electronics

MDPI

Article

Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism

Fan Wu [1], Guolian Chen [2,*], Junkuo Cao [3,4,*], Yuhan Yan [1] and Zhongneng Li [1]

[1] School of Information Science and Technology, Hainan Normal University, Haikou 571158, China; 202212083900013@hainnu.edu.cn (F.W.); 202212083900014@hainnu.edu.cn (Y.Y.); 202212083900005@hainnu.edu.cn (Z.L.)
[2] State-Owned Assets Management Office, Hainan Normal University, Haikou 571158, China
[3] Information Network and Data Center, Hainan Normal University, Haikou 571158, China
[4] Key Laboratory of Data Science and Smart Education, Ministry of Education, Haikou 571158, China
* Correspondence: cgl@hainnu.edu.cn (G.C.); cjk@hainnu.edu.cn (J.C.)

Abstract: Hateful memes are malicious and biased sentiment information widely spread on the internet. Detecting hateful memes differs from traditional multimodal tasks because, in conventional tasks, visual and textual information align semantically. However, the challenge in detecting hateful memes lies in their unique multimodal nature, where images and text in memes may be weak or unrelated, requiring models to understand the content and perform multimodal reasoning. To address this issue, we introduce a multimodal fine-grained hateful memes detection model named "TCAM". The model leverages advanced encoding techniques from TweetEval and CLIP and introduces enhanced Cross-Attention and Cross-Mask Mechanisms (CAM) in the feature fusion stage to improve multimodal correlations. It effectively embeds fine-grained features of data and image descriptions into the model through transfer learning. This paper uses the Area Under the Receiver Operating Characteristic Curve (AUROC) as the primary metric to evaluate the model's discriminatory ability. This approach achieved an AUROC score of 0.8362 and an accuracy score of 0.764 on the Facebook Hateful Memes Challenge (FHMC) dataset, confirming its high discriminatory capability. The TCAM model demonstrates relatively superior performance compared to ensemble machine learning methods.

Keywords: sentiment; multimodal; fine-grained; Cross-Masked; Cross-Attention

check for updates

Citation: Wu, F.; Chen, G.; Cao, J.; Yan, Y.; Li, Z. Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism. *Electronics* **2024**, *13*, 2780. https://doi.org/10.3390/electronics13142780

Academic Editors: Patrick Siarry and Junaid Rashid

Received: 19 June 2024
Revised: 11 July 2024
Accepted: 13 July 2024
Published: 15 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Multimodal memes are a specific form of communication on social media, narrowly defined as images with text exchanged between individuals [1]. While many internet memes typically convey humorous and harmless sentiment, a considerable portion of seemingly humorous content may contain hateful speech, escalating harmful information dissemination. The term "hateful memes" encompasses various types of discriminatory content, such as inciting violence or excluding a group through aggressive language, as well as derogatory language based on race, gender, religion, nationality, and disability. Content moderation on social media has become a significant societal challenge, as online platforms have been effectively weaponized for various geopolitical events and social issues [2–4]. A recent study revealed concerning patterns of online abuse, estimating that 1.1 million harmful tweets were directed at women within a year [5]. Studies also indicate that Black women are more likely to experience abuse on these platforms compared to White women. Together, these studies demonstrate that gender discrimination and hatred remain prevalent globally [6] despite initiatives such as the United Nations Sustainable Development Goals emphasizing the importance of gender equality, peace, and justice [7].

Given the vast scale of the internet, determining the harmfulness or appropriateness of content is not straightforward for humans, as social media interactions blend virtual

and real-world contexts. The existence or absence of content management policies can profoundly impact individuals, groups, and entire societies; therefore, manually detecting and preventing the spread of hateful memes is exceptionally challenging. However, due to the multimodal nature of the issue, the task of automatically detecting hateful memes also faces significant challenges. The FHMC [8] competition held at NeurIPS 2020 [1] has spurred research into hateful memes detection, focusing on identifying multimodal hateful memes. In this challenge, memes are carefully curated, requiring the successful integration of visual and textual information to accomplish the task.

For multimodal hateful memes containing both images and text, as depicted in Figure 1, the left image displays the hateful meme text "I like my women like I like my chickens: held in captivity until I behead them", where only "chickens" weakly correlates with the image features. The right image has almost no direct correlation between the text and image features. The high complexity of image content and the weak correlation between images and text may lead to divergence in feature space, making it challenging to learn their relationship. Adopting appropriate feature fusion methods is crucial in cases where image and text features are weakly correlated. Most existing feature fusion methods typically assume high similarity between images and text pairs to achieve high-quality semantic Alignment. However, for hateful memes, it is necessary to independently pre-train image and text encoders in a single modality to obtain specific information and semantic Alignment across different modalities.



Figure 1. A graphical representation of multimodal hateful memes in the FHMC competition. The (**left image**) signifies expressions of hatred, while the (**right image**) represents non-hateful expressions (Caution: This content may involve sensitive topics; please read objectively and rationally. This study aims to delve into the essence and impact of hateful memes, which are not intended for misleading or misinterpretation and do not constitute any specific recommendations or mandatory opinions).

Additionally, hateful memes detection requires representing complex interactions between image and text attributes. Thus, effective integration of features at different granularities is essential. This study contributes in several key aspects to achieve this goal:

- TCAM was introduced as an end-to-end multimodal hateful memes classification architecture, as shown in Figure 2. This architecture relies on two primary encoder components, CLIP [9] and TweetEval [10], extracting features using a specially fine-tuned Text Projection Layer Module (TPD) and an Image Projection Layer Module (CPD) to align image and text representations;
- Employing the BLIP model to generate image descriptions for textual data augmentation. By enhancing Cross-Attention and Cross-Mask Mechanisms and integrating textual and image features, along with textual features and image description features, the method facilitates the thorough interaction of multimodal information. This approach explicitly models the correlation between image and text feature spaces;
- Using transfer learning methods, introducing multiple fine-grained labels into the model training process to enrich and enhance the model's expressive capability with additional annotated data. As shown in Figure 3, researchers can achieve advanced

hateful memes classification performance on the FHMC benchmark dataset within a few training cycles due to the powerful representation capability of the encoders in TCAM.

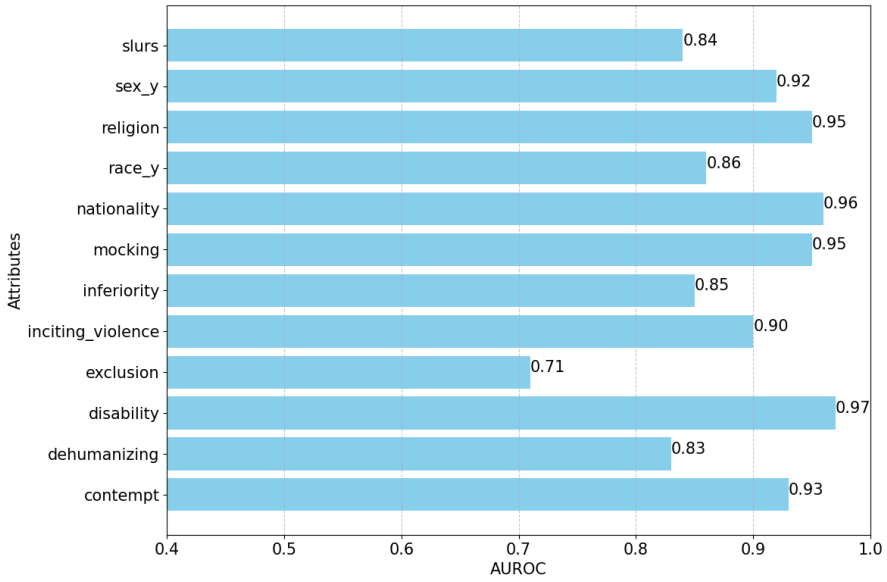

Figure 2. The AUROC of fine-grained features in the "hate" subset of the FHMC dataset.



Figure 3. Word clouds of the original text and image descriptions from the FHMC dataset ((**left image**) for the original text; (**right image**) for image descriptions).

This work is organized as follows: Section 2 extensively surveys hateful memes classification in deep learning. Section 3 introduces the TCAM architecture for memes classification. Section 4 provides a comprehensive comparative analysis of the results. Section 5 summarizes the paper.

2. Related Works

Early studies on malicious content classification primarily utilized machine learning (ML) techniques, focusing on character-level features. These methods typically relied on features such as word level, metadata, and part-of-speech tags, requiring greater contextual awareness. Due to the fast-paced, conversational nature of social media, coupled with platform-specific constraints (such as Twitter's character limit), modern natural language processing faces additional challenges, including part-of-speech tagging [11], lexical normalization [12,13], and named entity recognition [14,15]. Additionally, seemingly simple tasks such as sentiment analysis have proven challenging on Twitter data [16], partly due to the limited context clues in short texts [17].

Most social media hateful speech datasets include images, yet many studies have not fully utilized visual information. For multimodal research, it is crucial to leverage information from the images themselves and integrate textual transformations of image information to establish cross-modal interactions. Image description is an area under active

investigation. Extracting specific contextual information solely through visual observation is challenging. You et al. [18] proposed a solution using a deep learning encoder–decoder architecture to address these issues. Anderson et al. [19] introduced a similar approach for image description issues, utilizing a hierarchical attention mechanism from top-down and bottom-up, capable of computing important image parts. They employed Faster R-CNN for image parts, where each part comes with its feature vector, aiding in determining appropriate weights for required features according to architecture needs.

Hosseinmardi et al. [20] attempted to explore the correlation between visual content and textual features. However, due to pre-trained visual features, they achieved only a 0.01 improvement in classification. MMHS150K [21] is one of the earliest datasets composed of manually annotated text and images, introducing three multimodal methods, i.e., a Feature Concatenation Model, a Spatial Concatenation Model, and a Text Concatenation Model, yet their performance compared to single-modal models is inferior. Sharma et al. [22] introduced the Memotion Analysis dataset for multimodal classification tasks (such as emotion, humor, and semantic categories). However, models trained on these multimodal datasets often suffer from strong data priors, resulting in impressive model performance but lacking a true understanding of content.

The FHMC competition established a benchmark dataset for detecting hateful memes and evaluated the performance of both single-modal and multimodal machine learning models to foster research. Single-modal models included the Image-Grid model based on ResNet-152 features [23] and the Text-BERT model based on original BERT features [24]. Multimodal models encompassed Concat BERT, utilizing concatenated features from ResNet-152 (image) and original BERT (text) for multi-layer perceptron classification, the MMBT [25] model, integrating Image-Grid and Image-Region features, ViLBERT [26], and Visual BERT [27]. Late fusion methods were also considered based on average scores from Image-Region and TextBERT outputs. The challenge report [28] shows that all top five participants [29–33] achieved higher AUROC scores than baseline methods. These participants mainly achieved these higher scores by utilizing ensemble models, external data, and additional input features. For instance, Zhu [29] employed diverse ensemble models, including VL-BERT [34], UNITER-ITM [35], VILLA-ITM [36], and ERNIE-Vil [37], and utilized Cloud APIs and other models to extract additional information about entities, race, and gender.

3. Materials and Methods

Ensemble learning typically involves training and combining multiple base models, requiring additional computational resources and storage space, thus increasing computational costs and complexity. The machine learning model proposed in this paper integrates information from meme images, relevant text, and various fine-grained features to accurately predict hateful memes. We aim to develop a simple end-to-end model for hateful memes classification, thus avoiding using complex ensemble methods.

3.1. Transfer Learning

Detecting hateful memes has proven challenging, necessitating research beyond the simple binary classification of textual and image content. Among the top five participants in the FHMC competition, their approaches primarily incorporated additional features from external data and relatively straightforward additional image features. These external data were primarily gathered through manual searches for additional sources on the web, followed by annotation. Upon reviewing these additional data, it became apparent that they relied solely on OCR for text extraction without subsequent filtering, resulting in various forms of interference. Furthermore, manual annotation of the data did not achieve the accuracy level of the original dataset. Mathias et al. [38] extended protected and attack categories by adding fine-grained labels to the "hate" subset of the FHMC dataset. These fine-grained labels demonstrated excellent performance in auxiliary tasks, prompting us to incorporate them into the TCAM framework.

Table 1 illustrates the distribution of fine-grained features in the "hate" subset of the FHMC dataset. To ensure these fine-grained labels positively impact the model's learning, we employed the same encoding–decoding method to train each label individually and removed categories with AUROC scores below the human-recognizable threshold of 0.82, as shown in Figure 2. Combining Table 1 and Figure 2, it is evident that the "exclusion" category has an AUROC score of 0.71 for our model architecture, significantly lower than other categories. Also, the "exclusion" category has too few samples in the training and validation sets. Therefore, this experiment will no longer consider the "exclusion" category. Given the richness of fine-grained feature categories and the scarcity of samples in some categories, we adopted a weighted loss transfer learning method to handle these labels, with the specific algorithm as follows:

$$f_{BCE}(y_i, \hat{y}_i) = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \quad (1)$$

$$total_loss = f_{BCE}(y_{pred}, y_{true}) + \sum_{i=1}^{n} w_i \cdot f_{BCE}\left(y_{fine_pred_i}, y_{fine_true_i}\right) \quad (2)$$

Here, Equation (1) represents the binary cross-entropy loss function. In Equation (2), $total_loss$ denotes the overall loss value after combining the weighted losses of original and fine-grained labels. n signifies the number of fine-grained labels, y_{true} represents the true original labels, y_{pred} signifies the predicted original labels, $y_{fine_true_i}$ denotes the true fine-grained labels, $y_{fine_pred_i}$ indicates the predicted fine-grained labels, and w_i represents the loss weights for different class labels. This method effectively controls the importance of fine-grained labels in model learning, mitigating the impact of a few insignificant features on the most significant features.

Table 1. The distribution of fine-grained features in the "hate" subset of the FHMC dataset.

Fine-Grained Attributes	Train	dev Seen
dehumanizing	1318	121
inferiority	658	49
Inciting_violence	407	26
mocking	378	35
contempt	235	10
slurs	205	6
exclusion	114	13
religion	1078	95
race	1008	78
sex	746	56
nationality	325	26
disability	255	22

3.2. Image Caption

Inspired by Vlad Sandulescu et al. [33], this paper adopts the more specialized BLIP [39] model to generate image descriptions as supplementary text. As shown in Figure 3, to more intuitively observe the relationship between the original text and image descriptions, we generated word clouds for both. Observing the word clouds in Figure 3, it is evident that the original text in the left image targets characters as hateful subjects, while the image descriptions in the right image focus on men and women as the central features. This indirectly links the image descriptions to the original text and images, with many similar indirect connecting words. Therefore, as shown in Figure 4, this paper employs an improved Cross-Mask mechanism to integrate the original text and image descriptions, enhancing the potential correlation between text and image features. The specific algorithm is as follows:

$$F_{flat} = vec(CF^{\top} - diag(CF^{\top})) \quad (3)$$

Here, F_{flat} represents the vector obtained by flattening the diagonal elements of the tensor resulting from the multiplication of C and F, where C denotes the image caption features, F represents the text features, $\top$ denotes the transpose, *diag* signifies setting the diagonal elements of a tensor to zero, and *vec* denotes the flattening operation.

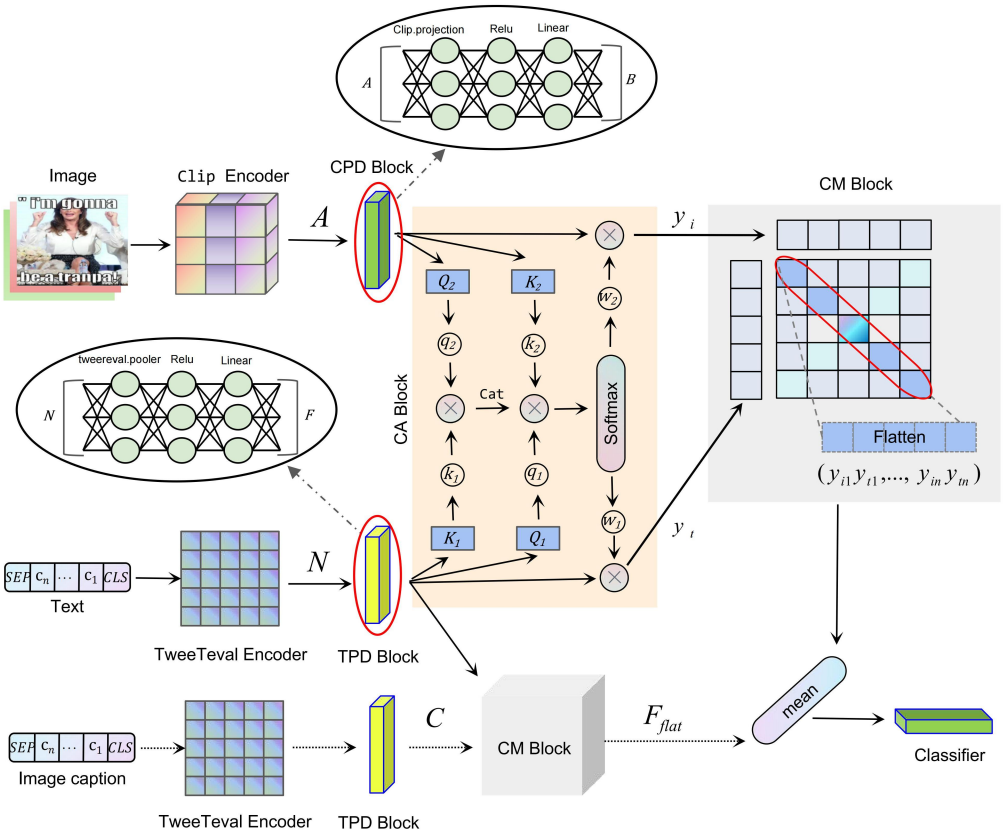

Figure 4. Word clouds of the original text and image descriptions from the FHMC dataset ((**left image**) for the original text; (**right image**) for image descriptions).

3.3. Encoding Layers and Alignment of Feature Spaces

The primary focus of the CLIP model lies in achieving a strong alignment between textual descriptions and image features for image generation from text. The visual representation of CLIP is pre-trained on a vast internet dataset containing 400 million images, providing a rich source of visual information. However, the CLIP text representation lacks emphasis on emotional factors such as hate and aggression. In contrast, TweetEval is pre-trained on a large-scale corpus from Twitter, providing text representations with higher emotional understanding. We can extract specific information from both modalities to a large extent by leveraging CLIP and TweetEval. Therefore, this paper adopts the CLIP-TweetEval model's image–text encoder combination for initial feature extraction in FHMC while simultaneously freezing all parameters of these encoders.

While the encoders are trained at a higher level to align these spaces, specific tasks, such as hateful memes, require further fine-tuning. Inspired by the projection layers in the CLIP model, we introduced enhanced projection layers at the outputs of both modal encoders. Specifically, the Text Projection Layer adopts the pooling layer pre-trained in the TweetEval model (TPD), and the image projection layer utilizes the pre-trained Image Projection Layer in the CLIP model (CPD). Unlike the approach of fine-tuning the entire encoder model with a smaller dataset, this paper adopts a more cautious strategy by adding pre-trained feature projection layers to align feature spaces and fine-tuning only these layers using the given dataset. The above approach, compared to fine-tuning the entire encoding layer, not only ensures robust feature extraction, prevents overfitting, and effectively

leverages pretrained knowledge but also significantly enhances the computational efficiency of the FHMC classification pipeline.

To better model the weak correlation between the feature spaces of these two modalities and enhance their semantic relationship, this paper introduces an improved Cross-Attention mechanism (CA) after the projection layers of both modalities. It applies simplified similarity calculations to weigh the attention between textual and image features, omitting the value v in the attention mechanism. The advantages of this approach include reducing the original attention mechanism's computational complexity and effectively promoting information interaction between the two modalities.

According to Figure 4, the text is encoded into N by the TweetEval Encoder module and then further processed through the TPD Block to obtain F; the image is encoded into A by the CLIP Encoder module and then further processed through the CPD Block to obtain B. Finally, the CA Block interacts with the multimodal information, resulting in the multimodal features y_i and y_t after the weighted attention mechanism. The specific algorithm is as follows:

$$q_1 = F \otimes Q_1 \tag{4}$$

$$k_1 = F \otimes K_1 \tag{5}$$

$$q_2 = B \otimes Q_2 \tag{6}$$

$$k_2 = B \otimes K_2 \tag{7}$$

$$w_1, w_2 = \sigma((k_1 \otimes q_2) \| (q_1 \otimes k_2)) \tag{8}$$

$$y_t = A \times w_1 \tag{9}$$

$$y_i = N \times w_2 \tag{10}$$

In Equations (4), (5), (6), and (7), q_1 and k_1 respectively represent query and key values for textual attention scores, while q_2 and k_2 represent query and key values for image attention scores. $(Q_1, K_1) \in R^{1024 \times hidden}$, $(Q_2, K_2) \in R^{768 \times hidden}$ are linear transformation modules of attention mechanisms used to compute attention scores. $\otimes$ denotes the matrix outer product. In Equation (8), w_1 and w_2 denote the attention similarity scores for image and text, σ denotes the softmax operation, and $\|$ represents concatenation. Finally, Equations (9) and (10) weight the similarity scores w_1 and w_2 onto the image feature A and text feature N, respectively. The design of TPD Block and CPD Block ensures that F and B have the same dimensions. In contrast to the architecture used by Zia et al. [40], this paper employs custom trainable projection layers as post-processing after the encoders, which is a key distinction.

3.4. Feature Fusion

For some memes, the textual modality may be more relevant, while images may play a crucial role for others. In many cases, understanding the meaning of text requires corresponding images. Achieving comprehensive interaction between text and its corresponding image is crucial. To address this issue, we utilize image descriptions to obtain image information that the original text cannot observe, enhancing the connection between text and images. It employs an improved Cross-Mask mechanism (CM) to integrate this feature information, as shown in Figure 4. The model utilizes a Cross-Attention mechanism to obtain feature tensors y_i and y_t from both modalities and performs matrix multiplication after reshaping them, resulting in a two-dimensional fused tensor. Through a masking mechanism, the diagonal part of this tensor is flattened and output. The algorithm is as follows:

$$output = \left(vec\left(y_i y_t^\top - diag\left(y_i y_t^\top \right) \right) + F_{\text{flat}} \right) / 2 \tag{11}$$

The tensor $Output$ here represents the fused bimodal features, obtained by averaging the fused text-image features and the fused text-image caption features. Its length is n^2.

In summary, the method based on an improved Cross-Attention machine and Cross-Masking mechanism preserves the bimodal similarity calculation from text to image and from image to text (attention weight calculation). In the Cross-Attention section, we did not directly merge the weighted similarity scores into values through concatenation [40] or other methods. On the contrary, we improved the feature fusion part of Cross-Attention by using a Cross-Mask mechanism to fuse features. This method not only retains the bimodal attention weights in the Cross-Attention mechanism but also uses a method more in line with the fusion of multimodal hate meme features. Because the TCAM architecture models both the feature space and attention of the data, it significantly enhances the correlation between text and images, thus outperforming Concatenation methods.

4. Results

4.1. Dataset and Setup

The primary dataset used for evaluation in this study is the FHMC dataset. Detailed information is provided in Table 2. This dataset comprises 12,140 unique meme images. In the FHMC report, the top five contributors utilized the "other" portion of the dataset to varying extents for training, which led to some improvement in experimental results on the test set. To ensure a fair comparison with baseline methods, this study uses only the dataset's first three categories (train, dev seen, test seen) and separately trains CAM and other fusion methods (e.g., concat). This study employs AUROC as the primary evaluation metric, detailed as follows:

$$AUROC = \int_0^1 \frac{TP}{TP + FN}\left(\frac{FP}{FP + TN}\right) d\left(\frac{FP}{FP + TN}\right) \tag{12}$$

Table 2. The official FHMC dataset.

Data	Data Segmentation	Data Volume	Positive	Negative
Experimental data	train	8500	5481	3019
	dev_seen	500	253	247
	test_seen	1500	510	490
Other	dev_unseen	540	340	200
	test_unseen	2000	1250	750

This study trained TCAM and all baselines using PyTorch on an NVIDIA RTX 3090 GPU with 32 GB dedicated memory. This setup includes installation of CUDA-11.8 and cuDNN-8.9.5. The hyperparameter values for all models are shown in Table 3. These hyperparameters were manually adjusted based on the performance on the validation set with respect to the target evaluation metrics, to select the optimal parameters. For the optimizer, we compared some commonly used optimizers such as AdamW, Adam, and SGD, and the result was that AdamW provided the model with a higher score. For the learning rate, we compared it in 1×10^{-4}–1×10^{-6} and found that the learning rate of 1×10^{-5} provided a more stable convergence for the model. Due to hardware memory limitations in this experiment, we limited the batch and hidden sizes to a certain range for optimization. The batch size had little impact on the model, while the hidden size with values limited below 768 showed better model performance with larger values. In the end, we set the hidden size to 512.

Table 3. Hyperparameter configuration.

Parameter	Value
Pretrained image model	Clip-vit-large-patch14
Pretrained text model	Twitter-roberta-base-hate
Optimizer	AdamW
Batch	16
Learning rate	1×10^{-5}
Epochs	20
hidden	512

4.2. Baseline Model

To evaluate the performance of the TCAM model in hateful memes classification tasks, we select several commonly used fusion methods as well as classic methods such as ViLBERT, Visual BERT, and TCAM ablation models as baselines for comparison:

- ViLBERT: adopts a dual-stream architecture that employs a shared attention mechanism to process visual and textual information separately, thereby learning and understanding their relationships;
- Visual BERT: integrates visual features directly into the BERT model, enabling unified sequence processing to capture relationships between visual and textual information;
- TCAM_no_cg: removes transfer learning and image description components;
- TCAM_no_cam: replaces improved Cross-Attention and Cross-Mask mechanisms with concatenation methods;
- TCAM_no_pd: removes CPD and TPD modules.

4.3. Key Findings

Figure 5 displays the four models' loss curves and AUC curves—TCAM, TCAM_no_cg, TCAM_no_cam, and TCAM_no_pd—on the FHMC test set. Figure 5a shows that the TCAM model achieved the optimal expected result (AUROC 0.8362), with its loss curve converging more quickly and stably and exhibiting lower loss. In contrast, the TCAM_no_cam and TCAM_no_pd models experienced certain oscillations in their loss values during the training process. This situation may be because the fine-tuning of the CPD and TPD modules has enabled the model to learn the common features of the data. Additionally, during the learning process, the improved attention mechanism and Cross-Mask mechanism have avoided most of the noise present in the data, allowing the model to concentrate more on identifying useful information within the data. Comparison of the AUROC values among the four models in Figure 5b reveals that fine-grained feature-based transfer learning and image description contributed to approximately a 0.01 improvement in model performance and was crucial in establishing feature relationships.

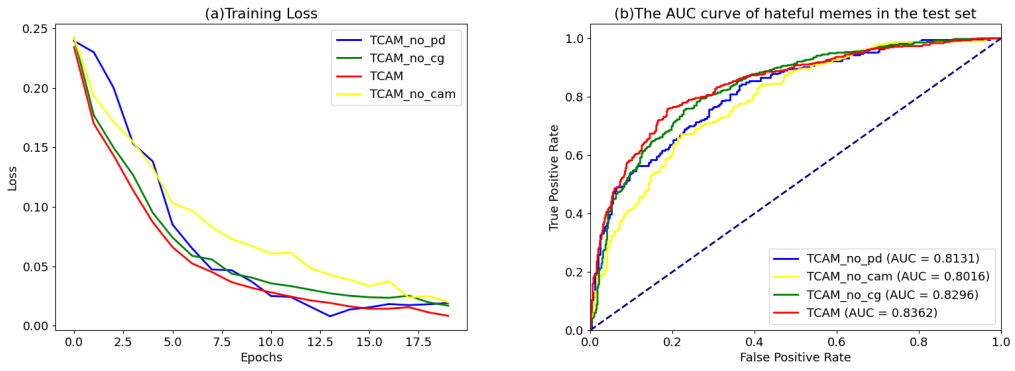

Figure 5. The loss curves and AUC curves on the FHMC test set. (**a**) The loss curves on the FHMC test set. (**b**) The AUC curves on the FHMC test set.

Figure 6 illustrates the confusion matrices of four models—TCAM, TCAM_no_cg, TCAM_no_cam, and TCAM_no_pd—on the FHMC test set. According to Figure 6, the TCAM model performs better in recognizing hateful memes than the other three models. The accuracy calculations yield 0.764 for TCAM, 0.747 for TCAM_no_cg, 0.732 for TCAM_no_cam, and 0.745 for TCAM_no_pd. This situation reaffirms the effective supportive role of transfer learning and image description in identifying weak-related hateful memes. Moreover, on the secondary metric of accuracy, the TCAM model achieved the best performance.

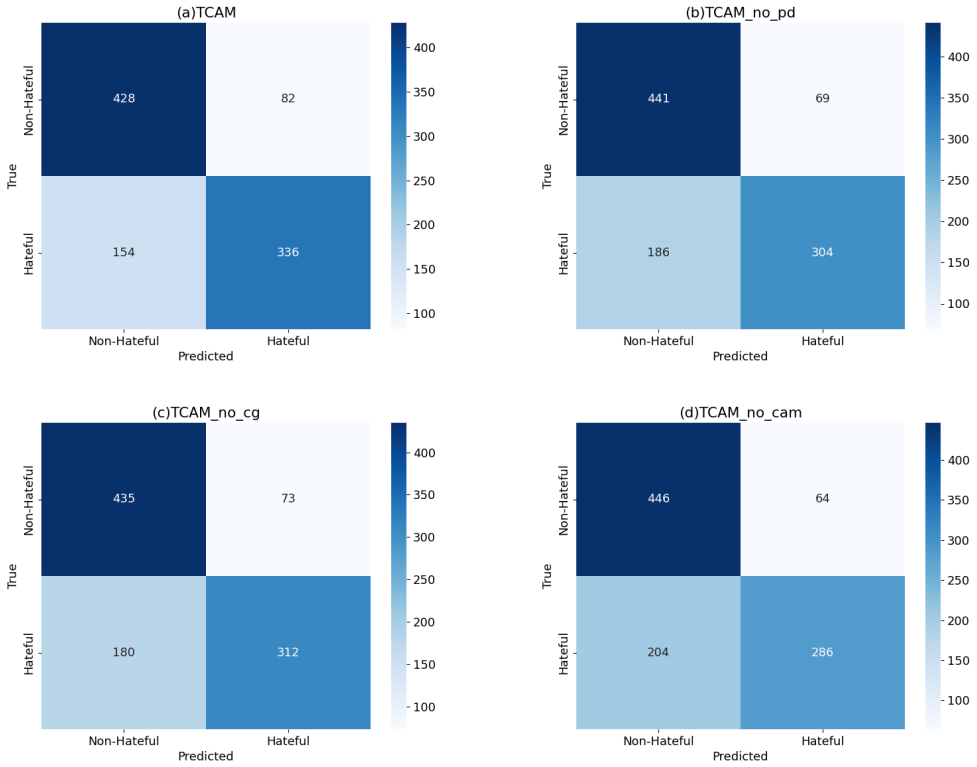

Figure 6. The confusion matrices on the FHMC test set. (**a**) The TCAM model. (**b**) The TCAM_no_pd model. (**c**) The TCAM_no_cg model. (**d**) The TCAM_no_cam model.

To further validate the effectiveness of the TCAM model, this study compares its final classification performance with various baseline models reported in the literature, as shown in Table 4. Upon observing Table 4, it is evident that TCAM outperforms several baseline methods, such as MMBT, ViLBERT, and VisualBERT. On the FHMC dataset, the best-performing baseline method (Visual BERT) achieves a test set AUROC of 0.77, significantly lower than TCAM's AUROC value (0.8362). Furthermore, the TCAM model also achieves excellent scores in secondary accuracy evaluation. In contrast, the highest accuracy achieved by the baseline models is only 0.6920. These results indicate that the TCAM model significantly enhances overall performance by integrating image–text and fine-grained feature information. Therefore, the TCAM model exhibits substantial potential in multimodal detection tasks.

Finally, the TCAM model was compared with the top five teams in FHMC, as shown in Table 5. The TCAM model ranked second in performance, trailing the first-place team by approximately 0.01 in AUROC. Despite not achieving the highest score, our streamlined end-to-end model demonstrates clear advantages over the high complexity, computational resources, and time consumption required by current ensemble models.

Table 4. Comparison of the TCAM Model with Published Baselines on the test set.

Models	Detecting Modality	AUROC	Accuracy
Image-Grid	Image	0.5300	0.5260
Image-Region	Image	0.6170	0.5580
Text BERT	Text	0.6450	0.6120
Late Fusion	Image, Text	0.7050	0.6300
ConcatBERT	Image, Text	0.6960	0.6360
MMBT-Grid	Image, Text	0.7000	0.6500
MMBT-Region	Image, Text	0.7350	0.6720
ViLBERT	Image, Text	0.7530	0.6680
Visual BERT	Image, Text	0.7570	0.6840
ViLBERT CC	Image, Text	0.7570	0.6740
Visual BERTCOCO	Image, Text	0.7700	0.6920
TCAM	Image, Text	0.8362	0.7640

Table 5. Comparison of the TCAM Model with FHMC's top five teams on the test set.

Team or Model	Architecture	AUROC
alfred lab	Ensemble	0.8450
Muennighoff	Ensemble	0.8310
HateDetectron	Ensemble	0.8108
kingsterdam	Ensemble	0.8108
burebista	Ensemble	0.7943
TCAM	End-to-End	0.8362

5. Conclusions and Future Works

Memes on social media are one of the most prevalent mediums for disseminating false and hateful information. This paper examines the challenge of detecting multimodal hateful memes. In response to this challenge, we propose an architecture capable of effectively detecting multimodal hateful memes. This architecture consists of three main components: Firstly, it enhances data with fine-grained features and image descriptions, serving as effective auxiliary inputs to the model, thereby improving performance through transfer learning. Secondly, it simplifies the fine-tuning process and aligns feature spaces by refining pre-trained models. Finally, it integrates modal features through improved Cross-Attention and Cross-Mask mechanisms, reinforcing modal relationships.

In this study, our contributions encompass three main aspects. Firstly, compared to ensemble learning methods, our end-to-end architecture TCAM avoids complex training processes, simplifying the learning workflow. Secondly, under minimal alteration to the inherent data structure, we convert image knowledge into textual knowledge to derive image descriptions, enhancing dual-modal information interaction through improved Cross-Attention and Cross-Mask mechanisms. Thirdly, we judiciously utilize fine-grained labels as auxiliary information via transfer learning. According to our experimental results, the TCAM model achieves an AUROC of 0.8362. Compared to existing baseline models in the literature, TCAM surpasses the human-recognizable threshold score of 0.82, demonstrating outstanding performance. Ablation experiments confirm TCAM's stable convergence, lower loss, and the capability of mitigating noise during learning, enhancing its ability to recognize hate memes. Overall, TCAM's robust integration of diverse information underscores the indispensability of each module.

Ethically, the concept of hateful speech is inherently subjective, making it challenging to clearly distinguish between hate and non-hate expressions. Technically, even on meticulously curated benchmark datasets, the accuracy of hateful memes classifiers remains unsatisfactory, which impedes their real-world deployment. Moreover, inadequate training data may not sufficiently encompass the diverse scenarios and variations of the target issue, leading to biased model evaluations.The multimodal dual-stream architecture combines separately extracted high-level features with cross-modal transformers to align cross-modal

representations at a higher semantic level. However, it often suffers from low parameter efficiency and may overlook fundamental feature-level correlations. Integrating these structures into a single model for hateful memes research could potentially enhance effectiveness. In the future, we plan to extend this work to other multimodal memes domains, exploring finer-grained features to improve training on specific datasets. Furthermore, given the rapid evolution of social media trends and patterns, as well as the ongoing development of hateful content definitions, real-time capturing of domain-specific memes is crucial for identifying influential content. This work can extend to capturing such real-time data and training deep learning models to identify memes containing sensitive information.

In conclusion, continuous research and innovation in this field are crucial, with a greater focus needed on effectively integrating and utilizing multimodal data. Building upon this foundation, exploring more precise and dynamic methods for hateful memes classification underscores the importance of interdisciplinary collaboration. Integrating diverse expertise from fields such as computer vision, natural language processing, and sociology can provide a comprehensive perspective and solutions to address this issue.

Author Contributions: Conceptualization, F.W., Y.Y., and Z.L.; methodology, F.W. and Y.Y.; software, F. W. and Z.L.; validation, J.C.; formal analysis, F.W.; investigation, F.W.; resources, F.W.; data curation, G.C.; writing—original draft preparation, F.W.; writing—review and editing, J.C.; visualization, F.W.; supervision, F.W. and J.C.; project administration, F.W., G.C., and J.C.; funding acquisition, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hainan Province Higher Education Teaching Reform Research Project: Hnjg2024ZD-19.

Data Availability Statement: The data presented in this study are available in the public domain: https://github.com/drivendataorg/hateful-memes/ (accessed on 18 June 2024).

Acknowledgments: The authors would like to acknowledge and thank all reviewers for their constructive and helpful reviews.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAM	Cross-Attention and Cross-Mask Mechanisms
CM	Cross-Mask mechanism
CA	Cross-Attention
FHMC	Facebook Hateful Memes Challenge
TPD	Text Projection Layer Module
CPD	Image Projection Layer Module

References

1. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2611–2624.
2. Pierri, F.; Luceri, L.; Chen, E.; Ferrara, E. How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. *EPJ Data Sci.* **2023**, *12*, 43. [CrossRef] [PubMed]
3. Nogara, G.; Vishnuprasad, P.S.; Cardoso, F.; Ayoub, O.; Giordano, S.; Luceri, L. The disinformation dozen: An exploratory analysis of COVID-19 disinformation proliferation on Twitter. In Proceedings of the 14th ACM Web Science Conference, Barcelona, Spain, 26–29 June 2022; pp. 348–358.
4. Chen, E.; Jiang, J.; Chang, H.-C.H.; Muric, G.; Ferrara, E. Charting the information and misinformation landscape to characterize misinfodemics on social media: COVID-19 infodemiology study at a planetary scale. *JMIR Infodemiol.* **2022**, *2*, e32378. [CrossRef] [PubMed]
5. Delisle, L.; Kalaitzis, A.; Majewski, K.; de Berker, A.; Marin, M.; Cornebise, J. A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter. *arXiv* **2019**, arXiv:1902.03093.
6. La Rue, F. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. *Hum. Rights Counc.* **2011**, *16*, 4–10.

7. Biermann, F.; Kanie, N.; Kim, R.E. Global governance by goal-setting: The novel approach of the UN Sustainable Development Goals. *Curr. Opin. Environ. Sustain.* **2017**, *26*, 26–31. [CrossRef]
8. Hamza, A.; Javed, A.R.; Iqbal, F.; Yasin, A.; Srivastava, G.; Połap, D.; Gadekallu, T.R.; Jalil, Z. Multimodal Religiously Hateful Social Media Memes Classification based on Textual and Image Data. *ACM Trans. Asian-Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–7. [CrossRef]
9. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8748–8763.
10. Barbieri, F.; Camacho-Collados, J.; Neves, L.; Espinosa-Anke, L. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv* **2020**, arXiv:2010.12421.
11. Derczynski, L.; Ritter, A.; Clark, S.; Bontcheva, K. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, Hissar, Bulgaria, 9–11 September 2013; pp. 198–206.
12. Han, B.; Baldwin, T. Lexical normalisation of short text messages: Makn sens a# twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 368–378.
13. Baldwin, T.; De Marneffe, M.-C.; Han, B.; Kim, Y.-B.; Ritter, A.; Xu, W. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In Proceedings of the Workshop on Noisy User-Generated Text, Beijing, China, 31 July 2015; pp. 126–135.
14. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: An experimental study. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 1524–1534.
15. Baldwin, T.; Cook, P.; Lui, M.; MacKinlay, A.; Wang, L. How noisy social media text, how diffrnt social media sources? In Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–18 October 2013; pp. 356–364.
16. Poria, S.; Hazarika, D.; Majumder, N.; Mihalcea, R. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.* **2020**, *14*, 108–132. [CrossRef]
17. Kim, S.; Weber, I.; Wei, L.; Oh, A. Sociolinguistic analysis of Twitter in multilingual societies. In Proceedings of the 25th ACM Conference on Hypertext and Social Media, Santiago, Chile, 1–4 September 2014; pp. 243–248.
18. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4651–4659.
19. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
20. Hosseinmardi, H.; Mattson, S.A.; Ibn Rafiq, R.; Han, R.; Lv, Q.; Mishra, S. Detection of Cyberbullying Incidents on the Instagram Social Network. *arXiv* **2015**, arXiv:1503.03909.
21. Gomez, R.; Gibert, J.; Gomez, L.; Karatzas, D. Exploring hate speech detection in multimodal publications. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1470–1478.
22. Sharma, C.; Paka, W.; Scott, D.B.; Das, A.; Poria, S.; Chakraborty, T.; Gambäck, B. Task report: Memotion analysis 1.0 @ SemEval 2020: The visuo-lingual metaphor! *arXiv* **2020**, arXiv:2008.03781.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
25. Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; Testuggine, D. Supervised multimodal bitransformers for classifying images and text. *arXiv* **2019**, arXiv:1909.02950.
26. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–8.
27. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv* **2019**, arXiv:1908.03557.
28. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Fitzpatrick, C.A.; Bull, P.; Lipstein, G.; Nelli, T.; Zhu, R.; et al. The hateful memes challenge: Competition report. In Proceedings of the NeurIPS 2020 Competition and Demonstration Track, Online, 6–12 December 2020; pp. 344–360.
29. Zhu, R. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv* **2020**, arXiv:2012.08290.
30. Muennighoff, N. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv* **2020**, arXiv:2012.07788.
31. Velioglu, R.; Rose, J. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv* **2020**, arXiv:2012.12975.
32. Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; Yannakoudakis, H. A multimodal framework for the detection of hateful memes. *arXiv* **2020**, arXiv:2012.12871.

33. Sandulescu, V. Detecting hateful memes using a multimodal deep ensemble. *arXiv* **2020**, arXiv:2012.13235.
34. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv* **2019**, arXiv:1908.08530.
35. Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 104–120.
36. Gan, Z.; Chen, Y.-C.; Li, L.; Zhu, C.; Cheng, Y.; Liu, J. Large-scale adversarial training for vision-and-language representation learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6616–6628.
37. Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; Wang, H. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *AAAI Conf. Artif. Intell.* **2021**, *35*, 3208–3216. [CrossRef]
38. Mathias, L.; Nie, S.; Davani, A.M.; Kiela, D.; Prabhakaran, V.; Vidgen, B.; Waseem, Z. 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), Bangkok, Thailand, 6 August 2021; pp. 201–206.
39. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 7–23 July 2022; pp. 12888–12900.
40. Zia, H.B.; Castro, I.; Tyson, G. Racist or sexist meme? classifying memes beyond hateful. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Bangkok, Thailand, 6 August 2021; pp. 215–219.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.