Multi Modal Learning - Hateful Meme Detection

Hawau Olamide Toyin, Muhammad Huzaifa, Ridwan Bello Salahuddeen, Sharon Chokuwa

Mohamed Bin Zayed University of Artificial Intelligence

Introduction

- **Multi-modal learning** - aims to build models that can process and relate information from multiple modalities.
- **Hateful meme detection** - The hate in a meme is conveyed through both the image and the text; therefore, these two modalities need to be considered, as singularly analyzing embedded text or images will lead to inaccurate identification.
- **Core Challenges**
  - **Representation** - learn computer interpretable descriptions of heterogenous data from multiple modalities
  - **Translation** - process of changing data from one modality to another
  - **Alignment** - identify relations between elements from two or more different modalities
  - **Fusion** - process of joining information from two or more modalities to perform a prediction task
  - **Co-learning** - goal of transferring knowledge between modalities and their representations

Project Objectives

The project's main objectives were as follows:

- Our goal in this project was to maximize our knowledge in multimodal learning when applied to text and images and to also comprehend its challenges
- We experiment on different kinds of fusion (simple concatenation, align fusion, cross fusion) to investigate which fusion yields the best results.
- We also experiment on the best way to project image and text embeddings so that they are very close in order for the model to learn the relations
- We evaluate our model's performance against the baselines

Dataset

We used the Hateful Meme Challenge dataset because it encourages and measures true multi-modal understanding and reasoning of the models, because of "benign confounders".

Hateful Meme Challenge Dataset					Memotion 7k Dataset
Train Set	Dev Seen	Dev Unseen	Test seen	Test Unseen	Train set
8500	500	540	1000	2000	328

Table 1. Hate Meme Challenge and Memotion 7k Dataset Distribution

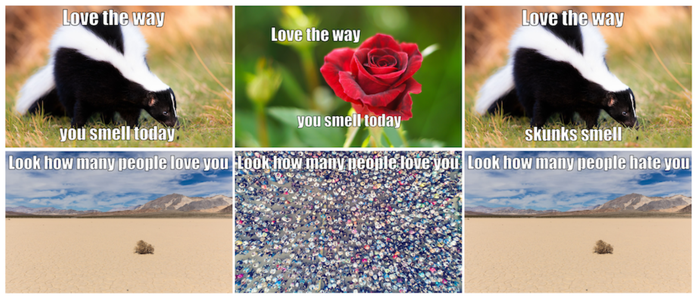

Figure 1. Illustrative examples of multimodal hateful memes. While the memes on the left column are hateful, the ones in the middle are non-hateful image confounders, and those on the right are non-hateful text confounders.

Method

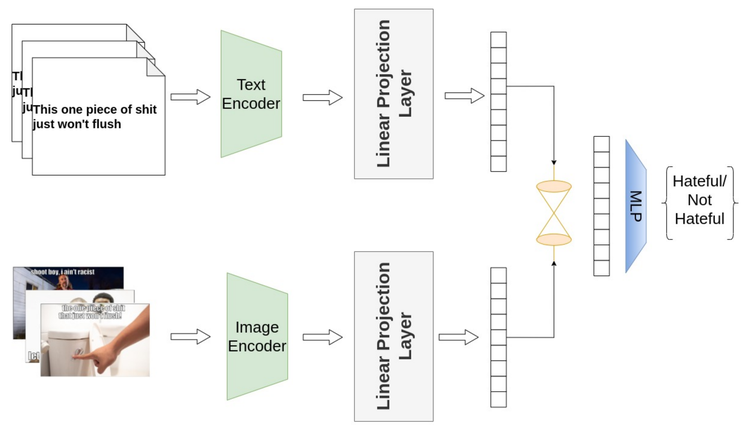

Figure 2. Hateful meme detection with CLIP for modality translation and alignment fusion of features

Mulitimodal Feature Representation
We used a pre-trained CLIP model (ViT-L/14@336px) which uses a vision transformer and a transformer for image and text encoding respectively.

Embedding Projection
We project the embedding vectors from the CLIP model to re-align embedding vectors with the classification objective.

Fusion

- **concat fusion**: stacking the image and text vectors to get a vector of dimension image_dim + text_dim. Fused dimension=1536
- **align fusion**: direct interaction (product) between corresponding features of the image and text vectors. Fused dimension=768
- **cross fusion**: models feature interaction between all features of the image and text vectors through variance and covariance. The number of features becomes squared as the interaction matrix is flattened. Fused dimension=529,824

More Data: We introduced 328 new data points from the Memotion-7K dataset.

Experiments and Results

The table below shows the set parameters for our experiments

	Parameters							
	Dropout	Batch size	Base model	Epochs	Learning rate	Weight decay	Activation	Linear Layers
Values	0.3	64	ViT-L14@336px	20	0.001	0.0001	Gelu	3

Table 2. Experiment Parameters

	Training Configuration			Dev Unseen		Test Unseen	
S/No	ALL	Add Memotion	Fusion Method	Accuracy	AUCROC	Accuracy	AUCROC
1	FALSE	FALSE	align	**74.07**	**80.05**	**76.25**	**83.04**
2	TRUE	TRUE	concat	72.04	74.66	74.05	79.48
3	FALSE	TRUE	align	72.04	76.60	74.90	82.13
4	TRUE	FALSE	concat	70.74	75.94	75.80	80.63
5	TRUE	FALSE	align	69.63	74.46	73.50	79.70
6	FALSE	FALSE	concat	69.81	76.01	72.80	79.50
7	TRUE	TRUE	align	72.59	75.53	73.40	79.39
8	FALSE	TRUE	concat	70.37	75.61	74.40	79.97
9	TRUE	TRUE	concat	72.04	74.66	74.05	79.48
12	TRUE	TRUE	align	72.59	75.53	73.40	79.39
13	FALSE	TRUE	align	72.04	76.60	74.90	82.13
14*	FALSE	TRUE	concat	68.52	73.47	73.35	79.99

Table 3. Experiments Results and configuration for our high performing experiments.

S/No 14* uses the OpenAI Clip's projection layers which are not fine-tuned on our dataset.
ALL : use an ensemble of the fused modalities (the text encoding and the image encoding)
Add memotion : add memotion datset to train set
Fusion method : fusion technique used

	Dev Unseen		Test Unseen	
Model	Accuracy	AUCROC	Accuracy	AUCROC
Human	-	-	84.70	86.25
VisualBert $CC_{Ensemble}$	74.23	79.26	76.50	81.08

Table 4. Baseline Results

Conclusion and Future Work

- Our model achieves $\sim$ **2%** improvement over the baseline
- Adding the Memotion dataset does not improve our model performance, as opposed to the baseline's conclusion.
- Training the projection layers achieves relatively better results.
- The align fusion technique performs better across all experiments we ran.
- Future work could be on implementing the cross fusion technique, other ensemble techniques, and fine tuning the base models on our data.

References

[1] Gokul Karthik Kumar and Karthik Nanadakumar.
Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features.
arXiv preprint arXiv:2210.05916, 2022.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al.
Learning transferable visual models from natural language supervision.
In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[3] Riza Velioglu and Jewgeni Rose.
Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.
arXiv preprint arXiv:2012.12975, 2020.