Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge

Riza Velioglu
Technical Faculty
Bielefeld University
Universitätsstraße 25, 33615 Bielefeld, Germany
rvelioglu@techfak.uni-bielefeld.de

Jewgeni Rose
Smart Data Analytics
Computer Science III, University of Bonn, Germany
jewgeni.rose@gmail.com

Abstract

Memes on the Internet are often harmless and sometimes amusing. However, by using certain types of images, text, or combinations of both, the seemingly harmless meme becomes a multimodal type of hate speech – a *hateful meme*. The Hateful Memes Challenge[1] is a first-of-its-kind competition which focuses on detecting hate speech in multimodal memes and it proposes a new data set containing 10,000+ new examples of multimodal content. We utilize VisualBERT – which meant to be the "BERT of vision and language" – that was trained multimodally on images and captions and apply Ensemble Learning. Our approach achieves 0.811 AUROC with an accuracy of 0.765 on the challenge test set and placed third out of 3,173 participants in the Hateful Memes Challenge[2]. The code is available at https://github.com/rizavelioglu/hateful_memes-hate_detectron

1 Introduction

Memes have gained huge popularity over the past years, resulting in over 180m posts on different social media platforms until 2018 [1]. Although memes are oftentimes harmless and generated especially for humorous purposes, they have also been used to produce and disseminate hate speech in toxic communities. Hate Speech (HS) is a direct attack on people based on race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, and serious disease or disability [2] – a growing problem in modern society. Giant tech companies, such as Facebook, own platforms where millions of users log in daily and they are obliged to remove a tremendous amount of HS to protect their users. 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 [3]. 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. Detecting hate speech in memes is challenging due to the multimodal nature of memes (usually image+text). Therefore, these techniques have to process the content the way humans

[1] https://www.drivendata.org/competitions/70/hateful-memes-phase-2/

[2] HateDetectron at https://www.drivendata.org/competitions/70/hateful-memes-phase-2/leaderboard/

The Hateful Memes Challenge at NeurIPS (2020), Vancouver, Canada.

do: holistically. When viewing a meme, a human would not think about the words and the picture independently; but understand the *combined* meaning. Moreover, while the visual and linguistic information of a meme is typically neutral or funny individually, their combination may result in a hateful meme.

A recent study shows that state-of-the-art methods for hate speech detection in multimodal memes perform poorly compared to humans: *64.73%* vs. *84.7%* accuracy [4]. To catalyze sophisticated research in this area, Facebook AI launched the Hateful Memes Challenge and published a dataset containing more than 10,000 newly created multimodal memes [4]. Multimodal tasks reflect many real-world problems, including how humans perceive and understand the world around them.

There has been a surge of interest in multimodal problems since 2015 in visual question answering [5, 6], image captioning [7, 8], speech recognition [9, 10] and beyond. But it is not always clear to what extent genuinely multimodal reasoning and understanding are needed to solve current challenges. For instance, for some datasets language can unintentionally impose strong priors, which might result in a remarkable performance, without any understanding of the visual content. The Hateful Memes challenge design and dataset are created to encourage and measure truly multimodal understanding and reasoning of the models. A key point to achieve this are the so-called "benign confounders" (also called *contrastive* [11] or *counterfactual* [12] examples) which addresses the risk of exploiting unimodal priors by models: for every hateful meme, there are alternative images or text that flip the label to not-hateful. Such image and text confounders require multimodal reasoning to classify the original meme and its confounders correctly. Thus, making the dataset challenging and appropriate for testing the true multimodality of a model.

In the following, we analyze the challenge dataset and describe our prize-winning solution that placed third among 3,173 participants in the Hateful Memes Challenge in detail. Our solution achieves *0.811* AUROC with an accuracy of *0.765* on the challenge test set, which improves all the benchmark models [4], including the state-of-the-art models at that time, such as ViLBERT [13] (trained on Conceptual Captions [14]) and VisualBERT [15] (trained on COCO [8]). Nevertheless, the accuracy is still behind humans with a mentionable gap, highlighting the need for progress in multimodal research.

2 Problem Statement

The Hateful Memes dataset is not created for training models from scratch, but to fine-tune and test large-scale, pre-trained multimodal models. Thus, the size of the dataset (10K images) is small compared to datasets such as Visual Genome (108K) [7], COCO (330K) [8], and Conceptual Captions (3.3M) [14]. The dataset is split into three sets: a train set of 8.500 samples, a dev set of 500 samples, and a test set of 1.000 samples. In addition to this "seen" test set, a new test set consisting of 2.000 samples has been published where the winners are determined according to their performance on this "unseen" test set. The area under the receiver operating characteristic curve (ROC AUC) [16] has been selected as the measure of performance, which is given by the following formula:

$$AUROC = \int_{x=0}^{1} \mathtt{TPR}(\mathtt{FPR}^{-1}(x))dx \tag{1}$$

The labels indicating whether a meme is hateful or not-hateful are provided within the dataset, hence the task can be cast as a binary classification problem.

3 Methods

The solution comprises VisualBERT [15], a multimodal BERT for vision-and-language approach. Figure 1 illustrates an overview of the architecture. The approach can be divided into four sections: dataset expansion, image encoding, training, and ensemble learning. Next, we will provide details of our solution.

3.1 Dataset Expansion

More data delivers stable learning and brings better scores. Thus, we searched for additional data sources in order to grow the dataset size and as a result, we expanded the training data by 428 additional memes.

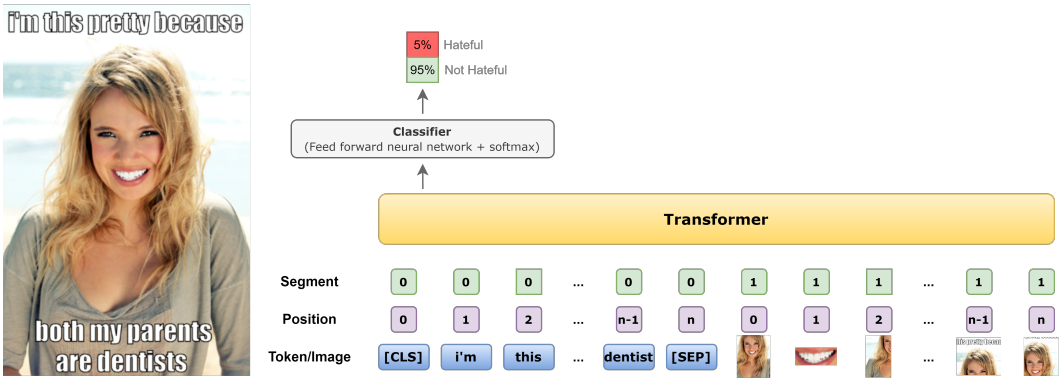

Figure 1: An example meme sampled from the dataset (left), and an illustration of the multimodal transformer architecture (right). Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision.

Unused data in dev seen We found that there are 500 samples in the seen dev set and 540 samples in the unseen dev set. By comparing the memes by their IDs, we identified 400 overlapping samples: $|\{dev\_seen\} \cap \{dev\_seen\}| = 400$, which means that there are 100 samples that are not in dev unseen: $|\{dev\_seen\} \backslash \{dev\_seen\}| = 100$. We added these 100 samples to the training data and evaluated the trained model on dev unseen.

Memotion Dataset The Memotion Dataset [17] 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). For instance, a meme could be annotated as *Not Funny, Very Twisted, Hateful Offensive, Not Motivational*. After an exploratory analysis of the dataset, we argue that the majority of the samples are wrongly labeled. Therefore, we manually re-labeled a part of the dataset. We picked memes that are similar to the ones in the Hateful Memes dataset considering the meme style and design of the challenge dataset. After cherry-picking the "similar" memes, we added 328 new memes to the training data.

3.2 Image Encoding

For every image we extract 100 boxes of $2048\mathcal{D}$ region-based image features from a `fc6` layer of a ResNeXT-152 based Mask-RCNN model [18], trained on Visual Genome [7] with the attribute prediction loss following [19]. Figure 2 shows an example of a processed image. We project the visual embeddings into the textual embedding space before passing them through the transformer layers. We learn weights $W_n \in \mathbb{R}^{PxD}$ to project each of the 100 image embeddings to D-dimensional token input embedding space:

$$I_n = W_n f(img, n), \tag{2}$$

where $P = 2048$, $D = 768$, and $f(\cdot, n)$ is the output of the n-th fully-connected layer in the image encoder.

3.3 Training

Pre-training VisualBERT is originally pre-trained on COCO image caption dataset [8], but in our experiments we noticed that the model pre-trained on Conceptual Captions [14] achieves noticeably better scores. Therefore, we conducted our research on the latter model which is provided by MMF: a framework for vision-and-language multimodal research from Facebook AI Research (FAIR) [20].

Fine-tuning We fine-tune the pre-trained VisualBERT model on the aggregated training set and evaluate it on dev unseen set.

Classification We use the first output of the final layer as the input to a classification layer $\texttt{clf}(x) = Wx + b$ where $W \in \mathbb{R}^{DxC}$, with D as the transformer dimensionality and C as the number of classes (also see Figure 1). We apply a softmax on the logits and train with binary cross-entropy loss.

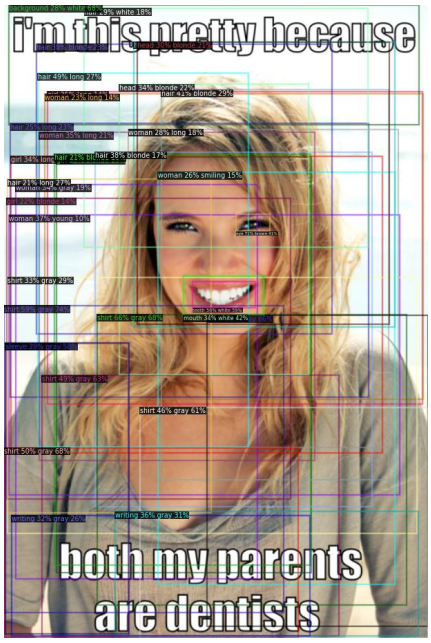

Figure 2: An example of a processed image where the boxes are extracted by Mask-RCNN. Originally 100 boxes are extracted per image but for plotting purposes only 36 boxes are shown.

Table 1: Ensemble models performances derived from VisualBERT CC

ID	Validation	
	Acc.	AUROC
1	70.93	**75.21**
2	69.63	75.16
3	70.74	75.02
...	...	...
25	70.56	73.76
26	70.93	73.75
27	69.81	73.68

3.4 Ensemble Learning

The idea of ensemble learning is to combine the predictions of multiple base models in order to improve generalizability and robustness over a single model. Specifically, we use Majority Voting technique (also known as Hard Voting or Voting Classifier) which combines different classifiers and use a majority vote to predict the class labels. The resulting classifier is oftentimes useful for a variety of equally well performing model as to balance out their individual weaknesses. Consequently, it achieves better performance than any single model used in the ensemble.

We constructed a hyper-parameter search that resulted in multiple models having different AUROC scores on dev unseen set. After sorting them by the AUROC score, the top 27 models are selected for ensemble learning as shown in Table 1 (the number of models is chosen arbitrarily). Then, predictions are collected from each of the models and the majority voting technique is applied: the class of a data point is determined by the majority voted class. Besides, in order to calculate AUROC, the probability that a data point is assigned to a class has to be determined: If the majority voted class is 1 (hateful), then the probability is the maximum among all the 27 models and minimum if it is class 0 (not hateful).

Table 2: Model performance

Type	Model	Validation Acc.	Validation AUROC	Test Acc.	Test AUROC
	Human	-	-	84.70	82.65
Baselines	ViLBERT	62.20	71.13	62.30	70.45
	VisualBERT	62.10	70.60	63.20	71.33
	ViLBERT CC	61.40	70.07	61.10	70.03
	VisualBERT COCO	65.06	73.97	64.73	71.41
Ours	Ensemble	-	-	**76.50**	**81.08**
	Best ensemble model	**70.93**	**75.21**	-	-

4 Experiments and Results

Majority Voting boosted both AUROC and accuracy by **2.5%**. We argue that this technique successfully applies ensemble learning and generates one strong model from multiple 'weak' models – in analogy to the idea of 'bringing the experts of the experts together'. Imagine that one model is very good at – in other words, an expert – detecting hate speech towards women, but might not be an expert in detecting hate speech towards religion. Then, we might have another expert whose expertise is just the opposite. By using the majority voting technique, we bring such experts all together and benefit from them as a whole. The results are shown in Table 2.

5 Conclusion

We proposed an approach detecting hate speech in internet memes multimodally, i.e. considering visual and textual information holistically. We took part in the Hateful Memes Challenge and placed third out of 3,173 participants. Our approach utilizes a pre-trained VisualBERT (a BERT of vision and language), fine-tuned on an expanded train dataset, finally applying Majority Voting over the 27 best models. Our approach achieves 0.811 AUROC with an accuracy of 0.765 on the challenge test set, which is a considerable result but also shows that we are still far from the accuracy of human judgement.

References

[1] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202, 2018.

[2] Facebook. Community Standards. `https://www.facebook.com/communitystandards/hate_speech`, 2020.

[3] Tekla S.Perry. Q&a: Facebook's cto is at war with bad content, and ai is his best weapon. `https://spectrum.ieee.org/computing/software/qa-facebooks-cto-is-at-war-with-bad-content-and-ai-is-his-best-weapon`, 7 2020.

[4] 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.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[6] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision

using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[9] Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. Multimodal speech recognition with unstructured audio masking. *arXiv preprint arXiv:2010.08642*, 2020.

[10] Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. Multiresolution and multimodal speech recognition with transformers. *arXiv preprint arXiv:2004.14840*, 2020.

[11] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

[12] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[14] 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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[15] 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.

[16] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[17] 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.

[18] 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.

[19] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[20] 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, 2020.