Hateful Meme Classification

Aditi Ramesh Athreya, Shanthi Lekkala
University of Colorado Boulder

University of Colorado Boulder

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation:

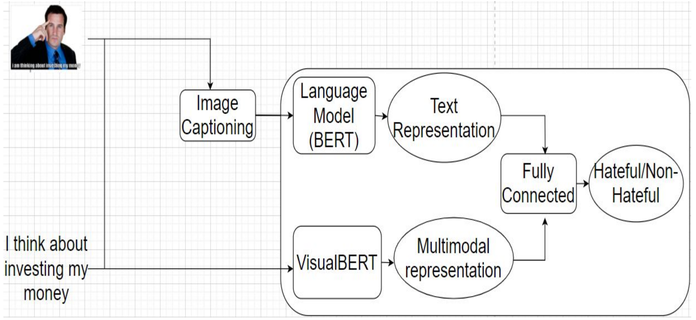
One of the most effective tools for promoting hatred online today are memes.Flagging hateful memes before they spread can prevent any impending acts of violence and harm, and can help reduce increasing community division caused by such hateful content online.

Related Work:

Trained Human achieved an auROC of 82.65 and the Baseline model with VisualBERT achieved an auROC of 71.4. Wiinning solutions of Facebook challenge are mostly ensemble models with auROC in range 78 and 84.2. Though these models exhibit high performance, they are computationally very expensive and cannot be used in real world applications.

Idea & Method:

Our idea is to create a model with good performance which is computationally less expensive. Figure illustrates the architectural framework of our model To create this, we plan to leverage we use Image captioning with BERT and VisualBERT models together for the classification. We used bottom up top down pre-trained image captioning model to generate captions for the meme images.



Model Architectural Diagram

Method :

Once the captions are generated, they are passed through a pre-trained language model BERT to get textual representation of the caption, which is a tensor of 768 dimension. While we generate the textual representation, we parallelly generate multimodal representation of the meme by passing both meme image and meme text to pre-trained VisualBERT model, which again is a tensor of 768 dimension. Finally these two representations are concatenated together and passed to a fully connected network which classifies if the meme is benign or hateful.

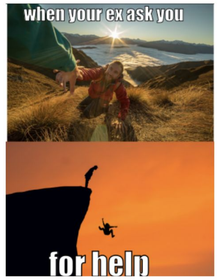
The model is trained on training dataset with approx. 8500 images for 30 epochs with 64 batch size, categorical cross entropy loss and adam optimizer. Used NVIDIA T4 GPU on google colab for training and testing. The performance of the model is evaluated in terms of AUROC and accuracy on test dataset

Results :

MODEL	AUROC	Accuracy
Baseline Model - VisualBERT	71.41	64.73
Top performing ensemble model by Zhu et al.	84.5	73.2
Our Model	**74.3**	**69**

- The results demonstrate that when image captioning is added, the model's auroc score rises by almost 3% in comparison to the baseline model.
- The accuracy of our model is almost 5% more than the baseline model
- Caption helps interpret the mage and its similarity with meme text improved the model performance
- If the memes have several sub-images, the captions generated can be incorrect, leading to wrong prediction. So, we have to look for efficient methods of image captioning in such cases.
- There is scope for performance enhancement

Caption: military commander
Actual Label: Hateful
Predicted Label: Hateful



Caption: digital art
Actual Label: Benign
Predicted Label: Hateful

Analysis:

- The hate meme and its benign confounders are better used to verify the performance of the model
- Here, 1st is a hate meme and 2nd, 3rd are its benign confounders and all three are correctly classified by the model showing that it uses both text and image to classify.

Correctly classified : **Hate**
Caption: actor at the premiere screening of film

Correctly classified : **Benign**
Caption: actor at the premiere screening of film

Correctly classified : **Benign**
Caption: a dumpster full of trash

Findings and Insights:

- Understanding memes is complex and the dataset provided contains only 12k+ images suggesting model underfitting.
- Image captioning helps interpret the image in better way thereby increasing the model performance

Summary/Conclusion

Using image captioning improved the model performance and the current architecture is simpler than ensemble model but there is scope for performance enhancement. Finally, if the performance of this model is improved, it can be used to flag hate memes online and stop hate spread.

References

- Kiela, D., Firooz, H et. al. The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems 33 (2020) 2611–2624
- Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring hate speech detection in multimodal publications. In: Proceedings of the IEEE/CVF on applications of computer vision. (2020)
- Zhu, R.: Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. arXiv:2012.08290 (2020)
- Muennighoff, N.: Vilio: state-of-the-art visio-linguistic models applied to hateful memes. arXiv preprint arXiv:2012.07788 (2020)