Hateful Memes Classification: Final Report

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

Presentation: Google Slides
Code Repository: GitHub

April 17, 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:

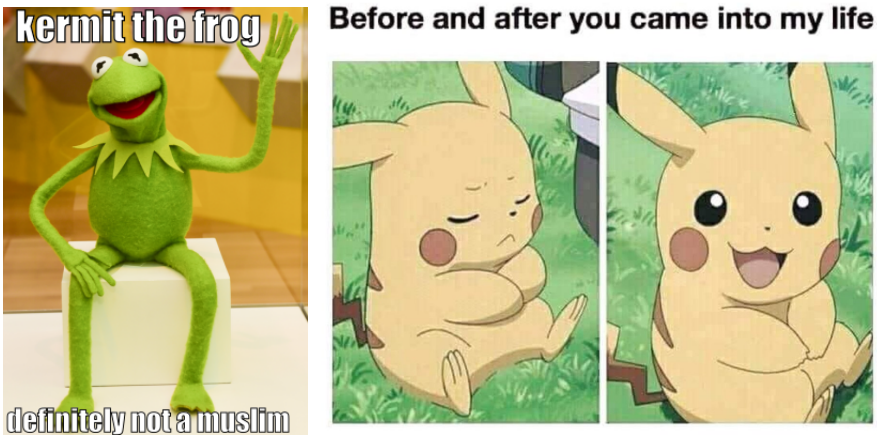

Figure 1: Both the images are fairly popular memes but clearly, one induces hate and negativity whereas the other one promotes love and positivity.

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 [7] 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 Our Project

2.1 Literature Survey

In the first phase of our project we started reading relevant literature available on Hateful Meme Classification. During our literature survey we found multiple papers that attempt to solve the problem. Since, this problem involves understanding both the modalities together, therefore we considered only those papers that propose a multimodal approach because unimodal approaches focus only on a single modality i.e., either image or text. Brief summary of some of the papers is listed as follows:

1. **Detecting Hate Speech in Multimodal Memes [2]**
   In this paper, it was observed that the majority of the data points in the Facebook Meme Dataset [7] which

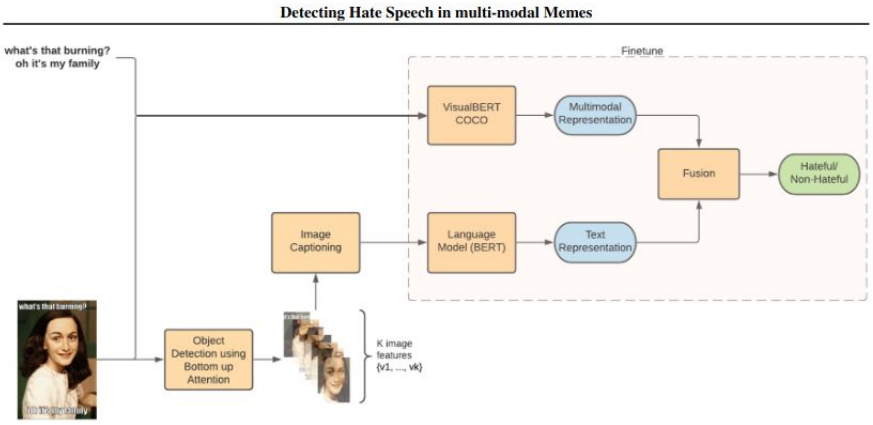

Figure 2: Classification Pipeline

   are originally hateful are turned into benign just by describing the image of the meme. This is done because in such a scenario, unimodal models that focus only on the textual or visual modality will fail and only the multimodal models will be able to learn true reasoning. Many multimodal models give more preference to hate speech (language modality). Therefore image captioning and object detection are used to focus more on the visual modality. Using image captioning and object detection, the actual caption of the meme is extracted and the features from the actual caption are combined with multimodal representation to make the final prediction. In addition to the above method, unimodal sentiment features (for both image and text) are also used along with the multimodal features from pre-trained networks. This is done to capture the context and relationship between the two modalities. The complete classification pipeline is shown in Figure 2.

2. **Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge [20]**
   In this paper, VisualBERT [10] is used to get multimodal representation. The approach is divided into 4 phases - dataset expansion, image encoding, training, and ensemble learning. In the dataset expansion phase, some memes from the Memotion Dataset are selected. The selection is based on the similarity of the memes with the ones in the Hateful Meme Dataset. In the Image Encoding phase, 2048 dimensional region-based image features are extracted using ResNeXT-152 based Mask-RCNN model. These features are then projected to the textual embedding space. A pre-trained visualBert model is used to get the multimodal representations. The VisualBert model is fine-tuned during the training phase. For the alignment between the textual modality and visual modality, the self-attention from the transformer model is used. Image regions and the language are combined and used as an input to the transformer. Classification is done using softmax over the output of a linear transformation (Wx + b) on the output of the transformer. For ensembling, different models are taken and a majority voting strategy is used to predict the final class label.

3. **Multimodal Learning for Hateful Memes Detection [24]**
   In this paper, the proposed model consists of an image captioner, an object detector, a triplet-relation network,

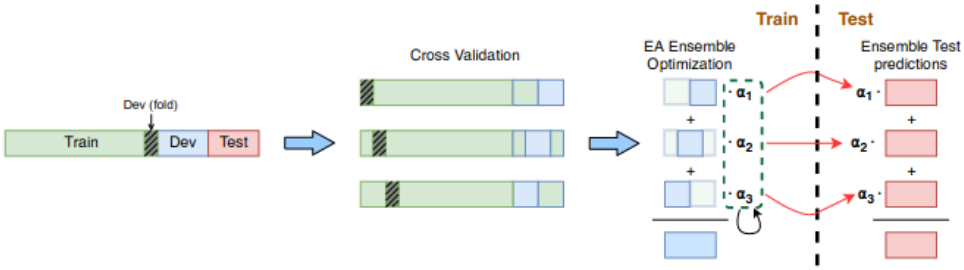

Figure 3: Classification Pipeline

and a classifier. They consider three different knowledge extracted from each meme: image caption, meme text, and visual features (using object detection). The proposed triplet-relation network models the triplet-relationships among caption, objects, and OCR sentences, adopting the cross-attention model to learn the more discriminative features from cross-modal embeddings. Sentence embeddings are generated by first extracting the image features with an image encoder and then visual features are decoded into a sentence using a sentence decoder. Finally, image caption embeddings and OCR text embeddings are concatenated. Image embeddings are generated using a pre-trained Faster R-CNN. The triplet-relation network is essentially a transformer network that is used to model the cross-modality relationships between image features and two textual features. A joint representation of the textual and visual content obtained from the transformer model is fed to a fully connected layer followed by a softmax layer to get the prediction probability. A binary cross-entropy (BCE) loss function is used as the final loss function for meme detection. The classification pipeline is shown in the Figure 3.

4. **Vilio: State-of-the-art Visio-Linguistic Models Applied To Hateful Memes [13]**
   In this paper, the proposed model uses an ensemble technique. In the first step, image features are extracted using the detectron2 framework. Next, meme text together with the features are fed into the ensemble models. The ensemble contains five models: ERNIE-ViL (small, large), OSCAR, UNITER, and VisualBERT.

2.2 Implementing current SoTA models

We will primarily use the Facebook AI's Hateful Meme Dataset [7] for training and testing our proposed models. The dataset is created in such a way that, uni-modal approaches that use only the image or text features will fail. Therefore the proposed approaches are multi-modal i.e., they use both image and text features. There were 11 baselines provided in the paper including both - unimodal and multimodal approaches, all of which we tried to replicate the results.

Among the 11 models used as baselines, three are unimodal approaches (Image-Grid, Image-Region, Text BERT) making use of either image or text modality to make the classification. There are five multimodal architectures (Late Fusion, Concat BERT, MMBT-Grid, MMBT-Region, ViLBERT, Visual BERT) but pretrained on unimodal objective and two multimodal architectures (ViLBERT CC, Visual BERT COCO) pretrained on multimodal objective. For image features in input, ResNET-152 and Faster-RCNN with ResNeXt-152 based features are used and for textual modality, BERT tokenizer is used. A quick idea of each of the baseline architectures is given below.

- Image-Grid
  Unimodal image-based classifier which uses convolutional features with average pooling from ResNet-152 architecture.
- Image-Region
  Unimodal image-based classifier which uses features from Faster-RCNN with ResNeXt-152 as the backbone network, and is pretrained on the Visual Genome dataset.
- Text BERT
  Unimodal text-based approach which uses BERT embeddings on the text given as a part of the hateful memes dataset.
- Late Fusion
  A simple multimodal approach where output of ResNet-152 as in Image-Grid and BERT based models is taken unimodally, and their mean is taken as model output.

- Concat BERT
  In this multimodal approach, an earlier fusion of the output of the unimodal ResNet-152 and BERT embeddings is performed by concatenation and an MLP is trained for classification.

- MMBT-Grid
  MMBT is a multimodal supervised bitransformer architecture that consists of individual unimodally pretrained components that is trained to map multimodal image embeddings to text token space. MMBT-Grid uses features from ResNet-152 for image embeddings.

- MMBT-Region
  In this approach, the MMBT transformer uses features from Faster-RCNN as in Image-Region for image embeddings.

- ViLBERT
  ViLBERT is a dual stream multimodal transformer architecture. Here, the VilBERT model without any multimodal pretraining is used. It has BERT initializations for the text stream and uses Faster-RCNN pretrained on Visual Genome dataset to extract image region features.

- Visual BERT
  Visual BERT is a multimodal single stream transformer architecture in which the text and image inputs are jointly processed by a stack of BERT-based transformer layers. It uses Faster RCNN for extracting image features. For this baseline, Visual BERT without any pretraining on multimodal tasks is used.

- ViLBERT CC
  ViLBERT architecture used here is pretrained multimodally on the Conceptual Captions (CC) dataset using two pretraining tasks - masked multi-modal modelling (masking 15% of text and image region inputs and reconstructing them with unmasked inputs) and multi-modal alignment prediction (given a pair of image and text, determine if the text describes the image).

- Visual BERT COCO
  For this baseline, Visual BERT architecture is pretrained multimodally on the Common Objects in Context (COCO) dataset. The two tasks the model is pretrained on are masked language modelling with image (some part of text is masked, and is to be predicted using image regions and unmasked text) and sentence-image prediction (given two captions for an image, while one of them is the right caption for the image, determine if the same holds for the remaining caption as well).

Pretrained models and the codebase and environment needed to replicate the results were provided in the mmf framework by Facebook AI Research. We first tested the performance of the models directly on the pretrained weights, after which we trained all the 11 models to see improvement in performance. The hyperparameters used were the same as those provided in the challenge paper. The models were trained or finetuned for a maximum of 22000 updates, independent of the batch size.

Type	**Model**	**Accuracy**	**AuROC**
Unimodal	Image-Grid	59.8%	0.54
	Image-Region	58.0%	0.55
	Text BERT	60.0%	0.58
Multimodal (Unimodal Pretraining)	Late Fusion	63.3%	0.62
	Concat BERT	63.9%	0.63
	MMBT-Grid	64.3%	0.64
	MMBT-Region	66.5%	0.70
	ViLBERT	69.6%	0.69
	Visual BERT	69.2%	0.71
Multimodal (Multimodal Pretraining)	ViLBERT CC	68.9%	0.70
	Visual BERT COCO	69.3%	0.72

Table 1: Results on finetuned weights of baseline models on *val* dataset

Table 1 shows the accuracy and AuROC scores after finetuning. Some of our observations from the process which are consistent with those in the challenge paper are stated below.

- The increase in accuracy and AuROC scores was too little after finetuning, and for worse was negative for some of the baseline models, implying there has to be a better way for training with multimodal objectives.

- Among multimodal architectures, the difference is not much between the performance of unimodally pretrained and multimodally pretrained models, implying the pretraining of multimodal models can be improved.
- The performance improves with more advanced fusion of the two modalities. For instance, after finetuning, the performance of ConcatBERT beats that of Late Fusion. It is in turn superseded by multimodal architectures like MMBT, ViLBERT, VisualBERT.
- Human AuROC of 0.8265 is still very far from the best baseline AuROC of 0.72. Hence, filling in this gap becomes a guiding force in trying out with further techniques in the project.
- Early occurrences of best update in many models indicates overfitting problem. Given the relatively small size of data as compared to the model architectures used, lesser no of epochs should be sufficient to produce good results.

Based on our literature survey and experimentation, we found many multimodal approaches that perform well on classification task. But many models were not able to achieve performance similar to humans, hence we decided to move forward with the following two approaches which achieve performance almost similar to humans.

1. **Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge [20]**
   The VisualBERT[10] architecture is very similar to that of BERT[19], that is, it essentially comprises of

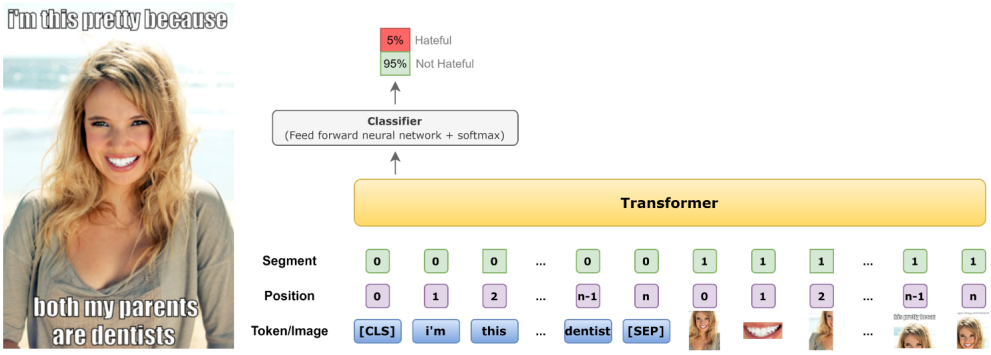

Figure 4: Multimodal transformer architecture, relationships between image regions and captions is learnt implicitly through self-attention.

a large stack of transformer based encoder layers focusing on self-attention. The only difference is the addition of image-region embeddings along with the word embeddings, which are similarly calculated using addition of a segment embedding, a position embedding and a token embedding as shown in Figure 4. 100 boxes of 2048D region-based image features are extracted from the fc6 layer of a ResNeXT-152 based Mask-RCNN[3], trained on the Visual Genome dataset[8].
For stable learning and better scores, we increased the training data by 328 additional memes. We used the memotion dataset [16] for increasing our training data, we picked those memes that were most similar to the memes already present in the Facebook AI's hateful memes training data.
The extracted image regions and text are the input to the pretrained visualBert model. Along with the image regions and text tokens, a special token - ["CLS"] is prepended to the input. For the image regions we used a ResNeXT-152 based Mask-RCNN [1] for extracting features. Using the self attention mechanism, the features of image regions, text and the special token are enhanced. From the final layer of the architecture, features of the ["CLS"] token (X) are passed through a classifier which outputs the probability ($\hat{Y}$) of the input meme being hateful and non-hateful. In our case, the classifier is just a fully connected Layer followed by softmax activation:

$$\hat{Y} = Softmax(WX + b) \tag{1}$$

Here, $X \in R^d$, $W \in R^{c \times d}$, $b \in R^c$ and c is the number of classes (2 in our case). W and b are learnable parameters.
Using the ground truth labels (Y) and the predicted probabilities ($\hat{Y}$), we calculate the cross entropy loss. The objective is to minimize the mean cross entropy loss computed over all the training examples in order to find the optimal parameters for the classification layer and fine tune the parameters of the pre-trained VisualBERT model. We perform the training for 5 epochs.

[1] https://dl.fbaipublicfiles.com/pythia/detectron_model/FAST_RCNN_MLP_DIM2048_FPN_DIM512.pkl

The VisualBERT model has a lot of hyperparameters, in order to enhance the performance of our model, we tuned few of the hyperaprameters such as batch size, learning rate, warmup steps, warmup type and warmup iterations. The total number of unique combinations for different values of these hyperparameters is 96. Since, we train a single model for only 5 epochs which takes only 40 minutes, therefore we performed a grid search over these 96 combinations of hyperparameters.
Out of these 96 models, we pick top 27 models based on the validation loss. Using these 27 models, we create an ensemble, where the final class label is computed using the outputs of individual models. Majority voting strategy is used to predict the final label.
The results in Table 2 are numerically similar to what the paper [20] claims to achieve.

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

Table 2: VisualBert Results

2. **Vilio: State-of-the-art Visio-Linguistic Models Applied To Hateful Memes [13]**
State-of-the-art Vision+Language models are of two types: Single-stream and Dual-stream. The former uses a single transformer to process the image and language input at the same time; some examples are VisualBERT, UNITER, OCSAR. The latter relies on separate transformers for vision and language, which are then combined towards the end of the model. LXMERT[18], ERNIE-ViL, DeVLBERT[23], VilBERT[12] are examples of Dual-stream models. Vilio uses a combination of Single-stream and Dual-stream architecture.
The models used features extracted from images using the detectron2 framework by Facebook. For extracting features, the no. of minboxes and max boxes was kept to the same number, either 36, 50, 72 or 100. Text extracted by OCR techniques from the original meme was used as is from the hateful memes dataset. The models were trained on the pretrained weights provided by the authors of the multimodal architectures. For each model, different strategies were used for pre-training phase, the details of which are provided in the paper. The models were trained on the pretrained weights provided by the authors of the multimodal architectures. For every model, finetuning was done using binary cross-entropy loss. For OSCAR, UNITER and VisualBERT, 3 different seeds with different extracted features and for ERNIE-ViL, 5 different seeds with different extracted features were used. The Adam optimizer is used with a learning rate of 1e-5 and 10% linear warm-up steps. Gradients are clipped at 5 for VisualBERT, OSCAR & UNITER and at 1 for ERNIE-ViL. VisualBERT, OSCAR & UNITER are trained for 5 epochs and Stochastic Weight Averaging is used during the last 25% of training. ERNIE-ViL models are trained for 5000 steps. The weights from the last step are taken for all models and used for inference on the test set. The results were processed in an ensembling loop applying simple averaging, rank averaging, power averaging and simplex optimization to produce the final predictions.
The results in Table 3 are numerically similar to what the paper claims to achieve.

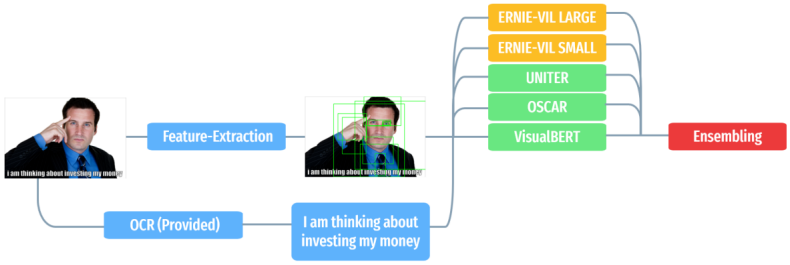

Figure 5: Vilio Ensemble Architecture

2.3 Training Additional Models

We trained two additional models - DeVLBERT[23] and LXMERT[18] on the hateful memes dataset from their implementations available in PyTorch[14] 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.

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

Table 3: Vilio Ensemble Performance

1. **DeVLBERT[23]**
   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.

2. **LXMERT[18]**
   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[11] or Visual Genome[9] datasets while also aggregating three image captioning datasets - VQA 2.0[1], GQA[5] and VG-QA[25].

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

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 decided to proceed with the original Vilio ensemble for further experimentation.

2.4 Further Enhancement - Mixture of Experts

Due to the much more superior performance of the Vilio architecture over VisualBert, we decided to do our our further experimentation on the Vilio architecture only. We tried out the Misture of Experts (MoE) ensembling strategy on the predictions from Vilio models.

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.

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

$$\alpha = MLP(O) \tag{3}$$

$$P = \sum_{i=i}^{N} (\alpha_i \times O^i) \tag{4}$$

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

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

Table 5: MOE Ensemble Performance, we can see the improvement in Dev Seen Accuracy from previous best

We can see from the results in table 5 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 performance 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.

2.5 Going beyond English

Figure 6: One of the memes(hateful) from our manually annotated dataset.

We also wanted to explore just how well would such a model perform when extended to different languages and we decided to evaluate the performance on the Hindi language, since abundant Hindi memes are available online. As there was no available dataset of Hindi memes, training the model on Hindi just wasn't feasible. So, we decided to evaluate the model on a manually created small dataset whose caption we will manually translate to English so that it is compatible with our model's input semantics.

Hindi Dataset	
Hateful memes	20
Non-Hateful memes	19
Total	39

Table 6: Hindi dataset details

Figure 6 shows one of the memes from our dataset, which the model was able to correctly classify as hateful. Overall our model classified only 22 correctly out of the total 39 memes which amounts to a mere **56.4%** accuracy. Since, this a very small dataset, it won't be exactly fair to generalize the results to Hindi memes but it does seem that the same model won't be performing well when extended to multiple languages. This could be due to many reasons like:

- The pre-training on English datasets which do not contain much of Hindi references.
- Simply because the Facebook memes dataset and the Hindi memes we've collected are very different in the kind of hateful content present.

Looking at the data, the latter seems like the prominent reason.

3 Analysis of Results

3.1 Visual Analysis

We conducted a visual analysis to find out what difficulty type of memes are being classified as either hateful or non-hateful for each of three classifier systems:

1. VisualBERT COCO
   We chose this model from the baseline models implemented, since this was numerically superior among all the other 11 baselines from the challenge paper.

2. VisualBERT Ensemble
   This was one of the two SoTA baselines as described in section 3, which gave better results than the baseline.

3. Vilio Ensemble
   Vilio ensemble gave the best results from all other models. We analyze the results of the other two models in direct comparison to Vilio. Finally, we also show some examples where Vilio failed.

For all the analysis below, we have focused on dev_seen dataset which consisted of 500 memes consisting of 247 hateful and 253 non-hateful memes.

3.1.1 VisualBERT COCO v/s. Vilio

Figure 7 shows some of the memes which were classified by both - VisualBERT COCO baseline and Vilio ensemble correctly as hateful. These memes typically have outright explicit content or use of derogatory comments such as references to women as "sandwich makers", and are easily detectable as being hateful. Many of these memes can also be unimodally classified correctly as hateful.

Figure 7: Memes classified as hateful by both: VisualBERT COCO and Vilio - correctly

Figure 8 shows some of the memes which were classified by VisualBERT COCO and Vilio ensemble correctly as being non-hateful. These memes are either purely non-hateful i.e. giving out joy or positivity, or fall under the problem definition as being non-hateful (e.g. memes targeting recognized terrorist organizations are not hateful).

Figure 8: Memes classified as non-hateful by both: VisualBERT COCO and Vilio - correctly

Figure 9 shows some of the memes which the VisualBERT COCO baseline was unable to classify as being hateful, but were correctly classified as hateful by Vilio. Unlike the previous set of hateful memes in figure 7, these memes may not have explicit content, but need some more context to be identified for their hatefulness. For instance, reference to nazis in the upper left meme, depiction of slavery in upper right meme, apparent refrence to a Muslim as a terrorist in the lower right meme, etc. The VisualBERT COCO baseline was not able to make the correct inference for these memes.

Figure 9: Memes classified as hateful by Vilio correctly, but incorrectly classified as non-hateful by VisualBERT COCO

Figure 10 shows some of the memes which the VisualBERT COCO baseline incorrectly classified as being hateful, but were non-hateful as detected rightly by Vilio. In these memes, the models can get tricked by use of phrases which usually appear in hateful contexts, but are not hateful in the present scenario due to the surrounding context.

Figure 10: Memes classified as non-hateful by Vilio correctly, but incorrectly classified as hateful by VisualBERT COCO

3.1.2 VisualBERT Ensemble v/s. Vilio

Figure 11 shows some of the memes which both - VisualBERT Ensemble and Vilio Ensemble were able to correctly classify as being hateful. Just like the previous set of hateful memes in figure 9, these memes require some contextual information for the text or image in the meme. These memes have a strong multimodal nature, as in no decision can be made without the presence of both the modalities together in most of the memes.

Figure 11: Memes classified as hateful by both: VisualBERT Ensemble and Vilio - correctly

Figure 12 shows some of the memes which VisualBERT Ensemble and Vilio Ensemble were both able to correctly classify as being non-hateful. Just like hateful memes, detecting these set of memes as non-hateful also requires information in multimodal sense. For instance, image of a bruised face can itself be suspicious of being hateful, but textual message corresponding it makes it non-hateful.

Figure 12: Memes classified as non-hateful by both: VisualBERT Ensemble and Vilio - correctly

Figure 13 shows some of the memes which the VisualBERT Ensemble was not able to detect as being hateful, but were correctly classified by Vilio.

Figure 13: Memes classified as hateful by Vilio correctly, but incorrectly classified as non-hateful by VisualBERT Ensemble

Figure 14 shows some of the memes which the VisualBERT Ensemble incorrectly classified as hateful, but were correctly classified as non-hateful by Vilio.

Figure 14: Memes classified as non-hateful by Vilio correctly, but incorrectly classified as hateful by VisualBERT Ensemble

These memes are incorrectly classified by VisualBERT Ensemble are increasingly subtle, in the sense a slight modification or misinterpretation can flip the label of the meme.

3.1.3 Cases where all the models failed

Figure 15 shows some of the memes which all the three models - VisualBERT COCO, VisualBERT Ensemble and Vilio Ensemble were unable to correctly classify as being hateful. From 500 memes in the dev_seen dataset, there are 62 such memes. The nature of the memes might indicate building more contextual information in the models is necessary for accurate detection of racial abuse, sexism, religious intolerance, etc.

Figure 15: Memes incorrectly classified as non-hateful by VisualBERT COCO, VisualBERT Ensemble and Vilio

Figure 16 shows the memes which all the three models were unable to identify as non-hateful memes. There are only three such memes from the dev_seen dataset, all of which are shown in the figure. Just like before, the models seem to be confounded by use of commonly used textual phrases or images due to lack of understanding of context.

Figure 16: Memes incorrectly classified as hateful by VisualBERT COCO, VisualBERT Ensemble and Vilio

3.2 Analysis of the superior performance of Vilio

Some of the reasons 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[4] or Faster-RCNN[15] based features for input, Vilio ensemble uses features extracted using the detectron2[21] 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[9] 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[22] dataset were used in addition to the pretrained models on Conceptual Captions[17] 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[6] 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.

4 Future Work

This project if needed, can be extended further in a lot of different directions. One could explore the *Hindi* memes regime further by finetuning the multimodal transformer architectures on a Hindi dataset rather than an English one. Or, efforts can be made at developing models that can generalize to memes from any contextual or linguistic domain, irrespective of the data distribution it is trained on. This would be rather beneficial for low-resource, poorly monitored scenarios in different languages around the world. It is because we have observed the models to not perform with an equivalent level of accuracy when tested on memes translated from a different language, and creation of an annotated dataset for every language would further be an expensive, manual task. There is also a possibility of trying out completely new ensembling and data augmentation techniques involving both image augmentation and text augmentation. As this is still a fairly new research topic, the possibilities are endless.

5 Conclusion

Even though the results we've achieved today are really encouraging, we still have a long way to go. The current SoTA accuracy is simply not enough to completely automate the process of removing hateful/distasteful memes from the vast network of modern day Social Media. The problem still remains one of the most complex multimodal problems simply because of the very nature of memes, which require a lot of contextual and a global current affairs knowledge to be comprehended in the right fashion. Memes themselves are evolving at a very rapid pace and hence this field of automated meme-interpretation is yet to evolve.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016.

[2] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*, 2020.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[5] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019.

[6] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2019.

[7] 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.

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[10] 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.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[12] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

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

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[16] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*, 2020.

[17] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

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

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv e-prints*, page arXiv:1706.03762, June 2017.

[20] 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.

[21] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[22] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2019.

[23] 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.

[24] Yi Zhou and Zhenhao Chen. Multimodal learning for hateful memes detection. *arXiv preprint arXiv:2011.12870*, 2020.

[25] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016.