METGen: Multimodal Emotional Text Generation

Nishant Raj
nishantraj
@umass.edu

Sowmya V Jallepalli
sjallepalli
@umass.edu

Himanshu Gupta
hgupta
@umass.edu

Abhishek Lalwani
alalwani
@umass.edu

1 Introduction

Several social-media-based platforms allow us to express our thoughts and emotions through various forms of multimodal inputs such as text, images, and audio. The existing multimodal Natural Language Generation space has been an active area of research in recent times. With this project, we seek to make some meaningful contributions to this field through our findings.

There could be different categories of emotions that can be represented using the images. This could range from emotions like happiness, neutral, sad to sinister or sarcastic. There has been extensive progress that has been made in the field of emotion detection but the research for multimodal text generation i.e. generation of text given an image is still in a nascent stage.

Sarcasm Detection is one of the topics in natural language processing field that has been widely explored. This task gets difficult when we add images to have a multimodal scenario. For example, an image of rainy and gloomy weather with text: "Weather seems to be amazing outside :(" might be sarcasm. This requires a multi modal approach to solve the problem. However, our problem statement is defined in a scenario where given an image, we would want to generate a sarcastic comment. This might be a challenging task both in terms of implementation and evaluation. The closest task to this problem that we could associate was image captioning and to the best of our understanding, there we have very little control around the type of text getting generated.

Text Generation has been one of the most fundamental problems in the field of Natural Language Processing. With the advent of large-language models such as BERT (Devlin et al., 2018) and GPT2 (Radford et al., 2019), it has been increasingly possible to generate random text which is indistinguishable from the human-written text. There have also been recent advances in the domain of specific text generation where the generated text is expected to meet and display certain characteristics such as a particular emotion or context (Sun et al., 2021).

In this paper, we leverage various such advances to develop a unique model for multimodal emotional text generation, based upon an input image. We also investigate what are the triggers and pitfalls to the existing approaches and perform certain experiments to improve upon the existing works. We finally conclude our research findings with the probable next steps that we feel could be relevant for our use case.

The final result from our model is a single generated phrase which is indicative of the pre-defined emotion and is contextually coherent with the image data. We also investigated with multiple approaches for this task, involving fine-tuned image captioning (Liu et al., 2021) and frameworks for dialogue modeling such as Parl-AI (Miller et al., 2017).

Finally, we explored various metrics as a part of our evaluation strategy and perform human as well as classification-based evaluations of our developed model. In this work, while we limit our emotion space to Sarcasm and Positive Emotions as a proof-of-concept, we strongly believe that our work can easily be extended over to other emotions with relevant data and minor model tweaks.

The paper outlines introduction to our problem statement, a comparison between what we proposed vs what we could accomplish, related work, dataset, baseline, experiments, results, error analysis, contribution of individual members and finally we end with a relevant conclusion and potential future explorations.

2 Changes from our initial proposal

Based on the feedback which we received on our proposal, along with our explorations and findings, we have made some modifications to our initial proposal. We present a list of these changes below. Detailed explanations for the same can be found in the respective sections.

- Baseline Model: ~~Bi-gram based text-only baseline~~. Instead of a completely text-based baseline as mentioned in our initial proposal, we have implemented RNN-based image captioning as our multimodal baseline. This was done to appropriately incorporate the visual data in our baseline.
- Choice of Dataset: ~~Omitted use of dataset from (Cai et al., 2019)~~ . During our exploration of the Parl-AI framework, we came across another relevant dataset which aligned with the requirements of our task in a better way as compared to the previous dataset. Thus, we pivoted to this new dataset and we detail the reasoning for this selection in Section 4.
- Explorations for only Sarcasm category: We felt a need to assess the issues and the working of our pipeline for other emotions as well. Apart from the proposed sarcasm generation model, we also experiment with and present results for the "Positive Emotion" category that we create. Details for the same can be found in Section 4.
- Augmentation with Back Translation: The amount of relevant data corresponding to sarcastic comment generation is highly limited in nature. To combat this, we perform data augmentation using Back Translation (Sennrich et al., 2015a) as detailed in Section 5.5.
- Intermediate Fine-tuning: Based on our explorations and literature review, we added an intermediate fine-tuning approach to our explorations (Phang et al., 2018) as detailed in Section 5.4.
- ~~Human Evaluation using 3-point scale:~~ We also switch to a more comprehensive Human Evaluation scheme based on the proposal feedback. More details for the same can be found in Section 5.7.1.

3 Related work

Extracting information as well as generating new information from multimodal data is extremely important in today's settings. Any relevant information can either be in the form of text, image, audio, video or a possible combinations of all of these. In some cases, it might be extremely difficult to capture the entire context with a single domain. Researchers have started focusing on these and as a result of which, there have been a large upticks on tasks and the involved datasets. Some of the most prominent applications of these include tasks like Visual Question Answering (VQA), Image Captioning, Emotion Detection and Visual Commonsense Reasoning.

Image Captioning is a task that closely matches multimodal text generation scenario. It is defined as a task for describing the contents of an image in the form of text. In our context, major researches like (Hazarika et al., 2018) focus on capturing emotions from only text. (Nezami et al., 2018) uses facial expression analysis for image captioning tasks. (Mathews et al., 2016) present another such dataset, which was curated with sentiment driven captions for images. (Lin and Parikh, 2016) combine two tasks meaningfully by leveraging VQA to rank the captions generated from image captioning task. MS COCO (Chen et al., 2015) is one of the most popular choice for image captioning tasks. Sentences generated with exiting image captioning approaches are mostly neutral in style. Our task i.e. multimodal emotional text generation involves generating captions of a target emotion for a given image. There are many works on multimodal emotion detection but quite a few focus on actually generating text.

In an extensive literature review, we found out that most of the computational work today with multimodal data revolves around emotion detection rather than emotion generation (Ghosh and Veale, 2017) (Riloff et al., 2013) (González-Ibánez et al., 2011) (Ghosh et al., 2015) (Muresan et al., 2016) (Ghosh et al., 2017) (Riloff et al., 2013). Research around generation specifically with multimodal data is still in an early stage.

(Cai et al., 2019) used data from twitter and its sarcastic comment to build a multimodal sarcasm detector. Their work also led to the creation of a new twitter-based multimodal sarcasm dataset. While the text provided in the dataset certainly provides the signals for exploration, we

suspect that the reverse signal might not be very helpful in our case. (Sun et al., 2021) used intent-based guided text generation where the intent can be specified by the author.

Facebook AI Research team recently released the 1.5.0 version of Parl-AI (pronounced as "par-lay") (Miller et al., 2017). It is the "framework for sharing, training and testing dialogue models, from open-domain chitchat, task-oriented dialogue, to visual question answering". One of the application areas is that, given a personality trait (like anxious, happy, dramatic etc.) and an image, the framework can generate a coherent sentence relevant to the personality trait and the image. We leverage the datasets that they provide for this task for our proposed approaches.

Apart from multimodal text generation, one of the most challenging areas is the evaluation of these outcomes. Due to the subjectivity, no single metric can capture full information. BLEU is a metric that is widely used for machine translation tasks. However, in our case for text generation, it is not a fair comparison with expectation of match between machine generated and ground truth gold outputs. (Chakrabarty et al., 2020) notes that for these scenarios, best evaluation can only be done using a human evaluation process and we agree to that point of view. However, in order to assess different progress and metrics and selection of final models, it is simply not feasible to use human evaluation at each stage and hence we also develop a classification-based metric.

4 Dataset

We had initially proposed to use (Cai et al., 2019) as the primary dataset for our task. As a part of our in-depth explorations of the same, we came across various issues which were present in this dataset. These issues did not compromise the capability of the dataset for training a sarcasm classification model, which was the original aim of (Cai et al., 2019). However, these issues made the dataset highly unsuitable for our task of sarcastic comment generation. Exact details of these issues can be found below.

- Extremely short captions. Example: *haha . # lol*.

- Random Noise in texts. Example: *: 'get some ! ..... # military # humor # friends # jokes # usmc # rifleman # arm ... url emoji_2210 emoji_2210 emoji_2210'*.

- Irrelevance to the content of the image/ Non-sarcastic caption. Example: *'sarcasm is just another service i offer shop @ url'*.

(a) Positive: I have been on this Disney ride before, it's a classic!

(b) Sarcasm: Looks like he is lost. Should have brought a sled

Figure 1: Sample Images from curated dataset obtained from ParlAI personality captions task

With these issues in mind, we decided to use another dataset whose source was Parl-AI. The authors used a specific dataset for the "Personality Captions" task. The dataset had the following characteristics:

- Images: Images in the "Personality Captions" dataset were a subset of Yahoo Flickr Creative Commons 100M dataset.

- Dataset Design and Annotations: (Shuster et al., 2019) note that existing image captioning sets like COCO and Flickr30k are factual, neutral in tone and (to a human) state the obvious (e.g., "a man playing a guitar"). These captions were not engaging and lacked control style and personality traits. Thus, they share a huge dataset with 201,858 captions conditioned on 215 unique personality traits that capture all human emotions

Reference Images for the Parl-AI dataset can be found in Figure 1.

4.1 Data Pre-processing

As explained above, we see that there are a total of 215 personality traits in the dataset. For our experiments, we decided to restrict our exploration to two categories which we decided to be: (1) Sarcasm (2) Positive

Our first task was to identify, what subcategories of personality traits could be clubbed together to form a dataset that we could leverage. Based on a common consensus among the group, we decided to use the club below personality to create a broader class. The sub categories chosen from the personality captions are shown in Table 1.

Emotion	Sub Categories
Sarcasm	Sarcasm,Witty
Positive	Fun-loving, Playful, Whimsical, Excitable, Charming, Cheerful, Colorful, Youthful, High-spirited, Earnest, Enthusiastic

Table 1: Categories formed from the corresponding sub-categories from the ParlAI personality traits

In addition to this, since we trained a module for multimodal classification, it also involved preprocessing and we go over these briefly in the section 5.8.1.

4.2 Data Statistics

We split the data into train, validate and test in the following numbers and throughout the experiments, we don't touch test set until our final evaluations. The statistics can be found in 2

Emotion	Train	Validation	Test
Sarcasm	1744	50	83
Positive	8675	216	476

Table 2: Dataset sizes used for corresponding emotions

5 Approach and Experiments

We explored two different approaches for multimodal emotion generation tasks: one is a baseline approach and another is fine-tuned image captioning approach. We also explored Parl-AI and briefly go over our findings from Parl-AI exploration in this section.

In addition to this, we not only explored emotion generation task but to have a guiding light for different experiments, we needed a metric which we could leverage for iterating and deciding our experiments and hence we also developed a multimodal emotion classification module. We detail the components of this module as well in this section.

5.1 Baseline Exploration

For a baseline model, we use an RNN-based image captioning module. The main idea for the baseline is to encode the image information into an embedding which we eventually use as an initial hidden state to our RNN. We extract the image features using Resnet50 (He et al., 2016) and use a fully connected layer to map it to the dimension of the hidden state used in our RNN. Following this, we train our RNN based upon the predicted outputs and the tokenized ground-truth captions. A sample architecture of our baseline approach can be found in Figure 2.

5.2 Captioning Transformer

Image Captioning has been a widely researched topic in approaches that have sought to integrate computer vision and natural language processing fields. Advent of attention mechanism based approaches have transformed the natural language space and has also started making inroads to the computer vision space with the introduction of vision-based transformers. Our attention based approach for image caption generation consists of three different stages:

1. Convolution-based feature extraction
2. Transformer architecture for leveraging attention mechanism over the image
3. Decoding step based on attention over the given image

Different backbone structures like VGG-16, Inception Nets, Efficient Nets have been used for feature generation and people have also experimented with different kinds of attention mechanisms like Adaptive Attention with Visual Sentinel, Semantic Attention and Local/Bhadanau's Attention.

In our work, we leverage ResNet101 architecture with frozen batch-norm as backbone to extract

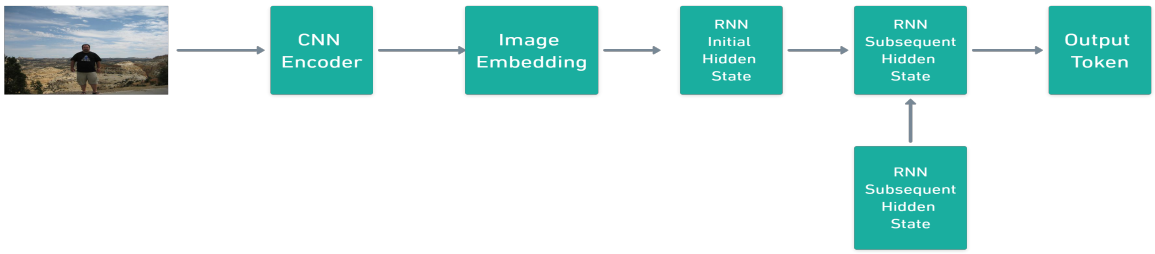


Figure 2: Baseline Architecture. As it can be seen, we encode an image using a CNN-based architecture and use the generated embedding as an initial hidden state for our RNN-based decoder.

image related features. Our transformer architecture that we leveraged consists of 6 layers of encoders and 6 layers of decoders. These layers use 8 heads each. The transformer used has 256 hidden dimensions with 2048 dimensions for each of the feed-forward layer. Architecture for this approach can be seen in Figure 3. Code used for reference can be found in this Github. This architecture has been leveraged for finetuning with our dataset.

5.3 Exploration of Parl-AI

Parl-AI by FAIR team take an retrieval based approach to tackle this problem. They generate probable candidates from the entire corpus with 201,858 captions and then score each of them to retrieve top possible candidates and finally select the best match. They also have a generative approach which we were initially planning to leverage as an upper bound but they have only open-sourced their retrieval model. Since in real-life scenario, we might not have a corpus for retrieval of possible candidates, we don't use this model as a benchmark. One of the most important outcomes of this exploration was that we were able to get most relevant dataset for our use-case.

5.4 Intermediate Fine-Tuning Experiment

As part of our experiment, we also work on intermediate fine-tuning before actual finetuning step. (Phang et al., 2018) in their work were amongst the first ones to demonstrate the effectiveness of an intermediate fine-tuning step in setting where the training examples are less.

There are experiments that have shown that conditions for good results for intermediate finetuning may vary and remain slightly unclear. (Pruksachatkun et al., 2020) demonstrate that intermediate tasks that require high-level inference help in improvement of performance.

For "sarcasm" use-case, we have close to 1.7k training examples and an effective strategy that we felt was to perform an intermediate finetuning with the data from "positive" use case and we perform this experiment and note the results in the outcome section. Our work tries to validate and build on the idea by (Vu et al., 2020) where they indicate that the similarity between the intermediate task and the target task is crucial for successful intermediate-task fine-tuning.

5.5 Data Augmentation using Backtranslation

Since, we had close to 1.7k examples only for our finetuning approach, a natural extension to our idea was to augment the data. However, we needed to make sure that the inherent meaning and style of the textual description remains same for the given image after augmentation. We draw inspiration of this extension from the works of (Sennrich et al., 2015b) (Ng et al., 2019) (Edunov et al., 2018). One of the reasons, we did not choose other augmentations like injection of noise, word embeddings based augmentation (non-context based) etc. because we suspected that this might interfere with the actual training process by altering the meaning. As a part of the augmentation strategy, we used Facebook's "wmt19-en-de" and "wmt-de-en" model, where we converted an english sentence to german and then from german to english. We layout our observations and results in the next section.

5.6 Model Parameters and Hyperparameter Selection

(Mosbach et al., 2020) in their work demonstrate through their work that instability in training while using small dataset can be attributed to lower iterations and suggest further to increase number of

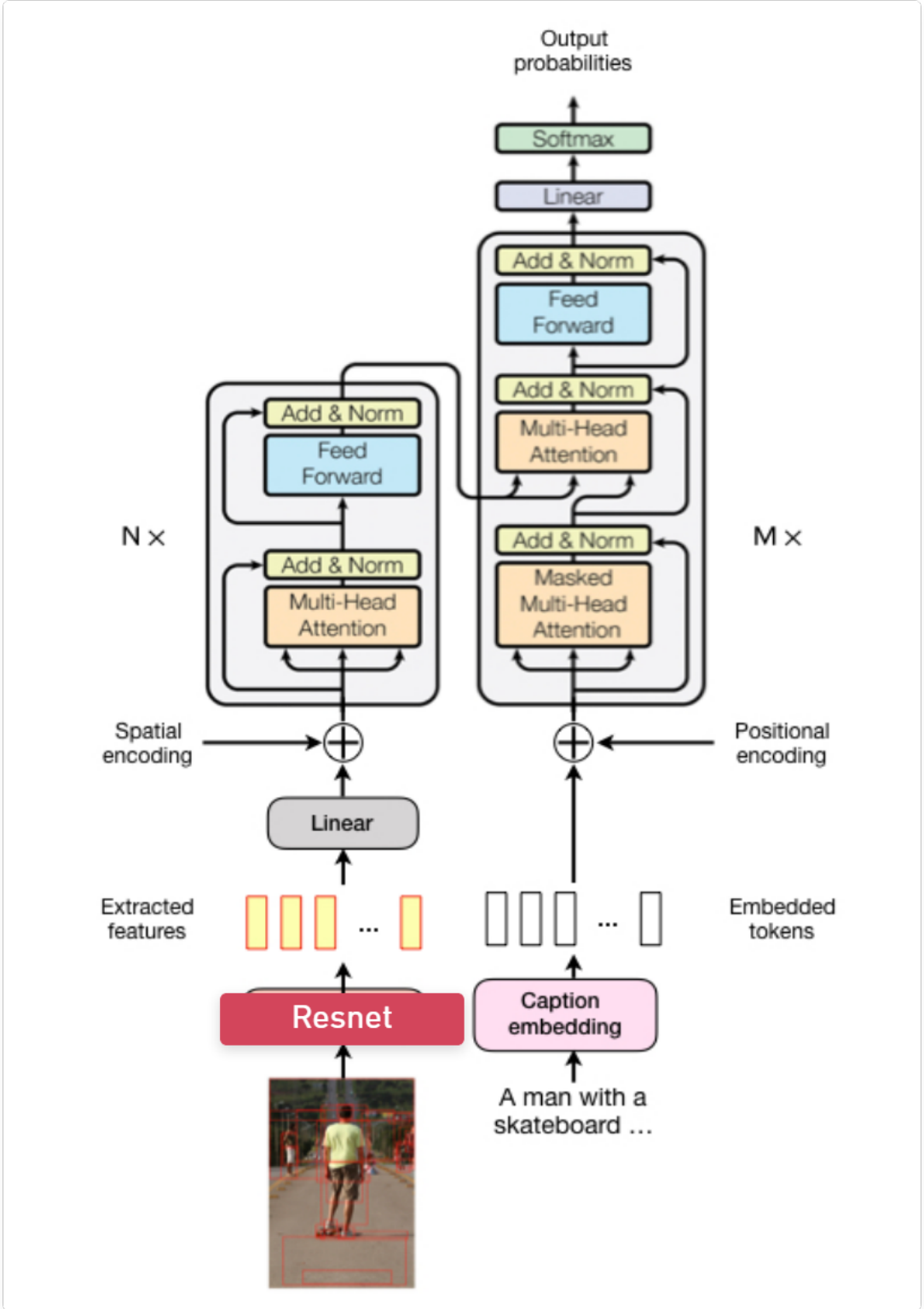
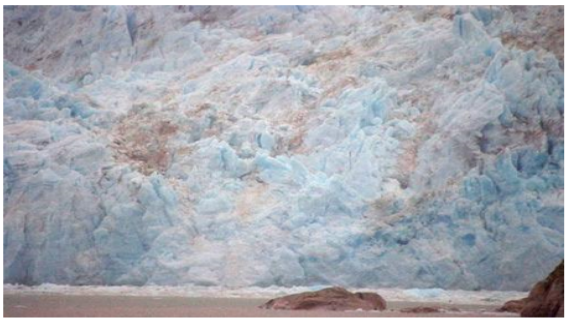


Figure 3: Captioning Transformer (CATR) Architecture. As it can be seen, we extract the image features using the ResNet Architecture which we then pass through our encoder. We then use the encoded image information and pass that to the decoder via cross attention to generate a suitable caption.

a) i love such a fun looking little girl!

b) i'm so excited to play soccer!



c) such a cool disconnect]

d) i love going to go snowball

Figure 4: Qualitative Results for Positive Emotion: (a) and (b) are the representative good results from the test data that capture grammatical correctness, are relevant to the image and represent emotions through text properly. (c) and (d) are the failure cases where they mostly fail to be either relevant to the image and or fail to capture the emotion on broader cases

iterations for a stable model and better predictions. We follow this recommendation and also note inferences from this experiment in the next section.

In case of sarcasm, our best performing model used a learning rate = 2e-5, batch size of 8 and lr decay after 10 epochs. For positive emotion, our best performing model used a learning rate = 5e-5, batch size of 8 and lr decay after 15 epochs.

Due to the system restrictions on Colaboratory, a number of different GPU models have been used to produce our results, including NVIDIA Tesla P100, T4, P4, and K80. For extracting image embeddings (used in classification module) through Vision Transformer (ViT) (Dosovitskiy et al., 2020), we used Colaboratory Pro+ for our experiments due to restricted access in Colab.

5.7 Model Evaluation

We believe that any single evaluation metric would not be able to capture, identify the presence of emotion in our generated text along with the context of the image. We can't use metrics like BLEU which leverage reference sentences in our case as the task of text generation with/without images is very subjective. Hence, we propose two major evaluation approaches namely Human Evaluation (5.7.1) , MAUVE Evaluation (5.7.2) and Classifier based Evaluation (5.7.3).

5.7.1 Human Evaluation

In order to perform human evaluation, each of us take the test sets corresponding to both "sarcasm" and "positive" data and perform evaluation based on the guidelines shared below. We then average out ratings from each of the individual to arrive at final value. In cases of confusion, we resolved that by discussing those examples amongst ourselves.

The major factors that we consider for ratings have been described below:

- Grammatical Correctness (GC) of Sentences: We evaluate on a scale of 1 to 5 how much grammatically relevant our sentences are.
- Sarcasm/Positive-Vibe Representation (ER): We also evaluate sarcasm/positive vibe represented from text with respect to the image in consideration on a scale of 1 to 5.
- Image Relevance: We define Image Relevance as even if the sentence might not represent emotion, does it still capture the thing

Figure 5: Qualitative Results for Sarcasm: (a) and (b) are the representative good results from the test data that capture grammatical correctness, are relevant to the image and represent sarcastic comments. A closer look at (a) reveals a different animal is in water. (b) is also sarcastic as makes a comment that pitcher won't throw a ball which is opposite to the expected nature. (c) and (d) represent our failure cases. In (c) the texture of the image does resemble that of giraffe but model is not fully able to understand the image. In (d) although "she" captures a lady, grammatical correctness is questionable but it does capture the idea of the image

represented in image. This is also represented on a scale of 1 to 5.

In all the above cases, 1 represents poor quality (lowest possible score) and 5 represents good quality (highest possible score). 7 shows the human evaluation results of the four evaluators for Sarcasm. 8 is the same table for Positive emotion.

5.7.2 MAUVE Evaluations

(Pillutla et al., 2021) in their paper propose a KL-divergence based metric for evaluating the quality of generated text with respect to the human translation. In their paper, they note that the relative metric comparison is more apt for MAUVE than absolute and in our case this would mean computing scores for (baseline model, reference test output) and (sarcasm finetuning, reference test output) and then making a relative comparison on MAUVE score as well as frontier integral score. A higher MAUVE score and a lower frontier integral score is more preferred for the model output (either baseline or finetuning approach). We also calculate this metric to demonstrate this as a way for comparison. However, we want to raise a word of caution and highlight the fact that this score might not make much sense in our case as they have highlighted in their work that MAUVE scores need 5000 examples for a meaningful comparison.

5.7.3 Classifier-Based Evaluation

While human evaluation is best suited for these generation tasks, we find that to make interim decisions having a classification based helps to move in the right direction. Hence, apart from a generation model that we developed in the previous sections, we also worked on creating a multimodal emotion classification module.

For this module, we use generated text embeddings using Sentence-BERT module. Text Embeddings generated using Sentence-BERT are 768 dimensional vectors.

For generating the image embeddings, we experimented with tiny DeIT (tiny data efficient Image Transformer) and ViT (Vision Transformer) and used their feature extractors to get last hidden state and for each of the image instances, we extracted 192 dimensional representation of [CLS] token in case of DeIT and 768 dimensions for [CLS] token in case of ViT.

We then create a concatenated representation as a composition function for using it as input representation. We use an extreme gradient boosting model (popularly known as XGBoost model) for the purpose of training a classifier. We use 500 n_estimators and a tree depth of 5 for the purpose of training. We carefully evaluate the model on the above mentioned test data and use a 5-fold cross validation approach to decide on final hyperparameters. We carry out this entire exercise to demonstrate the effectiveness and establish trust for our classification modules. Since the performance of ViT model is better than the DeiT models, we report the AUC-ROC, accuracy, precision and recall based metrics for ViT supported models in the next section where we discuss the results of our models.

We repeat the experiments for both "Sarcasm" and "Positive" datasets. Also, we would like to highlight the fact that ViT model is very heavy and thus to generate the embeddings, we had to extract features in a batch of 75 and then as had to clear RAM, restart Colab sessions' runtime. For DeIT model, this was done in a batch of 200 and then restarting the sessions. For Sarcasm dataset, we had selected 1000 positive examples representing sarcastic personality captions and for non-sarcasm, we had used 1000 randomly sampled non-sarcasm personality traits from the entire data out of the possible 213 personality traits (except Witty which was a close class to sarcasm). For the positive vibe model, we followed a similar approach but used 2000 instances for positive vibe class and randomly sampled 2000 instances for non-positive vibe class. We have mentioned the personality traits that make up positive vibe class in the dataset section of this papers.

6 Results and Error Analysis

In this section, we detail the results for all our models and all the experiments that we performed and detailed in the previous section. The Qualitative results from our models can be seen in 4 for Positive Emotion and 5 for Sarcasm. Each of the below sections explains the results and does an in depth analysis of the Error

6.1 Baseline

As mentioned in Section 5.1, we use a simple RNN-based captioning mechanism to generate baseline results for our task. We use top-k sampling to generate the final set of captions on the test data. A couple of such captions can be found below.

- *myself. Wohoo, cousin encourage encourage fixer-upper times. dip Despite Magic welcome explosions drawing*
- *Giant Smashing Sat Sat Sat giggled helps helps helps skipping tracks! tracks! kitties. romantic! romantic!*

As it can be seen, even without looking at the images, that the captions are highly unrealistic in nature. We believe this happens primarily because of two reasons.

- There is a large amount of diversity in the format of the captions, both for the positive and the sarcastic caption dataset. The baseline works well on a dataset like COCO, where all the examples follow a similar text format which the model is then able to learn by processing a large number of examples. In our case, the model fails to learn the text format due to a high amount of diversity in the same.
- In the case of simpler tasks such as COCO captioning, there is a direct relationship between the objects which occur in the images and the words which occur in the captions. However, within the context of our task, this does not hold true since there are multiple ways to generate positive/sarcastic captions corresponding to a given image. This makes it difficult for the model to learn this complex mapping which eventually shows up in the subpar results as shown above.

Classifier-based results for the baseline exploration can be found in Table 5 and 6. MAUVE evaluation for the baseline exploration can be found in 9 and 10.

Precision	Recall	AUC Score
72.09	76.01	80.64

Table 3: Classifier Performance on the Positive Emotion Dataset

6.2 Intermediate Finetuning

Quantitative results for the intermediate fine tuning task can be found in Table 5 and 6. In the intermediate finetuning experiment, we reach to a

Precision	Recall	AUC Score
66.71	69.66	67.08

Table 4: Classifier Performance on the Sarcasm Dataset

Model	Perplexity	Accuracy
Baseline	738.56	93.6
CATR	1055.82	96.03
Inter. Fine-Tuning	384.59	96.01

Table 5: Classifier-Based Evaluation for the experiments performed on Positive

conclusion that finetuning one "positive" first and then on "sarcasm" as well as vice versa does not lead to any improvement in results. One reason after the experiment that we can think of is that positive and sarcasm are very different emotions and hence intermediate fine-tuning might not be very helpful in this case.

6.3 Back Translation

Quantitative results for the augmented data generated using Back Translation corresponding to the sarcasm setup can be found in Table 6. We observe in the results section that back-translation does lead to a lower perplexity but the accuracy metric and human evaluation leads us to a conclusion that the model's performance is slightly lower or equal. In our experiment, we had used only one pivot language and hence we can't make a conclusive decision if back-translation is helpful or not. A good future experiment would be to use multiple and diversified pivot languages for back-translation.

6.4 Captioning Transformer (CATR)

Below sections describe the three modes of evaluations for the Captioning Transformer approach.

6.4.1 Human Evaluation

All the four annotators performed annotations independently on 50 test examples, each on "positive" and "sarcasm" data. We observe the grammatical correctness of the "positive" dataset is more nuanced than the "sarcasm" dataset. This could be attributed to the fact that the training data for "positive" data is greater in number, and hence the model can learn better representations. From *Relevance* metric, which denotes the relevance of the output with respect to the given image, we observe that the generated texts for both the cases

Model	Perplexity	Accuracy
Baseline	407.41	70.0
CATR	230.48	78.31
Inter. Fine-Tuning	229.20	75.23
Back Translation	165.77	72.28

Table 6: Classifier-Based Evaluation for the experiments performed on Sarcasm

are equally relevant. This leads us to believe that the image feature representation input is able to capture important aspects of the image. The same can be seen in 8 and 7, which has the congregated human evaluation data. One of the plausible reasons that we could comprehend was that since "sarcasm" is a more difficult emotion to capture, the image features alone are not sufficient. To increase the capability of capturing image context and generating sarcastic captions, we might need to induce commonsense reasoning capabilities of the model.

Also, while examining the examples individually, we did find some semantic commonalities. Fine-tuning on "emotion" dataset tends to use the word "love" more often than other words and for "sarcasm", "i bet" and "funny" were the phrases that were used often by the model.

Evaluator	ER	GC	Relevance	Total
1	2.37	2.85	3	2.74
2	3	3.68	3.42	3.37
3	2.45	2.96	3.1	2.83
4	2.57	3.1	3.2	2.95

Table 7: Human Evaluation Results for Sarcasm. ER indicates Emotion Representation, GC represents Grammatical Correctness and Relevance is the relevance of the caption w.r.t image

Evaluator	ER	GC	Relevance	Total
1	4.11	3.94	2.97	3.67
2	4.4	4.11	3.48	4
3	4.23	4	3.6	3.94
4	4.01	4.2	3.4	3.87

Table 8: Human Evaluation Results for Positive. ER indicates Emotion Representation, GC represents Grammatical Correctness and Relevance is the relevance of the caption w.r.t image

6.4.2 MAUVE Evaluation

We report MAUVE scores for comparing the performance of captioning transformer and the baseline approach for both the emotions (Positive and Sarcasm). While the results for the sarcasm dataset align with our understanding (Captioning Transformer is better than Baseline), the results for Positive dataset is counter-intuitive in nature. We believe the lack of data to be the major cause of these observations. Since our test examples are extremely low in number in comparison to what the requirement is (i.e. 5k), we observe that the results in Table 9 and 10 are not that intuitive and hence we don't report this value for other experiments that we have performed.

Model	MAUVE Score	Frontier Integral
Baseline	0.13	0.43
CATR	0.06	0.43

Table 9: Mauve Evaluation results for Positive Dataset

Model	MAUVE Score	Frontier Integral
Baseline	0.23	0.29
CATR	0.50	0.13

Table 10: Mauve Evaluation results for Sarcasm Dataset

6.4.3 Classifier Based Evaluation

We present a classification-based evaluation mechanism inspired by the Inception Score in the image generation literature (Salimans et al., 2016). The main idea revolves around feeding the text features of generated sentences in a classifier along with image features generated using vision transformers and measuring the performance of the said classifier on the entire set of the generated sentences. Within the context of our task, we focus on training 2 separate classifiers, one for each of the emotions with which we work (Positive and Sarcasm). We then feed our entire set of generated sentences corresponding to the test set in the trained classifiers. We measure the accuracy of the classifier on these generated sentences, and we report the same in Table 5 and 6. One interesting thing to note here is the high accuracy (93.6%) achieved on the sentences generated by the baseline approach. We believe this happens because the classification model does not take into account the order of words in the generated sentences and instead focuses on specific keywords which occur in the generated sentences corresponding to an emotion ("love" for Positive). This also indirectly compromises the perplexity of generated sentences and, thus, is a drawback of this evaluation mechanism. Possible solutions to this drawback can be to use a more robust classifier built on top of large language models such as BERT.

7 Contributions

All the members contributed equally to this research and we shared working on a lot of tasks together in the group. Wherever possible, we have demonstrated the split to the best of our understanding.

- Nishant: Dataset selection, creation and preprocessing for both training and evaluation models, finetune image captioning prediction module, Parl-AI exploration, augmentation (back-translation), human evaluation, MAUVE evaluations, classification model (image and text feature extraction), classification model design and development, report development
- Sowmya: Dataset selection, baseline model development, classification model (image and text feature extraction) design and development, human evaluation, report development
- Abhishek: Baseline model design and development, intermediate fine tuning approach, results compilation from these model outcomes, human evaluation, report development.
- Himanshu: Baseline model design and development, intermediate fine tuning approach, fine-tuning based hyper-parameter optimization, results compilation from these model outcomes, human evaluation, report development.

8 Conclusion

We present an emotion-based image captioning pipeline developed on top of transformer architecture. We contrast this with an RNN-based image captioning baseline. We also conduct experiments using intermediate finetuning and back-translation. We finally developed a rigorous evaluation scheme comprising of human evaluation,

MAUVE evaluation, & classification-based evaluation. We measure all our explorations against the evaluation schemes & highlight the shortcomings both qualitatively & quantitatively.

One important exploration that we feel is to incorporate "commonsense reasoning" element into our model. This could lead to a huge boost in performance of the pipeline corresponding to sarcasm dataset given the complex nature of the sarcasm emotion. Another extension of our work can be to incorporate more emotions such as negativity or confusion. It will also be worthwhile to train a single model capable of handling multiple emotions. Prompt-tuning inspired approaches can also be a potential subsequent exploration for this work.

References

Cai, Y., Cai, H., and Wan, X. (2019). Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Chakrabarty, T., Ghosh, D., Muresan, S., and Peng, N. (2020). r^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. *arXiv preprint arXiv:2004.13248*.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Ghosh, A. and Veale, T. (2017). Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics.

Ghosh, D., Fabbri, A. R., and Muresan, S. (2017). The role of conversation context for sarcasm detection in online interactions. *arXiv preprint arXiv:1707.06226*.

Ghosh, D., Guo, W., and Muresan, S. (2015). Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1003–1012.

González-Ibánez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586.

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., and Mihalcea, R. (2018). Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lin, X. and Parikh, D. (2016). Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer.

Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J. (2021). CPTR: full transformer network for image captioning. *CoRR*, abs/2101.10804.

Mathews, A., Xie, L., and He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Miller, A. H., Feng, W., Fisch, A., Lu, J., Batra, D., Bordes, A., Parikh, D., and Weston, J. (2017). Parlai: A dialog research software platform. *CoRR*, abs/1705.06476.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Muresan, S., Gonzalez-Ibanez, R., Ghosh, D., and Wacholder, N. (2016). Identification of nonliteral language in social media: A case study on sarcasm. *Journal of the Association for Information Science and Technology*, 67(11):2725–2737.

Nezami, O. M., Dras, M., Anderson, P., and Hamey, L. (2018). Face-cap: Image captioning using facial expression analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 226–240. Springer.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.

Phang, J., Févry, T., and Bowman, S. R. (2018). Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. (2021). Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34.

Pruksachatkun, Y., Phang, J., Liu, H., Htut, P. M., Zhang, X., Pang, R. Y., Vania, C., Kann, K., and Bowman, S. R. (2020). Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

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

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242.

Sennrich, R., Haddow, B., and Birch, A. (2015a). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2015b). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shuster, K., Humeau, S., Hu, H., Bordes, A., and Weston, J. (2019). Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526.

Sun, S., Zhao, W., Manjunatha, V., Jain, R., Morariu, V. I., Dernoncourt, F., Srinivasan, B. V., and Iyyer, M. (2021). IGA : An intent-guided authoring assistant. *CoRR*, abs/2104.07000.

Vu, T., Wang, T., Munkhdalai, T., Sordoni, A., Trischler, A., Mattarella-Micke, A., Maji, S., and Iyyer, M. (2020). Exploring and predicting transferability across nlp tasks. *arXiv preprint arXiv:2005.00770*.