SentiMeme - Sentiment Identification in Memes

Mario Arduz [1] **Florina Asani** [1] **Amisha Jindal** [1] **Atifa Sarwar** [1]

Abstract

From the past few years, a sudden increase has been witnessed in online content. Information is present in different modalities such as text, image, videos, and audio. Memes, most commonly referred to as internet memes represent an idea, style, or behavioral pattern, are a combination of text and images, and are primarily used for humorous purposes. However, can humor not be offensive or propagate hate speech? Yes, it can, and thus analyzing the sentiments of memes is a crucial thing to do. Most of the recent work in sentiment analysis has been done by considering only one modality. Internet memes require a hybrid approach where the decision can not be made by looking at text or image separately. These two modalities have to be considered together to analyze the context, thus making sentiment analysis from the internet a challenging task. This work aims to identify the context of memes by considering text and image together. We propose SentiMeme, an architecture that employs BoTNet to extract features from image and Bi-LSTM for textual context and combined both to predict the final output. Our model achieved the macro F1-score of 0.52, outperforming the baseline (0.21) and the best in the competition (0.35) so far.

1. Introduction

In the past few years, multimodal situations have seen a rise owing to the fact that many real-world problems are multimodal in nature. Memes belong to a very interesting subset of multimodal problems. While it is relatively easy to crack down on an emotion automatically when it is purely text through methods such as word filters, it can be very hard to deduce the context when it is in the form of a text picture. When there's just text, it becomes a unimodal scenario and is relatively easy to identify. But when combined with a picture, the meaning of the text can change considerably. While humans can look at a meme and understand the combined meaning of it, it's not easy for machines to do the same since they look at the image and text independently (Kiela et al., 2020). The main motivation behind this project is to employ and implement from scratch, the latest BoTNet architecture from Google Research (Srinivas et al., 2021), introduced as a backbone architecture that implements self-attention computer vision tasks. The overall goal is to evaluate the performance of BoTNet on context identification in Memes.

1.1. Research contributions

Our work focuses on the effect of BoTNet in image classification based on their sentiment. As an architecture proposed in Jan 2021, for image classification, instance segmentation and object detection, BoTNet has not been used for sentiment analysis tasks, and our work brings this novel architecture to use in this field.

2. Related Work

A lot of prior work has been about detection of hate speech and natural language processing. While hate speech identification has been prominent, there has been little work on multimodal hate speech with only some works having both images and text. Augmenting text with image embedding information has been said to boost performance in the detection. Some works collect on Instagram images and their associated comments and experiment with various features and classifiers to detect bullying (Kiela et al., 2020).

Earlier methods propose interpreting multimodal speech through modalities consisting of text and socio-cultural information instead of images. A larger and noisier Twitter-based dataset of memes has also been created for the purpose (Waseem, 2016) (Davidson et al., 2017) (Founta et al., 2018).

Multimodal hate speech detection involves vision and language both. A prominent number of such tasks focus on (autoregressive) text generation or retrieval objectives. Even though such situations are important to the interest of the community, they are not in sync with the real-world problems that are encountered in industry by companies

[1]Worcester Polytechnic Institute.

CS/DS 541 Deep Learning Final Project, Worcester Polytechnic Institute.

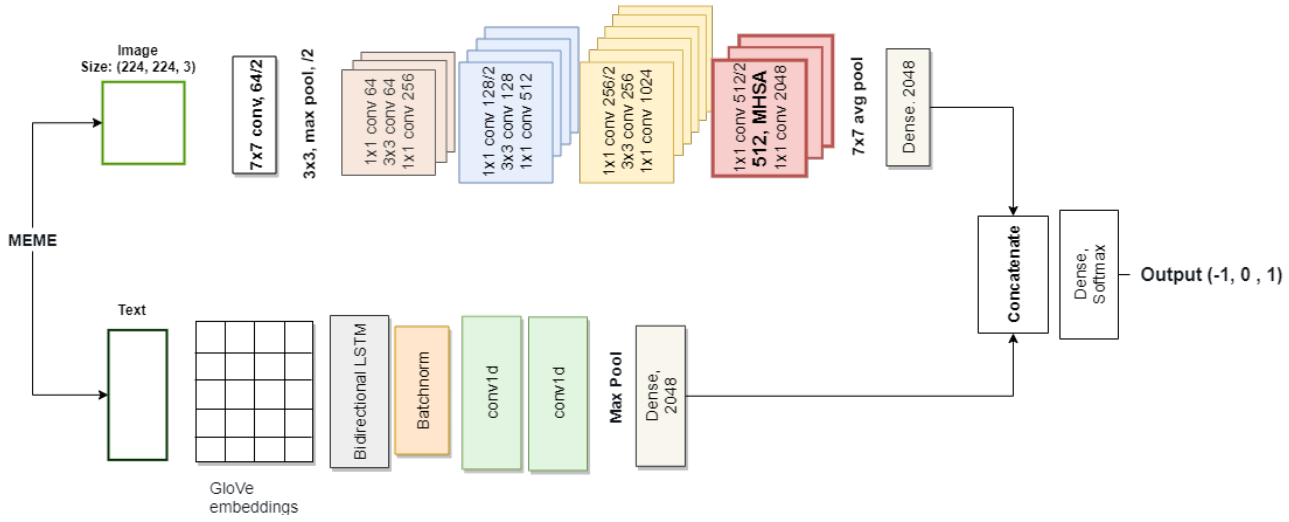

Figure 1. Proposed architecture

like Facebook or Twitter that focus on classification of multimodal posts, comments, etc. for a huge variety of classes. Such problems often require large scale multimodal classifiers akin to the one proposed by Facebook AI.

Another problem arises due to the lack of a standard dataset or benchmark task. Despite having multimodal classification tasks, issues arise due to differences in the qualities of datasets which vary substantially and unavailability of data to different organizations.

In (Kiela et al., 2020), Facebook AI has worked on the evaluation of various models belonging to one of the three classes: unimodal models, multimodal models that were unimodally pre trained and multimodal models that were multimodally pretrained. Image encoders such as ResNet-152 and Faster-RCNN were used along with BERT for textual modality. Transfer learning from pre-trained text and image models on multimodal tasks also formed a category for choice of models. As the final step, these methods were compared to models trained on a multimodal objective as an intermediary step before fine tuning on the current task.

3. Proposed Method

This project aims to conduct emotion analysis on memes based on natural language processing and computer vision. A proper classification would incorporate the extraction of information from the text and image contained in the memes to classify their context in terms of positive, negative and neutral sentiment. To achieve this end, we will base our project on the architecture presented at "SemEval-2020 Task 8: Memotion Analysis - The Visuo-Lingual Metaphor!" (Sharma et al., 2020). The task Memotion analysis releases 8K annotated memes - with human-annotated tags namely sentiment (positive, negative or neutral).

As a first step, we will be getting context from the text. This part of the architecture converts text into embedding, using GloVe. As a next step, these embeddings are passed to a Bidirectional LSTM and two 1D convonolutional layers which are passed through a dense layer for final classification. The next step is to get context from the image. For this part, the architecture proposed in (Sharma et al., 2020), will be modified to use the newly proposed BoTNet (BottleNeck Transformers), instead of VGG-16. In our work we use BoTNet 50. By design, BoTNet is simple: the last three spatial (3x3) convolutions in a Resnet are replaced with Multi-Head Self Attention (MHSA) layers that instead of performing the attention in a single space, splits the input dimension into multiple sub-spaces to compute attentions parallelly. These are then concatenated and after linear projection, final output is generated. (Srinivas et al., 2021). The difference between a Resnet and BoTNet block, is shown in figure 2.

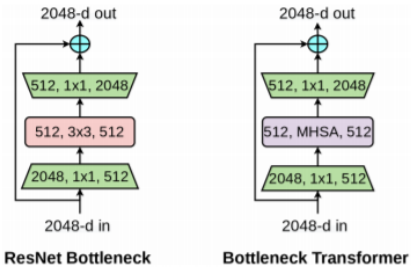

Figure 2. Left: Resnet Bottleneck Block, Right: BoTNet block with multi-head self attention (MHSA)

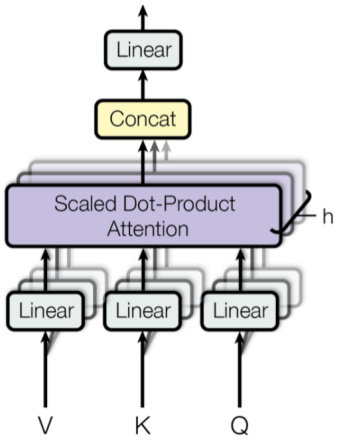

Figure 3. Multi-Head Self-Attention (MHSA)

The output of the BoTNet architecture is passed through a dense layer and concatenated with the text output. And finally, the image is classified as having one of the three sentiments: positive, negative or neutral. The entire architecture is shown in figure 1.

4. Experiments

The used dataset for this project consists of 6982 memes, which are divided into three categories with respect to their overall sentiment: 631 negative (9.02%), 2198 neutral (31.48%), and 4153 positive (59.50%) instances. Since these categories are unbalanced, the cost function had to be adapted to improve model performance. The weights are inversely proportional to their relative frequency of all the data set. Before training, examples were modified to provide more consistent information to the proposed models. Image sizes were formatted to have the same dimension and text content was ignored when it referred to digits or URL directions. The division between the training and testing sets had ratios of 80% and 20%. The training set was additionally divided into ratios of 85% and 15% to train the model and select the best set of parameters on a validation set. For all models, Adam was used as the optimized, along with the Categorical Cross-Entropy Loss.

5. Results

To evaluate the performance of our proposed architecture, we use Macro F1 Score as it is the metric that is being used on the reference work that we compare with. The results shown in table 1, suggest that using BoTNet for image sentiment classification outperforms VGG-16, by a score of 0.30. And using bidirectional LSTMs with 1D CNN's outperform a simple LSTM by a score of 0.28. The final results, combining Text and Image to perform final sentiment classification, outperform the baseline approach proposed in the referenced work (Sharma et al., 2020), as well as the best results obtained from the Sentiment Analysis Competition using the same dataset.

Table 1. Comparison of Macro F1-Scores

Data	**Macro F1-Score**		
	Ours	**Baseline**	**Best in competition**
Text	0.49	0.21	-
Image	0.48	0.18	-
Text + Image	**0.52**	0.21	0.35

The confusion matrices shown below, suggest that the hardest sentiment to distinguish is between neutral and positive sentiment. Using only images for sentiment classification shows positive sentiment to be easier to identify, as shown in table 2, whereas neutral and negative, being the underrepresented classes, were harder to identify and often confused.

Table 2. Confusion Matrix of Image model

	Negative	**Neutral**	**Positive**
Negative	**5**	19	102
Neutral	12	**81**	347
Positive	41	147	**643**

When using only text for sentiment classification, the model finds it a little easier to identify neutral and negative sentiment, as compared to when only using images. Shown in table 3

Table 3. Confusion Matrix of Text model

	Negative	**Neutral**	**Positive**
Negative	**7**	35	84
Neutral	20	**138**	282
Positive	46	267	**518**

Combining both text and images in one model, helps to further identify neutral and negative patterns, as shown in table 4 and achieves an F1 score of 0.52.

Table 4. Confusion Matrix of Text+Image model

	Negative	**Neutral**	**Positive**
Negative	**8**	31	87
Neutral	31	**140**	269
Positive	67	269	**495**

6. Discussion

Employing BoTNet into our architecture, increases the prediction accuracy and the sentiment of positive, negative or neutral is distinguished more clearly. Our work shows that

for image classification tasks, BoTNet is efficient and outperforms VGG-16, because of the fact that the image size used for our task is small, hence self-attention becomes more efficient for pooling context.
However, as with every other sentiment analysis task, the limitation of this work lies in the fact that the annotation of the labels was done by a group of workers at Amazon Mechanical Turk (AMT), therefore it is highly subjective. This also explains why most of the confusion for our models occurs between memes that are labeled as having a positive or neutral sentiment, as there is similar content for both classes. Figure 4 and figure 5 show sample outputs generated by our proposed model.

Figure 4. Correctly classified positive meme as having a positive sentiment

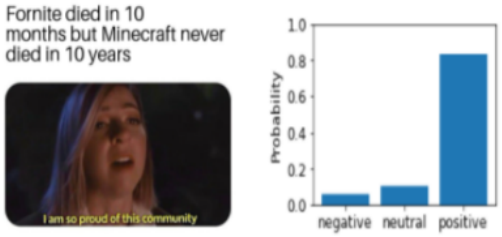

Figure 5. Incorrectly classified negative meme as having a positive sentiment

7. Conclusions and Future Work

Memes are ubiquitous and widely consumed means of communication. As such, they hold the potential to propagate negative ideas which can not be only limited by human intervention. Models that treat multimodal problems can support a responsible use of this type of entertainment by flagging the creation and spread of ill-intentioned content. As mentioned earlier, sentiment identification in memes includes an intrinsic challenge since they are composed of images and text. A human always attempts to analyze both the text and the image to understand the intended meaning behind a meme, so a model combining and weighting both modalities is needed for this task. The use of deep learning architectures has provided a useful approximation for meme sentiment classification. In particular, this project was able to enhance meme classification with respect to their labeled sentiment. We consider that the main improvement relies in the performance of the image classification task by the use of BoTNet, and that the success of the final classification tasks has the potential to address the type of content conveyed by memes.
Since for image classification, adding multi-head self attention (MHSA) to the last spatial convolution layers increased the performance significantly, for future work, the model can be further optimized, by employing an attention-based sentiment analysis model for the text extracted from the memes.

References

Davidson, Thomas, Warmsley, Dana, Macy, M., and Weber, Ingmar. Automated hate speech detection and the problem of offensive language. In *ICWSM*, 2017.

Founta, Antigoni-Maria, Djouvas, Constantinos, Chatzakou, Despoina, Leontiadis, Ilias, Blackburn, Jeremy, Stringhini, Gianluca, Vakali, Athena, Sirivianos, Michael, and Kourtellis, Nicolas. Large scale crowdsourcing and characterization of twitter abusive behavior, 2018.

Kiela, Douwe, Firooz, Hamed, Mohan, Aravind, Goswami, Vedanuj, Singh, Amanpreet, Ringshia, Pratik, and Testuggine, Davide. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *ArXiv*, abs/2005.04790, 2020.

Sharma, Chhavi, Paka, Scott, William, Bhageria, Deepesh, Das, Amitava, Poria, Soujanya, Chakraborty, Tanmoy, and Gambäck, Björn. Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep 2020. Association for Computational Linguistics.

Srinivas, A., Lin, Tsung-Yi, Parmar, Niki, Shlens, Jonathon, Abbeel, P., and Vaswani, Ashish. Bottleneck transformers for visual recognition. *ArXiv*, abs/2101.11605, 2021.

Waseem, Zeerak. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618. URL https://www.aclweb.org/anthology/W16-5618.