EMNLP 2022
Dec 7-11, Abu Dhabi

NLP for Social Good (NLP4SG)

AI / NLP Research — How to help? — Social Good

Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

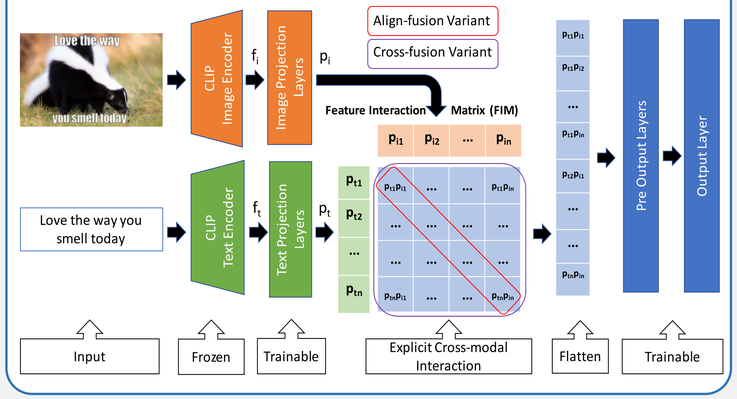
Gokul Karthik Kumar **Karthik Nandakumar**

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE

Abstract

- Hateful memes are a growing menace on social media. While the image and its corresponding text in a meme are related, they do not necessarily convey the same meaning when viewed individually.
- Multimodal pre-training can be beneficial for this task because it effectively captures the relationship between the image and the text by representing them in a similar feature space. Furthermore, it is essential to model the interactions between the image and text features through intermediate fusion.
- Most existing methods either employ multimodal pre-training or intermediate fusion, but not both.
- We propose the Hate-CLIPper architecture, which explicitly models the cross-modal interactions between the image and text representations obtained using Contrastive Language-Image Pre-training (CLIP) encoders via a feature interaction matrix (FIM).
- A simple classifier based on the FIM representation is able to achieve state-of-the-art (SOTA) performance on the Hateful Memes Challenge (HMC) dataset with an AUROC of 85.8, which even surpasses the human performance of 82.65.
- Experiments on Propaganda Memes and Tamil Memes also demonstrate the generalizability of the proposed approach.
- Finally, we analyze the interpretability of the FIM representation and show that cross-modal interactions can indeed facilitate the learning of meaningful concepts.

Methodology



- Work done for MSc course **"Trustworthy AI"** project @ MBZUAI
- Project Page: https://github.com/gokulkarthik/hateclipper
- Contact: gokul.kumar@mbzuai.ac.ae

Experimental Results

# proj. layers	# p.o. layers	Fusion	Model	Dev seen	Test seen	# t.params.
1	1	Concat	Baseline	76.72	79.87	3.9M
1	3	Concat	Baseline	79.02	83.73	6M
1	5	Concat	Baseline	78.6	83.8	8.1M
1	7	Concat	Baseline	78.63	83.29	10.2M
1	1	CMAF	MOMENTA	77.36	80.15	4.5M
1	3	CMAF	MOMENTA	76.85	82	6.6M
1	5	CMAF	MOMENTA	79.51	83.35	8.7M
1	7	CMAF	MOMENTA	78.88	82.4	10.8M
1	1	Cross	HateCLIPper	**82.62**	85.12	1.1B
1	3	Cross	HateCLIPper	82.19	82.66	1.1B
1	1	Align	HateCLIPper	81.18	85.46	2.9M
1	3	Align	HateCLIPper	81.55	**85.8**	5M
1	5	Align	HateCLIPper	80.88	85.46	7.1M
1	7	Align	HateCLIPper	81.09	84.88	9.2M

Table 1: AUROC of Hate-CLIPper variants and other fusion approaches on HMC dataset. Expansions: proj. -> projection; p.o. -> pre-output; t.params. -> trainable parameters; M -> million; B -> Billion.

# proj. layers	# p.o. layers	Fusion	Model	Dev	Test	# t.params.
1	1	Concat	Baseline	89.9	88.93	4M
1	3	Concat	Baseline	89.9	88.82	6.1M
1	5	Concat	Baseline	89.18	88.55	8.2M
1	7	Concat	Baseline	89.83	88.82	10.3M
1	1	CMAF	MOMENTA	89.11	88.34	4.5M
1	3	CMAF	MOMENTA	89.75	88.73	6.6M
1	5	CMAF	MOMENTA	89.11	88.34	8.7M
1	7	CMAF	MOMENTA	89.61	88.66	10.8M
1	1	Cross	HateCLIPper	**90.98**	**90.41**	1.1B
1	3	Cross	HateCLIPper	**90.98**	89.95	1.1B
1	1	Align	HateCLIPper	89.11	88.34	2.9M
1	3	Align	HateCLIPper	89.11	88.34	5M
1	5	Align	HateCLIPper	89.68	88.66	7.1M
1	7	Align	HateCLIPper	89.68	88.66	9.2M

Table 2: Micro F1 scores of Hate-CLIPper variants and other fusion approaches on Propaganda Memes dataset. Expansions: proj. -> projection; p.o. -> pre-output; t.params. -> trainable parameters; M -> million; B -> Billion.

Model	Dev Seen	Test Seen
Human	-	82.65
Image-Grid	52.33	53.71
Image-Region	57.24	57.74
Text-BERT	65.05	69
Late Fusion	65.07	69.3
Concat BERT	65.88	67.77
MMBT-GRID	66.73	69.49
MMBET-Region	72.62	73.82
ViLBERT CC	73.02	74.52
Visual BERT COCO	74.14	75.44
CLIP-ViT-L/14-336px	77.3	-
SEER-RG-10B	73.4	-
FLAVA w/o init	77.45	-

Table 4: AUROC of different models on the HMC dataset, compiled from Kiela et al. (2020); Goyal et al. (2022); Singh et al. (2021).

Model	Test
Random	7.06
Majority Class	29.04
ResNet-152	29.92
FastText	33.56
BERT	37.71
FastText + ResNet-152	36.12
BERT + ResNet-152	38.12
MMBT	44.23
ViLBERT CC	46.76
VisualBERT COCO	48.34
RoBERTa	48
RoBERTa + embeddings	58
Ensemble of BERT models	59

Table 5: Micro F1 scores of different models in Propaganda Memes dataset, compiled from Sharma et al. (2022); Dimitrov et al. (2021)

- **Intermediate fusion** with the CLIP encoders is better than early (MMBT, ViLBERT, and VisualBERT) and late fusion methods
- **Cross and Align fusion** variants of Hate-CLIPper achieve the best AUROC
- Results on **Propaganda Memes** confirm the same findings
- Hate-CLIPper also achieves SOTA on **Tamil Memes** with the micro F1 score of 59
- Ablation experiments (i) with unfrozen CLIP encoders (ii) Non-CLIP encoders (mBERT, VIT) resulted in significantly poor scores

Interpretability

Conclusion

- We emphasized the need for intermediate fusion and multimodal pretraining for hateful meme classification.
- We proposed a simple end-to-end architecture called Hate-CLIPper using explicit cross-modal CLIP representations, which achieves the state-of-the-art performance quickly in 14 epochs with just 4 trainable layers .
- Our model does not require any additional input features like object bounding boxes, face detection, text attributes, etc
- We also studied the interpretability of cross-modal interactions