Exploring Meme Recognition with GANs

Paolo Giaretta, Marco Giuliano, Luca Salvador
Group 6

Problem definition

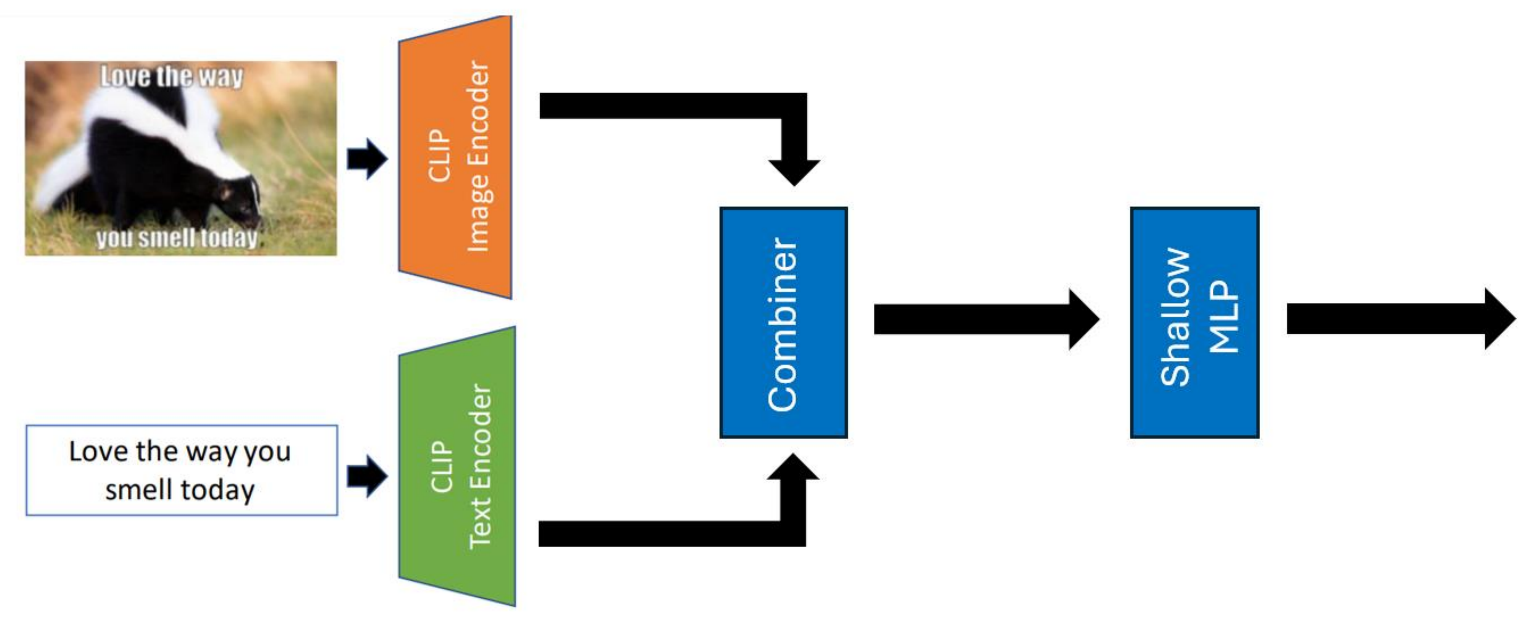
- Hate speech is difficult to detect online due to its **multimodal** nature.
- Classifiers struggle to understand hate speech arising from the **semantical fusion** of image and text.
- Datasets are often **unbalanced**, containing few samples of true multimodal hate speech.
- ***Objective*: Compare GAN augmentation with standard techniques on the same classifier architecture.**

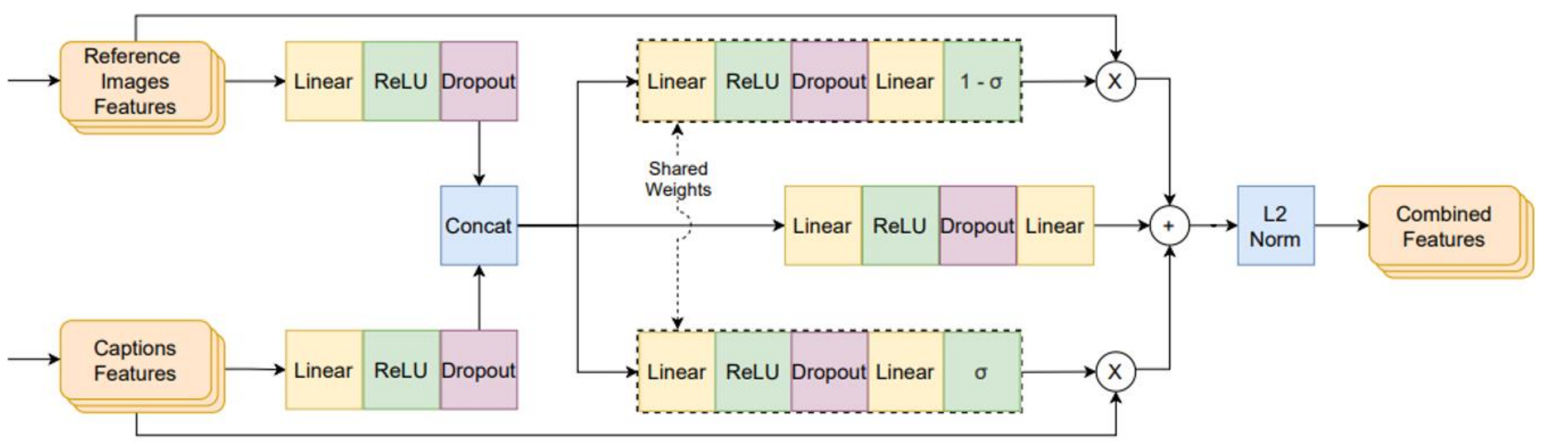
Key Related Works

- The *Hateful Memes Challenge* [1] set a benchmark for hateful meme detection, showing human accuracy at **84.7%.**
- Advances in multimodal models, such as Burbi et al. [2] significantly improved detection by leveraging the *shared* **CLIP embedding space** for image and text.
- **GANs** have also been used for unimodal hate speech, but with a *deep reinforcement learning* approach.

Method

- **CLIP**-based classifier architecture.
- Generation entirely within the CLIP embedding space. **Pre-processing** of the dataset with ViT-L14 encoders.

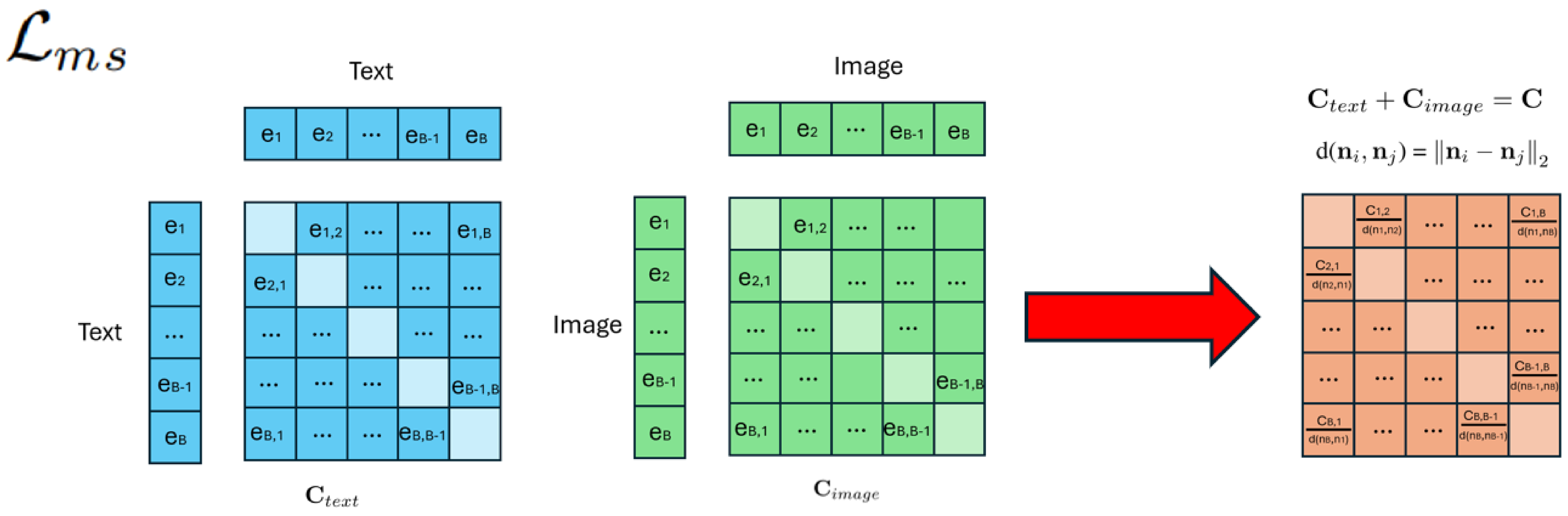




Combiner [2]

- Wasserstein GAN with **gradient penalty** to generate *hateful* text-image embedding pairs.
- Modified **mode-seeking loss** [3] to avoid mode collapse.
- Encourage low cosine similarity for distant noise vectors.

$$\mathcal{L}_{ms} = \frac{1}{N^2 - N} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{(\mathbf{C}_{text}(i,j) + \mathbf{C}_{img}(i,j))}{\|\mathbf{n}_i - \mathbf{n}_j\|_2}$$



- **Filtering** generated samples by:
  - *Similarity* with the dataset.
  - *Similarity* across themselves.
- **Few** *well-chosen* samples perform better than *fully-balancing* the dataset.

Dataset(s)

- **Facebook's Hateful Memes Challenge** [1].
- Confounding examples (non-hate speech)

Validation

- Same classifier trained with *four* methods:
  - On the **unbalanced** dataset.
  - **Weighted** BCE Loss.
  - **Over-sampling** of under-represented class.
  - **GAN** Augmentation.

Methods	Accuracy	AUROC	Better?
Unbalanced	0.752	0.825	
Weighted Loss	0.753	0.823	∼
Over-sampling	0.748	0.826	∼
GAN augmented	0.765	0.829	∼ (√)

Limitations

- Small dataset with features in a high dimensional space.
- High sensitivity of performance when deviating from *optimal* training of the GAN. (**± 1%**)
- **Hyperparameter** choice and **initialization** significantly impact GAN training and augmentation performance.

Conclusion

The small size and high-dimensional nature of the dataset presented **significant challenges**, often leading to **inconsistent** improvements of the classifier working with the GAN-augmented dataset.

The **mode-seeking loss** improves generation diversity. GAN training performance remains inconsistent.

A **larger set** of multimodal hate speech may be required to assess the **efficacy** of this approach.

References

[1] D. Kiela et al., 'The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes', arXiv [cs.AI]. 2021.
[2] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. D. Bimbo, 'Mapping Memes to Words for Multimodal Hateful Meme Classification', arXiv [cs.CV]. 2023.
[3] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, 'Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis', arXiv [cs.CV]. 2019.