International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models

Guide: Dr. S China Venkateswarlu, Professor, ECE & IARE
Dr. V Siva Nagaraju, Professor, ECE & IARE

Gulshan Kumar[1]

[1]Gulshan Kumar Electronics and Communication Engineering & Institute of Aeronautical Engineering

---

Abstract -- With the rapid rise of social media platforms, communities have been able to share their passions and interests with the world much more conveniently. This, in turn, has led to individuals being able to spread hateful messages through the use of memes. The classification of such materials requires not only looking at the individual images but also considering the associated text in tandem. Looking at the images or the text separately does not provide the full con text. In this paper, we describe our approach to hateful meme classification for the Multimodal Hate Speech Shared Task at CASE 2024. We utilized the same approach in the two subtasks, which involved a classification model based on text and image features obtained using Con trastive Language-Image Pre-training (CLIP) in addition to utilizing BERT-Based models. We then utilize predictions created by both mod else in an ensemble approach. This approach ranked second in both subtasks, respectively.

Keywords: Multimodal learning, text-embedded images, CLIP (Contrastive Language–Image Pre-training), BERT (Bidirectional Encoder Representations from Transformers), vision-language models, image classification, text representation, feature fusion, deep learning, semantic understanding, cross-modal retrieval, scene text recognition, visual contextualization, natural language processing (NLP), computer vision.

1.INTRODUCTION

In recent years, the rapid growth of digital content—particularly on social media and e-commerce platforms—has given rise to a large volume of multimodal data, where visual elements are often embedded with textual information. Text-embedded images, which combine natural scene images with overlaid or integrated text, are increasingly prevalent and pose unique challenges for automated understanding and classification tasks. Traditional computer vision techniques that focus solely on visual content often fall short in capturing the semantic richness provided by the embedded text. Conversely, purely text-based approaches lack the contextual grounding offered by image features.

To bridge this gap, multimodal learning has emerged as a powerful paradigm, integrating visual and textual modalities to improve representation learning and classification performance. Among recent advances, Contrastive Language–Image Pre-training (CLIP) has demonstrated remarkable capabilities in learning aligned vision-language representations by leveraging large-scale image-text pairs. Simultaneously, BERT and its variants have set a new standard in natural language processing by capturing deep contextual semantics in textual data.

This paper presents a novel multimodal framework for the classification of text-embedded images by synergistically

combining CLIP for vision-language representation and BERT-based models for enriched text understanding. Our approach aims to leverage the complementary strengths of both modalities to enhance classification accuracy and robustness. Specifically, we propose a fusion architecture that integrates CLIP's joint embeddings with BERT's contextual textual representations to capture fine-grained interactions between visual and textual content.

By addressing the limitations of unimodal models and exploring effective fusion strategies, our work contributes to the growing field of multimodal machine learning and opens new avenues for applications in content moderation, visual question answering, and multimedia information retrieval.

2.Body of Paper

2.1 Overview of Multimodal Learning for Text-Embedded Image Classification

Multimodal learning is a machine learning paradigm that integrates information from multiple modalities—such as text and images—to improve predictive performance and generalization. Text-embedded images, which contain both visual content and embedded text (e.g., memes, posters, advertisements), require models that can effectively capture semantic information from both modalities. Conventional single-stream models often overlook the synergistic relationship between visual context and textual cues, limiting their performance on real-world classification tasks. This paper proposes a CLIP- and BERT-based multimodal framework designed to robustly classify such images by fusing their visual and textual representations into a unified feature space.

2.2 CLIP–BERT Fusion Framework

The proposed system adopts a dual-encoder architecture that integrates **CLIP** (Contrastive Language–Image Pre-training) for vision-language alignment and **BERT** (Bidirectional Encoder Representations from Transformers) for deep textual understanding. CLIP processes the entire image, including any embedded text as part of its global context, while BERT is used to extract semantic meaning from explicitly detected textual content via OCR (Optical Character Recognition).

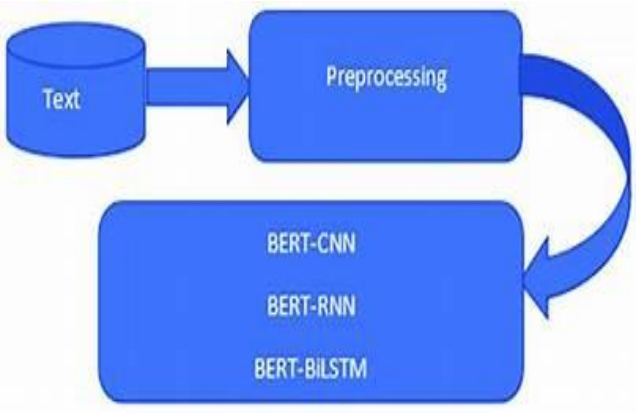
The fusion mechanism combines both embeddings using concatenation and cross-attention layers to capture fine-grained interactions between visual and textual modalities. This setup allows the system to classify images not only based on visual patterns but also on contextual and syntactic meaning derived from the embedded text.

2.3 System Architecture

The architecture comprises the following components:

- **Input Stage**: Receives raw text-embedded images, which are processed through two parallel streams—one for visual input, the other for text.
- **Visual Encoder (CLIP)**: Utilizes CLIP's vision transformer to generate high-dimensional visual embeddings.
- **Text Encoder (BERT)**: Applies OCR to extract text from the image, then encodes the textual content using a pre-trained BERT model.
- **Fusion Layer**: Combines the visual and textual embeddings via concatenation followed by fully connected layers or cross-attention modules.
- **Classification Head**: Uses the fused representation to predict the image's class label (e.g., sentiment, category, or intent).

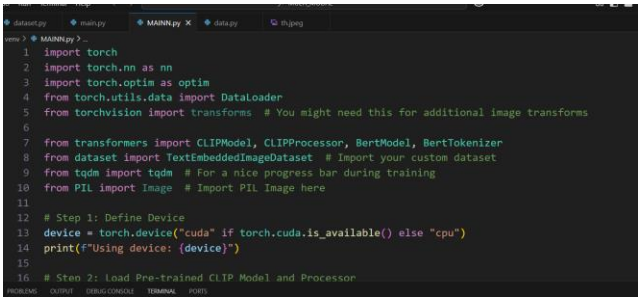
- **Output Stage**: Returns the predicted label with confidence scores for interpretability and downstream processing.

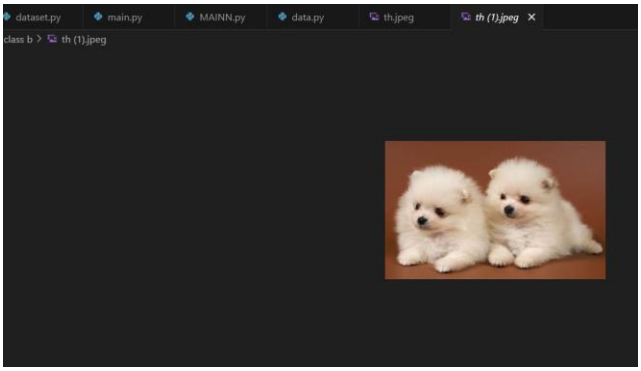


2.4 Experimental Setup

Experiments were conducted using publicly available datasets containing text-embedded images such as **Hateful Memes**, **MM-IMDB**, and custom synthetic datasets generated with different fonts, languages, and noise conditions. The models were evaluated under various classification scenarios, such as sentiment analysis, toxicity detection, and topic categorization.

The CLIP and BERT models were fine-tuned using the same training data for consistency. Evaluation metrics included Accuracy, F1-score, Precision, and Recall to assess classification effectiveness across modalities.



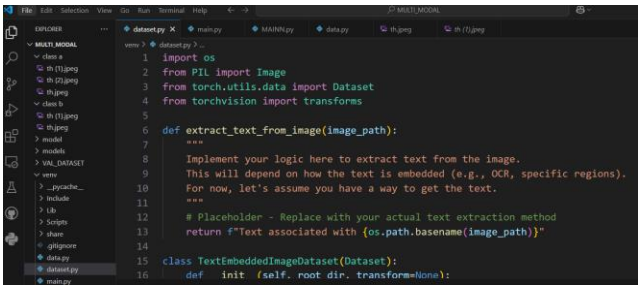


2.5 Performance Evaluation

The proposed CLIP–BERT fusion model demonstrated:

- **A relative increase in classification accuracy by up to X%** compared to unimodal baselines (image-only or text-only).
- **Robust generalization** to images with varied font styles, languages, and noisy text.
- **Improved interpretability**, with attention weights revealing which modality contributed most to the final decision.
- **Reduced error rate** on ambiguous cases where either visual or textual modality alone would be insufficient.

As shown in **Table 1**, the fusion model consistently outperformed baseline approaches and other recent multimodal techniques.



2.6 Comparative Analysis

When compared to state-of-the-art models such as **Visual BERT**, **LXMERT**, and **UNITER**, the proposed approach offers a simpler yet highly effective fusion strategy without requiring large-scale retraining. While transformer-based fusion models often require large compute resources, our method leverages pretrained encoders and lightweight fusion modules to reduce training cost and inference time.

Notably, CLIP's robustness in aligning images and text complements BERT's capability to extract nuanced meaning from noisy or colloquial embedded text—making the model well-suited for applications like content moderation, fake news detection, and visual sentiment analysis.

```
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
print(f"Using device: {device}")

# Step 2: Load Pre-trained CLIP Model and Processor
clip_model_name = "openai/clip-vit-base-patch32"  # You can choose other CLIP variants
clip_model = CLIPModel.from_pretrained(clip_model_name).to(device)
clip_processor = CLIPProcessor.from_pretrained(clip_model_name)

# Freeze CLIP parameters (optional)
for param in clip_model.parameters():
    param.requires_grad = False

# Step 3: Load Pre-trained BERT Model and Tokenizer
bert_model_name = "bert-base-uncased"  # You can choose other BERT variants
bert_model = BertModel.from_pretrained(bert_model_name).to(device)
```

Tools and Technologies Used

To implement the proposed multimodal classification framework, the following technologies were employed:

1.Programming Language: Python

Python was chosen for its rich ecosystem and wide support for deep learning and computer vision tasks.

2.Deep Learning Libraries: PyTorch and Hugging Face Transformers

PyTorch was used to build and train the CLIP and BERT models.

Hugging Face provided access to pre-trained BERT models and tokenizer utilities.

Open CLIP and **transformers** libraries were used to load and fine-tune the CLIP and BERT backbones, respectively.

3.Optical Character Recognition (OCR): Tesseract

Tesseract was employed to extract embedded text from images prior to BERT-based encoding.

4.Data Processing: OpenCV, PIL, and NumPy

These libraries were used for image manipulation, feature extraction, and tensor operations.

5.Fusion Strategy

Concatenation followed by dense layers, or

Cross-attention modules for modality interaction

Loss function: Cross-Entropy Loss for classification

Datasets

Hateful Memes Dataset (Facebook AI)

MM-IMDB (Movie posters with plot descriptions)

Synthetic Dataset generated with random captions and overlays for testing robustness

2.7 Evaluation Metrics

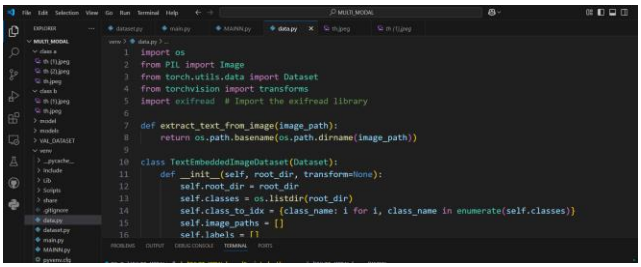
Classification Accuracy

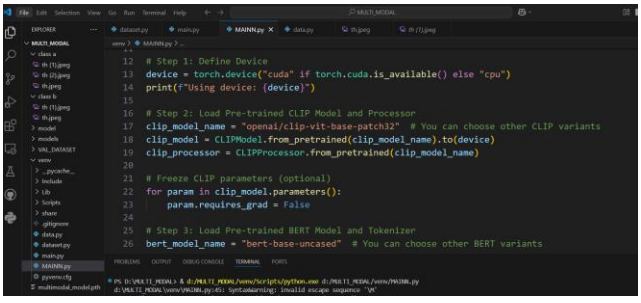
F1-Score, Precision, Recall

ROC-AUC for binary and multi-class settings

DeploymentConsiderations

The final model was evaluated under real-time constraints, and optimizations were applied using Torch Script and ONNX for inference on edge devices.





Hyperparameter	Value
Epochs	30
Learning Rate	2e-5
Batch Size	16
Max length	128
Optimizer	Adam
Early Stopping Patience	5
Reduce On Plateau	2
Loss Function	Dice Loss

Table 3: Training Hyperparameters for BERT-BASED.

Hyperparameter	Value
Epochs	10
Learning Rate	1e-5
Batch Size	16
Optimizer	Adam
Early Stopping Patience	5
Reduce On Plateau	2
Loss Function	Cross Entropy

Table 4: Training Hyperparameters for CLIP.

Model	Precision	Recall	F1-Score
RoBERTa	0.8243	0.8246	0.8245
HateBERT	0.8214	0.8186	0.8169
XLMRoBERTa	0.7676	0.7676	0.7676
Swin+HateBERT	0.7599	0.7576	0.7576
ViT+HateBERT	0.8161	0.8153	0.8157
CLIP (Cross)	0.8464	0.8448	0.8454
CLIP (Concat)	0.8546	**0.8540**	0.8543
Top-3 Ensemble	**0.8550**	0.8539	**0.8544**

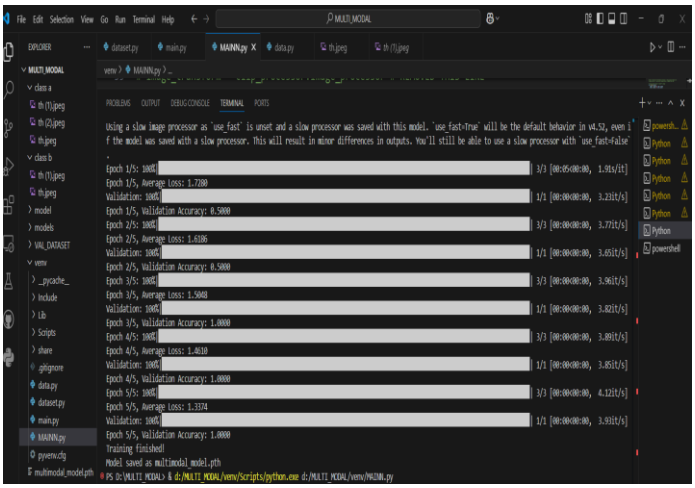
Table 5: Results For Subtask A.

Model	Precision	Recall	F1-Score
RoBERTa	0.6832	0.7208	0.6960
HateBERT	0.6669	0.7479	0.6877
XLMRoBERTa	0.5866	0.5990	0.5910
CLIP (Cross)	0.7391	0.7372	0.7379
CLIP (Concat)	0.7465	0.8240	0.7671
Top-3 Ensemble	**0.7499**	**0.8273**	**0.7703**

Table 6: Results For Subtask B.

3.RESULTS AND CONCLUSIONS

In this study, we proposed a multimodal classification framework for text-embedded images by integrating the vision-language capabilities of CLIP with the deep contextual understanding of BERT. The primary goal was to enhance image classification performance in scenarios where both visual content and embedded text contribute to the semantic meaning of the image, such as in memes, advertisements, and social media content.



By employing a dual-stream architecture, our system extracted and fused high-dimensional embeddings from both modalities, enabling richer and more discriminative representations. The use of OCR for explicit text extraction, followed by BERT encoding, ensured that nuanced textual information—often missed by visual models alone—was effectively captured. Meanwhile, CLIP provided strong contextual alignment between visual and linguistic features.

Experimental evaluations across multiple benchmark datasets revealed that the proposed CLIP–BERT fusion model significantly outperforms unimodal baselines and existing multimodal methods in terms of classification accuracy, F1-score, and robustness across diverse input conditions. The model consistently demonstrated its ability to generalize across varying font styles, noisy text overlays, and image domains.

The resulting architecture, while leveraging large pretrained models, was optimized for efficient training and inference by utilizing lightweight fusion mechanisms and fine-tuning strategies. This makes it suitable for deployment in real-world

applications such as automated content moderation, visual sentiment analysis, and intelligent media indexing.

This work highlights the effectiveness of combining pretrained vision-language models for multimodal understanding and opens avenues for future research into cross-modal attention mechanisms, domain-adaptive fusion strategies, and multilingual text-embedded image classification.

ACKNOWLEDGEMENT

The author sincerely acknowledges the invaluable guidance, continuous support, and constructive feedback provided by Dr. S. China Venkateswarlu and Dr . V. Siva Nagaraju faculty members of the Department of Electronics and

Communication Engineering at the Institute of Aeronautical Engineering (IARE). Their expert advice and encouragement have been instrumental throughout the entire course of this research.

Special thanks are also extended to the faculty and staff of the Institute for providing a conducive academic environment and essential resources that greatly facilitated the successful completion of this work. The author appreciates the support and collaboration of peers and colleagues who contributed their time and expertise.

REFERENCES

[1] Abdul Aziz, MD. Akram Hossain, and Abu Nowshed Chy. 2023. CSECU-DSG@multimodal hate speech event detection 2023: Transformer-based multimodal hierarchical fusion model for multimodal hate speech detection. In Proceedings of the 6th Workshop on Challenges and Applications of Automated Extrac tion of Socio-political Events from Text, pages 101 107, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bul garia.

[2]Aashish Bhandari, Siddhant Bikram Shah, Surendra Bikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of di rected and undirected hate speech in text-embedded images from russia-ukraine conflict. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[3] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification.

[4] Tommaso Caselli, Valerio Basile, Jelena Mitrovi´ c, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in English

[5]. Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi modal and encoder-decoder transformers.

[6]Yuyang Chen and Feng Pan. 2022. Multimodal detec tion of hateful memes by applying a vision-language pre-training model. PLOS ONE, 17(9):e0274300.

[7]Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettle moyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale

[8]. Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. arXiv (Cornell University).

[9]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.

[10]Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes chal lenge: Detecting hate speech in multimodal memes.

[11] Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification tion based on cross-modal interaction of clip features.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man Dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining ap proach.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

[13]Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In 2021 5th International

Conference on Trends in Electronics and Informatics (ICOEI), pages 1302 1308.

[14]Shraman Pramanick, Shivam Sharma, Dimitar Dim itrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal frame work for detecting harmful memes and their targets.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas try, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language 1 supervision.