Hateful Memes Classification: Deliverable-II Report

Akshay Goindani (20171108) Preet Thakkar (20171068) Vaibhav Garg (20171005)
Sagar Joshi (2020701007)
Team 6

April 1, 2021

1 Problem Recap

Problem Statement: The goal is to predict whether a meme is hateful or non-hateful. This is essentially a binary classification problem with multimodal input data consisting of the the meme image itself (the image mode) and a string representing the text in the meme image (the text mode).

Input and Output: Given a rgb meme image, and a string representing the english text in the meme image, the trained model will output the probability that the meme is hateful.

Motivation:

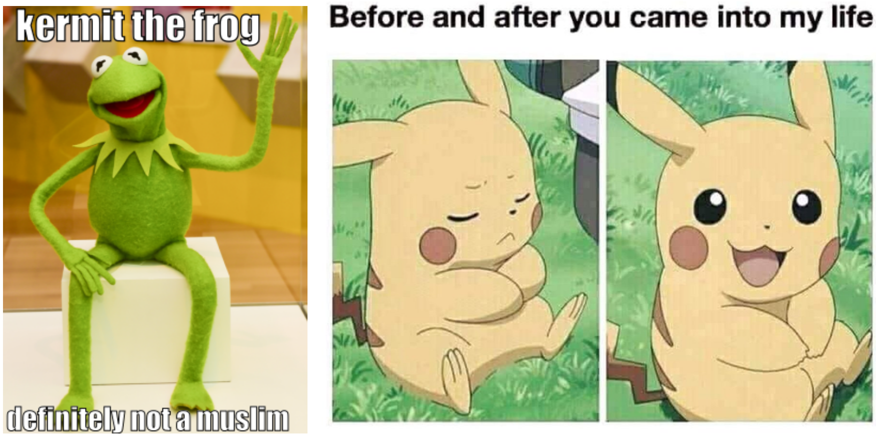

Figure 1: Both the images are fairly popular memes but clearly, one induces hate and negativity whereas the other one promotes love and positivity.

Internet memes have become a major form of communication and expression, they are an essential part of social media's popular culture. Modern-day social media platforms are full of memes as they are very easy to consume and also have the associated humor aspect. Figure 1 contains examples of such commonly shared memes. Moreover, these platforms have made them incredibly easy to share, two clicks are literally all it takes to make sure all your friends/followers have the same meme on their feeds. But like with any other type of content in social media, they also can lead to the spread of hate speech and propaganda. At the end of the day, these memes become an indirect source for our information consumption and play a significant role in shaping our belief systems.

The scale at which such content is produced makes it impossible for humans to go through all the memes and filter out the distasteful ones and hence there is an urgent need for fast and scalable automated solutions. Facebook AI's Hateful Memes Challenge [1] was one of the very first large-scale attempts to explore this problem and the results prove just how challenging this problem is, interpreting a meme requires a lot of cultural, political, and contextual knowledge and multimodal understanding since it is both the text and the image combined that brings out the real meaning of the meme. Another important thing to note here that even the human annotators themselves had a hard time classifying the memes and an average time of 27 minutes was spent per meme in the annotation process. And lastly, even though the results from previous research might look promising, but at the scale of the internet that still equates to a lot of hateful memes still being shared openly.

2 Work Done Till Previous Deliverable

In the previous deliverable, we had successfully managed to replicate results for the visualBert architecture [5] and the vilio architecture [3].

Dataset	Accuracy	AuROC
Dev Seen	71.2%	0.81
Dev Unseen	71.3%	0.75
Test Seen	66.2%	0.74

Table 1: VisualBert Model results, Accuracy and AuROC score computed on dev seen, dev unseen and test seen using the ensemble of top 27 models using majority voting strategy.

Dataset	Accuracy	AuROC
Dev Seen	73.2%	0.81
Dev Unseen	72.0%	0.80
Test Seen	73.6%	0.83

Table 2: Vilio Ensemble Performance

While doing our model comparisons, due to its much superior performance, we had finally decided to move ahead with the Vilio which uses a rich variety of V+L models: UNITER, OSCAR, VisualBERT (single-stream) and ERNIE-ViL(dual-stream). Some of the reasons we that can explain the noticeably superior performance of Vilio are as below:

1. Use of better feature extraction algorithm for images
   As compared to the baseline models and the models in VisualBERT ensemble which use ResNet-152 or Faster-RCNN based features for input, Vilio ensemble uses features extracted using the detectron2 framework from Facebook which is the state-of-the-art in object detection. The models used for feature extraction were pretrained on the Visual Genome dataset.

2. Diversity in feature set per model type
   Each model is trained on 3-5 different feature sets (3 for OSCAR, UNITER, VisualBERT, 5 for ERNIE-ViL small and large) which are essentially features with varying Regions of Interest (RoIs) having the no of min/max boxes set to either 36, 50, 72 or 100. The idea behind using different features per model was that using diverse features can improve the performance. After each individual model of a model type was trained on one particular feature set, the prediction probabilities of the model were simple-averaged for use in further ensembling.

3. Diverse model architecture
   As mentioned before, Vilio combines multimodal transformers with different architectures in the family. It includes three different single-stream transformers (OSCAR, UNITER, VisualBERT) and two dual stream transformers with different sizes (ERNIE-ViL - Small & Large) as mentioned before. Even among the 5 models trained individually for each variant of ERNIE-ViL, two of the models pretrained on the VCR dataset were used in addition to the pretrained models on Conceptual Captions dataset to further push diversity.

4. Task-adaptive pretraining
   VisualBERT and OSCAR models were pretrained on the hateful memes dataset before finetuning on the classification task. Image-text matching (given a text and the image, determine if they are related) and masked language modelling (predict the masked token with the help of unmasked tokens and image features) were used for OSCAR, while VisualBERT was pretrained using masked language modelling. In general, for different tasks across literature, task-adaptive pretraining has always found to boost the performance since it adapts the model better to the domain of the dataset, so this was an important step in optimizing the model performance on the task.

5. Better learning strategies & ensembling
   Apart from the above mentioned points highlighting the superior backdrop provided for training, the training process also employed different learning strategies for enhancing performance. One of the interesting ideas was the use of Stochastic Weighted Averaging for the last 25% of training. The final ensembling employs

superior ensembling strategies as compared to the other models as well, in which the simple-averaged predictions of models are run through an ensembling loop of simple averaging, power averaging, rank averaging and simplex optimization.

Overall, there are multiple aspects in the Vilio family of ensemble in various stages from feature selection to finetuning that are well-engineered as compared to previously encountered strategies. All these factors together explain the large gap in performance, which is closest to human benchmark on the hateful memes dataset.

3 Work Done after Previous Deliverable

In the previous deliverable, we implemented the 2 papers ([3], [2]) that we mentioned in our project scope document. Both the approaches used ensemble of different models, for predicting the final label. After the previous work, our major focus was on exploring different ensemble techniques that can be used to enhance the performance of the aforementioned models. We tried different ensemble methods and we found that the Mixture of Experts (MOE) ensemble method lead to a small improvement in the Accuracy and AUROC score for development but the score for test set decreased by a small margin.

3.1 Training Additional Models

We trained two additional models - DeVLBERT[6] and LXMERT[4] on the hateful memes dataset from their implementations available in PyTorch and added to the Vilio ensemble in order to attempt improvement in score. Both are dual stream multimodal transformer architectures. A brief description of their architectures is given below.

3.1.1 DeVLBERT[6]

DeVLBERT architecture attempts to de-confound visuo-linguistic representation by investigating the problem of out-of-domain pretraining and shows the effectiveness of its approach in direct comparison with VilBERT. Spurious correlations are found to exist in models because of confounders present in the data used for pretraining. They imply high conditional probability of one token given another in absence of any robust relationship between them, for instance, ViLBERT model shows an unusually high conditional probability for the visual object 'shirt' given the word 'instrument'. By introducing the idea of backdoor adjustment from causal inference, the architecture mitigates this issue of spurious correlations and improves the generalization capability for out-of-domain pretraining data.

3.1.2 LXMERT[4]

LXMERT is composed of three different transformer models in its architecture - language encoder for linguistic modality, object relationship encoder for image modality, and cross-modality encoder for the fusion of two modalities. Unlike the ViLBERT architecture which has both the transformer streams for both - image and text intertwined through the layers, in this architecture, the modalities are separated from each other for some layers, after which the cross-modality encoder enables multimodal representation learning. For pre-training, data from five vision-and-language datasets is used whose images come from COCO or Visual Genome datasets while also aggregating three image captioning datasets - VQA 2.0, GQA and VG-QA. Three types of pretraining tasks are performed:

1. Masked Cross-Modality LM for Language Representation
   This task is similar to BERT masked language modelling where 15% of the tokens are randomly masked, however, for LXMERT, the cross-modality encoder enables making use of image embeddings to predict the masked tokens.

2. Masked Object Prediction for Visual Representation
   Here, the model learns on two tasks: RoI-feature regression and Detected-label classification. In the former, random 15% of the objects in the image object detection input are masked by replacing the RoI feature values with zeros, and the model regresses the feature value. In the later task, the model has to predict the labels for the masked objects. Output of Raster R-CNN is used as the noisy ground truth for this pre-training.

3. Cross-Modality Tasks

(a) Cross-Modality Matching
Given a sentence and an image, the model has to predict if they match each other, the probability of which is 0.5.

(b) Image Question Answering (QA)
Given a question related to an image and the image itself, the model has to predict its answer. Ablation study shows an impactful contribution of this task in improving the accuracy on three different finetuning tasks - VQA, GQA, NLVR.

3.2 Mixture of Experts

In the MOE ensemble method, output probabilities from multiple models are used to make the final prediction. Let there be N models. For each model, we have the output probabilities for a given input. There is a gating network which is basically a Multi Layer Perceptron (MLP), whose objective is to predict N normalized scalars (α_i). The final probability of the input being hateful is a weighted sum of the probabilities from the individual models, where the weights (α_i) are calculated using the MLP. We use the final probability(P) to calculate the loss. The loss function that we used was Binary Cross Entropy with Logit Loss. To make the final prediction, we applied a sigmoid activation over the final probability(P) and used a threshold value of 0.5 to classify the input into hateful.

$$O = (O^1, O^2, ..., O^N) \tag{1}$$

$$\alpha = MLP(O) \tag{2}$$

$$P = \sum_{i=i}^{N} (\alpha_i \times O^i) \tag{3}$$

Here, O_i is the output probabilities from the i^{th} model and P is the weighted sum of individual probabilities.

4 Results and Analysis

Dataset	Accuracy	AuROC
Dev Seen	73.6%	-
Test Seen	79.5%	0.732

Table 3: MOE Ensemble Performance, we can see the improvement in Dev Seen Accuracy from previous best

Ensemble	Dataset	AuROC
VisualBERT, UNITER, OSCAR, ERNIE-ViL	*Dev Seen*	0.8119
	Test Seen	0.8254
VisualBERT, UNITER, OSCAR, ERNIE-ViL, ***LXMERT, DeVLBERT***	*Dev Seen*	0.8123
	Test Seen	0.8221

Table 4: AuROC comparison of Vilio model with an extended ensemble created by including DeVLBERT and LXMERT models

We can see from the results in table 3 that the MOE ensemble does not increase the performance by a large margin. For Dev Seen set, the accuracy improves by a very small margin (0.4%) while for test seen test, the accuracy decreases. From our experiments, we found that the optimal values for α_i are similar to each other, this implies that the gating network is giving equal weight to all the individual models. After analysing performance of individual models in the ensemble, we found that the accuracy and AUROC score for those models were almost similar. Due to the similar peformance of individual models, the gating network is assigning equal weight to all of them because no individual model outperforms the other by a high margin.

It can be inferred from table 4 that after including two new models in the original Vilio ensemble, the overall performance of the model deteriorates (marginally). Hence, we will proceed with the original Vilio ensemble for future deliverables.

5 Next Deliverable

As planned in our timeline, we'll try to explore the same problem space but with Hindi text captions for the remaining period of the course. We'll try to collect some Hindi memes from the internet, since, hateful meme detection models for other languages are not currently available therefore, we will first translate the caption to English and run through our vilio ensemble model and try to make some insights about the performance.

References

[1] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

[2] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[3] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020.

[4] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019.

[5] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020.

[6] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert. *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020.