Hate Meme Detection

Abhishek Goyal
ScAI,IIT Delhi,India
aib232073@iitd.ac.in

Mikshu Bhatt
ScAI,IIT Delhi,India
aib232067@iitd.ac.in

Bogam Sai Prabhath
ScAI,IIT Delhi,India
aib232079@iitd.ac.in

Abstract

A meme is "an image, video, piece of text, etc., typically funny. But, certain categories of images, text, or the combination of the two modalities could turn into a hateful meme. The Hateful Meme Detection presents a challenge due to its unique characteristics. The unimodal models fails to learn, and only multimodal models have shown to succeed. Non-expert human trained annotators have an accuracy of 84.7%. These State-Of-The-Art (SOTA) multimodal models are shown to perform poorly compared to humans on this dataset. Various competitions like Meta meme detection , act as a benchmark for both AUC score and multimodal modals. We propose here a multimodal modal using MMBT [2] as fusion modal and CLIP [5] as image encoder as a better multimodal modal.

1 Introduction

Specifically, we focus on hate speech detection in multimodal memes. Memes pose an interesting multimodal fusion problem: Consider a sentence like "love the way you smell today" or "look how many people love you". Unimodally, these sentences are harmless, but combine them with an equally harmless image of a skunk or a tumbleweed, and suddenly they become mean.Here we use VisualBERT (trained on COCO) [3] and BERT trained on Hateful text as our benchmark models. Our main model is MMBT based model (late fusion) which uses CLIP to obtain image embeddings. We show that this modal outperforms VisualBERT and , thus various multimodals can be used along with ensemble learning as done in competitions like Meta hate meme detection challenge.

1.1 Motivation

To participate in online safety prize challenge Singapore and detect multimodal multilingual harmful meme online.

1.2 Problem Statement

The Meta Hateful Memes Dataset consists of a training set of 8500 images, a dev set of 500 images & a test set of 1000 images. The meme text is present on the images, but also provided in additional jsonl files. To increase its difficulty, the dataset includes text & vision confounders. Such confounders change from being hateful to non-hateful or vice-versa by swapping either text or image only. They ensure that models must reason about both vision & language. A vision-only or language-only model cannot succeed in the task.The AUROC score was used as the competitions key metric. Intuitively, it penalizes models that are bad at ordering memes by hatefulness.

1.3 Challenges

1. Major challenge here is Multimodal datasets are typically much smaller in size than other visual datasets used to train deep image classification models. Training on different subsets of the rather small training set of the HM dataset can lead to considerable variation in model predictions. To stabilize the predictions and tackle overfitting, people utilize ensemble learning.

2. Another challenge is faced by special media apps.According to Mike Schroepfer, Facebook CTO, they took an action on 9.6 million pieces of content for violating their HS policies in the first quarter of 2020. This amount of malicious content cannot be tackled by having humans inspect every sample. Consequently, machine learning and in particular deep learning techniques are required to alleviate the extensiveness of online hate speech. To, tackle this problem, Facebook AI's releases the Challenge Set which encourages Researchers and scholars to build a model which classifies the memes into hateful or not

hateful.

1.4 Solution Methodology

We went through various kinds of approaches used in the competition. And found out that multimodal with late fusion were SOTA , and applying preprocessing like OCR removal of images , Object detection , face color detection , face sentiment detection and adding these as textual tags can increase the accuracy . Further various teams in the end increased their accuracy using ensemble learning.

1.5 Model-1 AUROC-0.84

This model developed by Alfred system used OCR detection and cleaned images of any text . Then used four different VL transformer architecture ensemble with VL-BERT, UNITER, VILLA, and ERNIE-Vil [8].

1.6 Model-2 AUROC-0.83

The approach pipeline is present in figure below [4].

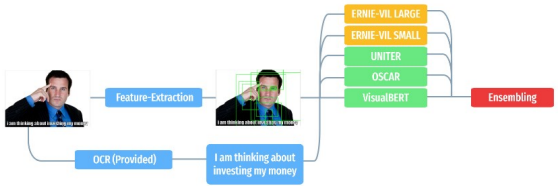

Figure 1: Model-2 Approach

1.7 Model-3 AUROC-0.81

The approach could be summed up as follows: Growing training set by finding similar datasets on the web, Extracting image features using object detection algorithms (Detectron), Fine-tuning a pre-trained V+L model (VisualBERT [3]) Hyperparameter search and applying Majority Voting Technique [7].

The approach pipeline is present in figure below .

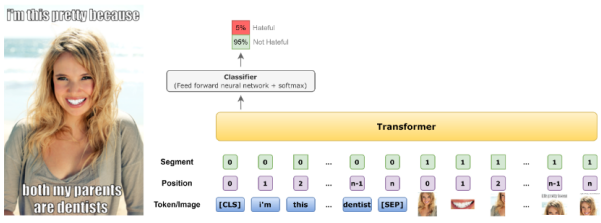

Figure 2: Model-3 Approach

2 Dataset

We obtain the dataset of Meta hate meme competition. The `img` folder contains all the images of the challenge dataset including train, dev, and test split. The images are named `<id>.png`, where `<id>` is a unique 5-digit number. `train.jsonl`, `dev_seen.jsonl`, and `dev_unseen.jsonl` are JSON files where each line has a dictionary of key-value pairs of data about the images.

Table 1: Distribution of the different types of memes in the Hateful Memes dataset.

Split	Multimodal hate	Unimodal hate	Benign confounders	Random benign	Dynamic adversarial benign confounders	Total
Train	1300	1750	3200	2250	–	8500
Dev-seen	200	50	200	50	–	500
Test-seen	400	100	400	100	–	1000
Dev-unseen	200	–	200	–	140	540
Test-unseen	729	–	597	–	674	2000

Figure 3: Dataset Distribution

2.1 Extending Dataset

We went through ways data is extended for memes. One of these ways is by the Memotion dataset.It is an open-sourced dataset containing 14K annotated memes with human-annotated labels, namely sentiment(positive, negative, neutral), type of emotion(sarcastic, funny, offensive, motivational).

3 Approach

We could deduce from the above used models in facebook competition that multimodal modal were SOTA and using object detection and ensemble of these models can increase the accuracy quite well.

3.1 Unimodal as benchmark- BERT trained on Hate speech

Unimodal learning focuses on building machine learning models using only one type of data – text, images, audio, or video. While specialized for single data types, unimodal learning has limitations as it does not consider other form of data image present in the meme. But for taking a benchmark for our multimodal modals we include it as a benchmark.

3.2 Multi modal single stream early fusion - VisualBERT

In simple terms, multimodal means combining information from different sources such as text, image, audio, and video to build a more complete and accurate understanding of the underlying data . We were only interested in early fusion modals since they have been proven to be better than late fusion modals . Different kinds of fusion are depicted in

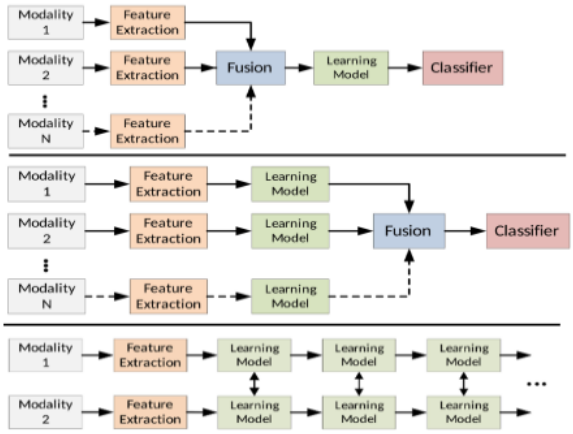

Figure 4: From top to bottom: early fusion, late fusion, cross- modality fusion

image below. One of these multimodel used is VisualBERT. In VisualBERT image regions and language are combined with a Trans-former to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling and sentence-image prediction task on caption data and then fine-tuned for different tasks.It is used as a benchmark for various new SOTA multimodals now.Image features for the model is derived from faster RCNN kind of modals. Image below depicts the modal.

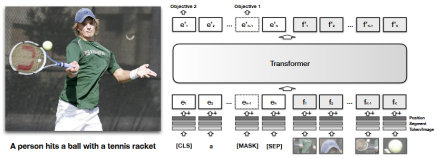

Figure 5: VisualBERT

3.3 Multi modal multiple stream early fusion - MMBT

In a paper Meta showed that super-vised bidirectional transformers with unimodally pretrained components are excellent at performing multimodal fusion, outperforming a variety of alternative fusion techniques. Moreoever, they found that their performance is competitive with, and can be extended to outperform, multimodally pretrained ViLBERT models on various multimodal classification tasks. Multimodal bitransformers provide what is effectively a deep fusion method . This method does not rely on a particular fea- ture extraction pipeline since it does not require e.g. region or bounding box proposals like VisualBERT. It works for any sequence of dense vec- tors. Hence, it can be used to compute raw image features, rather than pre-extracting them, and back-propagate through the entire encoder. A supervised multi-modal bitransformer jointly finetunes unimodally pretrained text and image encoders by projecting image embeddings to text token space. It works well on text heavy dataset .

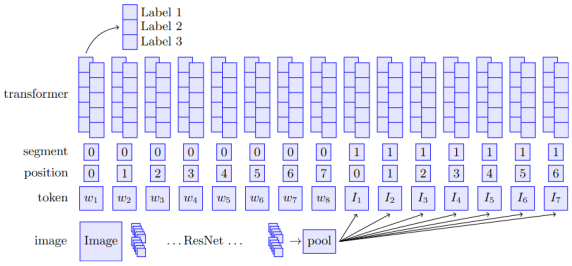

Figure 6: MMBT architecture

We took advantage of MMBT architecture flexibility and replaced ResNet with CLIP for image encoding. CLIP is a state of the art vision-language model that was originally developed by OpenAI. It's original pretraining task was to match a caption with an image. It achieves this by jointly learning image and text encodings and projecting them into a shared latent space.CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in a dataset. Our assumption was that features from CLIP are more versatile and better suited for a multimodal domain.

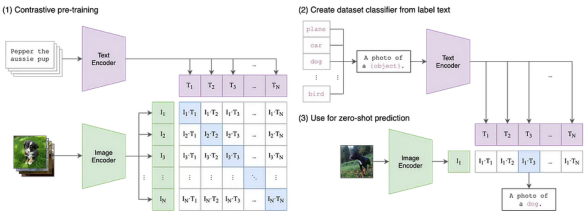

Figure 7: CLIP

For text encoding we used bert-base-uncased-hatexplain model which is available in Huggingface Hub. This model was created for hatespeech detection in English, so in our case features from it are better than from bert-base-uncased that was used in MMBT initially. Figure 8 shows the picture of what eventually the model looks like .

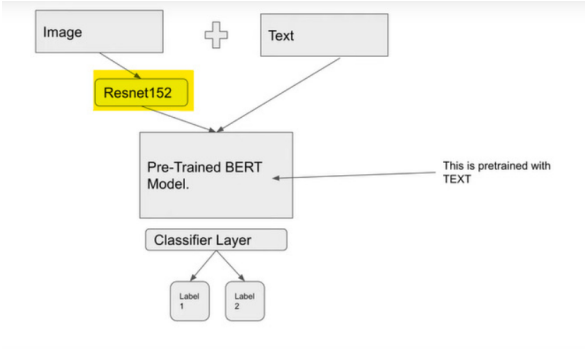

Figure 8: Final Modal - Resnet is replaced with CLIP

3.4 Augmentation And Pre-processing

We have not done much augmentation or pre-processing , but according to our study , this step improves accuracy a lot .

1. Removing OCR from images.
2. Face detection using Haar features.
3. Face sentiment detection.
4. Entity detection.
5. Getting similar images using Getty Images.
6. Inpainting images.

Below image shows how above augmented tags can be applied to extended multi-modal networks.

Figure 9: Extended Vilbert

3.5 Results

1. BERT trained on Hate speech

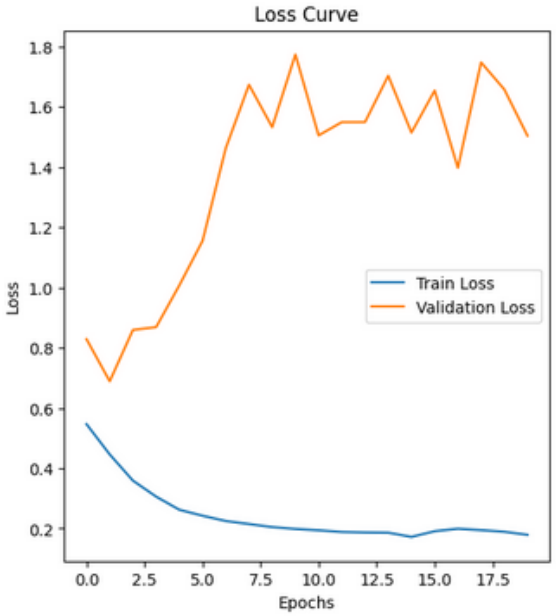

Figure 10: BERT loss curve

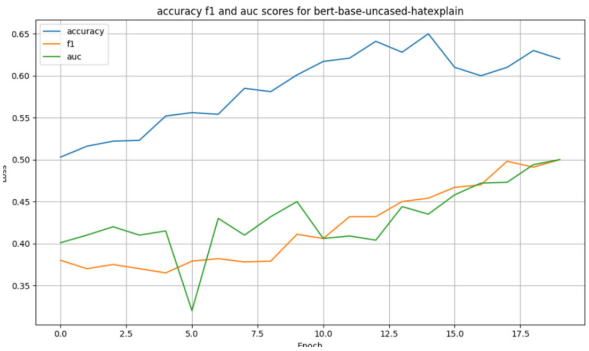

Figure 11: BERT AUC is 0.48

2. VisualBERT AUC score

```
training the model
Epoch 1/30, Training Loss: 0.6671, Validation Loss: 0.0430 ,Validation AUC: 0.5682
Epoch 2/30, Training Loss: 0.6919, Validation Loss: 0.0393 ,Validation AUC: 0.5115
Epoch 3/30, Training Loss: 0.6974, Validation Loss: 0.0402 ,Validation AUC: 0.4521
Epoch 4/30, Training Loss: 0.6945, Validation Loss: 0.0405 ,Validation AUC: 0.5467
Epoch 5/30, Training Loss: 0.6938, Validation Loss: 0.0385 ,Validation AUC: 0.5003
Epoch 6/30, Training Loss: 0.6962, Validation Loss: 0.0390 ,Validation AUC: 0.4853
Epoch 7/30, Training Loss: 0.6938, Validation Loss: 0.0391 ,Validation AUC: 0.4876
Epoch 8/30, Training Loss: 0.6944, Validation Loss: 0.0406 ,Validation AUC: 0.5032
Epoch 9/30, Training Loss: 0.6957, Validation Loss: 0.0394 ,Validation AUC: 0.4877
Epoch 10/30, Training Loss: 0.6927, Validation Loss: 0.0386 ,Validation AUC: 0.5077
Epoch 11/30, Training Loss: 0.6933, Validation Loss: 0.0405 ,Validation AUC: 0.4837
Epoch 12/30, Training Loss: 0.6937, Validation Loss: 0.0391 ,Validation AUC: 0.5277
Epoch 13/30, Training Loss: 0.6924, Validation Loss: 0.0403 ,Validation AUC: 0.4317
Epoch 14/30, Training Loss: 0.6925, Validation Loss: 0.0396 ,Validation AUC: 0.5016
Epoch 15/30, Training Loss: 0.6925, Validation Loss: 0.0394 ,Validation AUC: 0.5397
Validation loss did not improve. Stopping early.
```

Figure 12: VisualBERT AUC score is 0.53

3. MMBT and CLIP

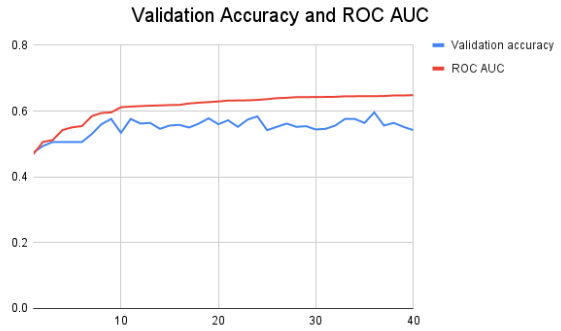

Figure 13: CLIP AUC is 0.65

Above results show that CLIP and MMBT have outperformed VisualBERT . So if we have an ensemble of weak classifiers and augmented data. This should do equally well with Meta competition toppers.

Table 1: Model AUROC Scores

Model Type	Model	AUROC
Unimodal	BERT	0.48
Multimodal single stream	VisualBERT	0.53
Multimodal double stream	MMBT + CLIP	0.65

Code was run on Kaggle using GPU P100.VisualBert took most time , approx 45 minutes for one epoch.

Limitations

Our work has a few limitations.

1. We conducted our project on English meme datasets and did not assess the model's capability for multilingual hate meme detection.

Figure 14: Non English Meme

2. The work presented sometimes suffers from the lack of real-world knowledge. It fails to detect certain symbolism and is not aware of real-word persons.

Figure 15: Real world famous people

3. The models finds it hard to understand the Cultural, Political, Societal References, Traditional attires, and Religious Practises.

Figure 16: Religious Meme

4. The task of hateful meme detection has only two classes - hateful and not hateful. It is hard to perform one-shot or few shot classification for two classes. We can incorporate more classes to facilitate active learning.

5. The fonts and text formats on the memes in reality are of varying sizes and shapes. In order to accommodate varying text formats, we need to integrate the OCR and make it trainable end-to-end with the VL model.

Figure 17: Hard to capture OCR

Future Work

We fill this section with important questions.

1. Can we extend this for hateful GIF detection ?

2. Can we integrate OCR and entity detection in our models ?

3. Can we integrate GenAI for better augmentation ?

4. We think RAG can improve performance of these models.

5. A better and more flexible source of knowledge that provides in-depth information about the entity that appears in the meme should help reduce the performance gap between machine and human . To extract common sense from large scale knowledge like wiki-data that usually have dozens of type of relation per entity . Thus building a subgraph that both includes enough information and has a reasonable size will be challenging.Modals like MHGRN and K-BERT can be used.

Related Work & Further Readings

1. Visual language modals are used in various prompt settings to focus on zero-shot classification of hateful/harmful memes. [6] Has observed that large VLMs are still vulnerable in zero-shot hate meme detection.

2. [1] External adversarial attacks can affect the inference of these multi modals leaving social media more prone to hatred.In order to counter such attacks Cited paper proposed two different methods ,contrastive learning and adversarial training and found that an ensemble of these two methods work well for a large majority of attacks for two of the three datasets.

Conclusion

In our project we found out that augmentation and ensemble learning were heavily used to increase accuracy for unseen data. Leaving that aside we focused on the mutimodal modals individually to increase the accuracy.Modals like CLIP trained on both text and image simultaneously are better encoders than BERT or CNN.Further we found out the limitations which can be solved as a RAG application , modals trained on Open QA can help reduce the gap between human level accuracy .

Ethical statement

we do not intend to harm any individuals or target communities. Objective of our project wad to identify hate memes only.

References

[1] Piush Aggarwal et al. "HateProof: Are Hateful Meme Detection Systems really Robust?" In: *Proceedings of the ACM Web Conference 2023*. WWW '23. ACM, Apr. 2023. DOI: 10.1145/3543507.3583356. URL: http://dx.doi.org/10.1145/3543507.3583356.

[2] Douwe Kiela et al. "Supervised Multimodal Bitransformers for Classifying Images and Text". In: *CoRR* abs/1909.02950 (2019). arXiv: 1909.02950. URL: http://arxiv.org/abs/1909.02950.

[3] Liunian Harold Li et al. "VisualBERT: A Simple and Performant Baseline for Vision and Language". In: *CoRR* abs/1908.03557 (2019). arXiv: 1908.03557. URL: http://arxiv.org/abs/1908.03557.

[4] Niklas Muennighoff. "Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes". In: *CoRR* abs/2012.07788 (2020). arXiv: 2012.07788. URL: https://arxiv.org/abs/2012.07788.

[5] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[6] Naquee Rizwan et al. *Zero shot VLMs for hate meme detection: Are we there yet?* 2024. arXiv: 2402.12198 [cs.CL].

[7] Riza Velioglu and Jewgeni Rose. “Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge”. In: *CoRR* abs/2012.12975 (2020). arXiv: 2012.12975. URL: https://arxiv.org/abs/2012.12975.

[8] Ron Zhu. “Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution”. In: *CoRR* abs/2012.08290 (2020). arXiv: 2012.08290. URL: https://arxiv.org/abs/2012.08290.