CLIPping at Hate Speech

Mattia Nino Barbiere **Jacopo Boscariol** **Michele Lanfranconi**

Group 19

Abstract

This project addresses the growing issue of hate speech in multimodal content, focusing on memes that combine text and images. A classification pipeline is proposed, integrating a fine-tuned `CLIP` model and a transformer-based classifier, with Attention Rollout applied for interpretability. This provides insight into how the model identifies and prioritizes elements of the image and text during classification, advancing understanding of the model's reasoning process.

Keywords: Multimodal classification, CLIP, transformer, interpretability, Attention Rollout, hateful memes.

1. Introduction

Nowadays, hate speech on social media is increasingly prevalent, appearing in posts, comments, messages, and especially in offensive memes. Identifying hateful memes is particularly challenging due to the strong interaction between text and images, making the task harder than traditional image classification tasks [1]. Furthermore, interpreting the results of machine learning models is a prominent topic of research and its important is heightened in contexts involving ethical concerns, such as hate speech [2].
This challenge is addressed by developing a classification pipeline that combines the text and image multimodal `CLIP` (Contrastive Language-Image Pre-training, [3]) architecture with a transformer-based classifier to detect hateful memes and provide some intuitive explanations behind the classifications given by the model.

2. Related Work

The analysis builds on previous research in memes classification, focusing on the combination of textual and visual inputs to influence the output. This approach has been explored in several works, such as [4] and [5], which primarily use `CLIP` encoders to model and incorporate multimodality. Another relevant contribution is `MOMENTA` [6], a multimodal model designed to detect harmful memes and their targeted entities by combining image and text modalities. These papers highlight the potential of multimodal approaches to improve single-mode classification but offer limited insight into how text-image interactions influence the final output. For instance, [4] attempts interpretability by clustering feature vectors and manually labelling the resulting groups.
A more general approach is proposed in [7], which uses a CNN-based method to highlight influential regions of the input meme. However, this restricts the model to CNN architectures. In contrast, Attention Rollout [8], offers a similar interpretability mechanism that is compatible with transformer-based models, expanding applicability.
This project focuses on the concept of interpretability, combining `CLIP` encoders with transformer classifiers and leveraging Attention Rollout for result visualization.

3. Method

As previously introduced, the proposed model builds on a pre-trained `CLIP` transformer. `CLIP`, developed by OpenAI and released under the MIT License, is trained on 400 million image-text pairs collected from the internet. It learns to associate images and text by pulling matching pairs closer together while pushing non-matching pairs apart [3].
As in [4], the strong performance of the pre-trained `CLIP` model is leveraged by using it as a baseline, with fine-tuning applied to the final layers using the project dataset. Fine-tuning `CLIP` not only yields sufficiently good results with limited training but also significantly reduces training time by decreasing the number of trainable parameters from approximately 154 million to 4.5 million. Two pretrained `CLIP` models are primarily used: `CLIP-vit-base-patch16` and `CLIP-vit-base-patch32`, both publicly released with [3]. These models differ in patch size for the visual transformer: smaller patches (e.g., 16×16) capture finer image details, while larger patches (e.g., 32×32) yield lighter models suited for analyzing larger objects.
The pre-trained architecture produces token embeddings for both text and image inputs. These are concatenated, along with a classification token (CLS token), and passed to a

EE-559: *Deep Learning*, 2025.

transformer architecture. The transformer's objective is to use self-attention to compute a CLS token value in $[0, 1]$, representing the input classification. The transformer classifier and the final layers of `CLIP` are trained using BCE loss to minimize classification error.

The Attention Rollout implementation follows the description in [8], summarized as follows. Let A_l denote the attention matrix of layer l, averaged across attention heads. Then compute:

$$\tilde{\mathbf{A}}_\mathbf{l} = \frac{1}{2}(\mathbf{A}_\mathbf{l} + \mathbf{I}) \Rightarrow \mathbf{R} = \prod_{l=1}^{L} \tilde{\mathbf{A}}_\mathbf{l}\,. \tag{1}$$

As the task is classification, the focus is on the effect of each token on the CLS token. $\mathbf{R}$ can be viewed as a contribution matrix where $\mathbf{R}_{ij}$ indicates the contribution of token j to token i, thus the contribution of each token is represented in the first row of matrix $\mathbf{R}$.

Visualizations are generated by performing attention rollout over both `CLIP` and classifier attention layers, with residual connections accounted for at each layer. Attention matrices are sequentially composed as in Equation 1 to compute the cumulative contribution of input tokens or image patches to the final CLS embedding. For the image modality, the resulting attention map is upsampled and optionally smoothed with Gaussian blur before being overlaid on the original meme image. For the text modality, the attention scores are used to modulate background colour intensity behind each token, visualized as heatmaps and colour-coded text overlays.

An additional consideration in the choice of the pre-trained models arises from this approach: since Attention Rollout assigns a weight to each token, smaller patches provide higher resolution in the rollout weights compared to larger patches. This motivates the use of different models in the implementation.

4. Data

A publicly available multimodal dataset, introduced and released under the MIT license as detailed in [6], is adopted for this project. The dataset incorporates both textual and visual components for various memes, together with the corresponding labels. The annotation process associated with the dataset is thoroughly documented in [6].

The *Harm-P* subset, which specifically targets harmful memes related to U.S. politics, is selected for this study. The dataset statistics are reported in Table 1, which shows a certain degree of imbalance for the *Very Harmful* category. In accordance with the recommendations explained in [6], the original labels *Very Harmful* and *Partially Harmful* are merged into a single *Harmful* label. This conversion allows to perform a binary classification task on a balanced dataset.

A data augmentation pipeline has been implemented to enhance generalization, applying mild transformations like colour jitter, affine distortion, and JPEG compression. However, the models achieved better results on the testing dataset without using augmentation.

SPLIT	*Very Harmful*	*Partially Harmful*	*Harmless*
TRAIN	216	1270	1534
VALIDATION	17	69	91
TEST	25	148	182

Table 1. Harm-P DATASET SPLITS AND STATISTICS, FROM [6].

5. Validation and Results

After fine-tuning both `CLIP` models (see Section 3) with an attention-based classifier, the resulting models are referred to as `V2_16` and `V2_32`, respectively. Performance metrics for the trained models are presented in Table 2.

MODEL	ACCURACY (%)	F1 SCORE (%)
`V2_16`	60.73 ± 2.00	59.50 ± 2.76
`V2_32`	57.18 ± 1.70	57.04 ± 1.77
`MOMENTA` [6]	89.84	88.26

Table 2. TEST ACCURACY AND AVERAGE F1-SCORE OF MODELS `V2_16` AND `V2_32`, COMPARED TO MOMENTA [6].

The main focus of this project is interpretability, which leads to lower performance compared to `MOMENTA`. It should be noted, however, that `MOMENTA` is a more advanced model, employing `VGG-19` and `DistilBERT` for classification, in contrast to the single attention-based classifier used here. Figures 1, 2 show the Attention Rollout results of `V2_16` for two different memes across both visual and textual modalities. A more intense red colour indicates higher rollout weight values, highlighting the most influential regions to the CLS token.

6. Survey

While the evaluation of the classifier component is straightforward and comparable with existing results, assessing the interpretability of the Attention Rollout component is less trivial. To provide a numerical indication of interpretability performance, an anonymous survey was conducted on 10 memes to be rated on a scale from 1 to 5.

To facilitate participants' understanding, attention rollout results were visualized by overlaying a colour map on each image, with highlighted regions indicating areas of high model attention. For textual inputs, coloured boxes were applied to individual tokens. These visualizations, previously described, were used in the survey. Examples are shown in Figures 1 and 2.

This procedure was reviewed and approved by the *Ethical Research Committee* at EPFL, which authorized the use of the anonymous responses for the present project.

Figure 1. Attention Rollout result of `V2_16` for a meme. A more intense red colour indicates influential regions to the CLS token.

Figure 2. Attention Rollout result of `V2_16` for a meme. The rollout weight map is upsampled to match the image dimensions and overlaid on the image.

A total of 35 people participated in the survey. The average rating was 3.67 (66.86%) for model `V2_16`, and 2.95 (48.68%) for model `V2_32`. Figures 3 and 4 display the votes distributions for `V2_16` and `V2_32` models, respectively.

7. Limitations

One of the main limitations of this study lies in the task itself. As noted in the Section 1, determining whether a meme is offensive is highly subjective. This issue was also addressed in the survey by asking participants whether they agreed with the dataset-assigned labels for each of the 10 presented memes. On average, only 51.12% of participants

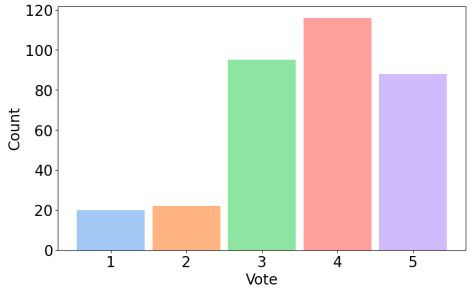

Figure 3. Anonymous survey vote distribution for model `V2_16`. Votes range from 1 to 5, where 5 is the best vote. The average vote was 3.67.

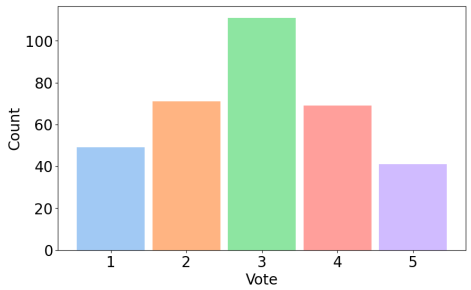

Figure 4. Anonymous survey vote distribution for model `V2_32`. Votes range from 1 to 5, where 5 is the best vote. The average vote was 2.95.

agreed with the provided labels.

A second major limitation, specific to the implementation, arises from the structure of the dataset. A more effective approach would involve extracting the text and its position from the image using *OCR* (Optical Character Recognition), then mapping text tokens to image tokens. This would enable a more unified and interpretable visualization that combines textual and visual highlights by incorporating information about the spatial location of the text within the image.

8. Conclusion

This project combines a fine-tuned `CLIP` encoder with a transformer-based classifier to detect hateful memes, with a primary focus on interpretability. By applying Attention Rollout, visual insights are provided into the elements that most influence the model's classification, making the decision process more transparent. Survey responses indicated moderate agreement with the model's explanations, while also highlighting the inherent subjectivity involved in interpreting hate speech. This work paves the way for more transparent and ethically aligned AI systems in the fight against online hate speech.

References

[1] 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.

[2] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, vol. 55, p. 1–42, July 2023.

[3] 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.

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

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

[6] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "MOMENTA: A multimodal framework for detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: EMNLP 2021* (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Punta Cana, Dominican Republic), pp. 4439–4455, Association for Computational Linguistics, Nov. 2021.

[7] A. Chhabra and D. K. Vishwakarma, "Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106991, 2023.

[8] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," 2020.