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

Gokul Karthik Kumar **Karthik Nandakumar**
Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE
{gokul.kumar, karthik.nandakumar}@mbzuai.ac.ae

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. Hence, detecting hateful memes requires careful consideration of both visual and textual information. 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. In this work, 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 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 other meme datasets such as Propaganda Memes and TamilMemes 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. The code for this work is available at https://github.com/gokulkarthik/hateclipper.

1 Introduction

Multimodal memes, which can be narrowly defined as images overlaid with text that spread from person to person, are a popular form of communication on social media (Kiela et al., 2020). While most Internet memes are harmless (and often humorous), some of them can represent hate speech.

Figure 1: Illustrative (not real) examples of multimodal hateful memes from Kiela et al. (2020). While the memes on the left column are hateful, the ones in the middle are non-hateful image confounders, and those on the right are non-hateful text confounders.

Given the scale of the Internet, it is impossible to manually detect such hateful memes and stop their spread. However, automated hateful meme detection is also challenging due to the multimodal nature of the problem.

Research on automated hateful meme detection has been recently spurred by the Hateful Memes Challenge competition (Kiela et al., 2020) held at NeurIPS 2020 with a focus on identifying multimodal hateful memes. The memes in this challenge were curated in such a way that only a combination of visual and textual information could succeed. This was achieved by creating non-hateful "confounder" memes by changing only the image or text in the hateful memes, as shown in Figure 1. In these examples, an image/text can be harmless or hateful depending on subtle contextual information contained in the other modality. Thus, multimodal (image and text) machine learning (ML) models are a prerequisite to achieve robust and accurate detection of such hateful memes.

In a multimodal system, the fusion of different modalities can occur at various levels. In early fusion schemes (Kiela et al., 2019; Lu et al., 2019; Li et al., 2019), the raw inputs (e.g., image and text) are combined and a joint representation of both modalities is learned. In contrast, late fusion approaches (Kiela et al., 2020), learn end-to-end

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 171 - 183

models for each modality and combine their outputs. However, both these approaches are not appropriate for hateful memes because the text in a meme does not play the role of an image caption. Early fusion schemes are designed for tasks such as captioning and visual question answering, where there is a strong underlying assumption that the associated text describes the contents of the image. Hateful memes violate this assumption because the text and image may imply different things. We believe that this phenomenon makes the early fusion schemes non-optimal for hateful meme classification. In the example shown in the first row of Figure 1, the left meme is hateful because of the interaction between the image feature "skunk" and the text feature "you" in the context of the text feature "smell". On the other hand, the middle meme is non-hateful as "skunk" got replaced by "rose" and the right meme is also non-hateful because "you" got replaced by "skunk". Thus, the image and text features are related via common attribute(s). Since modeling such relationships is easier in the feature space, an intermediate fusion of image and text features is more suitable for hateful meme classification.

The ability to model relationships in the feature space also depends on the nature of the extracted image and text features. Existing intermediate fusion methods such as ConcatBERT (Kiela et al., 2020) pretrain the image and text encoders independently in a unimodal fashion. This could result in the divergent image and text feature spaces, making it difficult to learn any relationship between them. Thus, there is a need to "align" the image and text features through multimodal pretraining. Moreover, hateful meme detection requires faithful characterization of interactions between fine-grained image and text attributes. Towards achieving this goal, we make the following contributions in this paper:

- We propose an architecture called Hate-CLIPper for multimodal hateful meme classification, which relies on an intermediate fusion of aligned image and text representations obtained using the multimodally pretrained Contrastive Language-Image Pretraining (CLIP) encoders (Radford et al., 2021).

- We utilize bilinear pooling (outer product) for the intermediate fusion of the image and text features in Hate-CLIPper. We refer to this representation as feature interaction matrix (FIM) which explicitly models the correlations between the dimensions of the image and text feature spaces. Due to the expressiveness of the FIM representation from the robust CLIP encoders, we show that a simple classifier with few training epochs is sufficient to achieve state-of-the-art performance for hateful meme classification on three benchmark datasets without any additional input features like object bounding boxes, face detection and text attributes.

- We demonstrate the interpretability of FIM by identifying salient locations in the FIM that trigger the classification decision and clustering the resulting trigger vectors. Results indicate that FIM indeed facilitates the learning of meaningful concepts.

2 Related Work

The Hateful Memes Challenge (HMC) competition (Kiela et al., 2020) established a benchmark dataset for hateful meme detection and evaluated the performance of humans as well as unimodal and multimodal ML models. The unimodal models in the HMC competition include: **Image-Grid**, based on ResNet-152 (He et al., 2016) features; **Image-Region**, based on Faster RCNN (Ren et al., 2017) features; and **Text-BERT**, based on the original BERT (Devlin et al., 2018) features. The multimodal models include: **Concat BERT**, which uses a multilayer perceptron classifier based on the concatenated ResNet-152 (image) and the original BERT (text) features; **MMBT** (Kiela et al., 2019) models, with Image-Grid and Image-Region features; **ViLBERT** (Lu et al., 2019); and **Visual BERT** (Li et al., 2019). A late fusion approach based on the mean of Image-Region and Text-BERT output scores was also considered. All the above models were benchmarked on the "test seen" split based on the area under the receiver operating characteristic curve (AUROC) (Bradley, 1997) metric. The results indicate a large performance gap between humans (AUROC of 82.65 [1]) and the best baseline using Visual BERT (AUROC of 75.44).

The challenge report (Kiela et al., 2021), which was released after the end of the competition, showed that all the top five submissions (Zhu, 2020; Muennighoff, 2020; Velioglu and Rose, 2020;

[1] https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/

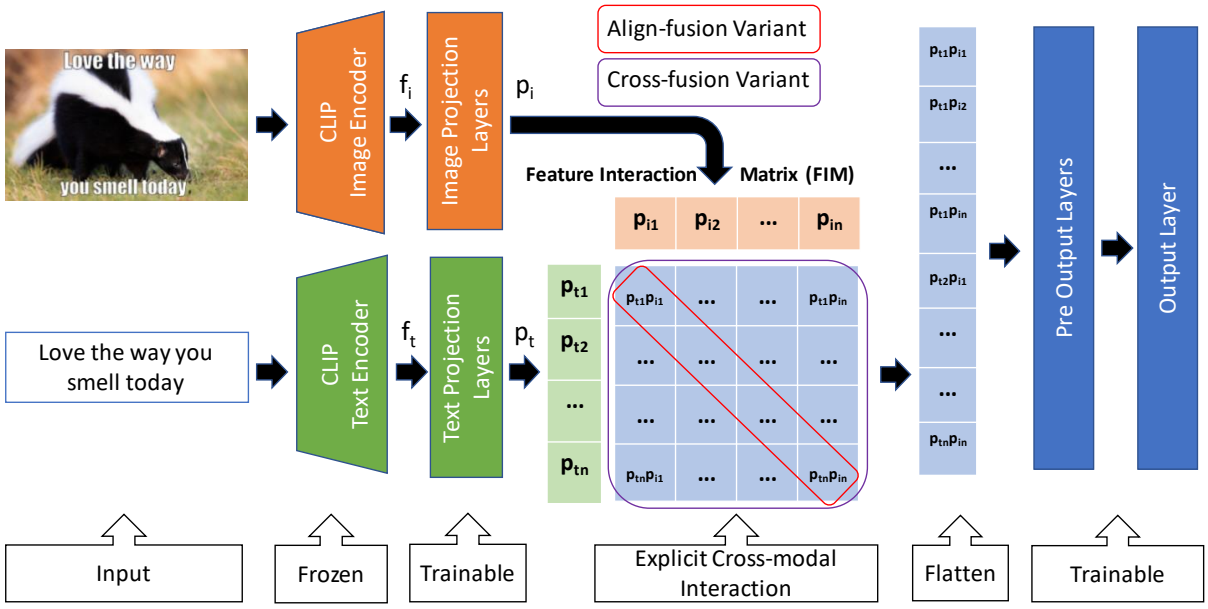

Figure 2: Proposed architecture of Hate-CLIPper for Multimodal Hateful Meme Classification.

Lippe et al., 2020; Sandulescu, 2020) achieve better AUROC than the baseline methods. This improvement was achieved primarily through the use of ensemble models and/or external data and additional input features. For example, Zhu (2020) used a diverse ensemble of VL-BERT (Su et al., 2019), UNITER-ITM (Chen et al., 2019), VILLA-ITM (Gan et al., 2020) and ERNIE-Vil (Yu et al., 2020) with additional information about entity, race, and gender extracted using Cloud APIs and other models. This method achieved the best AUROC of 84.50 on the "test unseen" split.

Mathias et al. (2021) extended the HMC dataset with fine-grained labels for protected category and attack type. Protected category labels include race, disability, religion, nationality, sex, and empty protected category. Attack types were labeled as contempt, mocking, inferiority, slur, exclusion, dehumanizing, inciting violence, and empty attack. Zia et al. (2021) used CLIP (Radford et al., 2021) encoders to obtain image and text features, which were simply concatenated and passed to a logistic regression classifier. Separate classification models were learned for the two multilabel classification tasks - protected categories and attack types. MOMENTA (Pramanick et al., 2021) also uses representations generated from CLIP encoders, but augments them with the additional feature representations of objects and faces using VGG-19 (Simonyan and Zisserman, 2014) and text attributes using DistilBERT (Sanh et al., 2019). Furthermore, MOMENTA uses cross-modality attention fusion (CMAF), which concatenates text and image features (weighted by their respective attention scores) and learns a cross-modal weight matrix to further modulate the concatenated features. MOMENTA reports performance only on the HarMeme dataset (Sandulescu, 2020).

Although bilinear pooling (Tenenbaum and Freeman, 2000) (outer product) of different feature spaces has shown improvements for different multimodal tasks (Fukui et al., 2016; Arevalo et al., 2017; Kiela et al., 2018), it is not well experimented with multimodally pretrained (aligned feature space) encoders like CLIP or for the Hateful Meme Classification task.

3 Methodology

Our objective is to develop a simple end-to-end model for hateful meme classification that avoids the need for sophisticated ensemble approaches and any external data or labels. We hypothesize that there is sufficiently rich information available in the CLIP visual and text representations and the missing link is the failure to model the interactions between these feature spaces adequately. Hence, we propose the Hate-CLIPper architecture as shown in Figure 2. In the proposed Hate-CLIPper architecture, the image i and text t are passed through pretrained CLIP image and text encoders (whose weights are frozen after pretraining) to obtain unimodal features f_i and f_t, respectively. We use

# 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.

pre-trained CLIP encoders from the original work (Radford et al., 2021), where the model is trained on Image-Text matching with 400 million <image, text> pairs collected from the Internet.

Trainable Projection Layers: Note that CLIP is pre-trained using contrastive learning on 400 million image–text pairs from the Internet. This multimodal pretraining encourages similarity between the feature spaces of the image and its corresponding text caption. However, in the dataset used for pretraining, the image and text pairs usually convey the same meaning, which is not always the case in hateful memes. Therefore, to better model the semantic relationship between the image and text feature spaces of memes, we further add trainable projection layers at the output of the CLIP image and text encoders. The main purpose of projection layers is not to ensure same dimensionality for both text and image embeddings, but to achieve better alignment between the text and image spaces. While CLIP is already trained to align the two spaces at a high-level, this needs to be further finetuned for the specific task/dataset at hand. Instead of finetuning the entire CLIP model using small datasets, it is more prudent to add projection layers and only learn these projection layers based on the given datasets. These projection layers map the unimodal image and text features f_i

and f_t to the corresponding image projection p_i and text projection p_t, respectively. The projection layers are designed such that both p_i and p_t have the same dimensionality n. The use of customized trainable projection layers after the CLIP encoders is one of the key differences between the proposed architecture and the one used in (Zia et al., 2021).

Modeling Full Cross-modal Interactions: The important component of the Hate-CLIPper architecture is the explicit modeling of interactions between the projected image and text feature spaces using a *feature interaction matrix* (FIM). The FIM representation $R \in \mathbb{R}^{n \times n}$ is obtained by computing the outer product of p_i and p_t, i.e., $R = p_i \otimes p_t$. The FIM can be flattened to get a vector r of length n^2 and passed through a learnable neural network classifier to obtain the final classification decision. This approach is different from the traditional concatenation (Concat) technique employed in the literature (Zia et al., 2021; Pramanick et al., 2021), which simply concatenates the two representations to obtain a vector of length $2n$. Since the FIM representation directly models the correlations between the dimensions of the image and feature spaces, it is better than the Concat approach, where the task of learning these relationships from limited data samples falls on the subsequent classification module. We refer to the fusion of text and image features using the FIM as *cross-fusion*.

Modeling Reduced Cross-modal Interactions: One of the limitations of the cross-fusion approach is the high dimensionality of the resulting representation, which in turn requires a classifier with a larger number of parameters. The diagonal elements of the FIM R represent the element-wise product between p_i and p_t and has a dimension of only n. Note that the sum of these diagonal elements is nothing but the dot product between p_i and p_t, which intuitively measures the alignment (angle) between the two vectors. Therefore, a vector representing the diagonal elements of R indicates the alignment between the individual dimensions of p_i and p_t, which can still be useful for classification as the encoders that we use are pretrained with the alignment task. Hence, we refer to the fusion of text and image features using only the diagonal elements of FIM as *align-fusion*.

Classification Module: The output of the intermediate fusion module is a vector r of dimension d (where $d = 2n$ for the baseline Concat technique, $d = n^2$ for the cross-fusion approach, and $d = n$ for the align-fusion method). We apply a shallow neural network on this feature vector r to obtain the final output o. The shallow neural network consists of a few fully-connected layers (referred to as pre-output layers) and a softmax output layer to produce the final output value o. The first layer of this classifier network maps an input r with d dimensions to a common pre-output dimension m and the rest of the pre-output layers have the same number of hidden nodes m. Each fully-connected layer is followed by ReLU activation and trained with dropout. For binary classification (hateful vs. non-hateful memes), we optimize the trainable (projection and pre-output) layers by minimizing the binary cross-entropy loss between the output o and the true label l. For fine-grained classification (protected category, attack type), we simply add auxiliary output layers and train the model using the total loss for all the classification tasks.

4 Experimental Results

4.1 Datasets

The primary dataset used in our evaluation is the HMC dataset (Kiela et al., 2020), which contains 8500 memes in the training set, 500 memes in the development seen split, 540 memes in the development unseen split, 1000 memes in the test seen split, and 2000 memes in the test unseen split. We also evaluate the proposed approach on the Propaganda Memes dataset Sharma et al. (2022), which is a multi-label multimodal dataset with 22 propaganda classes. Finally, to evaluate the multilingual generalizability, we test the performance on TamilMemes (Suryawanshi et al., 2020), which is a dataset for troll/non-troll classification of memes in the Tamil language. Similar to the HMC dataset, the TamilMemes dataset has meme images and corresponding meme texts that are transliterated from Tamil to English. However, unlike the HMC dataset, the TamilMemes dataset is not compiled with the motive of making only multimodal information useful for target classification.

4.2 Setup

We train Hate-CLIPper and other baselines (concat fusion and attention-based CMAF) based on the train split and evaluate them on the dev-seen and test-seen splits of the HMC dataset using AUROC as the evaluation metric. For the Propaganda Memes and the Tamil Memes dataset, the micro F1 score is used as the evaluation metric to en-

sure a fair comparison with results reported in the literature. We use TorchMetrics library[2] to compute all the evaluation metrics. For multi-label classification in Propaganda Memes dataset, we set 'mdmc_average' to 'global' in computing the micro-F1 score, which does the global average for multi-dimensional multi-class inputs. We use Pytorch on NVIDIA Tesla A100 GPU with 40 GB dedicated memory and CUDA-11.1 installed. The hyper-parameter values for all models are shown in Table 3, which are chosen based on the manual tuning with respect to the target evaluation metric of the validation set. We use ViT-Large-Patch14 based CLIP model consistently for all the experiments in Tables 1 & 2. The models experimented in Table 1 took around 30 minutes (median is 30 minutes and longest is 32 minutes) for the combined training and evaluation. To do a fair evaluation, we use the same evaluation metric as in the previous works for the corresponding datasets.

Hyperparameter	Value
Image size	224
Pretrained CLIP model	ViT-Large-Patch14
Projection dimension (n)	1024
Pre-output dimension (m)	1024
Optimizer	AdamW
Maximum epochs	20
Batch size	64
Learning rate	0.0001
Weight decay	0.0001
Gradient clip value	0.1

Table 3: Hyperparameter configuration for HateCLIPer and other baselines.

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)

4.3 Key Findings

When we interpret Tables 1 & 2 in conjunction with Tables 4 & 5 respectively, we can clearly see that the performance of intermediate fusion with the CLIP encoders is better that that of several early fusion approaches such as MMBT, ViLBERT, and VisualBERT as well as late fusion methods. For instance, on the HMC dataset, the best early fusion approach (Visual BERT) had an AUROC of 75.44 and late fusion method had an AUROC of 69.3 on the test set. These AUROC values are significantly lower than AUROC of the proposed align (intermediate) fusion scheme, which is 85.8. In fact, all the intermediate fusion methods considered in Table 1 clearly outperform the early and late fusion methods reported in Table 4. These results strongly support the claim that intermediate fusion is more suitable for hate classification.

From Tables 1 & 4, it is clear that cross-fusion and align-fusion variants of Hate-CLIPper achieve the best AUROC for both the evaluation sets of HMC dataset, which is also better than the reported human performance. This trend is also consistent when we replaced ViT-Large-Patch14 with ViT-Base-Patch32. Despite having only n multimodal features, align-fusion performs significantly better than concat-fusion with $2n$ multimodal fea-

[2] https://torchmetrics.readthedocs.io

tures and is closer to cross-fusion with n^2 multimodal features. This signifies the importance of pre-aligned image and text representations of CLIP. Hence, for low computational resource conditions, it would be appropriate to replace cross-fusion with align-fusion in the Hate-CLIPper framework.

Our results also show that a single projection layer for each modality and a shallow neural network (1 or 3 layers) for the classifier is sufficient to achieve good performance. This shows that the discriminative power of Hate-CLIPper is mainly a consequence of modeling the interactions between text and image features from CLIP encoders using cross and align fusion. The results on the Propaganda Memes dataset also confirm the same findings. Although the differences between the various configurations shown in Table 2 are marginal, the performance of the proposed approach is a significant leap compared to those reported in the literature (see Table 5).

As noted in Section 2, methods proposed in (Zia et al., 2021) and (Pramanick et al., 2021) are the closest to the proposed approach since both of them use CLIP encoders. Results in Table 1 show that under the same experimental setup, the proposed approach is better than the cross-modal attention fusion (CMAF) scheme used in MOMENTA (Pramanick et al., 2021), when no additional information is utilized. If additional information is available, our proposed approach can also leverage them in the same way (using intra-modal fusion) as MOMENTA. The work in (Zia et al., 2021) claims that train and development seen splits were used for training and development unseen split was used for evaluation. However, a careful analysis of the published code for (Zia et al., 2021) indicates that 400 out of 540 memes in development unseen split (74%) are also included in the development seen split. Since a fair comparison is not possible under these circumstances, we ignore all the results of Zia et al. (2021).

Our ablation experiments (i) with unfrozen CLIP encoders (AUROC of <63), and (ii) Non-CLIP encoders (mBERT (Devlin et al., 2018), VIT (Dosovitskiy et al., 2021)) (AUROC of <59) resulted significantly poor scores in the HMC dataset.

4.4 Multilinguality

The baseline evaluations on TamilMemes dataset (Suryawanshi et al., 2020) used only image based classifiers such as ResNet (He et al., 2016) and MobileNet (Howard et al., 2017) and their test set (300 memes) is different from the released test set (667 memes). Hence, they are not directly comparable to the proposed approach. (Hegde et al., 2021) proposed a multimodal approach for TamilMemes classification. They used pretrained ViT (Dosovitskiy et al., 2021) and BERT (Devlin et al., 2018) as encoders to get the image and text features, respectively. These features were concatenated and used for classification. This model achieved a micro F1 score of 47 on the test set. It is critical to note that the Hate-CLIPper also uses the same encoders of ViT and BERT but they are multimodally pretrained with the CLIP loss. Thus, the Hate-CLIPper achieves state-of-the-art performance with a micro F1 score of 59 on the test set.

For the TamilMemes dataset, both cross and align fusion had the same performance as concat fusion. This could be due to the fact that the pre-aligned text space of CLIP has never encountered Tamil-to-English transliterated text data. Consequently, the resulting text and image feature spaces are not well-aligned, making it difficult to model the relationship between the two feature spaces. Replacing the CLIP text encoder with multilingual BERT (Devlin et al., 2018) also did not lead to any further performance improvement, which can again be attributed to the feature space misalignment caused by the lack of multimodal pretraining using the image and corresponding Tamil text pairs.

5 Interpretability

To determine the interpretability of the feature interaction matrix (FIM), we employ the following simple approach. First, we compute a n^2-dimensional binary trigger vector for each hateful meme, where a value of 1 indicates that the specific element in the FIM R is salient for determining if the given input belongs to the hateful class. These trigger vectors are then clustered into groups using a K-means clustering algorithm. We manually examine these clusters to determine if most samples within a cluster have a common underlying pattern.

To compute the trigger vector, we first reset the feature interaction matrix R to zero values and evaluate the gradient of the loss function for the non-hateful class with respect to R. Let $D \in \mathbb{R}^{n \times n}$ denote the model-specific gradient matrix. Each element in the D matrix represents the direction (positive/negative) of the corresponding element in R matrix towards the hateful class. We then bi-

Figure 3: Hateful memes clustered by K-means clustering algorithm (number of clusters = 15) based on the trigger vector of Hate-CLIPper with cross-fusion. Featuring hateful examples with all the original text in this place would be distasteful; hence the meme text is masked with the exception of few words that are required for discussion. However, the reader can choose to look at the non-censored memes in the appendix

narize the D matrix by setting all elements in the top-20 and bottom 20 percentiles (based on magnitude) to value 1 and assigning 0 values to all the other elements. Then, for each hateful meme (i, t) in the training set, we perform one forward pass through the Hate-CLIPper and compute the meme-specific FIM R. Again we binarize the R matrix by setting all elements in the top-10 and bottom-10 percentiles (based on magnitude) to value 1 and assigning 0 values to all the other elements. Finally, the trigger matrix T is computed for each meme as the element-wise (Hadamard) product of binarized D and R matrices, i.e., $T = D \odot R$. This trigger matrix is then flattened to obtain the trigger vector corresponding to a meme. We apply K-means clustering algorithm available in Scikit-Learn (Pedregosa et al., 2011) on the trigger vectors to group the hateful memes. Samples from the resulting groups of memes are shown in Figure 4.

From Figure 4, we observe that clusters 5, 7, and 11 contain memes related to the same concept. It is interesting to note that Hate-CLIPper is able to produce similar features for the same concept expressed in different modalities. For example in cluster 5, which is characterized by the concept 'death', we can see some memes representing 'death' only in images (b) and other memes representing the same concept only in text (a, d, e). Furthermore, note that the memes (f) and (g), under the same cluster, do not directly relate to death, but the meme texts could hint toward death related events (blow -> blast; popcorn sounds -> bullet sounds) in different contexts. With the clustered memes, we can also identify the positions in FIM R, which get activated for the matching concepts. However, some of the clusters are ambiguous. For example, cluster 13 has memes from different concepts. Also, when the clusters have less than 3 memes or greater than 10 memes, they exhibit greater diversity in terms of the underlying concepts and are not useful for the explanation.

6 Conclusion

In this work, 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 (1 image projection, 1 text projection, 1 pre-output, and 1 output) in the Hateful Memes Challenge dataset, Moreover, our model does not require any additional input features like object bounding boxes, face detection, text attributes, etc. We also demonstrated similar performance in multi-label classification based on the Propaganda Memes dataset. Finally, we performed preliminary studies to evaluate the interpretability of cross-modal interactions.

7 Limitations

From an ethical perspective, the concept of hate speech itself is quite subjective and it is often difficult to draw a clear line between what is hateful and non-hateful. On the technical front, the accuracy of hateful meme classifiers is still far from satisfactory even on carefully curated benchmark datasets, which impedes real-world deployment. Apart from these general limitations, the proposed Hate-CLIPper framework for hateful meme classification also has several specific limitations. Firstly, handling the high dimensionality of the feature interaction matrix is a computational challenge. For $n = 1024$ and $m = 1024$, this requires a model with a billion parameters ($O(n^2m)$). Fortunately, the align-fusion approach performs quite close to the cross-fusion method and requires only $O(nm)$ parameters. The CLIP encoders used in Hate-CLIPper are well-trained on a massive dataset in English. Such models are rarely available for low-resource languages, limiting their direct applicability for such languages. While the multilingual experiment highlights the issues arising from misaligned text and image feature spaces, more thorough ablation studies are required to understand the ability of learnable projection layer(s) to overcome this misalignment. Furthermore, the proposed approach to judge interpretability is simple and ad-hoc and a more systematic evaluation of explainability is needed. The fine-grained labels Mathias et al. (2021) have not been utilized for FIM interpretation.

References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Andrew P. Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Universal image-text representation learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Uvce-iiitt@dravidianlangtech-eacl2021: Tamil troll meme classification: You need to pay more attention. *ArXiv*, abs/2104.09081.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. 2021. The hateful memes challenge: Competition report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient large-scale multimodal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? *ArXiv*, abs/2204.05454.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online. Association for Computational Linguistics.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visiolinguistic models applied to hateful memes.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shivam Sharma, Firoj Alam, Md Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, et al. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Joshua B Tenenbaum and William T Freeman. 2000. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *ArXiv*, abs/2202.03052.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: An efficient and accurate scene text detector. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online. Association for Computational Linguistics.

A Appendix

A.1 Variations of HateCLIPper

We experimented with several training/architectural modifications to the core Hate-CLIPper framework proposed in the main paper:

1. **Pretraining with captions:** We generated captions that describe each image in the Hateful Memes Challenge dataset with the state-of-the-art transformer based image captioning model, OFA (Wang et al., 2022). Then, we pretrained the image and text encoders of Hate-CLIPper using contrastive loss between the meme images and the generated captions like Radford et al. (2021). Then, finetuning on the target dataset using meme images and meme texts is done as usual.

2. **Finetuning with captions:** We incorporated the generated captions during finetuning of Hate-CLIPper in different ways: (1) replacing image with generated captions and image encoder with the same text encoder, (2) concatenating generated caption with the meme text (3) concatenating features from "meme image + meme text" flow and "meme image+ generated caption" flow.

3. **Unimodal losses:** Linear output layers, for hateful meme classification, are added on top of the image and text projection layers of Hate-CLIPper and the corresponding unimodal losses are jointly optimized with the original multimodal loss as recommended by Ma et al. (2022).

4. **Fine-grained losses:** Linear output layers, for fine-grained hateful meme classification, are added in parallel to the output layer of Hate-CLIPper, and the corresponding fine-grained losses are jointly optimized with the original loss. This is done using fine-grained labels provided by Mathias et al. (2021).

5. **Data augmentation:** We identify the text bounding box regions in the meme images using EAST (Zhou et al., 2017) and replace them with either average pixel value masks or inpainting using Navier-Stokes[3] based method and finetune the Hate-CLIPper as usual.

Although, the above mentioned variations are backed by some reasoning, they either produced the same results or slightly degraded the performance.

[3] https://docs.opencv.org/4.x/d7/d8b/group__photo__inpaint.html

Figure 4: Hateful memes clustered by K-means clustering algorithm (number of clusters = 15) based on the trigger vector of Hate-CLIPper with cross-fusion. This non-censored version is just for more understading and the reader can choose to skip this figure as it features distasteful content.