CAN DECOUPLING EMBEDDED TEXT FROM IMAGES IMPROVE MULTIMODAL LEARNING?

Siddhant Bikram Shah
Delhi Technological University
siddhantbikramshah@gmail.com

ABSTRACT

Multimodal models have widely been used to process text-embedded images on social media. However, the effect of embedded text on the image encoding process remains unexplored. In this work, we eliminated the text in text-embedded images and compared the intervention's effect on the performance of unimodal and multimodal models. We find that the image encoders of multimodal models utilize linguistic information in the pixel space to a considerable degree. Further, we observe that disentangling text and images can improve multimodal learning under certain circumstances.

1 INTRODUCTION

Multimodal foundation models have shown impressive performance in a multitude of downstream tasks (Li et al., 2023). Recently, frameworks harnessing such models have successfully been applied to social media image processing (Kumar & Nandakumar, 2022). These frameworks commonly use the CLIP model (Radford et al., 2021) to learn meaningful representations from images and text. A key aspect of images on social media is text overlay, where text is added on top of images to make them more informative, expressive, and easier to share. While the image encoders of multimodal models harness optical character recognition (OCR) to extract meaningful text information from pixel data (Burbi et al., 2023), they show subpar zero-shot OCR performance on out-of-distribution data (Meyer, 2023). Further, the visual and linguistic components of images may individually convey contrasting ideas, which can confound the image encoding process. These observations give rise to the following questions: To what extent do image encoders utilize the text embedded in images? Is linguistic content in the pixel space redundant in multimodal processing? Do vision-language models rely on one modality more than the other?

In this work, we attempt to answer the aforementioned questions by investigating the consequences of disentangling visual and linguistic content in text-embedded images. We remove text from images by combining text detection with image inpainting. We perform these interventions on the CrisisHateMM dataset (Bhandari et al., 2023) and the Hateful Memes Challenge dataset (Kiela et al., 2020), and evaluate the performance of CLIP under these interventions. Our results show that while CLIP does rely on linguistic information in the pixel space, this information may benefit or confound the learned image representations depending on the circumstances.

2 METHODOLOGY

We evaluate our method on two datasets, CrisisHateMM and HMC. The CrisisHateMM dataset consists of text-embedded images pertaining to the Russia-Ukraine war scraped from various internet sources. This dataset was chosen as it is representative of real-world social media interactions that may contain hateful visual content. The HMC dataset was released by Facebook as part of the Hateful Meme Challenge and contains hateful synthetic memes targeting religion, race, disability, and sex. This dataset was chosen as it contains seemingly benign images coupled with hateful text, challenging multimodal reasoning further. Dataset statistics for both datasets are presented in Appendix A.1. Sample images from both datasets are presented in Appendix A.2. We perform all our experiments by using a pre-trained frozen CLIP model with a ViT-B/32 backbone and a simple 2-layer classifier as a head. Multimodal representations are created by using a simple concatenation function on the corresponding text and image representations. Further implementation details are available in Appendix A.3. Figure 1 illustrates a summarized workflow for the preprocessing.

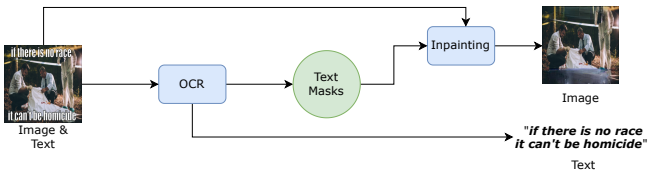


Figure 1: Methodology followed to disentangle text from images. We extract text by using OCR and inpaint over the text masks in the images to obtain disentangled image and text pairs.

3 Results and Discussion

Table 1: Classification performance of CLIP variants on the HMC and the CrisisHateMM datasets. The visual and textual encoders of CLIP were used for image-only and text-only experimentation respectively. TRI stands for Text Removed from Images.

Modality	Model	HMC Dataset			CrisisHateMM Dataset		
		Accuracy	F1-score	AUROC	Accuracy	F1-score	AUROC
Textual	Text-Only	0.71	0.65	0.63	0.78	0.76	0.75
Visual	Image-Only	0.57	0.55	0.54	0.63	0.60	0.60
	Image-Only (TRI)	0.54	0.54	0.53	0.60	0.59	0.59
Multimodal	Baseline	0.67	0.64	**0.65**	**0.80**	**0.80**	**0.79**
	Image + Text (TRI)	**0.74**	**0.66**	**0.65**	0.79	0.77	0.75

The results of our experiments are presented in table Table 1. For both datasets, the text-only model performs significantly better than the image-only model and only marginally worse than the multimodal baselines, which suggests that the text modality is able to capture cues for classifying hate speech better than the image modality. Further supporting this theory, removing text from images slightly hampers performance, as seen in the text-removed image-only models for both datasets. A qualitative analysis of model performance on misclassified memes and comparison with GPT-4V is presented in Appendix A.4. Below, we discuss insights into dataset-specific results.

HMC dataset: The HMC dataset is designed to challenge multimodal reasoning by providing contrasting cues from vision and language modalities. This explains model behavior in our multimodal experiments, as completely disentangling text from images leads to better performance. This also suggests that the CLIP image encoder utilizes semantic information from text present in the pixel space to a considerable degree, as removing this text causes a shift in model performance. Finally, the subpar multimodal performance on HMC without any intervention suggests that the CLIP image encoder is sensitive to images where the text and images convey opposing ideas, such as memes harboring sarcasm or misdirection.

CrisisHateMM dataset: The CrisisHateMM dataset contains real-world samples where the text and images are largely in congruence; therefore, this dataset does not challenge multimodal reasoning to the extent that HMC does, as shown by the superior performance on CrisisHateMM across all metrics. The multimodal model performance decreases when text is removed from images, suggesting that the model favorably leverages text in images despite the same text being supplied to the model as a different modality. This further suggests that text can support the image encoding process when the text and image are analogous to each other.

4 Conclusion

In this work, we investigated the effect of text in the pixel space when being processed by a multimodal model. We introduced several questions and presented possible hypotheses backed by empirical results. We conclude that decoupling these modalities may benefit models when text and image individually convey contrasting ideas. Further, text in the pixel space provides rich semantic cues that may not be redundant even when the same text is supplied to the model as a separate modality. Finally, for hate speech detection, multimodal models seem to rely on linguistic cues more than visual cues. This investigation can help uncover and improve the inner workings of foundation models and large language models as they delve further into multimodality.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1993–2002, 2023.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Mapping memes to words for multimodal hateful meme classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2832–2836, 2023.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

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

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2), 2023.

Fabian Meyer. On the potential and limits of zero-shot out-of-distribution detection. 2023.

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*, pp. 8748–8763. PMLR, 2021.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

A APPENDIX

A.1 DATASET STATISTICS

Dataset statistics for the HMC dataset and the CrisisHateMM dataset are given in Table 2 and Table 3 respectively. We split the data into train/test/validation splits in the ratio of 0.70/0.15/0.15.

Table 2: HMC Dataset statistics.

Split	Hate	Non-Hate
Train	3824	2126
Validation	813	462
Test	813	462

Table 3: CrisisHateMM Dataset statistics.

Split	Hate	Non-Hate
Train	1456	1851
Validation	304	404
Test	298	410

Figure 2: Examples of text-embedded images from the HMC dataset.

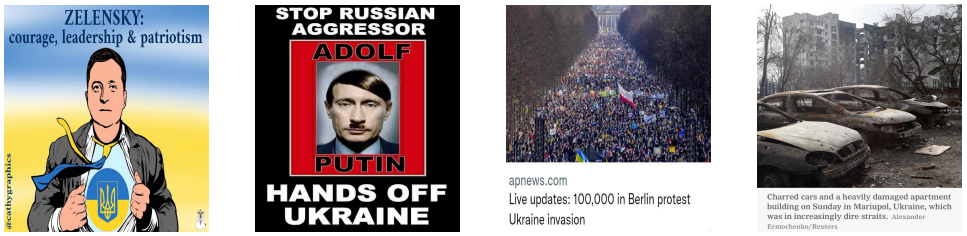


Figure 3: Examples of text-embedded images from the CrisisHateMM dataset.

A.2 Dataset Samples

Samples from the HMC and CrisisHateMM datasets are presented in Figure 2 and Figure 3 respectively.

A.3 Implementation Details

We train our models using Pytorch 2.1.0 and an Nvidia Tesla T4 GPU. The batch size for all experiments was 8. We anchor the random seed to 510 for all applicable libraries. The models were compared across three performance metrics: Accuracy, F-1 score, and AUROC. We use the EasyOCR[1] library for text detection and LaMa (Suvorov et al., 2021) for inpainting.

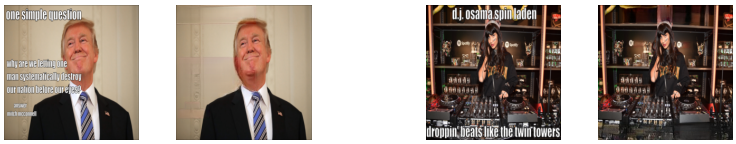
The text was preprocessed to remove non-English characters that the OCR software might have extracted. Further, we removed symbols and non-alphanumeric characters that might distort model predictions. After extracting text, all images were resized to 224x224 pixels.

A.4 Qualitative Analysis

We present examples of misclassified images[2] in Figure 4. We observe that GPT4-V tends to make conservative predictions, leading to false positives in hate speech detection. Further, it cannot make predictions based on a person's identity as faces are blurred for privacy when image inputs are given.

[1]EasyOCR library: https://github.com/JaidedAI/EasyOCR

[2]Disclaimer: Some readers may find the illustrative images disturbing or harmful.



Ground Truth	Non-Hate
CLIP Text-Only	Hate
CLIP Image-Only	Non-Hate
CLIP Image-Only (TRI)	Non-Hate
CLIP Text + Image	Hate
CLIP Text + Image (TRI)	Non-Hate
GPT-4V	Hate
GPT-4V (TRI)	Hate

Ground Truth	Hate
CLIP Text-Only	Hate
CLIP Image-Only	Non-Hate
CLIP Image-Only (TRI)	Non-Hate
CLIP Text + Image	Hate
CLIP Text + Image (TRI)	Hate
GPT-4V	Hate
GPT-4V (TRI)	Hate

Figure 4: Exemplary images from the HMC dataset misclassified by our methods. A red label indicates that the image was misclassified, and a green label indicates that the image was correctly classified. TRI stands for Text Removed from Images.