# Multimodal Deep Ensemble for Hateful Meme Identification


## Abstract

This paper delves into the utilization of machine learning techniques for identifying hate speech, while addressing the persisting technical challenges to enhance their performance to match human-level accuracy. We explore several current visual-linguistic Transformer models and suggest enhancements to boost their effectiveness for this task. The model we propose demonstrates superior performance compared to the established benchmarks, achieving a 5th place ranking out of over 3,100 participants.


## 1 Introduction

This paper addresses the critical influence of the internet on our daily lives, where our online presence showcases our personalities and beliefs, as well as our biases. Daily, billions of individuals engage with various forms of online content, and despite some of this content being valuable and informative, an increasing portion is harmful, including hate speech and misinformation. There is a growing need to quickly detect this content, improve the review process and automate decisions to rapidly remove harmful material, thereby reducing any harm to viewers.

Social media platforms are frequently used for interactions, sharing messages and images with private groups and the public. Facebook AI launched a competition to tag hateful memes that include both images and text. For this, a dataset of 10,000+ labeled multimodal memes was provided. The aim of the challenge is to develop an algorithm that identifies multimodal hate speech in memes, while also being robust to their benign alterations. A meme's hateful nature could stem from its image, text, or both. Benign alteration is a technique used by organizers to switch a meme's label from hateful to non-hateful, requiring modifications to either the text or the image.

The core assessment metric for this binary classification task is the area under the receiver operating characteristic curve (AUROC), representing the area under the ROC curve. This curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The primary objective is to maximize the AUROC.

$$AUROC = \int_0^1 TPR(T)dFPR(T) \tag{1}$$

Accuracy is the secondary metric, calculating the proportion of instances where the predicted class matches the actual class in the test set.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y}_i) \tag{2}$$

The aim is to maximize both metrics.

In brief, this paper makes three contributions:

.

- We conduct experiments using single-stream and dual-stream architectures such as VL-BERT, VLP, UNITER and LXMERT and compare their performance with the established baselines. These models were chosen because of their pre-training on diverse datasets.
- We put forward a novel bidirectional cross-attention mechanism that connects caption information with meme caption text, which increases performance in detecting hateful memes. This is similar to the cross-attention between images in other research.
- We demonstrate that deep ensembles greatly improve single model predictions.

## 2 Related Work

Transformer models pre-trained on extensive datasets have shown state-of-the-art results in numerous language processing tasks. BERT is one of the most popular due to its ease of use and strong performance. Recently, training these large models on combined visual-linguistic embeddings has shown very promising outcomes for visual-linguistic tasks such as visual question answering, reasoning, and image captioning. LXMERT uses dual networks to process text and images, learning cross-modality encoder representations by using a Transformer to combine the two streams of information. The images' features are derived using a Faster R-CNN feature extractor. This is also used in single-stream architectures, VL-BERT and UNITER, which employ a single Transformer on top of the combined image-text embeddings. A unified model for visual understanding and vision-language tasks has also been proposed.

Table 1: Pre-training datasets for each model

| | Books Corpus | CC | COCO | VG | SBU | GQA | VQA 2.0 | VG-QA |
|---|---|---|---|---|---|---|---|---|
| VL-BERT | X | | | | | | | |
| VLP | X | | | X | | | | |
| UNITER | X | X | X | X | | | | |
| LXMERT | X | X | X | X | X | | | X |

A dataset for multimodal hate speech detection was created by gathering data from Twitter, using particular hateful keywords. However, studies found that multimodal models did not do better than text-only models.

## 3 Methodology

One goal of this research is to leverage the fact that single and dual stream Transformer models have been pre-trained on a variety of datasets across various fields. Transformer attention models excel at NLP tasks, and the masked language modeling pre-training method in BERT is both powerful and versatile. Studies show that the pre-training process can better align visual-linguistic embeddings and help downstream tasks like visual question answering and reasoning. Given that pre-training a visual-linguistic Transformer architecture is helpful for downstream tasks, might ensembling different models pre-trained on different datasets yield better results?

Table 1 shows the pre-training datasets used for each model.

### 3.1 UNITER with Meme Text and Inferred Caption Cross-Attention

The Natural Language for Visual Reasoning for Real (NLVR2) is an academic dataset of human written sentences connected to pairs of photos. The dataset includes pairs of visually intricate images coupled with a statement and a binary label. UNITER was among the top models in this challenge by adding a cross-attention module between text-image pairs, dividing each sample in two and repeating the text. They then apply attention pooling to each sequence, concatenate them and add the classification head, a multi-layer perceptron. Similar to this, we propose to repeat the meme image in each half-sequence and add an inferred meme caption as the second text. We generate captions using the Show and Tell model. This way, the model could learn from both the original meme text and the new captions generated by a model trained on a different dataset.

## 4 Experiments

We carry out several experiments using LXMERT, VLP, VL-BERT, and UNITER. We apply bidirectional cross-attention using inferred captions for UNITER, VL-BERT, and VLP, but not for LXMERT due to its low performance on the dataset.

We also experiment with a dataset from previous research. We filter and balance it down to 16K samples by excluding cartoon memes and memes with little text. We fine-tune VL-BERTLARGE using the reduced dataset for four rounds, then fine-tune it using the hateful memes dataset for another four rounds. The results were lower than the majority of the other models.

The baselines for models trained on the Hateful Memes dataset are in Table 2.

## 5 Results

Our best performing solutions are derived from averaging probabilities using a single VL-BERTLARGE and one UNITERLARGE+PA (UNITERLARGE with extra attention). We used the default training parameters of the vanilla pre-trained UNITERLARGE model, but changed the training steps according to the dataset size. A deep ensemble of UNITERLARGE+PA models got the best performance. For this ensemble, we simply rerun training using various random seeds and average the predictions from each model. Table 2 displays the top results for the final competition phase as well as the improvements cross-attention brings to the UNITER model in the first phase. The final results are significantly better than the baselines.

The most important findings are as follows:

- Single-stream Transformer models pre-trained on the Conceptual Captions (CC) dataset give the best results, and deep ensembles improve the overall performance further. The choice of pre-training datasets matters in terms of domain similarity to the fine-tuning dataset.
- We believe that UNITER gets better results due to being pre-trained on the COCO dataset which has less noise. Similarly to the Hateful Memes dataset this is also high quality. Further work should investigate if pre-training VL-BERT on COCO would improve its results.
- Interestingly, the paired attention technique only works for UNITER and not for the other models.
- Training large models from scratch did poorly, which is expected due to the small dataset size.
- The dataset of multimodal hate speech is heavily skewed towards hateful text and the keywords used to collect it. The memes are less subtle compared to the ones in the Hateful Memes dataset, although they are perhaps more typical of what is seen online.

## 6 Conclusion

We present effective techniques to detect hate speech in a distinct dataset of multimodal memes from Facebook AI. The aim is to identify hate speech using a multimodal model, and to be robust to the "benign confounders" that cause the binary label of a meme to change.

We have performed tests on various large pre-trained Transformer models and fine-tuned state-of-the-art single-stream models like VL-BERT, VLP, and UNITER, and dual-stream models like LXMERT. We compare their performance against the baselines, showing that the single-stream models perform significantly better. Our choice for these models stems from their pre-training on a wide variety of datasets from different fields. We also adapt a novel bidirectional cross-attention mechanism that links caption information with meme text. This leads to increased accuracy in identifying hateful memes. Furthermore, deep ensembles can improve single model predictions. Training the models from scratch performed poorly due to the small dataset size. We also observed that the pre-training dataset influences results.

We conclude that despite the improvements in multimodal models, there is still a gap when comparing to human performance. This suggests considerable scope for the development of better algorithms for multimodal understanding.

Table 2: Baselines from previous research. For our final models, we report the top performance scores, specifying both Accuracy and AUROC results.

| Type | Model | Acc. | Validation AUROC | Acc. | Test AUROC |
|---|---|---|---|---|---|
| | Human | – | – | 84.70 | 82.65 |
| 3*Unimodal | Image-Grid | 52.73 | 58.79 | 52.00 | 52.63 |
| | Image-Region | 52.66 | 57.98 | 52.13 | 55.92 |
| | Text BERT | 58.26 | 64.65 | 59.20 | 65.08 |
| | Late Fusion | 61.53 | 65.97 | 59.66 | 64.75 |
| Multimodal 5* (Unimodal Pretraining) | Concat BERT | 58.60 | 65.25 | 59.13 | 65.79 |
| | MMBT-Grid | 58.20 | 68.57 | 60.06 | 67.92 |
| | MMBT-Region | 58.73 | 71.03 | 60.23 | 70.73 |
| | ViLBERT | 62.20 | 71.13 | 62.30 | 70.45 |
| | Visual BERT | 62.10 | 70.60 | 63.20 | 71.33 |
| Multimodal 2*(Multimodal Pretraining) | ViLBERT CC | 61.40 | 70.07 | 61.10 | 70.03 |
| | Visual BERT COCO | 65.06 | 73.97 | 64.73 | 71.41 |
| 3*(Phase 1) | UNITER | – | – | 68.70 | 74.14 |
| | UNITERPA | – | – | 68.30 | 75.29 |
| | UNITERPA Ensemble | – | – | 66.60 | 76.81 |
| 2*(Phase 2) | VL-BERT + UNITERPA | 74.53 | 75.94 | 73.90 | 79.21 |
| | UNITERPA Ensemble | 72.50 | 79.39 | 74.30 | 79.43 |