Hateful Memes Detection Proposal

Project Summary:

Social media is becoming a major part of our everyday life. It makes communication easier and spreads information faster than ever before. It can also be used for hate speech - used to "attack a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor"[1]. Online hate speech can encourage hateful crimes and social polarization. Some hate speech is represented in a multimodal way, a combination of image and text like memes. The text or image may seem neutral if looked at individually but becomes hate speech when viewed together with the others. In recent years, AI made significant progress in unimodal hateful speech detection, but multimodal hate speech detection remains a difficult technical challenge. Given that the current multimodal model performances applied in hateful speech memes detection are still relatively poor compared to humans, there is a lot of room for improvements. In this project, we will focus on (exploring a variety of ways to improve model performances in) detecting hateful speech in multimodal memes, (which is also helpful in solving a real-world problem.

Approach:

We are going to understand all baseline models and the pre-trained model provided by Hateful Memes Challenge, then we plan to implement and combine VisualBERT and UNITER models as the base model to see the performance, and analyze different models which have been pre-trained on datasets like COCO and try to transfer them to downstream Hateful Memes task to see if the performance can be improved. Then we are going to add an embedding algorithm to improve the accuracy. We will also explore applying different object detectors specifically trained for memes, like content, if possible. We are going to use the dataset of competition phase 1 to train and test our model. Finally, we will compare our model with the pre-trained model and existing model. We will use ROC and AUC, F1 score, and Accuracy as our evaluation metrics of model performances.

Alternatively, we propose an ensemble approach to multimodal sentiment analysis to detect hateful memes. The purpose of the work is to improve the sentiment prediction accuracy of the baseline implementations of the Hateful Memes Challenge by analyzing and ensembling the text-based sentiment and image-based sentiment.

Resources/Related Work:

[1] "UNITED NATIONS STRATEGY AND PLAN OF ACTION ON HATE SPEECH"
[2] Li et al., "VisualBERT: A Simple and Performant Baseline for Vision and Language."
[3] Chen et al., "UNITER: UNiversal Image-TExt Representation Learning."
[4] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications.
[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context.
[6] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining.