled Dot-Product Attention

The scaled dot-product attention function is described as mapping a query and a set of key-value pairs to an output. The inputs to the scaled dot-product attention are the query (Q), Keys (K) and Values (V) vectors. The scaled dot-product attention is given by the formulation described in Equation (4.1) and is illustrated in Figure 4.2 (left).

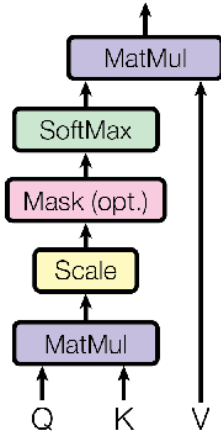
$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4.1)$$

Multi-Head Attention

The multi-head attention is composed of multiple scaled dot-product attention running in parallel over the inputs. The multi-head attention is illustrated in Figure 4.2 (right) and the outputs of the multiple heads are concatenated using the formulation described in Equation (4.2) before being passed on to a feed-forward layer.

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \dots, head_h)W^O, \\ \text{where } head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (4.2)$$

Scaled Dot-Product Attention



Multi-Head Attention

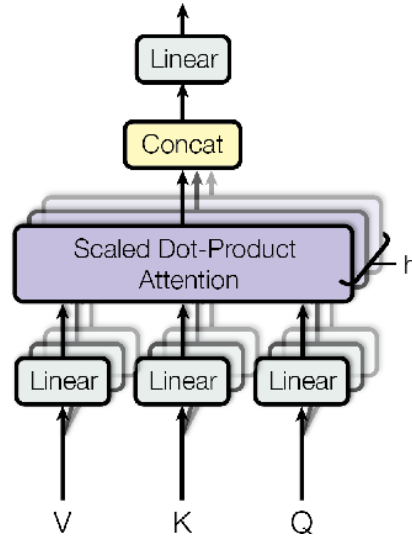


Figure 4.2: Attention Mechanisms in the Transformer (Vaswani *et al.*, 2017). (left) Scaled Dot-Product Attention. (right) Multi-Head Attention.

Position-wise Feed-Forward Networks

The position wise feed forward network is a two layered linear network with a ReLU activation between the layers.

Positional Embedding

The positional information of the words in a sentence is important in a natural language task. In the transformer architecture, the position of the words is given by position embedding which is given by the formulation described in the equation below,

$$\begin{aligned}
 PE(pos, 2i) &= \sin(pos/10000^{2i/d_{model}}) \\
 PE(pos, 2i + 1) &= \cos(pos/10000^{2i/d_{model}})
 \end{aligned}
 \tag{4.3}$$

where d_{model} represents the embedding representation size of the input, pos is the position of the word in the sequence and i is the dimension in the embedding space.

In the transformer model, the encoder is composed of N identical layers where each layer contains a multi-head attention followed by a feed forward network. The decoder is also composed of N identical layers similar to the encoder but each layer contains an additional multi-head attention over the output of the encoder. The learned embedding from the decoder are used to convert the input tokens to the output tokens. A linear layer with softmax activation is used on top of the decoder output to predict the words in the output sequence.

4.2 Meme Template Selection Module

We aim to learn the ideas or concepts associated with the meme templates by mapping the ideas and concepts present in the sentence to the meme template with similar ideas or concepts associated with it. We achieve this mapping using a trained meme template selection module as illustrated in Figure 1.2.

Pre-trained language representations from transformer based architectures BERT (Devlin *et al.*, 2019), XLNet (Yang *et al.*, 2019) and Roberta (Liu *et al.*, 2019) are being used in a wide range of natural language Understanding tasks. The intuition behind these pre-trained models is to learn the language representations from a big language corpus by training the models on tasks like language modelling, masked language modelling, next sentence prediction, etc.,. During pre-training, the models learn language representations capable of representing the intricacies and composition of sentences and how the language is constructed.

The pre-trained language representations are used in variety of different tasks and the architectures utilizing these representations have performed to near human level on many Natural Language tasks and have pushed the state of the art performance to new heights. Devlin *et al.* (2019), Yang *et al.* (2019) and Liu *et al.* (2019) show that these models can be fine-tuned specifically to a range of NLU tasks to create state-of-the-art models.

For our template selection module, we use a linear neural network on top of the pre-trained language representation models. In training, the probability of selecting the correct template for a given sentence is maximized by using the formulation given in Equation (4.4) and our meme template selection module is illustrate in Figure 4.3.

$$l(\theta_1) = \arg \max_{\theta_1} \sum_{(T,S)} \log(P(T|S, \theta_1)), \quad (4.4)$$

where θ_1 denotes the parameters of the meme template selection module, T is the template and S refers to the input sentence.

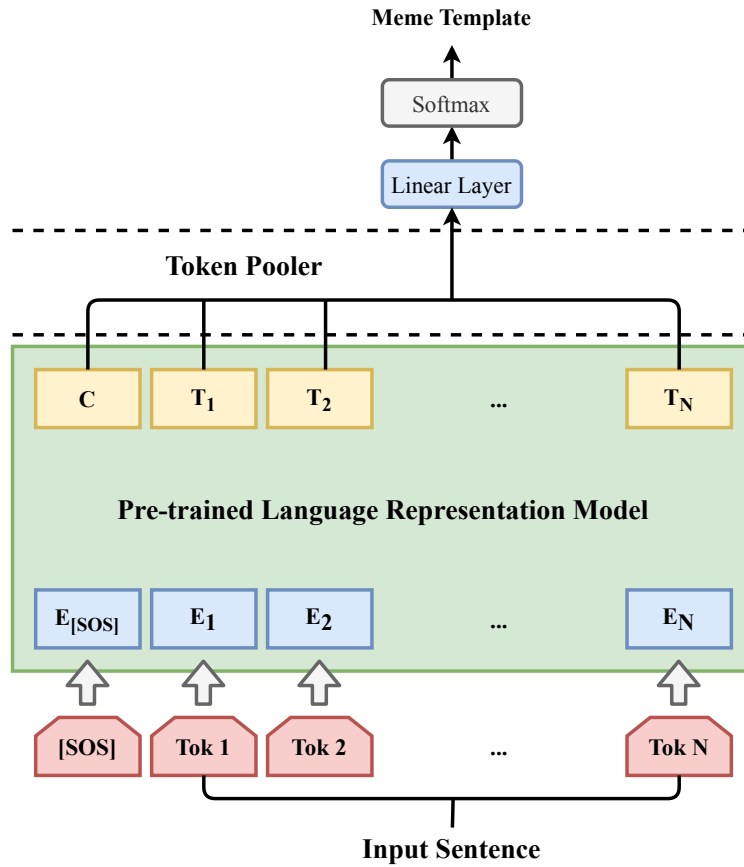


Figure 4.3: Meme Template Selection Module.

Pre-trained language representation models (BERT (Devlin *et al.*, 2019), XLNet (Yang *et al.*, 2019) and Roberta (Liu *et al.*, 2019)) provide different pooling strategies on the

encoded tokens from the final layer of the transformer models to be used for the classification task. The most popular strategies are to use the start of the string token encoding or to use the mean/sum of all the vectors representing each token from of the sentence sentence.

4.3 Meme Caption Generator

A transformer (Vaswani *et al.*, 2017) architecture relies entirely upon the attention mechanism and is capable of drawing the global dependencies between the input and output owing to its stacked Scaled Dot-Product and Multi-Head attention mechanisms. It also enables parallelization of computation in the network to speed up the training process. To this end and given the success of transformers in wide variety of language modelling and translation tasks, we adopt a transformer architecture inspired from the original work for meme caption generation in our meme generation pipeline.

4.3.1 Meme Template to Meme Caption

We design a baseline transformer model for generating a meme caption for the selected meme template. The goal of this architecture is to learn the dependency between a meme template and its meme captions. This architecture is a conditional language model and during training the transformer is optimized by maximizing probability of generating the meme caption token at step i given the meme template and the meme caption tokens generated before step i . We optimize the transformer by using the formulation given in Equation (4.5) and the model architecture is illustrated in the Figure 4.4.

$$l(\theta) = \arg \max_{\theta} \sum_{(C)} \log(P(C|T, \theta)), \quad (4.5)$$

where θ denotes the parameters of the transformer, C is the meme caption and T is the meme template.

We enforce the conditional caption generation by leveraging the scaled-dot product

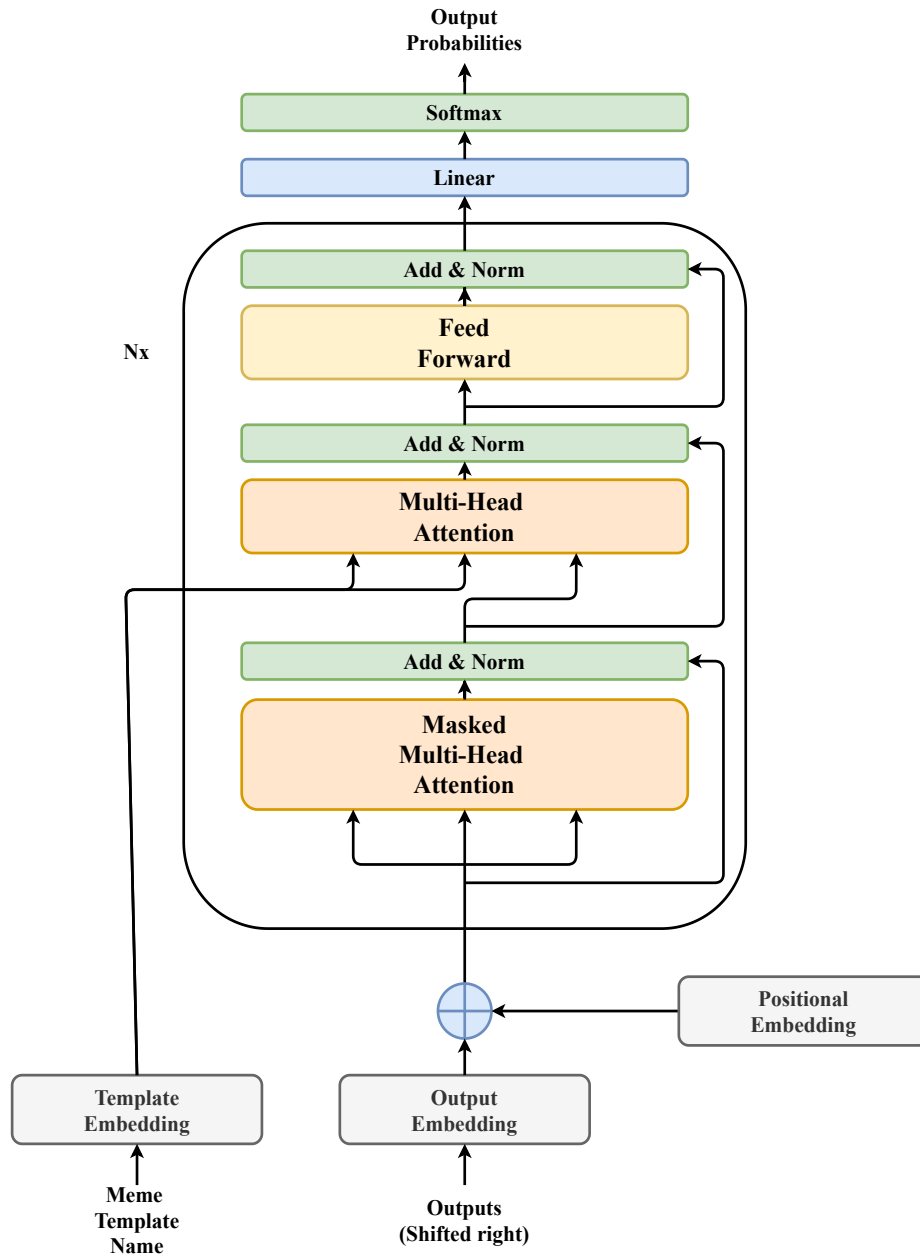


Figure 4.4: Meme Template to Meme Caption Generator - Model Architecture.

and multi-head attention mechanisms. We initially perform a masked multi-head attention between the expected captions and then perform a multi-head scaled dot-product attention between the template embedding and the output of the masked multi-head attention as illustrated in Figure 4.4 and Figure 4.5.

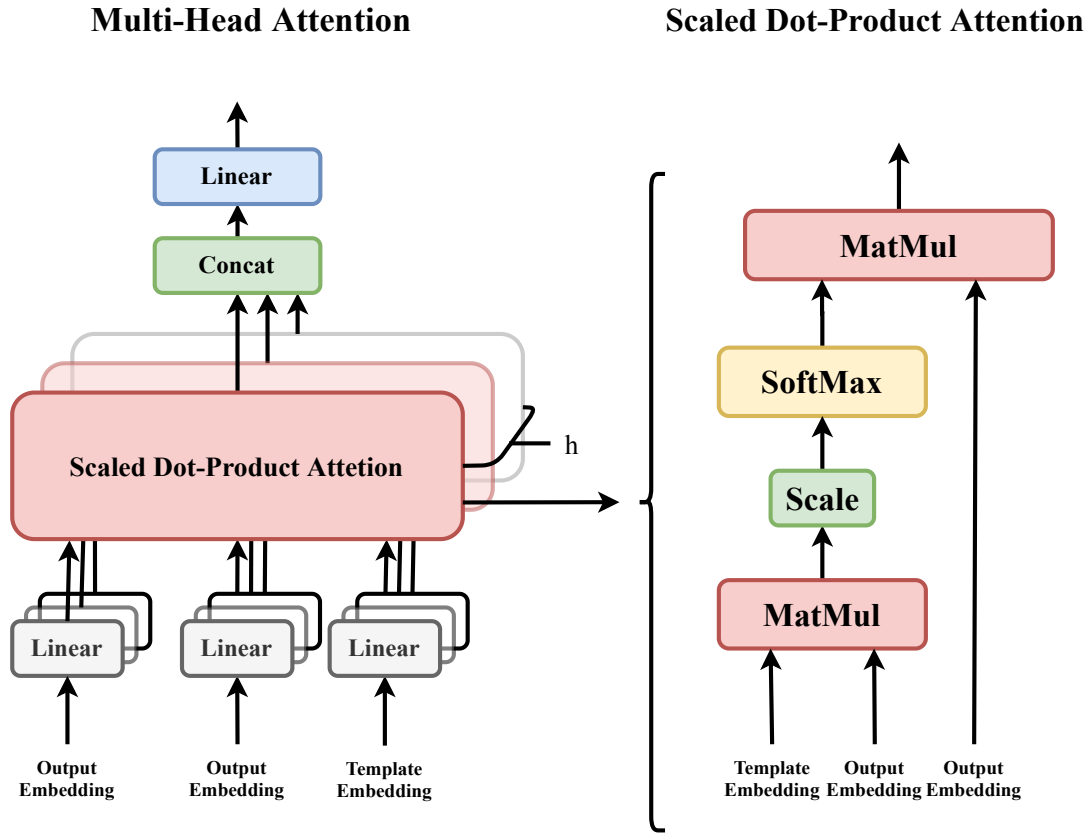


Figure 4.5: (left)Multi-Head attention. (right) Scaled Dot-Product Attention.

4.3.2 Sentence to Meme Caption

This transformer architecture is designed to generate a meme caption for the given input sentence and the selected meme template. The goal of this architecture is to learn the dependencies between the selected meme template and the output meme caption for a given input sentence and this dependency can be formulated as $P(C|S, T)$, where C is the meme caption, S is the input sentence and T is the meme template.

We train our transformer by corrupting the input caption, borrowing from denoising autoencoder (Vincent *et al.*, 2008). We extract the parts of speech of the input caption using a Part-Of-Speech Tagger (POS Tagger) (Honnibal and Montani, 2017). Using the POS vector, we mask the input caption such that only the noun phrases and verbs are passed as input

to the transformer. We corrupt the data to facilitate our model to learn meme generation from existing captions and to generalize the process of meme generation for any given input sentences during inference.

For a given sentence, we get the meme template from the meme template selection module. We extract the keywords from the input sentence using the POS Tagger mentioned above. We create a meme embedding for the sentence using the selected meme template and the extracted keywords using a transformer encoder. We use our decoder to decode the meme captions from the encoded meme embeddings. An illustrative figure of the model architecture is presented in Figure 4.6.

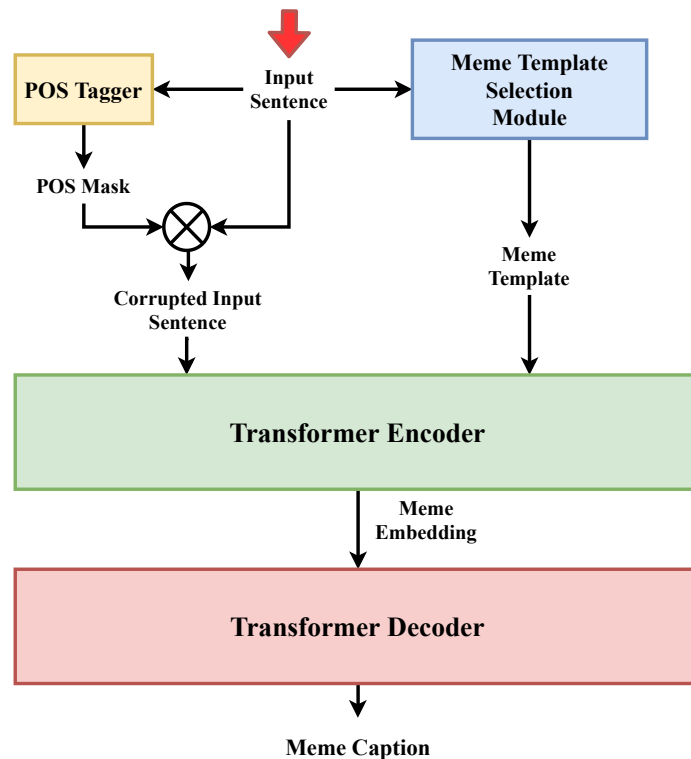


Figure 4.6: An Illustrative Figure of the Sentence to Meme Caption Generator Architecture.

The transformer encoder inputs the meme template embedding and the input keywords embedding. The encoder performs multi-head scaled dot product attention between the input embedding and the meme template embedding to create the meme embedding for the given input keywords. The transformer decoder utilizes the masked multi-head attention between the expected output captions and multi-head scaled dot-product attention between the meme embeddings from the encoder and the output of the masked multi-head embedding as shown in Figure 4.7.

Our model architecture to capture the conditional dependencies between the input keywords, selected meme template and the output caption where the generated meme captions are conditioned on the meme embedding which in turn is conditioned on the selected meme template and the input keywords.

During training, the model parameters are optimized by maximizing the formulation given below,

$$l(\theta) = \arg \max_{\theta} \sum_{(S,C)} \log(P(C|M, \theta)), \quad (4.6)$$

where θ denotes the parameters of the transformer, S is the sentence and M is the meme embedding.

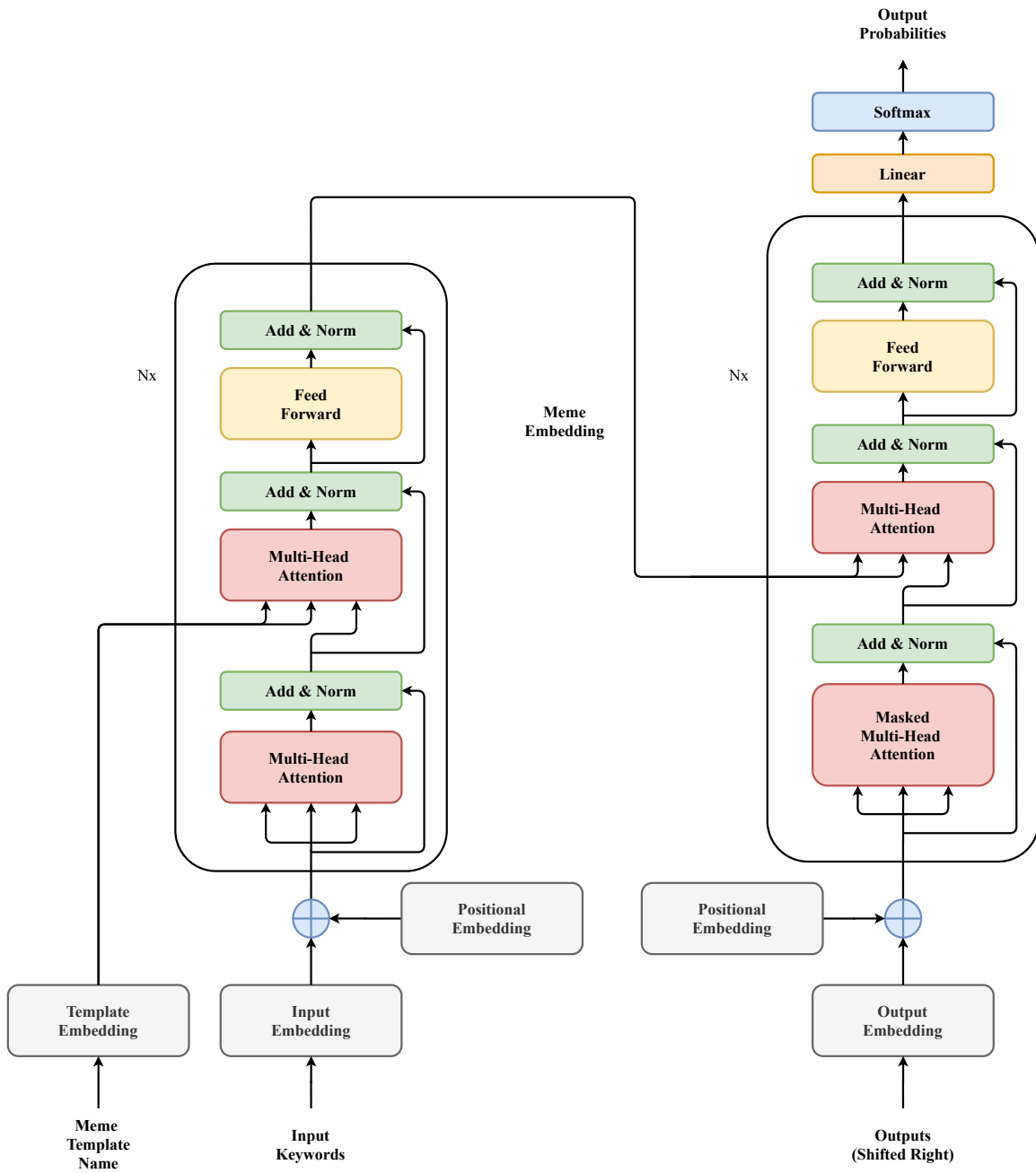


Figure 4.7: Sentence to Meme Caption Generator - Model Architecture.

EXPERIMENTS AND RESULTS

We train our model on the Meme Caption dataset (Chapter 3.1) to learn the dependency between a meme caption and a meme template. The statistic of the dataset used in our experiments is reported in Table 5.1. We evaluate the performance of our meme template selection module in selecting the compatible template and effectiveness of the transformer in generating captions that are similar to the input captions from the meme caption dataset using a set of automated metrics. We evaluate the efficacy of our model in generating memes for real world examples (Tweets) through human evaluation.

Train	Validation	Test
142341	17802	17799

Table 5.1: Dataset Statistics.

5.1 Training Details

5.1.1 Meme Template Selection Module

We use a single layer linear neural network with 768 units on top of pre-trained language representation models (BERT_{base} (Devlin *et al.*, 2019), XLNet_{base} (Yang *et al.*, 2019) and Roberta_{base} (Liu *et al.*, 2019)) as our meme template selection module. Performance of the meme template selection module on the meme caption test data using variants of pre-trained language representation models is reported in Table 5.2.

We adopt the best-performing and pre-trained Roberta_{base} model for the meme template selection module in the meme generation pipeline.

Model	Accuracy	Precision	Recall	F_1 score
BERT _{base}	0.7106	0.7190	0.7106	0.7137
XLNet _{base}	0.7099	0.7217	0.7099	0.7138
Roberta _{base}	0.7183	0.7280	0.7183	0.7210

Table 5.2: Meme Template Selection Performance on Meme Caption Test Dataset. Bold Font Highlights the Best Scores Obtained.

5.1.2 Meme Caption Generator

Both of our transformer based meme caption generator architectures follow the same denotations as of Vaswani *et al.* (2017). We report the hyper-parameters used in Table 5.3.

Architecture	N	d_{model}	d_{ff}	h	d_k	d_v
Meme Template to Meme caption	8	768	2048	12	64	64
Sentence to Meme Caption	6	512	2048	8	64	64

Table 5.3: Hyper-parameters Used in the Transformer Model.

We use residual dropout (P_{drop}) (Srivastava *et al.*, 2014) for regularization and Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 1e^{-9}$ and a scheduler using cosine annealing with warm restarts (Loshchilov and Hutter, 2016). We train for a maximum epoch of 250 iterations and the training process is guided by perplexity cost of the validation data.

During inference we generate meme captions using Beam search with a beam of size 6 and length penalty $\alpha = 0.7$. We stop the caption generation when a special end token or the maximum length of 32 words is reached.

Sample of memes generated by the Template to Caption architecture is presented in Figure 5.1.

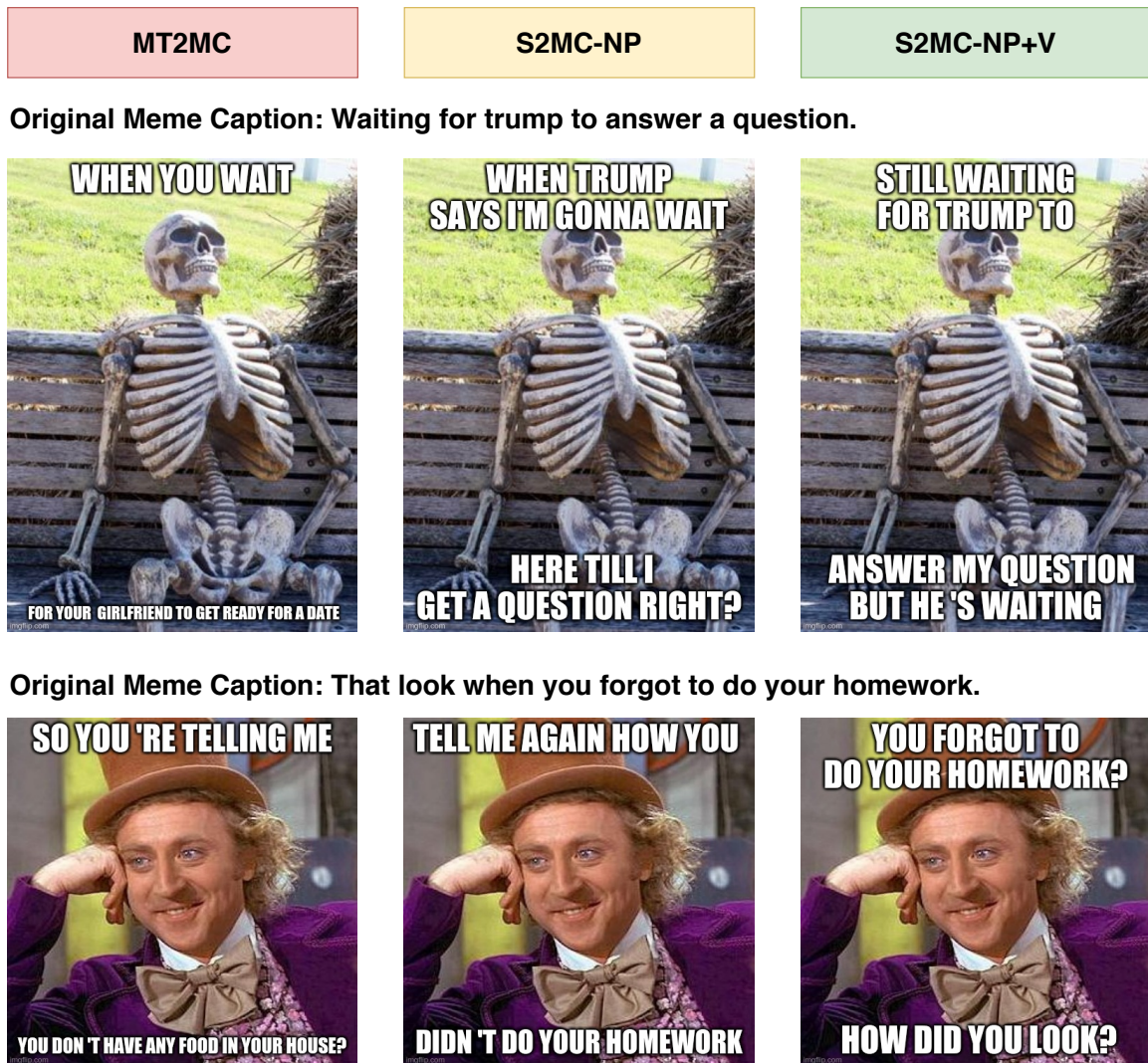


Figure 5.1: Memes Generated by Transformer Model Variants for the Given Input Sentence.

For sentence to meme caption architecture, we experiment with two different variants of the transformer. The first variant uses only the noun phrases from the input sentence while the second variant uses the verbs along with the noun phrases. We experiment only using the noun phrases in order to study to what extent the addition of verbs directs the context of the generated meme towards the context of the input sentence. A sample of memes generated

by the two transformer variants for a given set of input captions is presented in Figure 5.1 and it can be seen that the memes generated using noun phrases & verbs as inputs better represent the input caption.

5.2 Evaluation Metrics

5.2.1 Automated Metrics

We use BLEU score (Papineni *et al.*, 2002) to evaluate the quality of the generated captions.

Variant	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑
Base Model	13.93	6.85.86	4.45	2.72
Noun Phrases	38.71	24.67	14.56	8.89
Noun Phrases + Verbs	45.67	27.56	17.14	11.12

Table 5.4: BLEU Scores for the Transformer Variants. Bold Font Highlights the Best Scores Obtained. Higher the Score Is Better.

5.2.2 Human Evaluation Metrics

The perspective of good quality of a meme is subjective and varies among people. To the best of our knowledge, there are no known automatic evaluation metrics to evaluate the quality of a meme. A fairly reliable technique is to perform human evaluation by a set of raters to evaluate the quality of a meme on a subjective score.

In Machine Translation, adequacy and fluency (Snover *et al.*, 2009) are used to subjectively rate the correctness and fluency of a translation. Inspired from adequacy and fluency, we define 2 metrics - **Coherence and Relevance** to evaluate the generated memes, described as follows:

- **Coherence:** Can you understand the message conveyed through the meme (Image + text)?
- **Relevance:** Is the meme contextually relevant to the text?

The Relevance and Coherence metrics are scored on a range of 1 - 4. Coherence score captures the quality (fluency) of the generated meme and, Relevance score capture how well the generated meme represents the input sentence (correctness). We also ask the rater’s if they like the meme to evaluate if the meme is considered good by them.

To score these metrics, we set up an Amazon Mechanical Turk (AMT) experiment. Sample memes generated by our model for a given input tweet is presented in the Figure 5.2.



Figure 5.2: Memes Generated by the Transformer Variants for the Input Tweet - "Please save the world from Corona".

5.3 Results and Analysis

5.3.1 Automated Evaluation

We use BLEU (Papineni *et al.*, 2002) to evaluate the quality of the generated captions. The scores for the three transformer variants is reported in Table 5.4 and it can be observed

that the transformer variant which inputs verbs along with noun phrases has better score, and using verbs with noun phrases enables the transformer to generate relatively similar captions to that of the input captions when compared to the variant which inputs only the noun phrases.

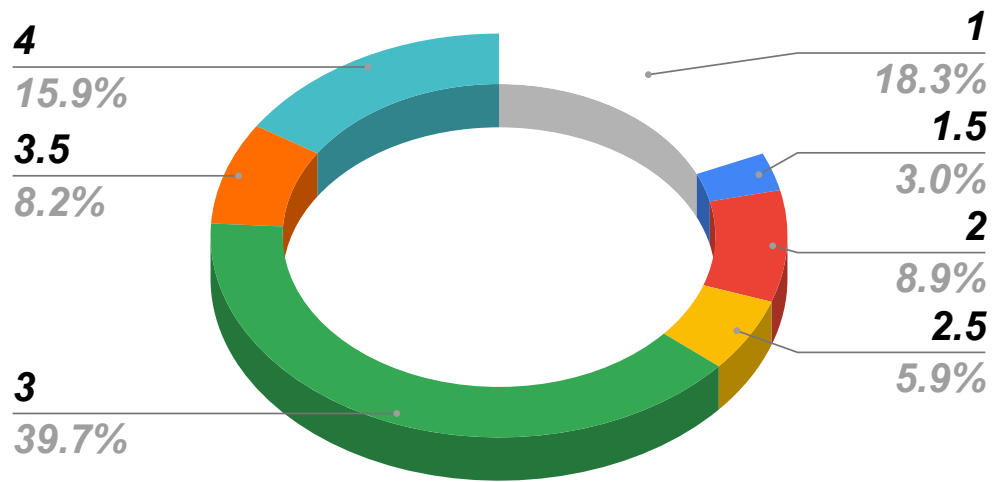
5.3.2 Human Evaluation Task Setup

We choose Amazon Mechanical Turk(AMT) for the evaluation of the generated memes due to its easy to use platform and the ready availability of a big worker pool with required skills. In our AMT evaluation setup, we design a two-stage process to evaluate the meme. An AMT example is in presented in Appendix B.

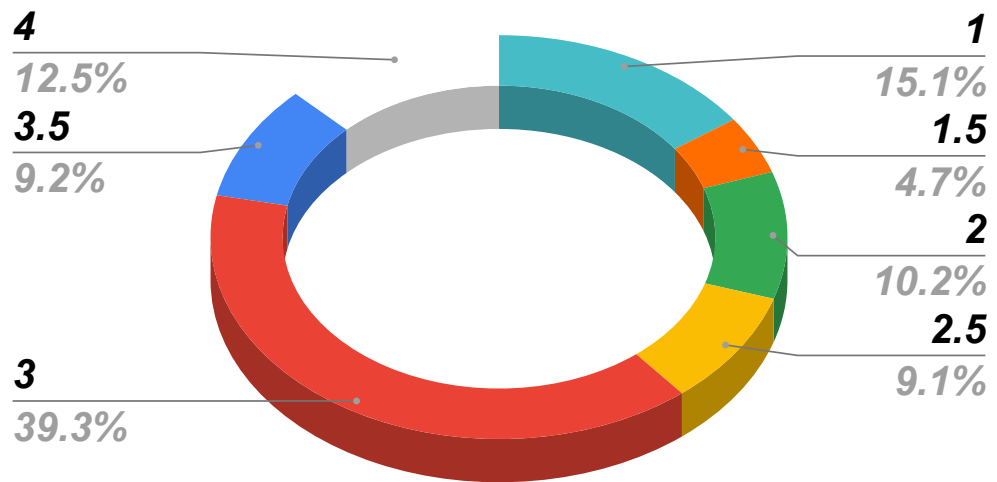
We first display the meme image and ask the workers to score the Coherence metric, only based on their understanding of the meme. Later we display the tweet and ask them to understand the text, and then ask them to score Relevance metric based on their comprehension of the tweet and the meme. Our expectation for the AMT workers merely is that they are capable of visually understanding an image, capable of semantically and contextually understanding a sentence and possess the reasoning ability to compare context from different information sources and we assume an adult human being is well qualified to meet our expectations. Each sample was rated by 2 workers. In case of disagreement among the raters, we consider their average score as the final score.

5.3.3 Human Evaluation

The model performance of the our model on the human evaluation metrics is reported in Table 5.5. The score distribution across the metrics Coherence and Relevance is presented in the Figure 5.3. User Likes score distrubution is presented in Figure 5.4. A qualitative comparison of memes grouped by rater scores across the two metrics in presented in Figures 5.5 and 5.6.



(a) Coherence Score Distribution



(b) Relevance Score Distribution

Figure 5.3: Human Evaluation Scores

Metric	Score
Coherence	2.66
Relevance	2.65
User Likes	0.65

Table 5.5: Human Evaluation Scores on Twitter Data. The Scores in the above Table Represent the Mean Scores of All Tweets. Relevance And Coherence are Scored on a Range of 1-4. User Likes Score Represents the Percentage of Total Raters Who Liked the Meme.

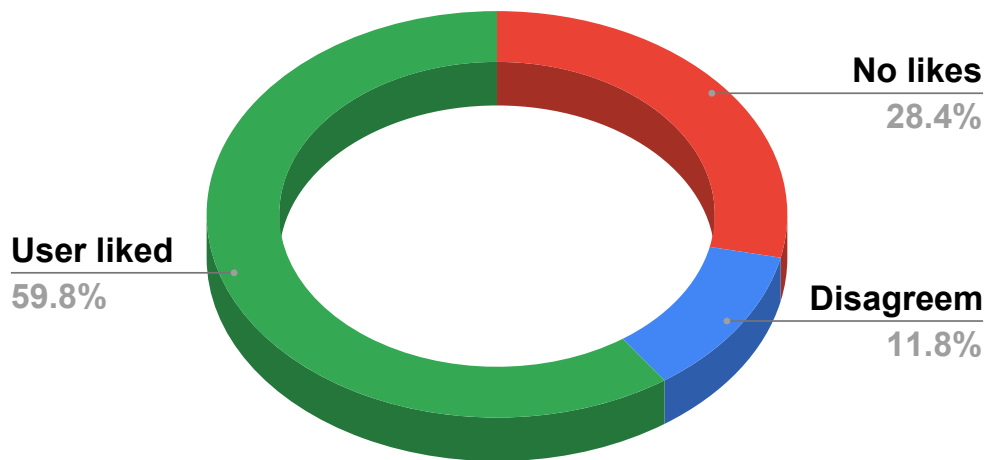


Figure 5.4: User Likes Score Distribution

Before interpreting the scores from Table 5.5, we review the meme generation task. It requires the ability to semantically and contextually understand the input sentence along with the contextual knowledge of the image memes. Even with the understanding of the input sentence and meme images, one has to possess a good fluency in natural language to generate a meme caption that is compatible with the meme image. The generated meme should also be relevant to the input sentence. We analyze the performance of our model by

assuming that a human generated meme would get perfect score across all the 3 metrics in generating a good quality meme that can represent a tweet.

From Table 5.5, we observe that using verbs along with the noun phrases performs better on Twitter dataset as well. It can be seen that our model is capable of generating a coherent meme that is relevant to an input tweet with 66.25% confidence. From the human evaluation scores across all the 3 metrics, we observe that our model to a good extent has learned to generate a quality meme which is relevant to a popular input tweet in an online social interaction.

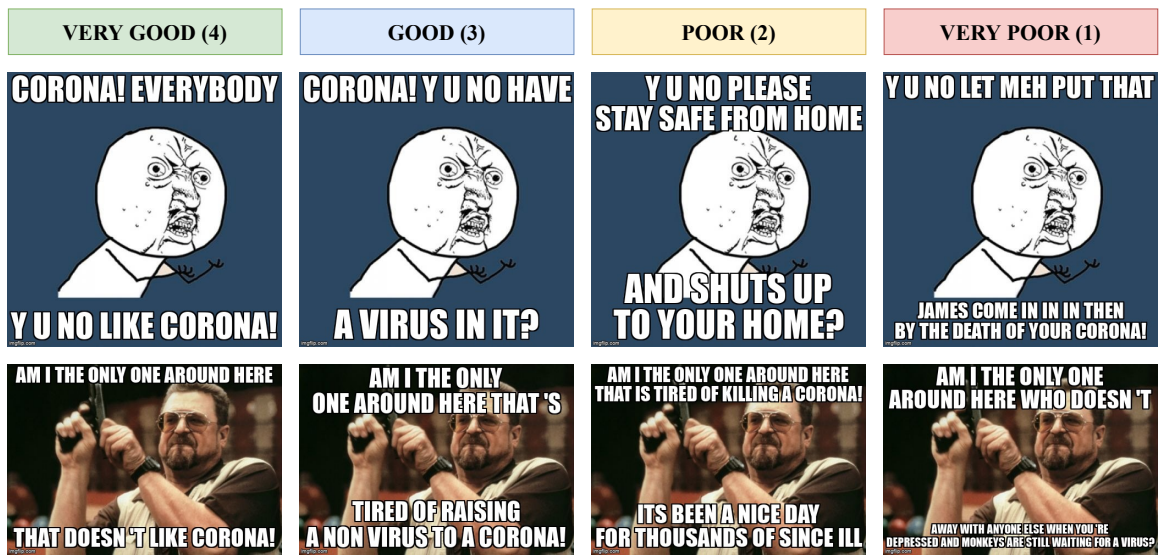


Figure 5.5: Qualitative Figure of Memes Grouped by Coherence Score.

5.4 Inter Rater Reliability

We use Cohen’s Kappa (κ) to measure the reliability among the raters. Cohen’s Kappa is defined in the formulation given below

$$\kappa = \frac{p_o - P_e}{N - P_e} \quad (5.1)$$

where p_o is the relative observed agreement among raters, p_e is the hypothetical proba-

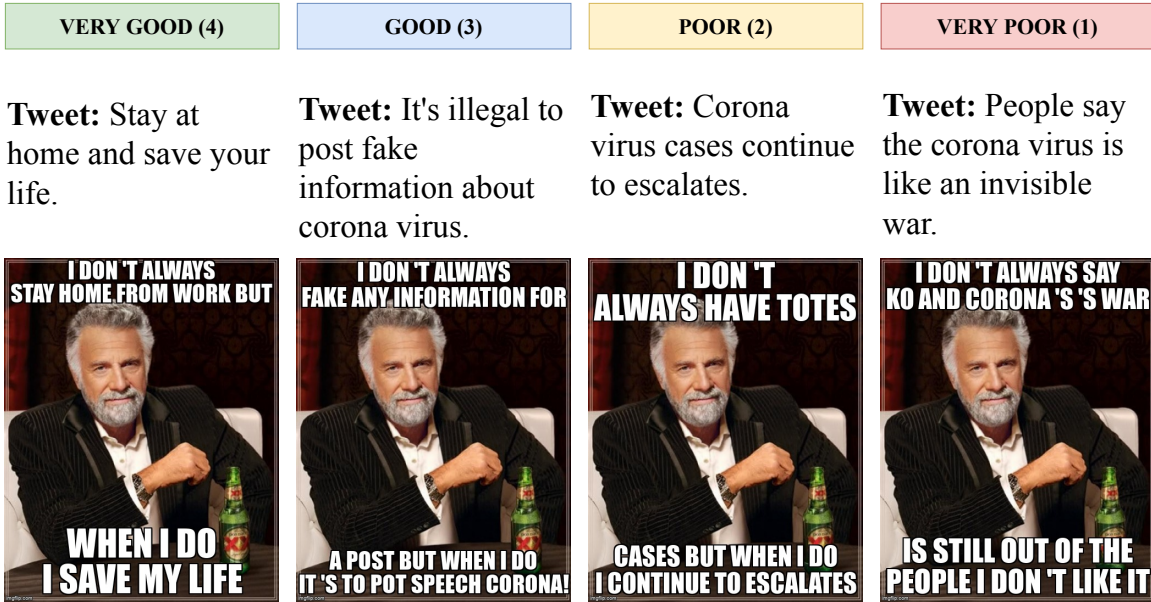


Figure 5.6: Qualitative Figure of Memes Grouped by Relevance Score.

bility of chance agreement and N is the number of samples.

The Inter Rater Reliability (IRR) score among the users on different metrics is reported in Figure 5.7.

5.5 Controlling Meme Generation

Corrupting the input data during training enables our model to learn from the meme caption dataset and scale our model for any input sentence during inference as shown in Figure 5.2. During the experiments, we observed that for any given sentence, the above mentioned information abstraction has enabled our architecture to create meme captions conditioned on any given meme template. We experiment further on this by forcing our transformer to generate captions for a input sentence conditioned on different meme templates and the generated memes are presented in Figure 5.8. It can be seen that our model has learned to generate captions using the learned dependency between the input sentence, meme template and the expected captions and is capable of generating a meme for

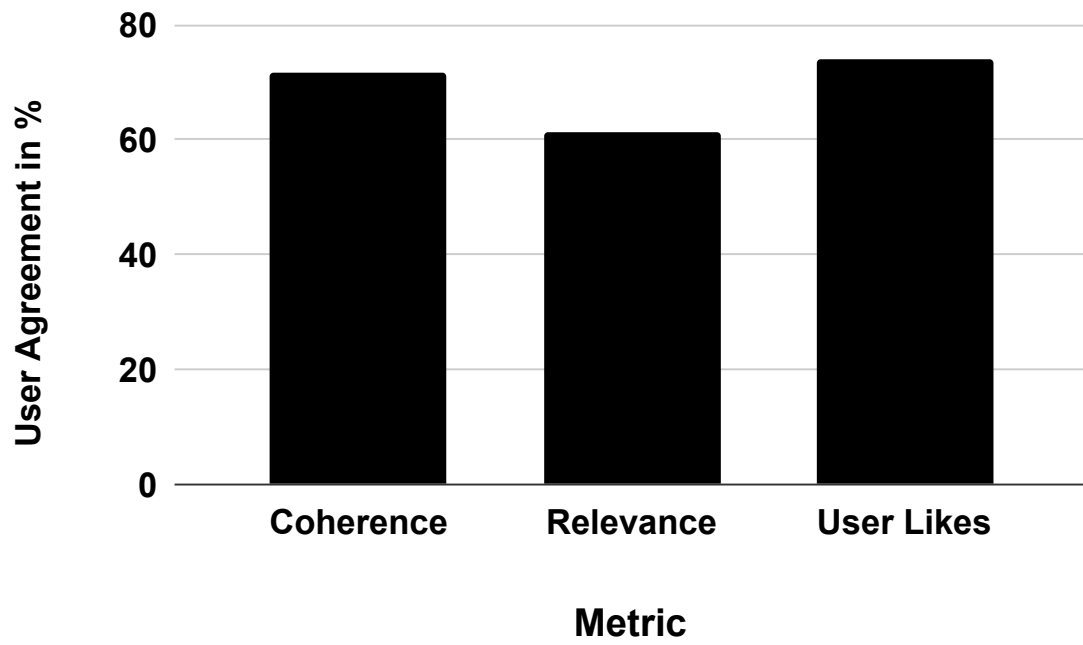


Figure 5.7: Inter Rater Reliability Scores for the Human Evaluation Metrics.

any input sentence conditioned on any meme template.



(a) Memes Generated for the Input Tweet



(b) Memes Generated for the Input Tweet - ""

Figure 5.8: Controlled Meme Generation. Memes Generated Using Different Meme Templates for a Given Input Sentence.

TWITTER BOT

We use a Twitter bot to generate memes for real world tweets. We use the Twitter API to scrape tweets for a given hashtag. We use sentiment filters to remove tweets with negative sentiments and pass the filtered tweets to the memeBot to generate memes. We use sentiment filters to remove generated captions with negative sentiments and post ti back tot he Twitter using the Twitter API. An illustrative figure of the pipeline is presented in the Figure 6.1.

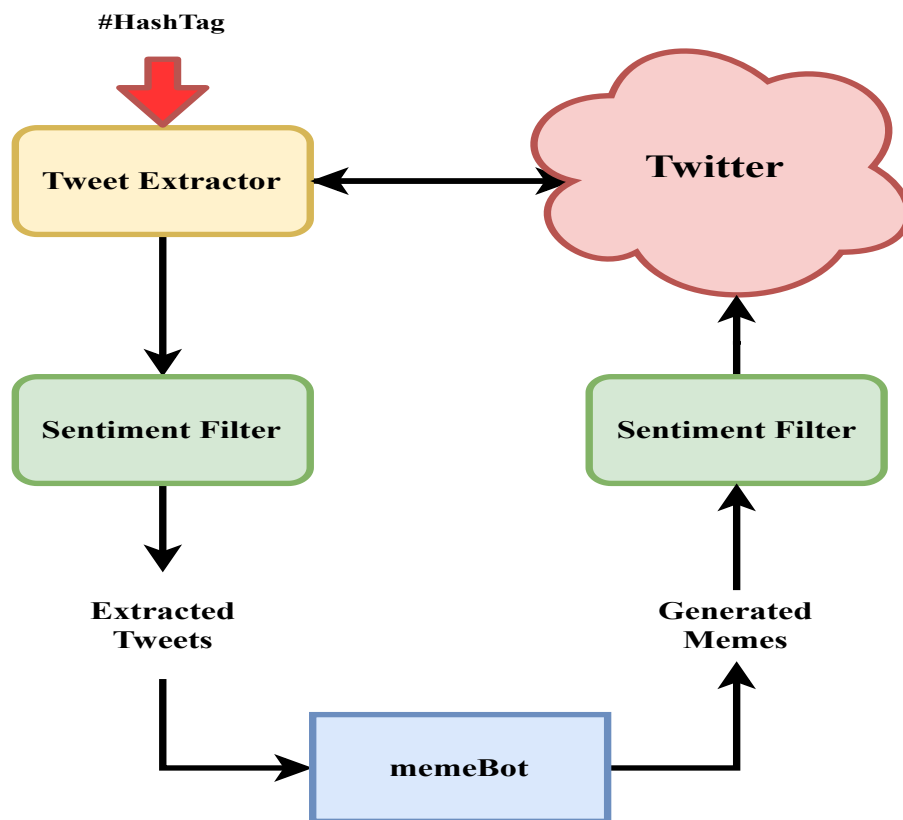


Figure 6.1: Twitter Bot - Illustrative Figure.

CONCLUSION AND FUTURE WORK

We have presented memeBot, an end to end architecture that can automatically generate a meme from a given sentence. memeBot is composed of two components, a module to select a meme template and an encoder-decoder to generate a meme caption. Our model learns the dependency between the input sentence, meme template and the meme caption by utilizing the multi-head scaled-dot product attention and mask multi-head attention mechanisms in the transformer architecture. The model is trained on a Meme Caption dataset to maximize the likelihood of selecting a template given a caption and to maximize the likelihood of generating a meme caption given the input sentence and the meme template. Automatic evaluation metrics on meme caption test data and human evaluation scores on Twitter data show promising performance in generating a meme representing a sentence in online social interaction.

The concept of quality of a meme highly varies among people and is hard to evaluate using a set of pre-defined metrics. In real-world scenarios, if an individual likes a meme, he or she shares it with others. If a group of individuals like the same meme then the meme can become viral or trending. Future work includes evaluating a meme by introducing it in a social media stream and rate the meme based on its transmission among the people. The meme transmission rate and the group of people it transmits across can be used as reinforcement to generate more creative and better quality meme.

As mentioned in Section 1.2, image memes are usually associated with ideas or concepts from some popular pop culture, TV shows or movies references. To understand the concept associated with the meme, one has to possess the cross domain reference associated with the reference. In our work on translating a natural language to an image meme, we use only the

linguistic features of the meme to learn the concepts and ideas associated with it. It requires good amount of training data to learn the concepts just from the linguistic features and we address the limited resource issue as the potential extension of our current work.

REFERENCES

- Bahdanau, D., K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, arXiv preprint arXiv:1409.0473 (2014a).
- Bahdanau, D., K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, arXiv preprint arXiv:1409.0473 (2014b).
- Chung, J., C. Gulcehre, K. Cho and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling”, arXiv preprint arXiv:1412.3555 (2014).
- Davison, P., “The language of internet memes”, *The social media reader* pp. 120–134 (2012).
- Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, in “Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)”, pp. 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), URL <https://www.aclweb.org/anthology/N19-1423>.
- Hochreiter, S. and J. Schmidhuber, “Long short-term memory”, *Neural computation* **9**, 8, 1735–1780 (1997).
- Honnibal, M. and I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”, To appear (2017).
- Hu, Z., Z. Yang, X. Liang, R. Salakhutdinov and E. P. Xing, “Toward controlled generation of text”, in “Proceedings of the 34th International Conference on Machine Learning-Volume 70”, pp. 1587–1596 (JMLR. org, 2017).
- Huang, C., O. Zaiane, A. Trabelsi and N. Dziri, “Automatic dialogue generation with expressed emotions”, in “Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)”, pp. 49–54 (2018).
- Karpathy, A. and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions”, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, pp. 3128–3137 (2015).
- Keskar, N. S., B. McCann, L. R. Varshney, C. Xiong and R. Socher, “Ctrl: A conditional transformer language model for controllable generation”, arXiv preprint arXiv:1909.05858 (2019).
- Kingma, D. P. and J. Ba, “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980 (2014).
- Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach”, arXiv preprint arXiv:1907.11692 (2019).

- Loshchilov, I. and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts”, arXiv preprint arXiv:1608.03983 (2016).
- Luong, M.-T., H. Pham and C. D. Manning, “Effective approaches to attention-based neural machine translation”, arXiv preprint arXiv:1508.04025 (2015a).
- Luong, T., H. Pham and C. D. Manning, “Effective approaches to attention-based neural machine translation”, in “Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing”, pp. 1412–1421 (Association for Computational Linguistics, Lisbon, Portugal, 2015b), URL <https://www.aclweb.org/anthology/D15-1166>.
- Miao, N., H. Zhou, L. Mou, R. Yan and L. Li, “Cgmh: Constrained sentence generation by metropolis-hastings sampling”, in “Proceedings of the AAAI Conference on Artificial Intelligence”, vol. 33, pp. 6834–6842 (2019).
- Oliveira, H. G., D. Costa and A. M. Pinto, “One does not simply produce funny memes! - explorations on the automatic generation of internet humor”, in “ICCC”, (2016).
- Papineni, K., S. Roukos, T. Ward and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation”, in “Proceedings of the 40th annual meeting on association for computational linguistics”, pp. 311–318 (Association for Computational Linguistics, 2002).
- Peirson, V., L. Abel and E. M. Tolunay, “Dank learning: Generating memnlp papes using deep neural networks”, arXiv preprint arXiv:1806.04510 (2018).
- Snover, M., N. Madnani, B. J. Dorr and R. Schwartz, “Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric”, in “Proceedings of the Fourth Workshop on Statistical Machine Translation”, pp. 259–268 (Association for Computational Linguistics, 2009).
- Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank”, in “Proceedings of the 2013 conference on empirical methods in natural language processing”, pp. 1631–1642 (2013).
- Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting”, *The Journal of Machine Learning Research* **15**, 1, 1929–1958 (2014).
- Su, J., J. Xu, X. Qiu and X. Huang, “Incorporating discriminator in sentence generation: a gibbs sampling method”, in “Thirty-Second AAAI Conference on Artificial Intelligence”, (2018).
- Sutskever, I., O. Vinyals and Q. V. Le, “Sequence to sequence learning with neural networks”, in “Advances in neural information processing systems”, pp. 3104–3112 (2014).
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention is all you need”, in “Advances in neural information processing systems”, pp. 5998–6008 (2017).

- Vincent, P., H. Larochelle, Y. Bengio and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders”, in “Proceedings of the 25th international conference on Machine learning”, pp. 1096–1103 (ACM, 2008).
- Vinyals, O., A. Toshev, S. Bengio and D. Erhan, “Show and tell: A neural image caption generator”, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, pp. 3156–3164 (2015).
- Wang, W. Y. and M. Wen, “I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions”, in “Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies”, pp. 355–365 (2015).
- Warstadt, A., A. Singh and S. R. Bowman, “Neural network acceptability judgments”, arXiv preprint arXiv:1805.12471 (2018).
- Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention”, in “International conference on machine learning”, pp. 2048–2057 (2015).
- Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding”, arXiv preprint arXiv:1906.08237 (2019).

APPENDIX A
MEME BOT DEMO

A demo of the meme bot is available at the [website](#).

APPENDIX B
AMAZON MECHANICAL TURK QUESTIONNAIRE

Instructions

There are 2 metrics which should be rated on a scale of 1 - 4. A value of 1 represents a very poor rating while a value of 4 represents a very good rating. Begin by understanding the image meme displayed on top of the page. Rate the metric **Coherence** just with your comprehension of the meme.

Coherence : Is the meme (text + image) coherent? (Explanation: Can you understand the message conveyed through the meme?)

Provide a rating of 1 if you cannot understand the meme. Provide a rating of 4 if you can understand the meme. Provide a rating of 2 or 3 if you feel the meme is understandable but ambiguous.

Metric	Rating
Coherence	very poor (1) <input type="radio"/> poor (2) <input type="radio"/> good (3) <input type="radio"/> very good (4) <input type="radio"/>

Figure B.1: Sample AMT Questionnaire - Coherence Metric.

Now read the text displayed under the "TEXT". Go through the meme once again and rate the metric **Relevance** based on your comprehension of the text and the meme.

Relevance : Is the meme contextually relevant to the text?

Provide a rating of 1 if the meme is not relevant to the text. Provide a rating of 4 if the meme is relevant to the text. Provide a rating of 2 or 3 if you feel the meme is relevant to the text but ambiguous.

TEXT

" Isn't it great how in most the world is going through the corona virus "

Metric	Rating
Relevance	very poor (1) <input type="radio"/> poor (2) <input type="radio"/> good (3) <input type="radio"/> very good (4) <input type="radio"/>

Do you like the meme?

yes no

Figure B.2: Sample AMT Questionnaire - Relevance and User Like Metric.

Disclaimer

The sentences listed in the "TEXT" are randomly selected from Twitter. The sentences are selected from a list of popular tweets and we have no role in selecting the input sentences. The meme displayed is automatically generated using AI without any conditioning. (No human factors or pre-made algorithms were involved in directing the contents of the generated meme). We do not intend to hurt or harm the feelings or beliefs of any individual. This is an experiment on exploring the capabilities of AI in automatically generating memes when prompted with a sentence. Please proceed only if you understand and accept the terms.

Submit

Figure B.3: Sample AMT Questionnaire - Disclaimer.