Team: Thirsty Learners
Facebook Project: Yes
Project Title: The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes
Project Summary:

The urgency of detecting hate speech in memes comes from a combination of increasing hate speed on social media and lacking effective machine learning tools to identify/stop them. Detecting hate speech in memes is complicated for AI as it requires reasoning about subtle cues in a multimodal setting (e.g. image and text). The Current state-of-the-art multimodal models perform much poorer compared to human performance, indicating that there is much room for research and improvement. The goal of this project will be construct some multimodal models to improve the hate speech detection task.

Approach:

- We first replicate the facebook unimodal and multimodal pretraining baselines: Visual BERT and Visual BERT COCO [1-5].
- Since the Facebook Hateful Memes Competition has been completed and top winners solutions have been publicly available on Github https://github.com/drivendataorg/hateful-memes/ [5], we will further replicate the third winner's solution which is applying grown training set, extracting image features using Detectron algorithm, fine-tuning pretrained Visual BERT model, and applying majority voting [6].
- With a finely tuned Visual BERT model, we plan to further improve the model performance by applying the following techniques:
  - Find more similar datasets on the web, or employ data augmentation techniques to further grow the datasets [6].
  - Apply the top winner's approach which is to remove the text from the image first before applying the object detection (Detectron) algorithm [7].

Resources/Related Work:

[1] Kiela, D., Firooz, H., Mohan A., Goswami, V., Singh, A., Ringshia P. & Testuggine, D. (2020). *The* Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv preprint arXiv:2005.04790
[2] 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.
[3] Facebook's MMF framework:
https://colab.research.google.com/github/facebookresearch/mmf/blob/notebooks/notebooks/mmf_hm_example.ipynb#scrollTo=1mB-z-6XWdBd (Links to an external site.)
[4] Facebook's Hateful Memes Baselines Reproduction:
https://github.com/facebookresearch/mmf/tree/master/projects/hateful_memes (Links to an external site.)
[5] Top Github submission of Hateful Memes Challenging:
https://github.com/drivendataorg/hateful-memes/ (Links to an external site.)
[6] Riza Velioglu, Jewgeni Rose. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge: https://arxiv.org/abs/2012.12975 (Links to an external site.)
[7] Ron Zhu. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. https://arxiv.org/abs/2012.08290 (Links to an external site.)
[8] Introduction to Multimodal Deep Learning :
https://heartbeat.fritz.ai/introduction-to-multimodal-deep-learning-630b259f9291(Links to an external site.)
[9] 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.
[10] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In The IEEE Winter Conference on Applications of Computer Vision, pages 1470–1478, 2020.
[11] Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. Interpretable multi-modal hate speech detection. In AI for Social Good Workshop at the International Conference on Machine Learning, 2019.

[12] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. Exploring deep multimodal fusion of text and photo for hate speech classification. In Proceedings of the Third Workshop on Abusive Language Online, pages 11–18, 2019.

[13] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950, 2019.

[14] J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, and A. Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In Proceedings of EMNLP, 2019.

Datasets: Facebook Hateful Memes Dataset

DrivenData:

https://www.drivendata.org/competitions/64/hateful-memes/page/214/ (Links to an external site.)

Team Members:

Guolin Yao

Jisheng Chen

Yao Xu

Looking for more members:

No