Extending multimodal models for the Hateful Memes Challenge

Rodrigo Alejandro Chávez Mulsa **Błażej Dolicki**
Msc Artificial Intelligence
University of Amsterdam
{rodrigo.alejandro.chavez.mulsa, blazej.dolicki}@student.uva.nl

Abstract

Multimodal learning can yield substantial benefits in tasks such as hateful memes detection. In this work we extend state-of-the-art multimodal models with multiple improvements such as denoising the visual input, upsampling most challenging memes or additionally providing gender and ethnicity scores. Our best combined score improves the baseline AUCROC by 0.0119. Our code is publicly available [1].

1 Introduction

Nowadays, a vast majority of work in the community focuses on tasks which include a single modality - be it text, vision or speech. However, many real-life problems are inherently multimodal and some believe that multimodality is the only way to achieve a true, human-like understanding in artificial intelligence (Baltrušaitis et al., 2018). Detecting hateful memes on the internet is certainly one of those problems and its importance continues to increase as more and more people have access to internet and use it for malicious purposes. In this work we use state-of-the-art multimodal models in attempt to improve their ability to detect hateful memes by implementing multiple extensions, followed by careful evaluation and error analysis. We aim to answer the following research questions. How is the model affected by:

- adding additional linear layers
- removing bounding boxes around text?
- upsampling memes that are particularly difficult to classify?
- providing gender and ethnicity scores as additional input?

[1] https://github.com/Noixas/Multimodal-NLP

2 Related work

2.1 Pretrained multimodal models

In recent years, pretrained models proved to be effective in many domains such as natural langauge processing (Devlin et al., 2018) and computer visio (Krizhevsky et al., 2012). Eventually, they were also applied to Vision-and-Language tasks. LXMERT (Tan and Bansal, 2019) was one the earliest attempts which obtained state-of-the-art results on several datasets. The proposed model contained two separate transformer encoders for the vision and language modalities that were later combined to a third transformer encoder whose goal was to combine visual and textual information using cross-modality attention. (Chen et al., 2020) took a different approach and replaced separate modules with a single transformer encoder. It seems to outperform LXMERT (Tan and Bansal, 2019) thus we decided to use it as our baseline model.

2.2 Pretraining tasks

One of the main ingredients for successful multimodal learning is the right choice of pretraining tasks. LXMERT uses 5 different tasks and UNITER uses 6. Given that UNITER already covers a vast majority of pretraining tasks discussed in literature, we decided against implementing more of them and focused on different aspects.

3 Dataset

The Hateful Memes dataset (Kiela et al., 2020) is a multimodal, binary classification task where the aim is to classify whether a meme is hateful based on the image (visual modality) and its caption (language modality). The authors define a *benign confounder* - "a minimum replacement image or replacement text that flips the label for a given multimodal meme from hateful to non-hateful" (Kiela et al., 2020). Essentially, when a meme becomes

non-hateful after changing its image, it is called a *image confounder*. Conversely, a meme that became non-hateful after changing the text is a *text confounder*. The whole dataset comprises of exactly 10k memes where for the training, development and testing data we use respectively 85%, 5% and 10% of all examples. Development and test set are fully balanced and consist of different kinds of memes in the following proportions (Kiela et al., 2020): 40% - multimodal hate, 10% - unimodal hate, 20% - benign text confounders, 20% - benign image confounders and 10% - random non-hateful.

4 Methodology

In this section, we describe the baseline and all our modifications with their respective motivations.

4.1 Baseline

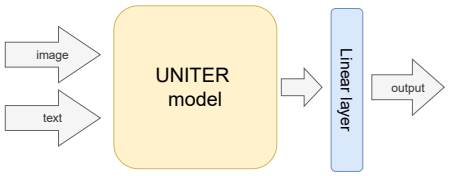
We use UNITER (Chen et al., 2020) as our baseline which is a multimodal transformer model that obtains strong results on many multimodal tasks. This model leverages the Faster R-CNN model (Anderson et al., 2018) to extract object representations from the input images which are then fed to the network, together with word embeddings. For each image, there are 100 bounding boxes extracted and for each bounding box the class label and the confidence are provided. On top of the UNITER model, there is a single linear which is used for finetuning. Figure 1a shows a high-level visualization of the model.

4.2 Additional linear layer

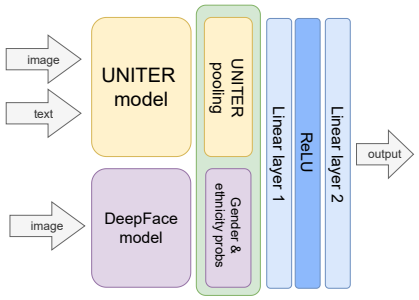
Using more linear layers improves the expressiveness of a network and often leads to increased performance. Many advances in the field were obtained by simply scaling the architecture. Therefore we experiment with adding another linear layer on top of the model with a ReLU non-linearity in between the two layers as depicted in Figure 1b.

4.3 Text bounding boxes

The first, quite intuitive extension is to ignore bounding boxes detecting the captions (the text) as the model already uses the text in the language modality and almost certainly the captions in the visual form (i.e. pixel values) do not provide any information, but rather introduce unnecessary noise. This change quite certainly will provide a small improvement.



(a) baseline architecture



(b) architecture including gender and ethnicity predictions and an additional linear layer

Figure 1: Visual representations of the baseline and modified architectures.

4.4 Upsampling difficult memes

As discussed in Section 3, the dataset contains many memes that are non-trivial to classify. To mitigate this problem, we upsample those examples so that the model is more exposed to them which hopefully will result in better recall. More precisely, we upsample benign (text and image) confounders and their corresponding hateful memes (i.e. we only upsample a hateful meme if there exists a non-hateful meme that contains the same image and different text or the same text and different image).

4.5 Including gender and ethnicity probabilities

When visually inspecting the data, we realized that many hateful memes target specific genders and ethnicities. This is confirmed by (Kiela et al., 2020) who show that Race and Ethnicity, Religion and Gender are the top 3 most frequently targeted protected categories. We use an additional model (Serengil and Ozpinar, 2020) which was specifically trained to detect those categories. We believe that by sending our model stronger signal about gender and ethnicity present on a particular image, it will better understand potential hatefulness.

Modification	Dev. AUCROC	Test AUCROC
Baseline (with normalization, 1 linear layer)	0.7650 (0.0069)	0.7598
Baseline + 2 linear layers (LL)	**0.7765** (0.0036)	0.7501
Baseline + Text Filtering (TF)	0.7663 (0.0142)	0.7544
Baseline + Upsampling (image and text) (Up)	0.7696 (0.0121)	**0.7712**
Baseline + Gender & Race probs	0.7710 (0.0038)	0.7526
Best: Baseline + 2 LL + TF + Up	0.7623 (0.0051)	**0.7717**

Table 1: Absolute improvements over baseline for each modification and the best combined score. For the development set, we take the mean AUCROC over 3 seeds and we evaluate the median seed on the test set.

5 Experiments

5.1 Experimental details about gender and ethnicity probabilities

To obtain gender and ethnicity probabilites, we use the *deepface* library (Serengil and Ozpinar, 2020) together with our own, custom modifications [2]. The whole procedure is as follows: we feed images to the *deepface* model which returns probabilities for each class label. For gender we have two labels - "female" and "male" and for ethnicity we have six - "asian", "indian", "black", "white", "middle eastern", "latino hispanic". These eight values are then concatenated in one layer together with the output of the UNITER model pooling i.e. the multimodal representation of input image and text. Such concatenated layer is then passed through the top layer(s) to achieve the final prediction, as shown in 1b. We explain our custom modifications in detail in the Appendix A.1

6 Results

Table 1 compares absolute performance of each technique. We use AUCROC for evaluation as this metric is agnostic to changes in the predicted probabilities threshold. We realized in our study that the scores are quite dependent on the random seed, therefore we repeated each experiment with 3 different random seeds and we present their average, together with standard deviation. Moreover, it is vital to notice that there is a large discrepancy between development and test AUCROC, for example adding linear layers (Row 2) obtains the best score on development set (0.7765) and the worst on test set (0.7501). Results of text filtering are barely above the baseline for the development set and worse than the baseline for the test set. Upsampling seems to be the best extension as it obtains decent results on development set (0.7696) and substantially outperforms the rest on the test set. Probably the biggest disappointment is the last extension - gender and ethnicity probabilities. Despite the second highest results on the development set, it is below the baseline on the test set. In the last row, we present the best result obtained using a combination of multiple modifications - 2 linear layers, text filtering and upsampling - which is relatively small improvement over the score of upsampling alone.

[2] https://github.com/blazejdolicki/deepface

7 Discussion

In this section, we discuss the results and conduct additional analysis in order to understand both positive and negative results and answer our research questions posed in Section 1.

7.1 Impact of text filtering

The intuition behind text filtering was that bounding boxes containing text were introducing nothing but noise, so removing them would help the model. By extension, the more text bounding boxes were included in the image, the more the performance would increase after removing them. To measure the performance increase we consider the cross-entropy loss difference per image between the baseline and the text filtering setup. Details can be found in the Appendix A.

Pearson correlation between the loss improvement L_{diff} and number of text boxes in the image is equal to -0.068 (with p-value 0.176) which shows that there is absolutely no correlation between the two variables. This finding definitively shows that our intuition was incorrect and explains why text filtering obtained only slight improvements or even degraded the performance.

	% actual	% predicted
female	16.06	17.81
male	83.94	82.19
asian	7.77	9.59
indian	2.59	2.05
black	17.10	13.70
white	58.03	62.33
middle eastern	10.36	8.22
latino hispanic	4.15	4.11

Table 2: Actual and predicted % hateful memes by gender and ethnicity classes

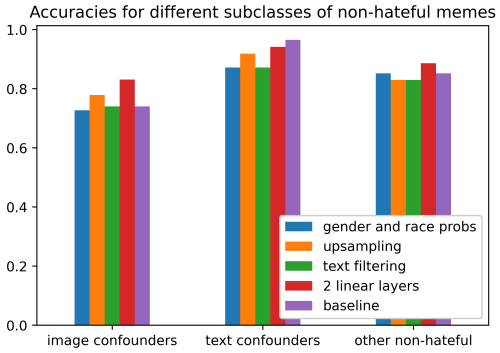


Figure 2: Subclass accuracies of non-hateful memes

7.2 Impact of gender and ethnicity probabilities

Visual inspection of predictions leads to two observations. Firstly, the gender and ethnicity recognition model is biased towards male gender and white ethnicity. Secondly, our multimodal architecture seems to overfit on particular gender or ethnicity classes. We examine the second insight quantatively by comparing the actual percentage of hateful memes in each gender and ethnicity class with the predicted percentage. Large discrepancies between actual and predicted percentages would be a sign of bias/overfitting towards some class. Table 2 shows that the actual and predicted percentages are rather similar for gender probabilities. However, it does seem that the model is biased towards classifying images with people of white ethnicity as hateful (% predicted is 4.30 points higher than % actual) and, on the other hand, more prone to classify images with people of black ethnicity as non-hateful.

7.3 Subclasses of memes

In Figure 2 we break down the non-hateful class of memes into 3 subclasses: image confounders (memes that changed from hateful to non-hateful after replacing their image), text confounders (memes that changed from hateful to non-hateful after replacing their text) and finally other non-hateful images. The latter should be the easiest to classify, the first two are especially difficult to classify. Upsampling is superior at classifying image confounders (bar group 1) which is in line with our expectations, since this was the group of memes that were upsampled. On the other hand, even though text confounders were also upsampled, for that group none of our modifications improves over the baseline. This clearly shows that our modifications focused on improving the visual modality rather than text.

8 Conclusion

We implemented a number of extensions of our baseline. Adding additional layer seemed to help based on AUCROC on the development set, but the test set results showed that it was merely overfitting. Our extensive analysis showed that text filtering does not have the effect that we expected and it might even decrease the baseline results. Upsampling confounders and their corresponding hateful memes seems to be the best improvement on its own, yet it still doesn't match the baseline in classifying text confounders. Finally, adding gender and ethnicity probabilities suffers from incorrect predictions of the external classifier which implements noise and amplifies bias towards certain classes.

9 Future work

Despite negative results we believe that using gender and race probabilities could still prove to be useful after some improvements such as using a better model, trained on more diverse data. In this work, we focused on improvements for the visual modality which is reflected by our results on text confounders in Figure 2. Therefore, an interesting research direction would be to focus on the language aspect, e.g. using a model for sentiment classification to simplify detecting unimodal hate.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Sefik Ilkin Serengil and Alper Ozpinar. 2020. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

A Appendices

A.1 Custom changes in the 'deepface' library

Deepface library performs gender and ethnicity classification only based on the first face detected in the image. However, there are often multiple people appearing in the memes, therefore we modify it so that it computes probabilities based on not one, but 10 detected faces which are then averaged to provide the final probabilities of the image. Another modification was performed to return gender probabilities instead of a hard label.

A.2 Cross-entropy loss difference between text filtering and baseline

To measure performance improvement per image we need a metric that increases when the probability of an example becomes closer to the ground truth label i.e. closer to 1 for a positive example and closer to 0 for a negative example. This requirement is exactly satisfied by the difference in cross-entropy loss for each example between the text filtering and the base probability. We express it in Equation 1. L_{tf} is the cross-entropy loss with probability of the positive class $\hat{p}_{tf}$ obtained using the text filtering modification, L_{base} is the loss with probability of the positive class $\hat{p}_{base}$ obtained using the baseline. The ground truth binary label is denoted with y.

$$
\begin{aligned}
L_{diff} &= L_{tf} - L_{base} \\
L_{tf} &= y \log \hat{p}_{tf} + (1-y) \log (1-\hat{p}_{tf}) \\
L_{base} &= y \log \hat{p}_{base} + (1-y) \log (1-\hat{p}_{base})
\end{aligned}
\tag{1}
$$