

memeBot: Towards Automatic Image Meme Generation

Aadhavan Sadasivam, Kausic Gunasekar, Hasan Davulcu, Yezhou Yang

Arizona State University, Tempe AZ, United States

{asadasi1, kgunase3, hdavulcu, yz.yang }@asu.edu

Abstract

Image memes have become a widespread tool used by people for interacting and exchanging ideas over social media, blogs, and open messengers. 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, an image 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 and a decoder is used to decode the meme caption from the meme embedding. The generated natural language meme 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 and human evaluation. An experiment is designed to score how well the generated memes can represent the tweets from Twitter conversations. Experiments on Twitter data show the efficacy of the model in generating memes for sentences in online social interaction.

1 Introduction

An Internet meme commonly takes the form of an image and is created by combining a meme template (image) and a caption (natural language sentence). The image typically comes from a set of popular image candidates and the caption conveys the intended message through natural language.

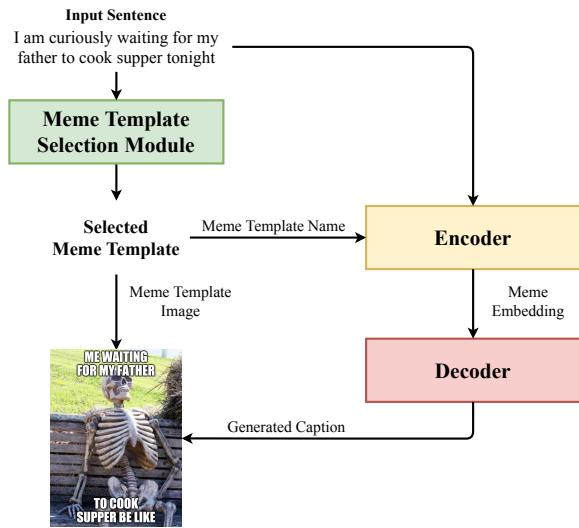


Figure 1: An illustrative figure of memeBot. It generates an image meme for a given input sentence by combining the selected meme template image and the generated meme caption.

Over the internet, information exists in the form of text, images, video, or a combination of these. However 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. The image memes are 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 specific demographics of people operate. For example, in Figure 2, 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 popularity of image-based 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 2: Memes used by the online deep learning community on social media to ridicule the state-of-the-art pre-training models.

Figure 2. The key role played by the image memes in shaping the popular culture of the internet community makes automatic meme image generation a demanding research topic to delve into.

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. Their approach of image meme generation is limited to selecting the most appropriate meme caption or generating a meme caption for the given image and template name. In this work, we intend to automatically generate an image meme to represent a given input sentence (Ideal) as illustrated in Figure 1, which is a challenging NLP task with immediate practical applications for online social interaction.

By taking a deep look into the process of image meme generation, we propose to co-relate image 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 for meme generation, we develop a model that operates beyond the known approaches and extend the capability of image meme generation to generate memes for a given input sentence. We summarize our contributions as follows:

- We present an end to end encoder-decoder architecture to generate an image meme for

any given sentence.

- We compiled the first large-scale Meme Caption dataset.
- 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.

2 Related Work

There are only a few studies on automatic image 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 based on a ranking algorithm. 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., 2015) mechanism. Although there is not much work on automatic meme generation, the task of meme generation can be closely aligned with tasks like Sentiment Analysis, Neural Machine Translation, Image Caption Generation and Controlled Natural Language Generation.

In Natural Language Understanding (NLU), researchers have explored classifying a sentence based on their sentiment (Socher et al., 2013). We extend this idea to classify a sentence based on its compatibility with a meme template. The idea of creating an encoded representation and decoding it into a desired target is well establish in Neural Machine Translation and Image Caption Generation. Sutskever et al. (2014), Bahdanau et al. (2014) and Vaswani et al. (2017) use an encoder-decoder model to encode and decode a sentence from a source to a targeted language. Vinyals et al. (2015), Xu et al. (2015), Karpathy and Fei-Fei (2015) encode the visual features of an image and use a decoder to generate natural language description of the image. Fang et al. (2020) generate natural language description containing common sense knowledge from the encoded visual inputs.

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

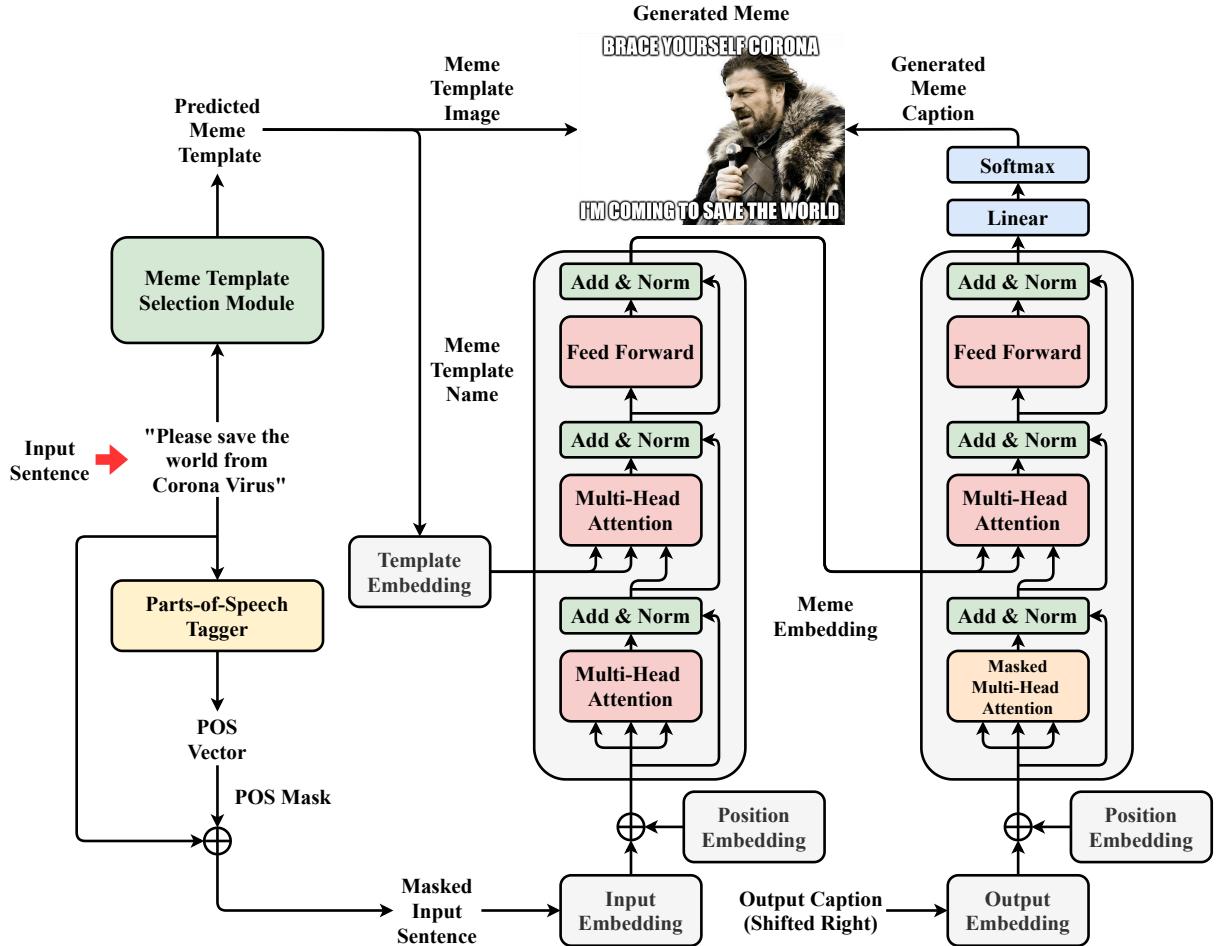


Figure 3: memeBot - model architecture. For a given input sentence, a meme is created by combining the meme image selected by the template selection module and the meme caption generated by the caption generation transformer.

sentence. However, the generated meme caption should represent the input sentence through the selected meme template, making it a conditioned or controlled text generation task. 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. Su et al. (2018) and Miao et al. (2019) generate a sentence with desired emotions or keywords using sampling techniques. While these approaches of controlled text generation involve relatively complex conditioning factors, we implement a transformer (Vaswani et al., 2017) based encoder-decoder model to generate a meme caption conditioned on both the input sentence and the selected meme template.

3 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 as illustrated in Figure 3.

3.1 Meme Template Selection Module

Pre-trained language representations from transformer based architectures like 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. 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 meme template selection module, we fine

tune the pre-trained language representation models with a linear neural network on the meme template selection task. In training, the probability of selecting the correct template for a given sentence is maximized by using the formulation given below:

$$l(\theta_1) = \arg \max_{\theta_1} \sum_{(T,S)} \log(P(T|S, \theta_1)), \quad (1)$$

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

3.2 Meme Caption Generator

We train the meme caption generator 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 meme caption generator. 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.

The meme cation generator model uses a transformer architecture inspired from Vaswani et al. (2017). Our transformer encoder creates meme embedding for a given sentence by performing multi-head scaled dot-product attention on the selected meme template and the input sentence. The transformer decoder initially performs masked multi-head attention on the expected caption and later performs multi-head scaled dot-product attention between the encoded meme embedding and the outputs of the masked multi-head attention as shown in Figure 3. This enables the meme caption generator to learn the dependency between the input sentence, selected meme template and the expected meme caption. We optimize the transformer by using the formulation given below:

$$l(\theta_2) = \arg \max_{\theta_2} \sum_{(S,C)} \log(P(C|M, \theta_2)), \quad (2)$$

where θ_2 denotes the parameters of the meme caption generator, C is the meme caption and M is the meme embedding obtained from the transformer encoder.

4 Dataset

4.1 Meme Caption Dataset

To make possible and validate the aforementioned technical framework, we collect a dataset that would enable us to learn the dependency between a meme template and a meme caption. Here we adopt the open online resource imgflip¹ which is one of the most commonly used meme generators. To automatically crawl the data, we developed a web crawler to collect the memes.

We observe that only a few meme templates dominate the collection. We investigate this dominating memes along with factors that can make a meme popular. 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 make a content shareable and are replicated frequently because of their capability to a make content viral. To this end, we experiment on image meme generation using the popular meme templates.

Our dataset has 177, 942 meme captions from 24 templates. The distribution of meme captions across the meme templates is presented in Figure 4. The dataset consists of meme template (image & template name) and meme caption pairs. A sample from the dataset is illustrated in Table 1. To add diversity to the generated memes, we use various images for the same meme template. The meme template figures used and a sample of the additional images used for the meme templates are presented in Appendices A and B.

4.2 Twitter Dataset

We collect tweets from Twitter to evaluate the efficacy of our model in generating memes for sentences used in online social interaction. We randomly sampled 6000 tweets for the query “Corona virus”. We prune the sampled tweets to 1000 tweets by selecting only those tweets with non negative sentiment using VADER-Sentiment-Analysis (Hutto and Gilbert, 2014). Twitter is an open domain and may contain tweets that could affect the beliefs and sentiments of people and to have a control over our model, we remove the tweets with negative sentiments. The goal is to prompt our model to generate an image meme by inputting a tweet and evaluate if the generated meme is relevant to the tweet.

¹<https://imgflip.com>

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 • when you see your cousin at a family gathering 	
Success Kid	<ul style="list-style-type: none"> • carries the laundry didn't drop a single sock • when you win your first fortnite game • late to work and boss was even later • when she gives you her phone number 	

Table 1: Sample examples (Template name, Captions and Meme Image) from the meme caption dataset.

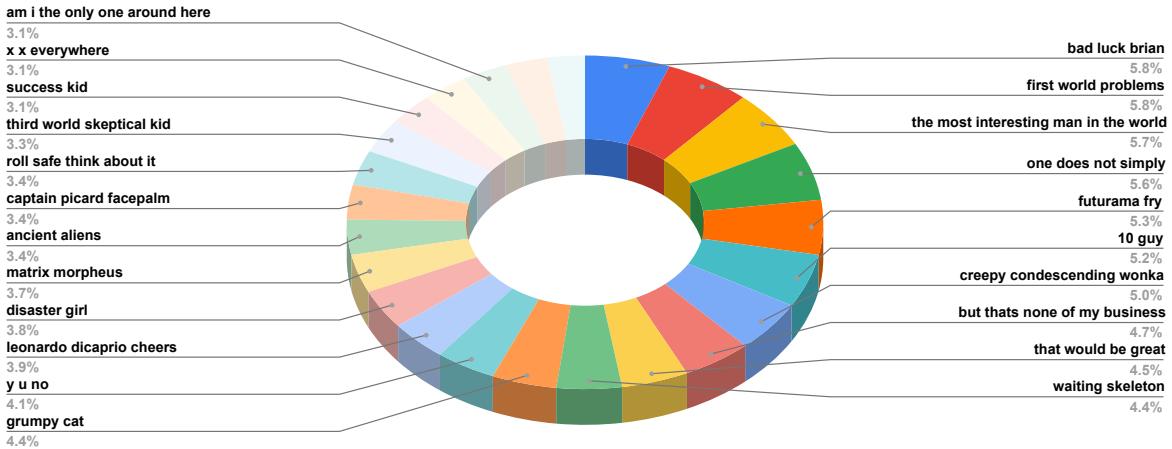


Figure 4: Distribution of meme caption count for the meme templates.

5 Experiments and Results

We train our model on the meme caption dataset (Section 4.1). The train, validation and test dataset contains 142341, 17802 and 17799 samples respectively. We evaluate the performance of the meme template selection module in selecting the compatible template, the effectiveness of the caption generator in generating captions that are similar to the input captions from the meme caption test dataset and the efficacy of the model in generating memes for real-world examples (Tweets) through human evaluation.

5.1 Meme Template Selection Module

We fine tune the pre-trained language representation models ($BERT_{base}$ (Devlin et al., 2019), $XLNet_{base}$ (Yang et al., 2019) and $Roberta_{base}$ (Liu et al., 2019)) using a single layer linear neural network with 768 units on the meme template selec-

tion task using the meme caption dataset. 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 2.

Model	Accuracy \uparrow	F1 \uparrow	Loss \downarrow
$BERT_{base}$	68.18	68.71	1.15
$XLNet_{base}$	68.74	69.06	1.17
$Roberta_{base}$	69.31	69.87	1.10

Table 2: Meme template selection module performance on the meme caption test dataset using the fine tuned language representation models. **Bold font** highlights the best scores obtained. For accuracy and F1, higher the score is better. For loss, lower the score is better.

We adopt the best-performing model with a fine tuned $Roberta_{base}$ model as the meme template selection module in the meme generation pipeline.

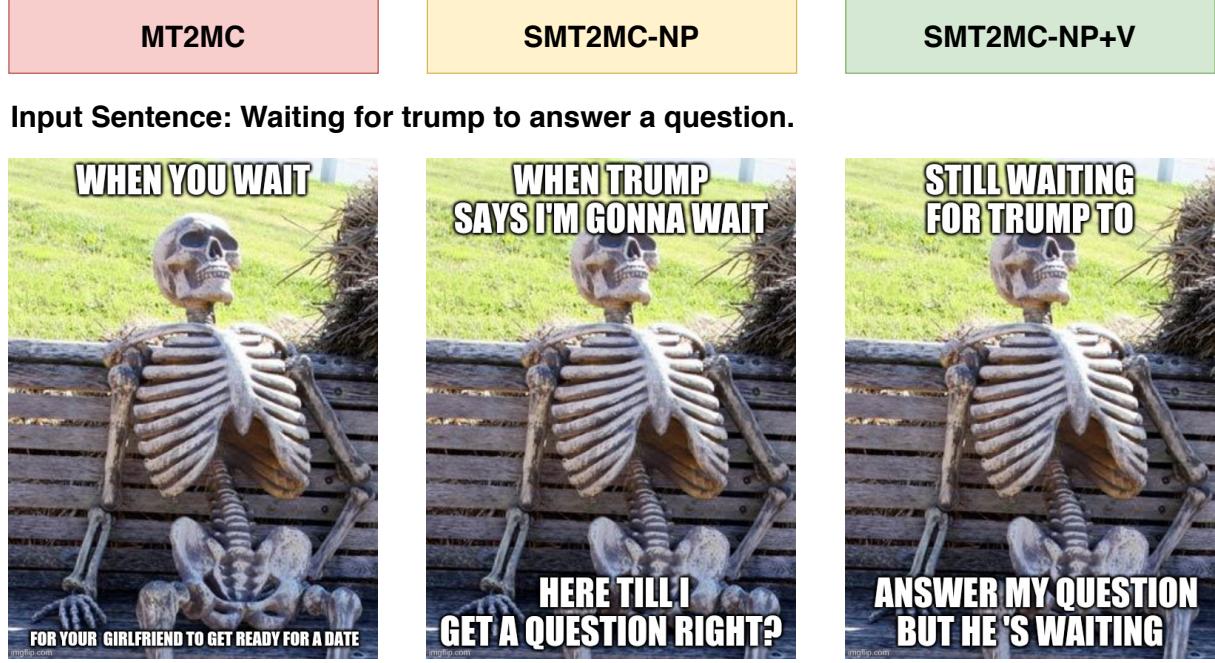


Figure 5: Memes generated by the caption generator variants for the given input sentence.

5.2 Meme Caption Generator

For meme caption generation, we experiment with two different variants. The first variant - Meme Template to Meme Caption (MT2MC), inputs the selected meme template and generates a meme caption. The second variant - Sentence and Meme Template to Meme Caption (SMT2MC), inputs the input sentence along with the selected meme template and generates a meme caption. We use two variants as a part of ablation study to demonstrate the usage of the input sentence features enabling our model to generate memes relevant to the input sentence.

We also experiment with two variants of SMT2MC. 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. SMT2MC and MT2MC architectures follow the same denotations as of Vaswani et al. (2017). We report the hyper-parameters used in Table 3.

Architecture	N	d_{model}	d_{ff}	h
MT2MC	8	768	2048	12
SMT2MC	6	512	2048	8

Table 3: Hyper-parameters used in the caption generation models.

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). 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 tokens is reached.

A sample of memes generated by the caption generator variants are presented in Figure 5. It can be seen that MT2MC generates a random caption for the given meme template which is irrelevant to the input sentence while the meme captions generated by the variants of SMT2MC are contextually relevant to the input sentence. Among the variants of SMT2MC, it can be seen that the caption generated using noun phrases & verbs as inputs better represent the input sentence.

5.3 Evaluation Metrics

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

Variant	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MT2MC	13.93	6.85.86	4.45	2.72
SMT2MC-NP	38.71	24.67	14.56	8.89
SMT2MC-NP+V	45.67	27.56	17.14	11.12

Table 4: BLEU scores for the caption generator variants. **Bold font** highlights the best scores obtained. NP - Noun Phrase and V - Verb.

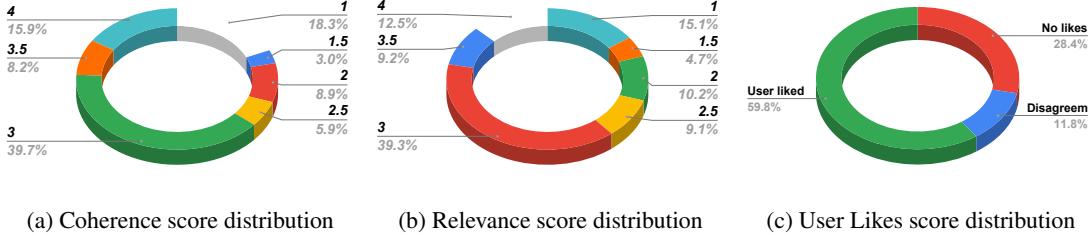


Figure 6: Human evaluation scores.

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?

Coherence score captures the quality (fluency) of the generated meme and the Relevance score captures how well the generated meme represents the input sentence (correctness). We also ask the raters if they like the meme to evaluate if the generated memes are good. The Relevance and Coherence metrics are scored on a range of 1 - 4. **User Likes** score represents the percentage of total raters who liked the meme. To score these metrics, we set up an Amazon Mechanical Turk (AMT) experiment.

5.4 Evaluation Results

5.4.1 Caption Generation Results

The scores for the caption generator variants are reported in Table 4. We see that the SMT2MC variants produce meme captions textually similar to that of the input sentence. Among the SMT2MC variants, the variant which inputs verbs along with noun phrases has better score, and using verbs with noun phrases enables the caption generator to generate relatively relevant captions to that of the input sentence when compared to the variant which

inputs only the noun phrases. We use the best performing SMT2MC-NP+V to generate memes for Twitter data.

5.4.2 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. An example for AMT questionnaire is in presented in Appendix C. Each sample was rated by 2 raters and in case of disagreement among the raters, we consider their average score as the final score.

In our AMT evaluation setup, we design a two-stage process to evaluate the meme. 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 the Relevance metric based on their comprehension of the tweet and the meme. Our expectation for the AMT workers 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. We assume an adult human being is well qualified to meet our expectations.

5.4.3 Human Evaluation Results

The performance of the SMT2MC-NP+V model on the human evaluation metrics is reported in Table 5 and the score distribution across the evaluation metrics is presented in the Figure 6. A qualitative



Figure 7: Qualitative comparison of memes grouped by Coherence score.



Figure 8: Qualitative comparison of memes grouped by Relevance score.

comparison of memes grouped by rater scores is presented in Figures 7 and 8.

Metric	Score
Coherence	2.66
Relevance	2.65
User Likes	0.65

Table 5: Human evaluation scores on twitter data. Relevance and Coherence metrics are scored out of 4. User Likes score represents the percentage of total raters who liked the meme.

Before interpreting the scores, we review the image 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 metrics in generating a good quality

meme for a tweet.

Observing the score distribution from Figure 6, we infer that more than 60% of the generated memes are coherent and relevant to the input tweets. From Table 5, we see that 65% of the raters like the meme shown to them and the like percentage correlate with the coherence and relevance scores. We infer that the raters have liked the meme if they understood the information conveyed through the meme and if the meme is relevant to the input tweet. Quantitatively, our model is capable of generating coherent memes with 66.5% confidence and relevant memes with 66.25% confidence. Our model performs with good confidence on the challenging image meme generation task using only the language features of the image meme during training.

6 Inter Rater Reliability

We use Cohen’s Kappa (κ) to measure the reliability among the raters. Cohen’s Kappa is defined as

$$\kappa = \frac{p_o - P_e}{N - P_e} \quad (3)$$

where p_o is the relative observed agreement and



Figure 9: Memes generated by the SMT2MC-NP+V variant for the input tweet -“Please save the world from Corona” by conditioning caption generation using different meme templates.

p_e is the hypothetical probability of chance agreement among the raters, and N is the number of samples. The Inter Rater Reliability (IRR) score among the users on different metrics is reported in Table 6.

Metric	Agreement Score
Coherence	71.68
Relevance	61.39
User Likes	73.85

Table 6: Inter Rater Reliability scores [%] on the human evaluation metrics.

The raters have higher than 60% agreement on all the metrics which establishes a good consistency among the raters for evaluating the quality of the generated image memes.

7 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 8. In an ideal scenario, the user might want to select the meme template. During the experiments, we observed that for a given sentence, information abstraction during training has enabled our model to create a meme caption conditioned on any given meme template. We experiment further on this by forcing the caption generator to generate captions for a input sentence conditioned on different meme templates. The generated memes are presented in Figure 9 and our model is capable of generating a meme for an input sentence conditioned on a selected or a given meme template.

8 Conclusion and Future Work

We have presented memeBot, an end to end architecture that can automatically generate a meme for 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. 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 on meme caption test data and human evaluation scores on Twitter data show promising performance in generating an image for sentences 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.

Acknowledgment: The Office of Naval Research (award #N00014-18-1-2761) and the National Science Foundation under the Robust Intelligence Program (#1750082) are gratefully acknowledged.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Patrick Davison. 2012. The language of internet memes. *The social media reader*, pages 120–134.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *arXiv preprint arXiv:2003.05162*.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Zhiteng Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR.org.
- Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. 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)*, pages 49–54.
- Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. **Effective approaches to attention-based neural machine translation**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- V Peirson, L Abel, and E Meltem Tolunay. 2018. Dank learning: Generating memnlp papes using deep neural networks. *arXiv preprint arXiv:1806.04510*.
- Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.
- Jinyue Su, Jiacheng Xu, Xipeng Qiu, and Xuanjing Huang. 2018. Incorporating discriminator in sentence generation: a gibbs sampling method. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

William Yang Wang and Miaomiao Wen. 2015. 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*, pages 355–365.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

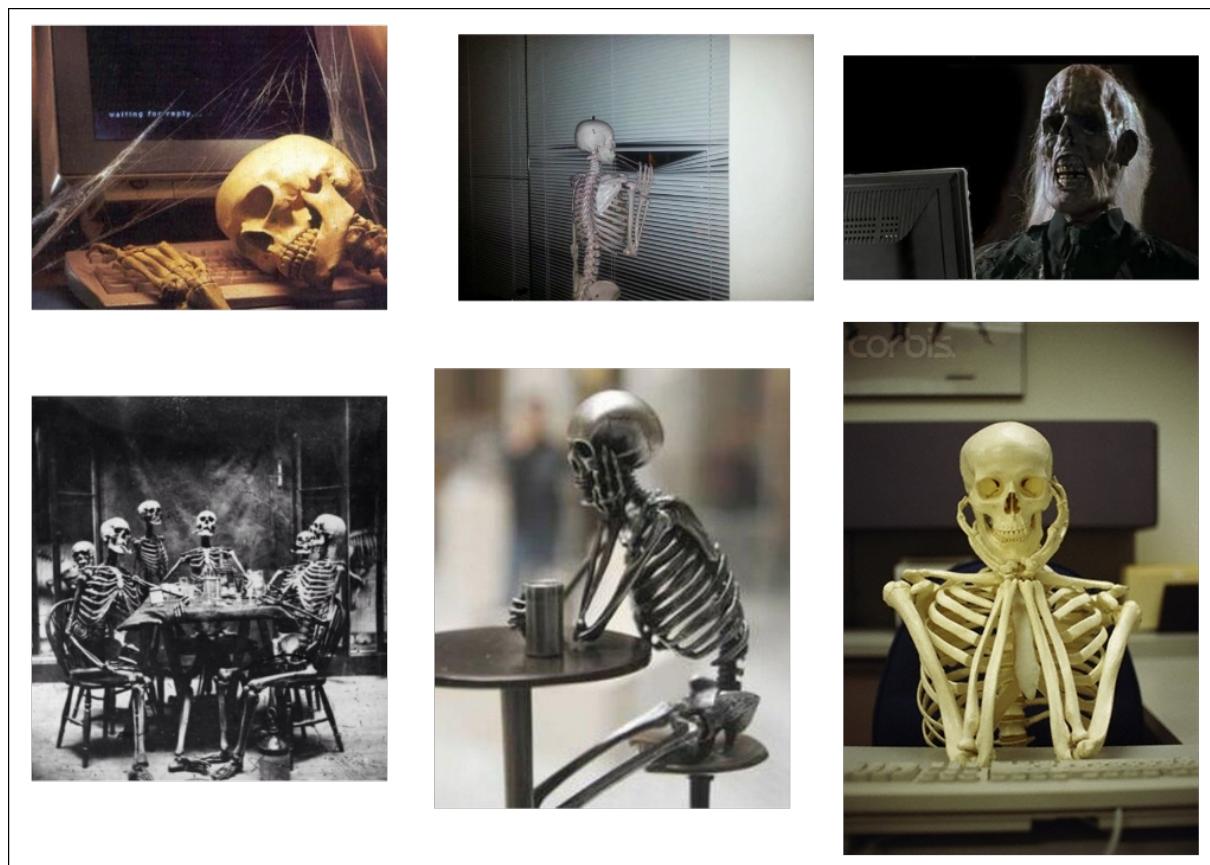
Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

A Meme Templates

Template Name	Template Image	Template Name	Template Image	Template Name	Template Image
Bad Luck Brian		Leonardo Dicaprio Cheers		Success Kid	
X X Everywhere		Ancient Aliens		Disaster Girl	
One Does Not Simply		Third World Skeptical Kid		Futurama Fry	
10 Guy		Am I The Only One Around Here		Captain Picard Facepalm	
Brace Yourselves X is Coming		Creepy Condescending Wonka		Matrix Morpheus	
Y U No		But Thats None Of My Business		Roll Safe Think About It	
Waiting Skeleton		The Most Interesting Man In The World		First World Problems	
Hide the Pain Harold		That Would Be Great		Grumpy Cat	

Meme templates and meme images from the meme caption dataset used in our experiments.

B Sample Additional Images



A Sample of additional images used for Waiting Skeleton meme template.

C 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.

**THIS ISN'T IT
GOING TO BE A POPULAR**

**BELIEF WHAT AM I IN THE WORLD?
POINTS I'M GOING TO GET A VIRUS?**

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"/>

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

Sample AMT questionnaire - Relevance and User Likes metrics.

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

Sample AMT questionnaire - disclaimer.

D memeBot Demo

A live demo of the presented memeBot is available in the [website](#).