Multimodal Sarcasm Generation

Aishwarya Dev
Department of Computer Science
UCLA
aishwaryadev30@g.ucla.edu
305515097

Abstract

Multimodality in Natural Language Processing is an interesting area of research which deals with the processing and interactions between textual data and visual information such as images/videos in tasks such as image captioning Yu et al. (2019), sentiment analysis Maynard et al. (2013), sarcasm and irony detection Schifanella et al. (2016). Sarcasm as a literary device is very commonly used by humans in everyday communication however surprisingly small amount of work has been done on sarcasm processing in NLP. With the already scarce work available, much of the existing work is done on Sarcasm Detection leaving little to no work in the area of sarcasm generation. This work extends the idea of Sarcasm generation to a multimodal setting through an unsupervised approach for sarcasm generation from images. A combination of techniques have been employed to enhance the quality and sarcasticness of the output. The quality of generated sarcastic text has been evaluated by the most reliable evaluation metric for creative language generation namely, human evaluation.

1 Introduction

Multimodality has been rapidly evolving as an interesting area of research in the field of Natural Language Processing. The inter-relationship between textual and image processing has prompted further exploration into the idea of multimodality for Natural Language Generation tasks such as story generation with images and text as in Bensaid et al. (2021), image caption generation Yu et al. (2019), sentiment analysis Maynard et al. (2013), sarcasm and irony detection Schifanella et al. (2016) among several others. A lot of work has been done on the text-only, image-only aspect of NLG but multimodality is relatively new.

Sarcasm as defined by the oxford dictionary is *"the use of irony to mock or convey contempt."* It is a literary device most commonly used by humans in everyday communications be it in conversations, social media comments, news headlines etc. Sarcasm is an effective way of conveying something with an intent that is different from the literal meaning. While sarcasm is a powerful literary device, it is challenging to be sarcastic even for humans for two reasons: 1.The process of sarcasm generation entails creativity to create the desired effect. 2. It is easy to mistake sarcasm for related literary devices such as humor or satire. The challenge of sarcasm generation becomes even more pronounced when an AI is tasked with it because in addition to the above, there are some other challenges such as 1. coherence, 2. grammatical correctness and 3. contextual relevance. Given the extremely small number of sarcasm datasets available, it is very difficult to train generative models on a sufficiently large corpus. Lastly, there is an absence of a standard evaluation metric for a lot of creative language generation tasks because the existing metrics such as BLEU are far less superior than human evaluation for the aforementioned tasks Chakrabarty et al. (2020).

This work extends the idea of Sarcasm generation to a multimodal setting through an unsupervised approach for sarcasm generation from images. Since it is difficult to contextualize the subjects and objects in an image through simple object detection, I leveraged image caption generation for a more holistic view and enhanced context description. The generated image captions are enhanced with 'sarcastic add-ons' which are a custom set of hyperboles and exaggerated responses that compliment the sarcastic appeal of the sentence by evoking the element of surprise.

Sentence 1: A night at the museum.

Sentence 2: A night at the museum. Today must be my lucky day! **More sarcastic**

These captions/enhanced captions are used as prompts for a GPT-Neo generative model with 125M parameters fine tuned on a combination of sarcastic and meme databases including sarcasmv2, reddit meme dataset, memotion 7k dataset. In the absence of a rich corpus of sarcastic data and very few sarcasm datasets, I resorted to training on meme corpus which encompasses the multimodal sarcasm aspect fairly well. The generated sarcastic captions are then refined further by retaining the best quality generated sentences based on their respective sentiment analysis scores. The idea here being, quality of sarcasm as a measure of negativity of the sentiment it entails.

There is an absence of a standard evaluation metric for a lot of creative language generation tasks because the existing metrics such as BLEU are far less superior than human evaluation.chakrabarty2020r Hence I chose human evaluation as the most reliable and unbiased metric for evaluating the sarcastic quality of generated output. A total of 5 human subjects compared the AI generated responses for a given image against human generated sarcastic captions as baseline. They scored the AI generated captions on the following sarcastic/non-sarcastic qualities: Coherence, Contextual Relevance, Humor and Sarcasticness. The evaluation was done on a total of 15 different random images and their associated generated sarcasms.

2 Related Work

Joshi et al. (2015) created a sarcastic chatbot that would generate a sarcastic output from user input. It selects a generator based on input analysis such as Tenses, offensive language, POS etc. Sarcasm is generated based on 8 generators including Offensive word response generator and Sentiment-based sarcasm generator.

Desai et al. (2021) have proposed a novel approach towards Sarcasm Explanation Generation (MuSE). They created their own dataset (MORE) with over 3500 sarcastic posts containing images annotated with reference captions. Their model uses two encoders for text and images along with a cross-modal module for encoding images and associated captions followed by a BART-based decoder that generates contextual explanation. This approach has performed fairly well on evaluation metrics such as BLEU, ROUGE, METEOR, BERTScore Zhang et al. (2019), and SentBERT including ablation study and human evaluation comparison. However, it is unclear if this model is generalizable across different sarcasm datasets and contemporary satire detection models.

Chakrabarty et al. (2020) have used the following sarcastic factors Burgers et al. (2012) for the task of sarcasm generation from a non-sarcastic text input: Reversal of Valence between literal and intended meaning, Addition of commonsense context for explicit incongruity and enhanced humor and Ranking of semantic incongruity. The authors identify the evaluative words in a non-sarcastic text to generate a new sentence through reversal of valence using lexical antonyms / plain negation using not and n't. They also generate commonsense context words for these evaluative words and derive sentences from this commonsense context. The sentence with the highest semantic incongruity with the reversed valence output is concatenated with the reversed valence output to generate the final output. Human evaluation based on Creativity, Sarcasticness, Humor, Grammatical correctness is used as the evaluation metric.

Figure 1: A random image from internet and its associated generated caption

Figure 2: Another random image and its associated generated caption

3 Methodology

Image Captioning

In order to generate sarcasm from the image, it is important to understand what's happening in the image. Since it is difficult to contextualize the subjects and objects in an image through simple object detection, I leveraged image caption generation for a more holistic view and enhanced context description. Most of the state of the art image captioning APIs offered by Azure and Google require a subscription so I decided to use my own image caption generator.

I used an LSTM neural network that leverages transfer learning to avoid training from scratch. The architecture uses InceptionV3 for image feature extraction and vectorized glove embeddings for textual feature extraction. The architecture was trained on flickr8k dataset which is a collection of 8000+ everyday images with no intended sarcasm. The idea is to generate sarcasm from seemingly non-sarcastic regular image inputs. This model did not do well towards caption generation. Since the generated sarcasm depends on the quality of the image caption, it was important to generate high quality image captions so I decided to use an alternate architecture.

The second image captioning technique used an attention-based model Xu et al. (2015) trained on MS-COCO dataset which preprocesses images using Inception V3 and trains an encoder-decoder model architecture. The model was trained on 20 epochs over a batch size of 60000+ images from MS-COCO and gave a far superior generated output that the first architecture on new images.

Sarcastic Add-Ons

Sarcastic Add-Ons is a custom array of some of the most common hyperboles/exaggerations. I decided to enhance the sarcastic appeal of the generated text by concatenating it with some common hyperboles/exaggerations that are used widely by humans such as "Oh my God!", "This is amazing!"

Figure 3: Generated caption with sarcastic Add-On: *'a little boy with a cell phone on a pair of tennis racket..Just what I needed! I'm surprised you don't have a computer.'*

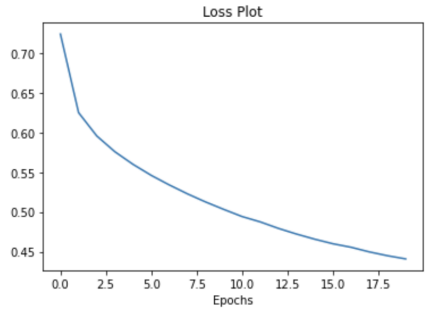


Figure 4: Loss plot for training

etc. The idea here is that adding exaggeration increases the sarcastic quality of a sentence due to the element of surprise, shock it entails. Results showed that sentences with Sarcastic Add-Ons got a better score than the original sarcastic output for about 50 per-cent of test images. However, this approach seemed to compromise the contextual relevance and coherence aspect for other outputs.

Sarcasm Generation

I used the happy transformers library for text generation and used it to train a GPT-Neo pre-trained generative model. GPT-Neo is an open-source version of GPT-3 by EleutherAI. The model I used has been trained on 125M paramaters. I fine-tuned the pre-trained GPT-Neo model on various sarcasm and meme datasets. Since there are a very small number of usable sarcasm datasets available, I tried to train using using a small corpus initially but did not receive a good performance. Hence I decided to widen the training corpus by adding some meme datasets which capture the essence of multimodal sarcasm. The initial training was done on a combination of Twitter sarcasm and sarcasm v2 datasets. For sarcasm v2, I collated sarcastic texts, hyperbole and rehetorical questions sarcasm into one joint training corpus.

The second round of training was done on a joint corpus comprising sarcasm v2(sarcasm, hyperbole, rhetorical question sarcasm), Facebook hateful memes, reddit memes, memotion 7k and sarcastic news headlines datasets respectively. I decided to exclude twitter sarcasm dataset for second round of training as it was noisy and requires a lot of cleanup and data preprocessing.

I employed three sampling algorithms on each prompt in addition to the default generated output: **1. Greedy sampling, 2. Generic sampling** and **3. Top k sampling.** Greedy algorithm surprisingly did better than the other two in the final evaluation.

Prediction enhancement

Verb Scraping

I used a Parts of Speech(POS) extractor on the image caption using Python's spacy library. Then from the extracted POS, I created a list of verbs in the caption. These verbs were then scraped

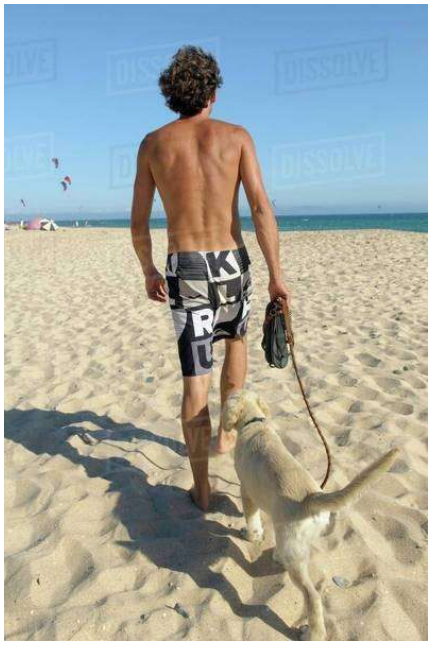


Figure 5: Verb scraped caption with sarcastic add on: *'a man and dog are walking on the beach. Walking causes back-pain. How great is that?'*

across the web to generate the most common causes associated with each verb in a commonsense context. The idea behind verb scraping stemmed from Chakrabarty et al. (2020) where they established that understanding the causes behind commonsense words/phrases can produce context-relevant sarcasm. They used a pre-trained model fine-tuned on ConceptNet to retrieve the causes for commonsense context words in a non-sarcastic text input. Instead of fine-tuning another model, I used a naive web scraping implementation which gave somewhat convincing results but the generated sentences were not intrinsically sarcastic. I tried to enhance this through sarcastic add-ons but struggled with the generalizability of this approach.

Retaining good quality sentences

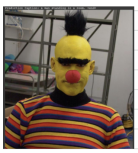
The generated sarcastic output contains a number of sentences. In order to filter generated sentences to retain only the best quality sarcastic sentence, I used sentiment-roberta-large-english model for sentiment analysis of each sarcastic sentence and filtered out the sentences with positive sentiments. Among the sentences with negative sentiments, I sorted the sentence with the highest negative sentiment and retained that for the final output.

4 Results

A total of 5 human volunteers were used for evaluating the quality of generated sarcasm and scoring was done with equivalent human generated sarcastic captions as baseline. The evaluation was done on the following qualifiers of a sarcastic sentence: 1. Coherence, 2. Contextual relevance, 3. Humor and 4. Sarcasticness. For every image, volunteers were shown the generated sarcastic caption and an equivalent human generated sarcastic caption for the same image. Then they were asked to rate the AI generated sarcasm on a scale of 1-10 based on: 1. How coherent is the sentence, 2. How contextually relevant it is, 3. How humorous it is and 4. How sarcastic it is. The averages for each evaluation are summarized in the table below.

It is observed that almost all methods produced fairly coherent sentences. Verb scraping did not do well on contextual relevance while greedy sampling did the best. Humor element was found to

Generated Caption: a man in green shirt in a sauna
Generated Sarcasm: a man in green shirt in a sauna. They are obese and overweight and are in need of a change



Generated Caption: a man standing in a room
Generated Sarcasm: a man standing in a room I guess you have a point

Generated Caption: two girls standing near a water tank.
Generated Sarcasm: two girls standing near a water tank. no one does

Figure 6: Results for some images

	Coherence	Contextual Relevance	Humor	Sarcasticness
Generated sarcasm	8.3	5.46	3.12	3.45
Generated sarcasm (greedy)	8.32	5.93	3.7	4.54
Generated sarcasm (generic sampling)	7.88	5.34	4.21	4.89
Generated sarcasm (top k sampling)	8.42	5.66	3.8	4.32
Generated sarcasm with sarcastic add-ons	8.24	4.56	6.2	5.24
Generated sarcasm with sarcastic verb scraping	8.2	3.23	1.12	2.2
Human generated sarcasm	**10**	**10**	**10**	**10**

Figure 7: Evaluation results

be highest for generated captions with sarcastic add-ons in evaluation. Verb scraping generated the least sarcastic output whereas add-ons, greedy sampling and generic sampling did fairly well.

5 Conclusion

This work explored multimodal sarcasm generation by generating sarcastic text from images. This was done through a three-fold approach: 1. Image caption generation, 2. Sarcasm generation, 3. sarcasm enhancement through sarcastic add-ons, verb scraping, quality sentence retention. Human evaluation was used as a metric with human generated sarcastic captions as baseline. While I was able to get reasonable success with the generated sarcasm, I was limited to conventional training with a huge corpus for both images as well as text processing. I tried to explore zero-shot generation and data augmentation to a limited extent but that is largely the focus for future work. I believe that some of the ideas explored by me in this work can be refined for related work in the future.

6 Future Work

Zero-shot sarcasm generation is an interesting direction of research which is particularly useful under limited resource settings such as unavailability of sufficient training data.

Data Augmentation is another interesting extension to help augment existing datasets. I tried to feed back some of the generated sarcasm data to the corpus in a feedback closed-loop approach but

failed due to error propagation and overfitting. Figuring out an efficient data augmentation approach would be a good future work.

References

Eden Bensaid, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. Fairytailor: A multimodal generative framework for storytelling. *arXiv preprint arXiv:2108.04324*, 2021.

Christian Burgers, Margot Van Mulken, and Peter Jan Schellens. Verbal irony: Differences in usage across written genres. *Journal of Language and Social Psychology*, 31(3):290–310, 2012.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. r^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. *arXiv preprint arXiv:2004.13248*, 2020.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. *arXiv preprint arXiv:2112.04873*, 2021.

Aditya Joshi, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Mark James Carman. Sarcasmbot: An open-source sarcasm-generation module for chatbots. In *WISDOM Workshop at KDD*, 2015.

Diana Maynard, David Dupplaw, and Jonathon Hare. Multimodal sentiment analysis of social media. 2013.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1136–1145, 2016.

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

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.