Hateful Memes Project

Proposal Guideline

- Motivation:
  - Detecting hateful speech in multimodal memes
  - Hateful content detection is one of the areas where AI can have significant impact. Due to the surge of internet content it is difficult for humans to classify or filter out hateful content. Hence leveraging AI can be useful in these scenarios.
- Approaches:
  - Current approaches for identifying hateful multimodal content include unimodal models(BERT, XLM-R, Roberta), multimodal models(BERT+Resnet) that were unimodally pretrained (where for example a pretrained BERT model and a pretrained ResNet model are combined in some way), and multimodal models that were multimodally pretrained(MMBT, VisualBERT, ViLBERT).
  - The Hateful Memes paper report some strong baselines using different types of models for Hateful Meme classification https://arxiv.org/abs/2005.04790
    - All the baselines are available and reproducible using MMF framework (https://github.com/facebookresearch/mmf/tree/master/projects/hateful_memes)
    - MMF can also be extended easily to run experiments with new models
- Metrics:
  - Datasets: Hateful Memes dataset (https://arxiv.org/abs/2005.04790)
    - Metrics to measure are AUC-ROC, F1 Score and Accuracy. Details about metrics and the dataset : https://www.drivendata.org/competitions/64/hateful-memes/page/206/
    - Leaderboard on Hateful Memes : https://www.drivendata.org/competitions/64/hateful-memes/leaderboard/
  - Baseline model performing best on Hateful Memes dataset is VisualBERT model pretrained on COCO Captions. The model is pretrained using self-supervised pretraining proxy-tasks and then fine-tuned on Hateful Memes dataset. Other SOTA models are MMBT, Vilbert etc.
- Scope:
  - Possible directions:
    - Implementing new classification layers for measuring model confidence.
    - Analyse pretraining on datasets like COCO and explore techniques to improve transfer to downstream Hateful Memes task.
    - Use external sources to add knowledge about personalities and objects present in an image like Wikipedia.
    - Use object detectors specifically trained for meme like content.
    - Use better different text encoders like Roberta, XLM-R etc replacing BERT.
    - Use data augmentation to generate more memes automatically.
  - How much work in each direction would justify a good grade?

        - It is difficult to be precise / formulaic about this, but a couple of example scenarios (not an exhaustive list) that would justify a good grade:
            - Students use pre-train models, qualitatively and quantitatively analyze the results, propose a concrete, reasonable idea that they believe should help make the results better, and successfully implement the idea. Note that whether the idea improves results or not should not play a big role in the final grade.
            - Students propose a concrete, reasonable approach to train a model on the dataset, implement the idea, and successfully train the model. Reasonable approaches will be better than the SOTA baselines.
    - Teams of 2-3 people would be ideal for this project.
- Resources:
    - Hateful Memes Dataset : https://arxiv.org/abs/2005.04790
    - Hateful Memes Challenge : https://www.drivendata.org/competitions/64/hateful-memes/
    - 2/4/8 GPUs training for 6-12 hours for each experiment
        - SOTA model used 2 V100 GPUs for 6 hours to complete training
    - Colab Notebook : https://colab.research.google.com/github/facebookresearch/mmf/blob/notebooks/notebooks/mmf_hm_example.ipynb
    - Tutorial : https://mmf.sh/docs/challenges/hateful_memes_challenge