CS 7643 Project Report: Hateful Memes Detection

Jinjun Liu
Georgia Institute of Technology
jliu788@gatech.edu

Chen Zhang
Georgia Institute of Technology
czhang613@gatech.edu

Ji Shen
Georgia Institute of Technology
jshen321@gatech.edu

Yi Wang
Georgia Institute of Technology
ywang3720@gatech.edu

Abstract

Hateful Memes Challenge, proposed by Facebook Inc in 2020, aimed to develop novel deep learning models to detect memes with hateful meanings. We applied an ensemble learning approach to a group of models generated by utilizing VisualBERT model to attempt to improve Hateful Memes detection accuracy. Our approach achieves 0.7675 AUROC with an accuracy of 0.7111 by exploring several potential approaches including data augmentation, hyperparameters tuning, and new assemble techniques.

1. Introduction

1.1. Background and Motivation

Memes have gained huge popularity over the past years. Although memes are oftentimes harmless and generated especially for humorous purposes, they have also been used to produce and disseminate hate speech to make communities more toxic, such as direct attacks on people based on race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, and serious disease or disability [4]. Recently it becomes a growing problem in modern society on social media platforms like Facebook, Twitter and Instagram. Due to the surge of the amount of malicious content it cannot be addressed by human inspection of every sample. In 2020, Facebook AI initiated a competition with focus on hateful memes classification.

Detecting hate speech in memes is challenging due to the nature of memes (usually image+text), in which multiple data modalities need to be analyzed together. In this project, we focus on hateful content detection in multimodal memes. We work on the Facebook Hateful Meme Challenge and aims to solve a binary classification problem of predicting whether a meme is hateful or not [5].

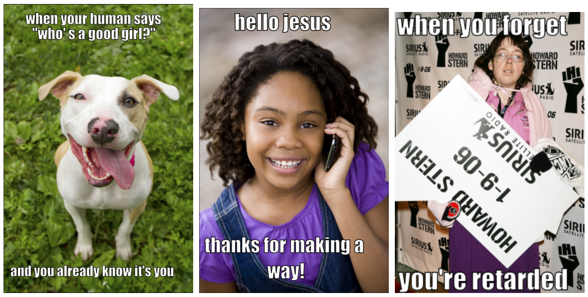
when your human says "who's a good girl?"
and you already know it's you
hello jesus
thanks for making a way!
when you forget
HOWARD STERN 1-9-06 SIRIUS
you're retarded

Figure 1. Examples of Hateful Memes Dataset [2]

1.2. Application

There are numerous contents posted on social media platforms like Facebook every single day. The hateful contents could potentially poison the community. Developing technologies using deep learning approaches is a feasible way to automatically check those contents reported by users. If our methods could improve the accuracy of automatic detection of hateful memes, it will benefit the society both in economic way and humanistic way. AI detection technology for hateful memes could save human resources and also make the online environment better.

1.3. Data

The dataset we focus on is provided by Facebook AI Hateful Memes Challenge [2]. The dataset provides over 10,000 memes (both image and text information) labeled with hateful or not, including 8500 training samples, 500 validation samples, and 1000 test samples. A few examples are shown in Figure 1. In this project, only the training samples and original validation samples are applied to this study. We use the phase 2 validation dataset *dev_undseen* for validation purpose. Besides, we augmented the datasets by applying image modifications and expanded data from Memotion dataset for training purpose, see section 4.2 for details.

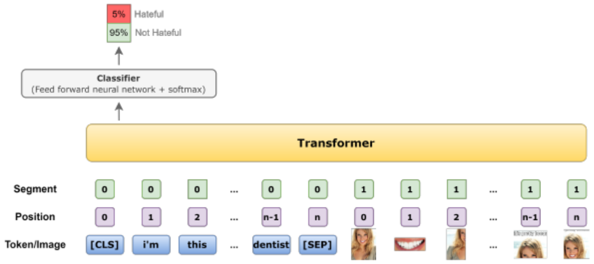
5% Hateful
95% Not Hateful
Classifier
(Feed forward neural network + softmax)
Transformer
Segment
0 0 0 ... 0 0 1 1 1 ... 1 1
Position
0 1 2 ... n-1 n 0 1 2 ... n-1 n
Token/Image
[CLS] i'm this ... dentist [SEP] ...

Figure 2. Representation of Multimodal VisualBERT Model taken from reference paper [9]

2. Related Work

Given the nature of the problem, it is widely accepted in the research community that unimodal models struggle and only multimodal models can succeed [5]. Most of the state-of-art methods were far behind from non-expert human performance on the hateful memes dataset, until we have seen an ROC-AUC of 0.845 from the winner of the last Hateful Memes Challenge [3]. With the emergence of more sophisticated tools such as XLM-R and self-supervised models, there is still room for improvement on this unique challenge.

A representative chart of the VisualBERT model [6], one of the multimodal models, is shown in Figure 2.

3. Approach

The overachieving goal of this project is to fine-tune a pre-trained model with new approaches to seek higher ROC-AUC scores for the Facebook Hateful Memes Detection challenge. This study is based on a framework called MMF [8]. We selected the VisualBERT model from MMF to fine-tune. The details about comparisons with other multimodal models are in Section 4.1. Section 4.2 displays data augmentation by conducting image modification techniques and introducing a new image dataset. Then, the Grid Search method with hyper-parameter tuning is applied to the pre-trained VisualBERT model to generate multiple models. The details of conducting the Grid Search process are described in Section 4.3. We applied alternative ensemble learning methods in Section 4.4. We selected the top 20 models by ranking all the generated models based on ROC-AUC and Accuracy Scores. All selected models are ensembled with different methods to predict the probability for classification in the dataset. Therefore, each image is mapped to 20 classification results. The final classification result is generated by weighted voting and simple averaging over their prediction probabilities.

MMF is a popular framework that contains multiple pre-trained multimodal models for vision-and-language research. Its ability to outperform other models in previous studies is one reason that we selected a pre-trained model from MMF. Data augmentation is a common approach for improving training performance. Furthermore, we expanded the training data to obtain better stable learning. Different models having different skills in encoding texts and figures. By implementing ensemble learning, the overall prediction model could take advantage of the expertise from different models.

The new approach in this study has two aspects. One is the classical computer vision techniques are applied to the data augmentation for noisy training. Another new point is that we created a new ensemble learning method to determine the hateful memes. By experimenting them, we aim to reach to a higher ROC score in prediction compared to the original model.

We anticipate a number of challenges such as setting up the development environment to run experiments in parallel, finding suitable additional datasets to augment the given data, avoiding overfitting, understanding the baseline models and seek effective ways for training and fine-tuning, etc. The details on some of solved challenges are described in each subsection of Section 4.

4. Experiments and Results

4.1. Pre-trained Model Selection

Facebook MMF provides various pre-trained model that can be fine-tuned for hateful memes detection (e.g., ConcatBERT, VisualBERT, VilBERT, etc.) The architecture of a deep learning model has a large impact on the performance of a certain task. Velioglu's [9] experiments show that VilBERT pre-trained on Conceptual Captions (CC) achieves better score than that on COCO images caption dataset. But they haven't tested on different architectures like VilBERT. We compared the accuracy of fine-tuned model based on pre-trained VilBERT on CC and VisualBERT on CC (as shown in figure 3.) The results show that VisualBERT has better ROC-AUC scores at all updates. Therefore, we decided to use VisualBERT on CC as our pre-trained model for further fine-tuning.

4.2. Data Augmentation

Data augmentation are widely-used technologies in data expansion as state-of-the-art deep learning models typically have parameters in the order of millions, especially in image processing. In this project we carry out two data augmentation experiments and compare the result with established model by Velioglu [9].

4.2.1 Images Modification

In the real world scenario, we may have a dataset of images taken in a limited set of conditions such as different orien-

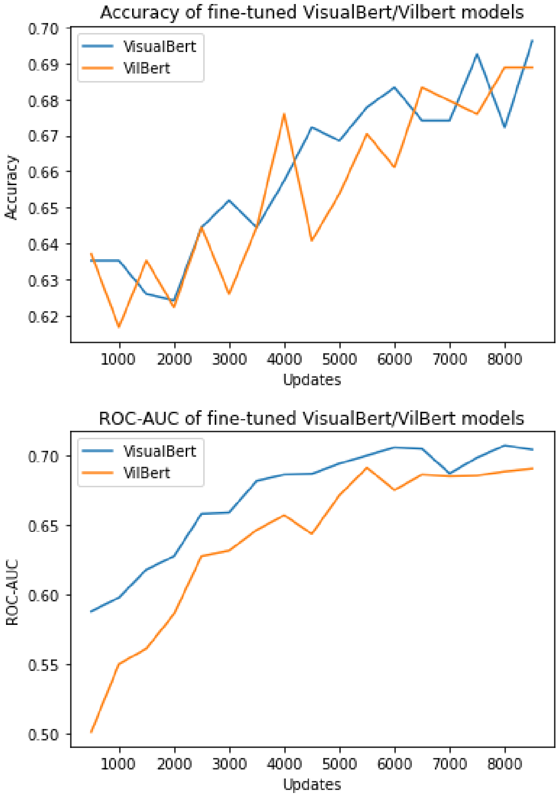

Figure 3. Accuracy (upper) and ROC-AUC (bottom) for VisualBERT and VilBERT pre-trained models on Conceptual Captions dataset

tation, location, scale, brightness etc. Here we account for these situations by training our neural network with additional synthetically modified data, i.e. image flipping.

We randomly pick 1k images from the Hateful Memes dataset followed by flipping and feature extraction [1]. Then we train the best models pre-trained for hateful memes detection with the 1k flipped memes. The trained models are evaluated by validation memes provided by the dataset. The result shows that the accuracy and ROC-AUC decreases slightly after training with flipped images. This result is not surprising, as the meme images often contain both visual and lingual information. The flipping technology may retain the visual information whereas the words on the meme loss their meaningful contents. Note that we feed in the same words to the transformer system after flipping. However, the 100 extracted features from image also have fair amount of lingual information, the loss of which causes the slightly drop in accuracy and ROC-AUC.

We further apply a Gaussian blur filter to those flipped images considering the words contents cause the loss of accuracy. Our hypothesis is that a blurred version of image may smooth out the word contents, and thus reduces the loss of information. See Figure 4 as an example. Result in Table 1 shows that flipped and blurred images performs slightly worse than flipped alone. Unfortunately our blurring method fails to improve the model.

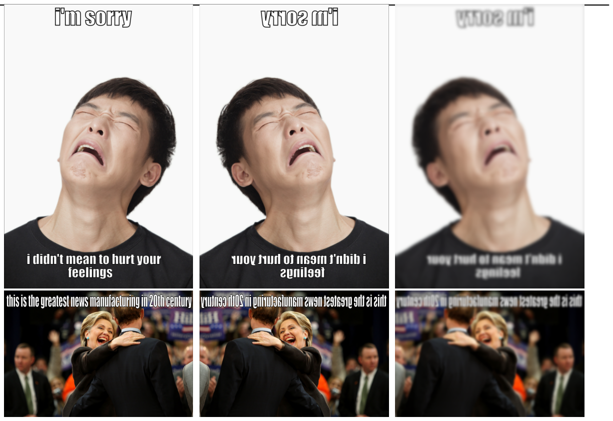
i'm sorry
i didn't mean to hurt your feelings
this is the greatest news manufacturing in 20th century

Figure 4. Image Modification example: left original, middle flipped, right flipped and blurred

4.2.2 Memotion Dataset

Memotion Dataset is an open-sourced dataset containing 7K annotated memes with human-annotated labels. The dataset is publicly available [7]. One of the sentiment classification is offensive language with four classes: not offensive, slight, very offensive, and hateful offensive.

We classify *hateful offensive* and *very offensive* as hateful memes, while *not offensive* and *slight* as non-hateful memes. Preliminary analytic result of the dataset indicates that the samples are labeled incorrectly. We pick memes that are assigned with same label by both initial dataset and the best-model prediction. In total 2173 images satisfy this criteria. We then select 700 samples with about 4:3 ratio of non-hateful/hateful, which mimics the ratio in Facebook Hateful Memes training set. The idea is similar to pseudo label in semi-supervised learning where unlabelled data are predicted and then used as training set.

With the 700 additional memes in hand, we conduct feature extraction using Mask R-CNN technology developed by He and coworkers [1]. 100 rectangular areas are extracted from each image. Those features are passed into the best models pre-trained for hateful memes detection developed by Velioglu [9] for another round of training. The result shows that The results are summarized in Table 1. The ROC-AUC achieves 0.7675, which is actually better than the original model 0.7654.

We notice that training using only newly generated data may introduce bias. Combine and shuffle the original 8500 training samples along with the 700 new training samples would be ideal, but our computational capability is limited to the GPU resources given on Google Cloud Platform. We would expect better results with unbiased training set.

Measurement	Original Results	Flip	Flip and Blur	Memotion Dataset
Accuracy	0.7232	0.7056	0.6981	0.7111
ROC-AUC	0.7654	0.7521	0.7497	0.7675

Table 1. Accuracy and ROC-AUC on Validation Dataset with Data Augmentation.

4.3. Hyperparameters Tuning and Overfitting

In this section, we generated multiple models used for ensemble learning based on the improved pre-trained VisualBERT model. The core method of generating the model is implementing the Grid Search approach in the model production. In order to improve the model performance and save training time, hyper-parameter tuning is applied in the Grid Search process.

The hyper-parameter Tuning process is a good approach to achieve a balance between underfitting and overfitting. When the deep learning model could not able to reduce the error for either the test or training set, most likely the model hits the underfitting issue. The learning rate and batch size are 2 parameters we mainly focus to find the optimal hyper-parameters.

The learning rate controls how quickly the model is adapted to the problem. Commonly, overfitting occurs if the learning rate is too small. If the learning rates are too large, the training may hit diverge. In this experiment, we worked on finding learning rates that converge or diverge by using a grid search of short runs. The loss function is used to detect converge and diverge. Cross entropy is selected as the loss function in this model. Figure 5a displays an example of the ROC-AUC for running the model for several epochs which the learning rate value is from 1e-5 to 9e-5. The ROC-AUC roughly increases until 5e-5 and then decreases. This could explain as the diverge appears when learning rate larger than 5e-5. Based on the experiments, we selected 1e-5, 3e-5, and 9e-5 as the learning rate values in the Grid Search process.

Larger batch sizes may have larger gradient steps that resulting in more rapid convergence of the model parameters. So the batch size is also in conjunction with the training time. Besides, we should note that the batch size is restricted by the GPU memory/RAM. The results are shown in figure 5b. Similarly, all other hyper-parameter are also tuned and then are used in generating models for ensemble learning prediction.

Overfitting is a common issue for deep learning projects. Overfitting refers to the model that fits the training set too well, but not for the test set. During the training, we implemented 2 approaches when detecting overfitting. The first one is that training with more data. The second method is adjusting dropout rate during training the model. Adding dropout layer is a common regularization approach to avoid overfitting in deep learning.

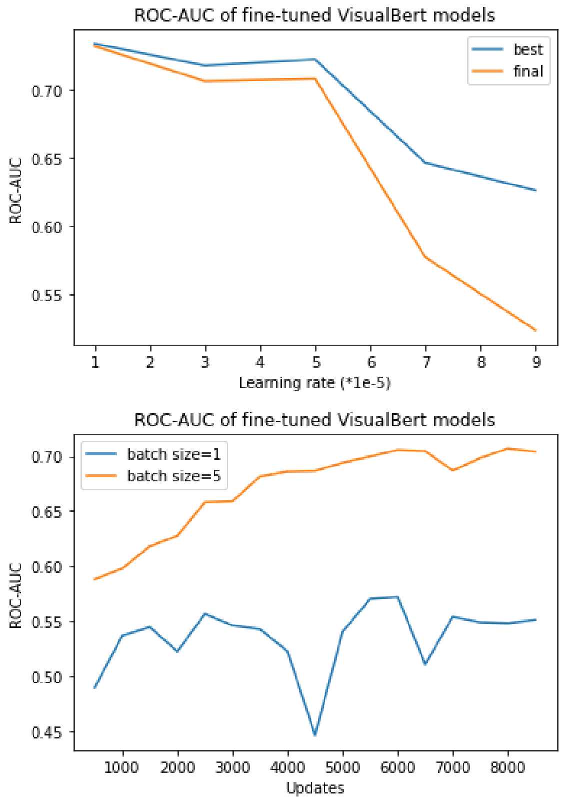

Figure 5. Hyper-parameter Tuning: upper is for learning rate; bottom is for batch size

In addition to the above two approaches, we also would like to avoid overfitting by conducting cross-validation of K-fold. The K-fold means the data is partitioned into k subsets and iterative train the model on k-1 folds while using the remaining fold as the test sets. Besides, implementing regularization of adding additional dropout layer is also suggested to avoid overfitting. However, we did not apply these additional approaches due to limited time. We believe that by implementing K-fold and tuning dropout rate, better results in avoiding overfitting would be expected.

4.4. Alternative Ensemble Methods

The core idea of this section is to detect hateful memes by a group of individual models. Weighted voting and simple averaging are conducted to seek a higher accuracy for the model prediction. In the end, the final score of Accuracy and ROC-AUC is produced by leveraging the optimal

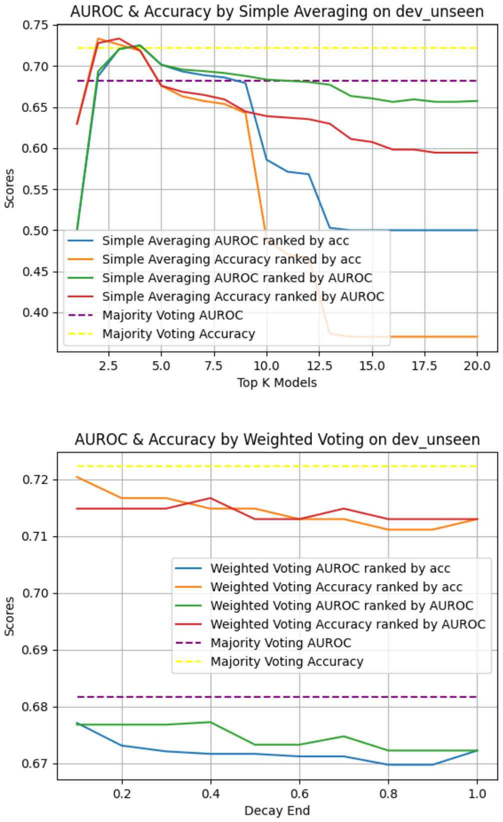

Figure 6. Accuracy and ROC-AUC with Alternative Ensemble Methods and Model Pre-trained on Augmented Data

combination of weighted values for each model and selecting a certain number of top-ranked models.

4.4.1 Weighted Voting

Not all models are created equal. In the majority voting applied by Velioglu et al. [9], the final classification label is determined by the dominant number in the 27 outputs. We claim that this methods although being effective might be over-simplified. It is also based on the assumption that these models should be treated equally. We argue that their votes should be assigned different weights based on the fact that they score differently on a number of evaluation metrics. We propose to sort our 20 generated models based on ROC-AUC and accuracy scores. We then assign decayed weights based on their ranking, with 1.0 assigned to the top model, and the Nth model would have decay ** N as its weight. The weights are normalized to represent the importance of each model's vote. Finally, the weighted sum of these models' probabilities is used to calculate the classification label.

Figure 6 illustrates the results of this weighted voting technique on the dev_unseen dataset. We observe that this method does not give better results when decaying from 1.0 for the top model to decay_end, probably due to not finding the reasonable set of weights. Another reason could be the Accuracy/ROC-AUC rankings base on other datasets do not generalize well to unseen data.

4.4.2 Simple Averaging

In this subsection, we propose another ensemble method that takes the average prediction probabilities from top K models. Using the same two rankings from previous experiments, the classification labels are predicted based on this simple averaging, which presumably averages out the noises introduced by different models. The results from different evaluation metrics are plotted in Figure 6. It is noticeable that the accuracy from both rankings scores higher than the benchmark at K=2 or 3, whilst the ROC-AUC scores above the majority voting from K=2 to 12. One obvious observation is that the average probabilities from more models does not necessarily give better results, most likely due to the lack of confidence from lower ranked models. And the results from the single top models are not great either, probably due to high biases caused by considering only one model.

5. Conclusion

In this project, we conducted multiple approaches to detect hateful memes based on pre-trained multimodal models. In order to reach a higher score, we compared multiple pre-trained models from MMF, trained models with hyperparameter tuning, expanded the training dataset by applying image manipulation (e.g. flipping and blurring), and trained with more images from Memotion dataset. Furthermore, we introduced two new ensemble techniques for post-processing the classifier probabilities to generate labels. Our final model pre-trained on visualBERT and augmented data with single averaging ensemble shows the ROC-AUC is 0.7675 and Accuracy is 0.7111, which is slightly higher than the benchmark.

6. Work Division

The delegation of work among team members has been provided in Table 2 at the end of the report.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[2] https://www.drivendata.org/competitions/64/hateful memes/. Hateful memes: Phase 1. *Facebook AI*, 2020. 1

Student Name	Contributed Aspects	Details
Chen Zhang	Data Cleaning and Data Augmentation Experiments	Carried out image modification, processed Memotion 7k data, and implemented training/testing using expanded datasets.
Ji Shen	Design and Implementation of Ensemble Techniques	Applied two ensemble methods, post-processed predicted data and fine-tuned key parameters for better results.
Jinjun Liu	Pre-trained model selection and hyper-parameters tuning	Fine-tuned VisualBERT and VilBERT and compared them. Tried different hyper-parameters like batch size.
Yi Wang	Hyper-parameter Tuning and Ensemble Learning	Trained and generated models with hyper-parameter tuning and avoiding overfitting for Ensemble Learning. Improved post-processing classifier probabilities.

Table 2. Contributions of team members.

[3] https://www.drivendata.org/competitions/70/hateful-memes-phase 2/leaderboard/. Hateful memes: Phase 2 leaderboard. *Facebook AI*, 2020. 2

[4] https://www.facebook.com/communitystandards/hate_speech. Community standards. *Facebook*, 2020. 1

[5] 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. 1, 2

[6] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and https://www.overleaf.com/project/6053c78eb5a9aa2d9f842a20language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[7] Chhavi Sharma, William Paka, Scott, Deepesh Bhageria, Amitava Das, Soujanya Poria, Tanmoy Chakraborty, and Björn Gambäck. 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. 3

[8] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020. 2

[9] 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. 2, 3, 5