Extending multimodal models for the Hateful Memes Challenge

Rodrigo Alejandro Chávez Mulsa, Błażej Dolicki

Natural Language Processing 2, University of Amsterdam 2021

University of Amsterdam

Introduction

There is an increasing number of applications where multimodal learning (i.e. using more than one modality such as vision and text) is necessary. One of them is the classification of hateful memes. Often considering only text or only the image is not enough to correctly judge if a meme is hateful, and both of them need to be taken into account. The Hateful Memes dataset [2] concerns exactly this task with 10k carefully collected memes.
We use UNITER [1] as our baseline which obtains strong results on many multimodal tasks.

Fig. 1: Multimodal "mean" meme and benign confounders [2]

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?

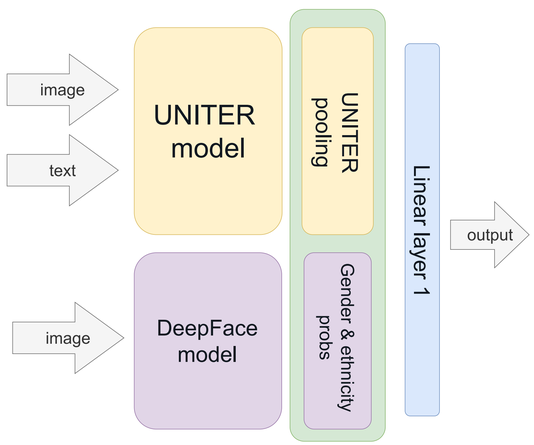

Fig. 2: Model architecture including gender and ethnicity scores

Methodology

Removing bounding boxes around text

The input images contain captions. However, the model already uses the text in the language modality and almost certainly the captions in the visual form (i.e. pixel values) only provide noise. Therefore, we decide to remove bounding boxes containing text.

Upsampling difficult memes

The dataset often contains pairs of memes that share the same text or image but have different labels. We believe that upsampling these cases can enhance the model's ability to combine modalities.

Gender and ethnicity scores

Many hateful memes target gender or ethnicity, therefore we hypothesize that adding gender and ethnicity scores (obtained using the DeepFace model [3]) will send a stronger learning signal.

Experiments and Results

We performed a number of experiments combining multiple modifications together and the results were often surprising. For clarity, we present below the development set score for each modification separately, and then the best result we obtained by combining multiple modifications.

Modification	AUCROC
Baseline (with normalization, 1 linear layer)	0.7582
Baseline + 2 linear layers (2 LL)	0.7743
Baseline + Text Filtering (TF)	0.7620
Baseline + Upsampling 3x	0.7704
Baseline + Gender & Ethnicity probabilities	0.7719
Best: Baseline + 2 LL + TF + Upsampling 3x	**0.7844**

All modifications yield improvement in isolation.

Discussion

The impact of adding an additional linear layer fluctuates depending on the combination. We have seen that it is decremental for gender and ethnicity scores while it boosts the upsampling results.

Text Filtering performs as expected as it obtains a small improvement which is quite stable for different combinations.

Upsampling yield considerable improvements in development and test sets. We believe that the former benefits more from linear layers because it contains more training data.

Adding gender and ethnicity probabilities also obtains good results however, it has more limitations. Firstly, the same gender or ethnicity often occurs in hateful memes and in text confounders, which means that they do not directly determine the target label. Secondly, during error analysis, we realized that the gender and ethnicity detection model [3] is biased towards white men.

Conclusions

- Adding layers is not stable, and its benefits are questionable.
- Text Filtering slightly improves the results and is a recommended modification.
- Upsampling difficult memes improve the results and benefits from the expanded architecture.
- Adding gender and ethnicity probabilities helps but suffers from several limitations.
- All our modification improved the result and the best score is obtained by combining text filtering, 2 linear layers and upsampling image confounders with their corresponding hateful memes.

References

[1] Yen-Chun Chen et al. "Uniter: Universal image-text representation learning". In: *European Conference on Computer Vision*. Springer. 2020, pp. 104–120.

[2] Douwe Kiela et al. "The hateful memes challenge: Detecting hate speech in multimodal memes". In: *arXiv preprint arXiv:2005.04790* (2020).

[3] Sefik Ilkin Serengil and Alper Ozpinar. "LightFace: A Hybrid Deep Face Recognition Framework". In: *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE. 2020, pp. 23–27. DOI: 10.1109/ASYU50717.2020.9259802.