Background & Data

Problem:
Given a meme, can our model correctly classify it as either hateful or non-hateful based on the combined input of the text and image?

Motivation:

- This task involves multimodal classification, which requires the model to effectively combine representations from drastically different modalities.
- Three main methods that have been studied:
  1. Unimodal models
  2. Multimodal models with unimodal pre-trainings
  3. Multimodal models with multimodal pre-trainings

Goal:
Implement and improve on the vanilla VisualBERT model.

Data:
- From Meta’s Hateful Meme challenge in 2020
- 11,040: 8,500, 540, 2,000 in train, dev, and test
- Each set contains: 10% unimodal hate, 40% multimodal hate; 20% benign text confounder, 20% benign image confounder, and 10% non-hateful - Challenging Set!

Analyses

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Non-Hateful</th>
<th>Hateful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None-Hateful</td>
<td>1202 (60.1%)</td>
<td>48 (2.4%)</td>
</tr>
<tr>
<td>Hateful</td>
<td>705 (34.2%)</td>
<td>45 (2.25%)</td>
</tr>
</tbody>
</table>

Confusion matrix of VisualBERT Fairface

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Non-Hateful</th>
<th>Hateful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None-Hateful</td>
<td>870 (43.5%)</td>
<td>380 (19.0%)</td>
</tr>
<tr>
<td>Hateful</td>
<td>424 (21.2%)</td>
<td>326 (16.3%)</td>
</tr>
</tbody>
</table>

Confusion matrix of Baseline Model

- Each model wrongly classifies a meme as non-hateful more often than wrongly classifies a meme as hateful.
- From Model 1 to 3, more likely to classify a hateful meme as non-hateful – which explains why AUC curve is lower for Model 3 (VisualBERT Fairface)

Model Architecture and Training

Baseline
- ResNet50 Pretrain for Images
- Sentence-BERT Pretrain for text
- Concatenation Fusion
- Hidden Size: 1200

VisualBERT
- Detectron2 R-CNN for Images
- Uses ResNet101 Backbone Model
- BERT Tokenizer for Text
- VisualBERT Pretrain
- NLYR2-COCO Pretrain Weights
- Hidden Size: 1000

General Feed-Forward Block
- Linear Layer (Kaiming)
- LayerNorm
- Leaky ReLU

Dropout (tuned)

Feature Extraction
- FairFace
  - Face in Image?
  - If Face, what race/gender/age?
  - Concatenated after embedding extraction

Additional Details
- All Models trained for 10 epochs
- Training set manually balanced

Images that all models wrongly classified as hateful.
Images that the vanilla VisualBERT classified correctly, but not the VisualBERT Fairface.
VB Fairface relies too heavily on facial tags, and these tags could be wrong in the first place.
Models are trained to be biased because of biases in our society.
Images that VisualBERT Fairface classified correctly but not the vanilla VisualBERT. All are actually non-hateful.
The bottom left may demonstrate that racial tags may be helpful.

Results

Accuracy

AUC

Future Work

Due to time and compute limitations, there is still room for improvement. These include:

1. Train on more data if available
2. Explore more advanced fusion techniques, since we use early fusion with simple concatenation (e.g. CNN or RNN-based fusion)
3. Explore other feature extractions (e.g. Web Entity Detection)
4. Improve the current model architecture to decrease probability of wrongly classifying memes as non-hateful
5. Use the current feature tags but fuse it prior to the VisualBERT pre-training

References