Enhancing Meme Recognition with GANs: Exploring GAN augmentation on multimodal classification

Paolo Giaretta [1] **Marco Giuliano** [1] **Luca Salvador** [1]

Abstract

Hate speech online poses a significant challenge due to its multimodal nature. This study aims to improve multimodal hate speech detection by augmenting a dataset using Generative Adversarial Networks (GANs) within the CLIP embedding space. We explore Auxiliary Classifier GANs (AC-GAN), Wasserstein GANs, and the use of a modified mode seeking loss to encourage generation diversity. Data augmentation using GANs is compared with standard balancing techniques to assess its efficacy. *Despite these enhancements, the small size and high-dimensional datasets limited performance gains. Our findings indicate that improvements may be possible, but larger datasets are essential for consistent and reliable GAN training.*

Keywords: Multimodal hate-speech, Wasserstein GAN, mode-seeking loss.

1. Introduction

Hate speech is growing rapidly online, fueled by the spread of easily shared disinformation enabled by digital tools [1]. Detecting and mitigating such content is a significant challenge due to its multimodal nature, requiring the understanding of both textual and visual content. Initial approaches to hate speech detection have mainly focused on unimodal, text-only or image-only analysis, which often fails to capture the interaction between the two modalities. Recent advancements in multimodal models, such as CLIP (Contrastive Language–Image Pre-training), have shown promise in bridging this gap by learning joint representations of text and images. [2] Notably, a recent study leveraging CLIP has achieved state-of-the-art results in multimodal hateful meme classification.[3]

To improve classification performance, we propose to augment our dataset from the Facebook HMC competition [4] using a GAN architecture [5]. Since state-of-the-art classifiers use the pre-trained frozen CLIP encoders, we decided to generate samples directly in the continuous CLIP embedding space, instead of working directly with text and images. In our work we used both an AC-GAN [6], and a standard WGAN architecture. This augmentation technique is compared with standard balancing techniques, such as oversampling, and the use of weighted losses to evaluate improvements on the same classifier architecture.

2. Related Work

The Hateful Memes Challenge (HMC) competition [4] established a benchmark dataset for hateful meme detection and evaluated the performance of humans as well as unimodal and multimodal machine learning models, arriving at an accuracy of 64.73% compared to the accuracy of humans in recognising hate speech of 84.7%.

Recent advancements in multimodal models have significantly enhanced detection capabilities by leveraging both textual and visual data. One notable study, "Hate-CLIPper" [7] utilised an explicit cross-modal interaction of CLIP image and text features to improve the classification of hateful memes. This approach demonstrated substantial improvements in the accuracy of hateful content detection by effectively modelling the correlations between the image and text features.

Additionally, the approach by Burbi et al. [3] leverages a pre-trained CLIP vision-language model and a textual inversion technique to capture the multimodal semantic content of memes. By mapping visual elements to textual descriptions, their method, called ISSUES, enhances the model's ability to understand and classify the content of hateful memes. This study achieved state-of-the-art performance on the Hateful Memes Challenge and HarMeme datasets, demonstrating the effectiveness of CLIP for multimodal hate speech detection.

GANs were introduced by Goodfellow et al. [5] in 2014. Since then, numerous developments and variations of GANs have been explored to address a variety of problems

In the context of hate speech, GANs have also been employed to enhance detection systems. For example, Cao and Lee (2020) [8] utilised GANs to generate synthetic hate

[1] Group 6.

EE-559: *Deep Learning*, 2024.

speech data via a deep reinforcement learning model. This data was used to augment training datasets and improve the robustness of hate speech detection.

Apart from this approach of GAN in the context of hate speech we did not find similar architectures in the literature, especially not in the CLIP embedding space.

3. Method

To combine the CLIP embeddings of image and text, we used the combiner proposed by Burbi et al. [3], which is a simple combiner that combines the two features. After the combiner, the features are passed through a shallow MLP to be classified.
Our focus is on improving this classifier performance by augmenting the dataset through data generation with a GAN architecture. We will work only in the CLIP embedding space for efficiency, so the data will be pre-processed and encoded with the ViT-L14 transformer.

3.1. AC-GAN

Based on the work from [3] and [7] introduced in section 2, we developed a classifier with a similar architecture to recognise hate speech.

Initially, we implemented an AC-GAN that generated text embeddings conditioned on the image embedding. The architecture can be seen in Figure 1. This is done to determine if the auxiliary classifier within the AC-GAN could benefit from both the effect of augmented data and the adversary training process. AC-GANs generate data with associated labels, which can be useful for training classifiers. Our method involved taking images, adding noise to generate text, and creating paired image-text data. This aimed to improve our classifier's performance by providing additional, diverse training examples. This approach was logical for our goal, as it addressed the need for more and better-labelled data for hate speech recognition.

Despite these efforts, we observed that the accuracy of our hate speech classifier did not improve and stable training was difficult to achieve. One reason for this could be that the classifier architecture led the discriminator to learn too well from the dataset, preventing the generator from improving and producing samples resembling the original distribution. We believe the auxiliary classifier learns to ignore data excessively far from the real distribution, thus showing no improvement. It is possible to reach more stable training by reducing the discriminator's capacity, at the cost of diminishing its classification abilities. We believe this method is not well suited for this dataset as for every image embedding there are too few training text examples in the sparse CLIP feature space for the generator to learn effectively.

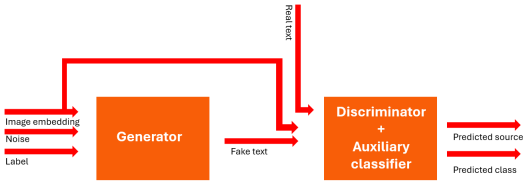

Figure 1. AC-GAN architecture: from CLIP image embedding, noise, and a label we generate text embeddings. The same image embeddings are passed to the discriminator together with the fake or real text embeddings. In the AC-GAN setting, the discriminator predicts both the source and class of the image-text input pair.

3.2. GAN

To address the issue encountered with the AC-GAN, we transitioned to using a Wasserstein GAN (WGAN) to generate (image, text) pairs of the underrepresented class (bimodal hate speech). We employed a loss with gradient penalty during training, as suggested by Arjovsky et al. (2017) [9] for better training stability and more reliable convergence.

Additionally, we experimented with L1 and L2 regularisations, increased the dropout probability, and incorporated batch normalisation. These methods were used to reduce discriminator overfitting and were essential to ensure training convergence.

The first trials showed that generated samples have much higher cosine similarities than the original dataset. We believe this mode-collapse tendency is due to the generator learning a subset of the clusters in the sparse CLIP space. We introduced a consistency loss to encourage the generator to produce samples with similar cosine similarities to the original dataset. We found more diverse results by employing a modified mode-seeking loss to stem the generator in mapping distant noise vectors to distant samples.

3.3. Loss

The loss functions for training the WGAN consist of the discriminator loss and the generator loss. The discriminator loss, adapted from [10], is designed to differentiate real data samples from generated samples. The loss function for the discriminator is given by Equation 1. To these, we added either L1 or L2 regularisation for training.

$$\mathcal{L}_D = \underset{\tilde{\boldsymbol{x}}\sim P_g}{E}[D(\tilde{\boldsymbol{x}})] - \underset{\boldsymbol{x}\sim P_r}{E}[D(\boldsymbol{x})] + \lambda_{GP} \underset{\hat{\boldsymbol{x}}\sim P_{\hat{\boldsymbol{x}}}}{E}\left[\left(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1\right)^2\right] + (\lambda_1 \mathcal{L}_{1,reg} + \lambda_2 \mathcal{L}_{2,reg}) \tag{1}$$

The generator loss is given by Equation 2

$$\mathcal{L}_G = -\underset{\tilde{\boldsymbol{x}}\sim P_g}{E}[D(\tilde{\boldsymbol{x}})] + \lambda_{ms}\mathcal{L}_{ms} + (\lambda_1\mathcal{L}_{1,reg} + \lambda_2\mathcal{L}_{2,reg}) \tag{2}$$

The modified mode-seeking loss $\mathcal{L}_{ms}$ is implemented to avoid mode collapse in the GAN training and the generation of images and text similar to one another. This loss was initially introduced by Qi Mao et al. [11]. We modified the loss by replacing the original distance metric between generated samples (layer activations in the discriminator) with the cosine similarity among generated embeddings.
We define $\mathbf{C}_{\text{text}}(i,j)$ as the cosine similarity between the i-th and j-th text embeddings, computed as the matrix product between the normalised text embedding $\mathbf{E}_{\text{text}}$ tensor and its transpose: $\mathbf{C}_{\text{text}} = \mathbf{E}_{\text{text}}\mathbf{E}_{\text{text}}^{\top}$. Similarly, we compute similarities for generated images. The two cosine similarities are then summed and averaged among all non-diagonal elements.

$$\mathcal{L}_{ms} = \frac{1}{N^2 - N}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}\frac{(\mathbf{C}_{\text{text}}(i,j) + \mathbf{C}_{\text{img}}(i,j))}{\|\mathbf{n}_i - \mathbf{n}_j\|_2} \tag{3}$$

4. Validation

The same classifier architecture is trained with four different methods: the primitive unbalanced dataset, using a weighted BCE loss, oversampling the underrepresented class, and augmenting the dataset with generated samples. We decided to add generated samples to the training dataset by ensuring that the additional samples are sufficiently diverse among each other and semantically diverse to the original dataset to represent different regions in the space. To ensure this, the generated samples are filtered based on the cosine similarity with the original dataset, previously added samples, and the classifier's predicted probability of hate speech. We have found that, on average, adding fewer but more diverse samples performs better than fully balancing the dataset with generated data.
We compared the classification accuracy and AUROC metrics on the test unseen dataset while using the dev unseen as validation. The results are reported in Table 1.

We notice the augmentation performance with generated data is very sensitive to the GAN training process. Slight deviations from convergence to an optimal Nash equilibrium point often lead to worse performance of the GAN-augmented dataset. Furthermore, the GAN training process is very sensitive to hyperparameter choice and initialisation, often leading to no classification improvements. However, for successful GAN training, the performance improvements

METHODS	ACCURACY	AUROC	BETTER?
UNBALANCED	0.752	0.825	
WEIGHTED LOSS	0.753	0.823	$\sim$
OVER-SAMPLING	0.748	0.826	$\sim$
GAN AUGMENTED	0.765	0.829	$\sim(\surd)$

Table 1. ACCURACY AND AUROC FOR DIFFERENT METHODS

are consistently higher than all other strategies, regardless of their initialisation. We have not been able to achieve a set of hyperparameters that leads to consistent improvements for all GAN initialisation. Furthermore, even though working entirely within the CLIP embedding space streamlined the generation process, there is no possibility of human feedback on the quality of generated samples and, thus on the GAN training process.

5. Conclusion

Our study aimed to improve multimodal hate speech detection by augmenting datasets using GANs within the CLIP embedding space. We experimented with various GAN architectures, including AC-GAN and WGAN with gradient penalty, and a modified mode-seeking loss to encourage generation diversity. The small size and high-dimensional nature of the dataset presented significant challenges, often leading to inconsistent improvements of the classifier on the GAN-augmented dataset. We believe, to characterise the efficacy of the proposed approach a larger corpus of multimodal hate speech may be required.

References

[1] U. Nations, "What is hate speech?," 2024.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[3] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. D. Bimbo, "Mapping memes to words for multimodal hateful meme classification," 2023.

[4] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," 2021.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[6] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," 2017.

[7] G. K. Kumar and K. Nandakumar, "Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features," 2022.

[8] R. Cao and R. K.-W. Lee, "HateGAN: Adversarial generative-based data augmentation for hate speech detection," in *Proceedings of the 28th International Conference on Computational Linguistics* (D. Scott, N. Bel, and C. Zong, eds.), (Barcelona, Spain (Online)), pp. 6327–6338, International Committee on Computational Linguistics, Dec. 2020.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017.

[11] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," 2019.