Article

Not peer-reviewed version

Toxic Memes Recognition Through with Multimodal Bidirectional Cross-Attention

Alex Hashim [*] , Natalie Coleman , Jannat Roy

Posted Date: 14 November 2024

doi: 10.20944/preprints202411.1071.v1

Keywords: Hateful Memes Detection; Multimodal Learning; Transformer; Bidirectional Cross-Attention; Deep Ensembles

Article

Toxic Memes Recognition Through with Multimodal Bidirectional Cross-Attention

Alex Hashim *, Natalie Coleman and Jannat Roy

Brandeis University

* Correspondence: alex.hashim@brandeis.edu

Abstract: Despite the considerable advancements achieved with machine learning techniques in identifying hate speech, numerous technical obstacles persist that hinder these models from reaching human-level accuracy. Challenges such as understanding nuanced context, detecting sarcasm, and effectively interpreting the interplay between visual and textual elements in memes complicate the detection process. This study delves into the comprehensive evaluation of several cutting-edge visual-linguistic Transformer architectures, including VL-BERT, VLP, UNITER, and LXMERT, to assess their capabilities and limitations in handling the multifaceted nature of hateful content within memes. Building upon these evaluations, we introduce significant enhancements aimed at boosting their efficacy in this domain by developing a novel bidirectional cross-attention mechanism. This mechanism facilitates a more seamless integration of visual and textual information, enabling the model to better capture the subtle cues that distinguish hateful memes from benign ones. In addition to the architectural improvements, we leverage deep ensemble strategies to aggregate predictions from multiple model instances, thereby enhancing the robustness and reliability of the detection system. By combining the strengths of diverse models, the ensemble approach mitigates individual weaknesses and reduces the likelihood of false positives and negatives. Our proposed framework not only addresses the existing shortcomings of single-model approaches but also markedly surpasses existing baseline performances by a large margin, achieving higher AUROC and accuracy scores. The refined model demonstrates superior capability in discerning hateful content within multimodal memes, offering a more robust and reliable tool for mitigating the proliferation of harmful online material. Furthermore, the scalability of our approach ensures its applicability to evolving online threats, providing a sustainable solution for automated hate speech detection. These advancements signify a meaningful step towards enhancing the effectiveness of machine learning models in creating a safer and more respectful online environment.

Keywords: Hateful Memes Detection; Multimodal Learning; Transformer; bidirectional cross-attention; deep ensembles

1. Introduction

The pervasive influence of the internet has fundamentally transformed various facets of our daily existence, shaping not only how we communicate and access information but also reflecting and reinforcing our personal identities, beliefs, and, unfortunately, our inherent biases and prejudices. In today's digital age, billions of individuals engage with a vast array of online content every single day. This content spans a broad spectrum, encompassing highly informative and educational materials that significantly enhance our knowledge and understanding of the world around us. However, alongside these beneficial resources, there is a growing surge in harmful content that poses substantial risks to individuals and communities alike. This detrimental content includes, but is not limited to, hate speech, the dissemination of misinformation, and various other forms of online abuse that can have severe psychological and societal impacts [3,4].

As the volume of harmful online content continues to escalate, the demand for effective detection and mitigation strategies becomes increasingly urgent. The rapid proliferation of such content necessitates the deployment of advanced technological solutions capable of swiftly identifying and addressing these issues. This involves scaling up content review processes and developing automated systems that can make real-time decisions to remove or flag harmful media. The primary objective is to minimize the potential harm inflicted upon readers by ensuring that offensive and abusive content is promptly taken down, thereby fostering a safer and more respectful online environment [8].

A significant portion of our online interactions transpires on social media platforms, which serve as conduits for sharing messages, images, and a multitude of multimedia content with both private communities and the general public. These platforms have become integral to modern communication, allowing for the instantaneous exchange of information and ideas across the globe. However, the very features that make social media platforms so effective for communication also make them susceptible to misuse. The ease of sharing content can be exploited to propagate hateful memes—multimodal content that combines images and text to convey messages that may be offensive or harmful [4].

In response to this challenge, the task of hateful meme recognition has been proposed, aimed at enhancing the detection of hateful memes that integrate both visual and textual elements. To facilitate this endeavor, the competition organizers have curated a unique and meticulously labeled dataset comprising over 10,000 high-quality multimodal memes [4]. This dataset serves as a critical resource for developing and benchmarking algorithms designed to identify hate speech within the intricate interplay of images and text that characterize memes. The primary goal of this competition is to devise an algorithm capable of accurately detecting multimodal hate speech in internet memes while maintaining robustness against benign alterations.

A meme can be classified as hateful or offensive based on its visual content, the accompanying text, or the synergistic combination of both. The concept of benign flipping plays a pivotal role in this context. Benign flipping is an augmentation technique employed by the competition organizers to invert the hateful classification of a meme, effectively transforming it from hateful to non-hateful or vice versa. This process involves altering either the textual component or the visual imagery of the meme, thereby changing its overall sentiment and impact. By incorporating benign flipping, the competition ensures that the developed algorithms are not only adept at identifying hate speech but are also resilient to subtle modifications that may otherwise obscure or alter the meme's intent [4].

The detection of hateful memes is framed as a binary classification task, wherein each meme is categorized as either hateful or non-hateful. The primary metric for evaluating the performance of classification models in this task is the Area Under the Receiver Operating Characteristic Curve (AUROC). The AUROC quantifies the model's ability to discriminate between the two classes by measuring the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve itself plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various classification thresholds T. The objective is to maximize the AUROC, thereby enhancing the model's discriminative power and ensuring it can effectively distinguish between hateful and non-hateful content under varying conditions [16].

$$\text{AUROC} = \int_{-\infty}^{\infty} \text{TPR}(T)\, d(\text{FPR}(T))$$

In addition to AUROC, Accuracy serves as a supplementary metric to evaluate the performance of the classification models. Accuracy measures the proportion of correctly predicted instances out of the total number of instances in the test set. It is calculated as the ratio of the number of correct predictions $\hat{y}$ to the actual class labels y:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} I(y_i = \hat{y}_i)$$

The ideal model aims to optimize both AUROC and Accuracy, ensuring high discriminative capability and reliable classification performance.

In this study, we present a novel framework named *BCA-MemeNet* (Bidirectional Cross-Attention MemeNet), which introduces significant advancements in the detection of hateful memes through sophisticated multimodal deep learning techniques. Our contributions to the field are threefold:

- We undertake a comprehensive evaluation and fine-tuning of both single-stream and dual-stream state-of-the-art Transformer architectures, including VL-BERT [9], VLP [13], UNITER [2], and LXMERT [11]. By meticulously comparing these models against established baselines, we leverage their extensive pre-training on diverse datasets [8,10] to enhance their applicability and performance in the context of hateful meme detection.
- We introduce a novel bidirectional cross-attention mechanism within *BCA-MemeNet* that effectively integrates inferred caption information with the original meme caption text. This mechanism facilitates a more nuanced and dynamic interaction between the visual and textual modalities, thereby improving the classifier's ability to discern subtle and complex forms of hate speech embedded within memes. Our approach draws inspiration from the cross-attention techniques utilized in [2], extending them to better capture the bidirectional dependencies between image and text features [1].
- We demonstrate that employing deep ensemble strategies [5] significantly enhances the predictive performance of *BCA-MemeNet* compared to single-model predictions. By aggregating the outputs of multiple model instances, the ensemble approach mitigates individual model biases and variances, leading to more robust and reliable classifications. This ensemble methodology not only improves overall accuracy and AUROC but also contributes to the model's resilience against overfitting and its ability to generalize across diverse and unseen data distributions [16].

Through these contributions, *BCA-MemeNet* represents a substantial advancement in the automated detection of hateful memes, addressing both the technical challenges and the practical requirements of maintaining a safe and respectful online environment. By leveraging cutting-edge Transformer architectures [2,9], innovative cross-attention mechanisms [1], and robust ensemble strategies [5], our framework sets a new benchmark for performance and reliability in the realm of multimodal hate speech detection [3,8].

2. Related Work

The landscape of natural language processing (NLP) and computer vision has been significantly transformed by the advent of Transformer architectures, which have set new benchmarks across a myriad of tasks. These architectures, particularly those pre-trained on expansive datasets, have demonstrated unparalleled performance in various language processing challenges [16]. The integration of visual and linguistic data through these models has opened new avenues for tackling complex multimodal tasks, including visual question answering, visual reasoning, and image captioning [2,9].

2.1. Pre-Trained Transformer

Transformer-based models have revolutionized the field of NLP by enabling deep bidirectional understanding of language. BERT (Bidirectional Encoder Representations from Transformers) [16] stands out as one of the most influential models due to its ability to capture context from both directions in a sentence. BERT's architecture allows it to generate rich, contextualized word embeddings, which have been pivotal in enhancing the performance of downstream tasks such as sentiment analysis, named entity recognition, and machine translation. The success of BERT has spurred the development of numerous variants and extensions, each aiming to refine and expand its capabilities [16].

Beyond BERT, other Transformer-based models like GPT (Generative Pre-trained Transformer) and RoBERTa have further pushed the boundaries of language understanding and generation. These models leverage large-scale pre-training on diverse corpora, enabling them to generalize effectively across different domains and tasks [16]. The robustness and versatility of these models make them indispensable tools in the modern NLP toolkit, providing a strong foundation for more specialized applications.

2.2. Visual-Linguistic Transformer

The fusion of visual and linguistic modalities has given rise to a new class of Transformer architectures designed to handle multimodal data. Models such as VL-BERT [9], UNITER [2], and LXMERT [11] exemplify this trend by integrating image and text processing within a unified framework. These models are typically pre-trained on large-scale datasets that encompass both visual and textual information, enabling them to learn joint representations that capture the intricate relationships between images and their corresponding descriptions.

LXMERT [11] employs a dual-stream architecture, consisting of separate encoders for processing text and images. The model leverages a Transformer-based cross-modality encoder to integrate information from both streams, facilitating a comprehensive understanding of the input data. Image features are extracted using a Faster R-CNN [1] feature extractor, which provides rich object-level representations that are crucial for tasks like visual question answering and image captioning. This separation of modalities allows LXMERT to effectively handle complex multimodal interactions, making it a powerful tool for tasks that require nuanced understanding of both visual and linguistic information [11].

In contrast to dual-stream models, single-stream architectures like VL-BERT [9] and UNITER [2] integrate visual and textual data within a single Transformer framework. These models concatenate image and text embeddings, allowing the Transformer to process them jointly. This unified approach facilitates seamless interaction between modalities, enabling the model to capture dependencies and correlations more effectively. VL-BERT, for instance, builds upon the BERT architecture by incorporating visual embeddings, thereby enhancing its capability to perform multimodal tasks [9]. Similarly, UNITER introduces a universal image-text representation that can be fine-tuned for a variety of downstream applications, demonstrating state-of-the-art performance in several benchmarks [2].

2.3. Multimodal Hate Speech Detection

The detection of hate speech in multimodal content, such as memes, presents unique challenges due to the interplay between visual and textual elements. Traditional hate speech detection methods have primarily focused on textual analysis using NLP techniques. However, memes often convey messages through a combination of images and text, necessitating the integration of computer vision and NLP to effectively capture their semantics [3].

One of the pioneering efforts in this domain is the Hateful Memes Challenge [4], which introduced a large-scale dataset specifically designed for multimodal hate speech detection. This dataset comprises over 10,000 memes annotated for hatefulness, incorporating both visual and textual content. The challenge highlights the difficulty of detecting hate speech in memes, as the hateful intent can stem from the image, the text, or the combination of both. Additionally, the concept of benign flipping—where a meme's hateful classification is inverted by altering either the image or text—adds another layer of complexity, requiring models to be robust against such manipulations [4].

Research by [3] explored hate speech detection in multimodal publications harvested from Twitter using specific hateful seed keywords. Their findings revealed that while multimodal models hold promise, they did not consistently outperform unimodal text-based models. This underscores the inherent challenges in effectively integrating and interpreting visual and textual data for hate speech detection. Factors such as the subtlety of hateful intent, cultural context, and the diversity of meme formats contribute to these challenges [3].

The integration of Transformer architectures into multimodal hate speech detection has led to significant improvements. Models like VL-BERT and UNITER have been leveraged to better understand the context and semantics of memes by capturing the relationships between images and text. The bidirectional cross-attention mechanisms inherent in these models allow for a more nuanced interpretation of how visual and textual elements interact to convey hateful messages [2,9]. Despite these advancements, there remains a need for more sophisticated approaches to fully harness the potential of multimodal data in hate speech detection.

Ensemble methods have long been recognized for their ability to improve model performance by combining the strengths of multiple models. Deep ensembles, in particular, offer a scalable approach to uncertainty estimation and performance enhancement [5]. By aggregating the predictions of several independently trained models, ensembles can mitigate individual model biases and variances, leading to more robust and reliable classifications.

The primary advantage of deep ensembles lies in their ability to capture diverse model perspectives, which enhances the overall predictive accuracy and stability. This diversity helps in reducing the likelihood of overfitting, as the ensemble can generalize better across different data distributions [5]. Additionally, deep ensembles provide a natural mechanism for uncertainty estimation, which is crucial for applications requiring high reliability and interpretability [5].

In the context of multimodal hate speech detection, ensemble methods can significantly enhance the performance of models like *BCA-MemeNet*. By leveraging multiple instances of Transformer-based architectures, ensembles can aggregate diverse interpretations of the same meme, leading to more accurate and reliable hate speech classifications. This approach not only improves metrics such as AUROC and Accuracy but also enhances the model's resilience against adversarial manipulations like benign flipping [4,5].

Implementing deep ensembles involves training multiple instances of the same model with different initializations or subsets of the training data. The predictions from these models are then aggregated, typically through averaging or majority voting, to produce the final output [5]. While this approach increases computational overhead, the benefits in terms of performance and robustness make it a valuable strategy for critical applications like hate speech detection.

Cross-attention mechanisms play a pivotal role in enhancing the interaction between visual and textual modalities within Transformer architectures. By allowing the model to focus on relevant parts of the input data from both modalities, cross-attention facilitates a more comprehensive understanding of the content.

In *BCA-MemeNet*, a bidirectional cross-attention mechanism is employed to synergize inferred caption information with the original meme caption text. This bidirectional approach ensures that the model can effectively capture dependencies and contextual relationships from both the image and text simultaneously [1,2]. Unlike unidirectional attention, which processes one modality at a time, bidirectional cross-attention enables a more holistic integration of multimodal information, thereby improving the model's ability to detect subtle and complex forms of hate speech embedded within memes.

The incorporation of bidirectional cross-attention significantly enhances the classifier's performance by allowing for dynamic interactions between the visual and textual components of a meme. This mechanism enables the model to identify and interpret nuanced cues that may indicate hateful intent, which might be overlooked by models relying solely on unimodal analysis [1,2]. Consequently, models equipped with advanced cross-attention mechanisms demonstrate superior capability in discerning hateful content, as evidenced by improved AUROC and Accuracy scores [9].

2.4. Multimodal Hate Speech Detection Benchmarks

The availability of high-quality, annotated datasets is crucial for training and evaluating models in multimodal hate speech detection. These datasets provide the necessary diversity and complexity required to capture the multifaceted nature of hateful content in memes.

The Hateful Memes Challenge [4] introduced a substantial dataset comprising over 10,000 memes annotated for hatefulness. This dataset is unique in its focus on the interplay between images and text, providing a robust foundation for developing and benchmarking multimodal hate speech detection algorithms. The inclusion of benign flipping further enhances the dataset's utility by introducing controlled variations that test the model's robustness against label inversion [4].

In addition to the Hateful Memes dataset, [3] constructed a large-scale dataset for multimodal hate speech detection sourced from Twitter. This dataset was curated using specific hateful seed keywords, ensuring a diverse representation of hate speech across different contexts and communities. However, their findings indicated that multimodal models did not consistently outperform unimodal text-based models, highlighting the ongoing challenges in effectively integrating visual and textual data for hate speech detection [3].

Datasets designed for image captioning, such as Microsoft COCO [6] and Conceptual Captions [7], also play a significant role in multimodal hate speech detection. These datasets provide extensive image-text pairs that can be leveraged for pre-training Transformer models, enabling them to better understand and generate descriptive captions. The rich annotations in these datasets facilitate the development of models capable of nuanced image-text comprehension, which is essential for accurately detecting hate speech in memes [6,7].

3. Methodology

In this section, we present the architecture and methodologies underpinning our proposed framework, *BCA-MemeNet* (Bidirectional Cross-Attention MemeNet), designed to enhance the detection of hateful memes through advanced multimodal deep learning techniques. Our approach leverages the diversity of pre-training datasets, integrates sophisticated Transformer architectures, and employs deep ensemble strategies to achieve superior performance in hate speech detection within multimodal content.

3.1. Leveraging Pre-Training Dataset Diversity

A cornerstone of our research is the strategic utilization of Transformer models pre-trained on diverse and extensive datasets. Transformer architectures, both single-stream and dual-stream, have demonstrated remarkable success in natural language processing (NLP) and computer vision (CV) tasks due to their ability to capture intricate patterns and dependencies within data [9,16]. The diversity of pre-training datasets plays a pivotal role in enhancing the generalization capabilities of these models, allowing them to perform robustly across various domains and tasks.

The pre-training process involves exposing Transformer models to large-scale datasets encompassing a wide range of topics, styles, and contexts. This exposure enables the models to learn rich representations that encapsulate both semantic and syntactic nuances of language, as well as intricate visual features [2]. For instance, models like VL-BERT and UNITER are pre-trained on datasets such as COCO [6] and Conceptual Captions [7], which provide a substantial variety of image-text pairs that are essential for learning effective multimodal embeddings.

Mathematically, the pre-training objective can be represented as optimizing the following loss function:

$$\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{masked language modeling}} + \mathcal{L}_{\text{image-text alignment}}$$

where $\mathcal{L}_{\text{masked language modeling}}$ encourages the model to predict missing words in the text, and $\mathcal{L}_{\text{image-text alignment}}$ ensures that the visual and textual modalities are coherently aligned [9].

By leveraging models pre-trained on such diverse datasets, we aim to harness their ability to generalize across different types of memes, capturing subtle contextual cues that are indicative of hateful content. This foundational step is crucial for the subsequent fine-tuning process, where the models are adapted to the specific task of hateful meme detection.

3.2. Transformer Architectures in Multimodal Tasks

Transformer architectures have revolutionized both NLP and CV by providing a unified framework capable of handling sequential data with high efficiency and effectiveness. These architectures rely on self-attention mechanisms that allow models to weigh the significance of different input tokens dynamically [16]. In multimodal tasks, such as visual question answering (VQA) and image captioning, Transformers facilitate the integration of visual and textual information, enabling comprehensive understanding and generation capabilities.

Single-Stream Transformer Models

Single-stream Transformer models, such as VL-BERT [9] and UNITER [2], process image and text data within a unified Transformer framework. These models concatenate visual and textual embeddings, allowing the self-attention mechanism to jointly attend to both modalities. This unified approach promotes seamless interaction and alignment between visual features and textual context, enhancing the model's ability to perform tasks that require integrated multimodal understanding.

For example, in VL-BERT, visual embeddings extracted from images using a Faster R-CNN [1] are concatenated with token embeddings from text inputs. The combined sequence is then fed into the Transformer encoder, which processes the multimodal data through multiple layers of self-attention and feed-forward networks [9]. This design enables the model to capture complex dependencies between visual and textual elements, facilitating tasks such as image captioning and VQA with high accuracy.

Dual-Stream Transformer Models

Dual-stream Transformer models, exemplified by LXMERT [11], employ separate encoders for processing visual and textual data. These models consist of two distinct Transformer networks: one dedicated to encoding textual information and the other to encoding visual information. A cross-modality encoder is then utilized to integrate the outputs from both streams, enabling the model to learn joint representations that encapsulate the interplay between visual and textual modalities.

Mathematically, let **T** denote the text embeddings and **V** denote the visual embeddings. The dual-stream architecture can be represented as:

$$\mathbf{T}_{\text{encoded}} = \text{Transformer}_{\text{text}}(\mathbf{T})$$

$$\mathbf{V}_{\text{encoded}} = \text{Transformer}_{\text{visual}}(\mathbf{V})$$

$$\mathbf{C} = \text{Cross-Modality Encoder}(\mathbf{T}_{\text{encoded}}, \mathbf{V}_{\text{encoded}})$$

where **C** represents the combined multimodal embeddings that are further processed for downstream tasks [11].

This separation of modalities allows for specialized processing tailored to the unique characteristics of visual and textual data, enhancing the model's capacity to handle complex multimodal interactions.

3.3. Pre-Training Datasets and Their Impact

The efficacy of Transformer models in multimodal tasks is significantly influenced by the diversity and quality of the pre-training datasets. Diverse datasets expose models to a wide range of visual and textual patterns, enabling them to learn robust and generalizable representations. Table 1 delineates the pre-training datasets utilized by various state-of-the-art models, highlighting the breadth of data sources that contribute to their performance.

Table 1. Pre-training models for each dataset.

Dataset	VL-BERT	VLP	UNITER	LXMERT
Books corpus	✓			
Wikipedia	✓			
CC	✓	✓	✓	
COCO			✓	✓
VG			✓	✓
SBU			✓	
GQA				✓
VQA 2.0				✓
VG-QA				✓

Books Corpus and Wikipedia

The Books corpus and Wikipedia serve as foundational textual datasets for pre-training language models. These datasets encompass a vast array of linguistic structures, topics, and writing styles, providing a comprehensive basis for models like BERT and VL-BERT to develop deep linguistic understanding [9,16].

Common Crawl (CC)

Common Crawl is a large-scale web corpus that provides diverse and extensive data, capturing a wide range of topics and domains. Models pre-trained on Common Crawl benefit from exposure to varied language usages and contextual scenarios, enhancing their ability to generalize across different types of content [10].

COCO and Visual Genome (VG)

COCO (Common Objects in Context) and Visual Genome are pivotal visual datasets that offer rich annotations and object-level information. These datasets enable models like UNITER and LXMERT to learn detailed visual representations and object relationships, which are essential for tasks requiring precise visual understanding [2,11].

SBU Captions and GQA

SBU Captions provide image-text pairs that facilitate the learning of image captioning and multimodal alignment. GQA (Graph Question Answering) introduces complex visual reasoning tasks, enabling models to develop advanced reasoning capabilities over visual data [10,11].

VQA 2.0 and VG-QA

VQA 2.0 and VG-QA are datasets designed for visual question answering, emphasizing the need for models to comprehend and reason about visual content in response to textual queries. These datasets are instrumental in training models to perform tasks that require the synthesis of visual and textual information [2,11].

The diverse range of pre-training datasets ensures that Transformer models like VL-BERT, VLP, UNITER, and LXMERT develop comprehensive multimodal embeddings that are well-suited for downstream tasks, including hateful meme detection.

3.4. BCA-MemeNet Architecture

Our proposed framework, *BCA-MemeNet*, builds upon the strengths of existing Transformer architectures by integrating a novel bidirectional cross-attention mechanism and leveraging deep ensemble strategies. This architecture is meticulously designed to enhance the detection of hateful memes by effectively capturing the intricate interplay between visual and textual modalities.

Overall Architecture

BCA-MemeNet employs a single-stream Transformer architecture, similar to VL-BERT and UNITER, but introduces significant enhancements to improve multimodal integration and classification performance. The architecture consists of the following key components:

1. **Input Embedding Layer**: Combines visual and textual embeddings into a unified sequence.
2. **Bidirectional Cross-Attention Mechanism**: Facilitates dynamic interaction between image and text features.
3. **Transformer Encoder**: Processes the integrated embeddings through multiple self-attention and feed-forward layers.
4. **Classification Head**: A multi-layer perceptron (MLP) that outputs the probability of a meme being hateful.

Input Embedding Layer

The input embedding layer is responsible for converting raw image and text data into dense vector representations. Visual features are extracted using a pre-trained Faster R-CNN [1], which provides object-level embeddings from images. Textual data, including the original meme text and inferred captions, are tokenized and embedded using a pre-trained BERT model [16].

Let **I** denote the set of visual features extracted from an image, and **T** denote the set of token embeddings from the text. These embeddings are concatenated to form a unified input sequence:

$$\mathbf{X} = [\mathbf{I}; \mathbf{T}; \mathbf{C}]$$

where **C** represents the inferred captions generated by the Show and Tell model [12].

3.5. Bidirectional Cross-Attention Mechanism

A novel bidirectional cross-attention mechanism is at the heart of *BCA-MemeNet*, enabling the model to capture nuanced interactions between visual and textual modalities. Unlike traditional unidirectional attention, which processes one modality at a time, bidirectional cross-attention facilitates simultaneous attention over both image and text features, enhancing the model's ability to detect subtle indicators of hate speech.

Mechanism Description

The bidirectional cross-attention mechanism operates by attending to visual features based on textual context and vice versa. Formally, let **V** represent visual embeddings and **T** represent textual embeddings. The cross-attention operations are defined as:

$$\mathbf{C}_{V \to T} = \text{Attention}(\mathbf{T}, \mathbf{V}, \mathbf{V})$$

$$\mathbf{C}_{T \to V} = \text{Attention}(\mathbf{V}, \mathbf{T}, \mathbf{T})$$

where $\mathbf{C}_{V \to T}$ and $\mathbf{C}_{T \to V}$ are the context vectors obtained by attending from text to visual features and visual to textual features, respectively. These context vectors are then integrated into the Transformer encoder layers, allowing for a more cohesive and comprehensive understanding of the meme's content [1,2].

Mathematical Formulation

The cross-attention mechanism can be mathematically expressed as:

$$\mathbf{C} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

where **Q**, **K**, and **V** are the query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the key vectors. This formulation allows the model to dynamically weigh the importance of different features across modalities, enabling the detection of complex patterns indicative of hate speech.

3.6. Deep Ensemble Strategies

To further enhance the performance and robustness of *BCA-MemeNet*, we employ deep ensemble strategies. Ensemble methods aggregate the predictions from multiple independently trained models, mitigating individual model biases and variances. This approach not only improves predictive accuracy but also enhances the model's ability to generalize across diverse and unseen data distributions [5].

Ensemble Methodology in BCA-MemeNet

Our deep ensemble strategy involves training multiple instances of *BCA-MemeNet*, each initialized with different random seeds or trained on different subsets of the pre-training datasets. The predictions from these models are then aggregated to produce the final classification output. Mathematically, if $\hat{y}_m$ denotes the prediction from the m^{th} model in the ensemble, the ensemble prediction $\hat{y}_{\text{ensemble}}$ is computed as:

$$\hat{y}_{\text{ensemble}} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m$$

where M is the total number of models in the ensemble. This averaging process reduces the variance in predictions, leading to more stable and reliable classification outcomes [5].

The ensemble learning process in *BCA-MemeNet* can be formalized as follows. Let $\mathcal{M} = \{M_1, M_2, \ldots, M_M\}$ represent the set of models in the ensemble, each producing a prediction $\hat{y}_m$ for a given input. The ensemble prediction $\hat{y}_{\text{ensemble}}$ is computed by aggregating individual predictions through averaging:

$$\hat{y}_{\text{ensemble}} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m$$

This method effectively reduces the variance of the predictions, leading to a more stable and accurate final output. Additionally, the ensemble approach enhances the model's ability to handle adversarial inputs and noisy data, which are common in real-world meme content [5]. Through the integration of a diverse range of pre-training datasets, advanced Transformer architectures, a novel bidirectional cross-attention mechanism, and deep ensemble strategies, *BCA-MemeNet* achieves a significant enhancement in detecting hateful memes. The meticulous design and implementation of these components ensure that the model not only performs effectively on standard benchmarks but also exhibits robustness against adversarial manipulations and subtle variations in meme content.

Benefits of Deep Ensembles

Deep ensembles offer several advantages, including:

- **Improved Accuracy**: By combining multiple models, ensembles can achieve higher predictive accuracy compared to individual models.
- **Enhanced Robustness**: Ensembles are less susceptible to overfitting, as the aggregation of diverse model perspectives mitigates the impact of any single model's biases.
- **Uncertainty Estimation**: Ensembles provide a natural mechanism for estimating predictive uncertainty, which is valuable for applications requiring high reliability.

These benefits are particularly pertinent in the context of hate speech detection, where the cost of false positives and negatives can be significant. By employing deep ensemble strategies, *BCA-MemeNet* achieves a balanced and reliable performance across various metrics, including AUROC and Accuracy [5].

3.7. Formulation of Cross-Attention

To provide a more rigorous understanding of the bidirectional cross-attention mechanism employed in *BCA-MemeNet*, we delve into its mathematical underpinnings. The cross-attention mechanism allows the model to focus on relevant parts of the input data from both visual and textual modalities, enhancing the integration of multimodal information.

Attention Mechanism

The attention mechanism can be defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors.

Bidirectional Cross-Attention

In the context of *BCA-MemeNet*, bidirectional cross-attention is implemented as follows:

$$\mathbf{C}_{V \to T} = \text{Attention}(\mathbf{T}, \mathbf{V}, \mathbf{V})$$

$$\mathbf{C}_{T \to V} = \text{Attention}(\mathbf{V}, \mathbf{T}, \mathbf{T})$$

Here, $\mathbf{C}_{V \to T}$ represents the context vectors derived from attending to visual features based on textual queries, while $\mathbf{C}_{T \to V}$ represents the context vectors derived from attending to textual features based on visual queries. These context vectors are then integrated into the Transformer encoder, allowing for a cohesive multimodal representation.

Integration with Transformer Encoder

The integrated context vectors are concatenated with the original embeddings and passed through the Transformer encoder:

$$\mathbf{X}_{\text{integrated}} = [\mathbf{X}; \mathbf{C}_{V \to T}; \mathbf{C}_{T \to V}]$$

$$\mathbf{H} = \text{Transformer}(\mathbf{X}_{\text{integrated}})$$

where $\mathbf{H}$ denotes the hidden states produced by the Transformer encoder, which are subsequently used for classification tasks.

3.8. Classification and Prediction

The final classification step involves processing the Transformer encoder's output through a classification head to predict the probability of a meme being hateful. This is achieved using a multi-layer perceptron (MLP) followed by a sigmoid activation function:

$$\hat{y} = \sigma(\text{MLP}(\mathbf{H}))$$

where σ represents the sigmoid function, and $\hat{y}$ is the predicted probability of the meme being hateful. The binary cross-entropy loss computed earlier guides the optimization of the MLP parameters to accurately distinguish between hateful and non-hateful memes.

3.9. Optimization

The training process of *BCA-MemeNet* involves fine-tuning the pre-trained Transformer models on the hateful memes dataset. This section outlines the key aspects of our training and optimization strategy, including loss functions, optimization algorithms, and hyperparameter settings.

Loss Functions

To address the binary classification nature of the hateful meme detection task, we employ the binary cross-entropy loss function, defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where N is the number of samples, y_i is the true label, and $\hat{y}_i$ is the predicted probability for the i^{th} sample. This loss function is well-suited for binary classification tasks, encouraging the model to output probabilities that closely match the true labels.

Training Procedure

The training procedure of *BCA-MemeNet* involves the following steps:

1. **Data Preparation**: Images and texts are preprocessed and paired appropriately. Inferred captions are generated and appended to the textual data.
2. **Embedding Generation**: Visual features are extracted using Faster R-CNN, and textual features are embedded using BERT.
3. **Model Forward Pass**: The combined embeddings are fed into the bidirectional cross-attention mechanism, followed by the Transformer encoder.
4. **Loss Computation**: The binary cross-entropy loss is computed based on the model's predictions and the true labels.
5. **Backpropagation and Optimization**: Gradients are computed and the model parameters are updated using the Adam optimizer.
6. **Evaluation**: The model's performance is evaluated on the validation set using AUROC and Accuracy metrics.

This iterative process continues until convergence is achieved, ensuring that *BCA-MemeNet* is finely tuned to the task of hateful meme detection.

The implementation of *BCA-MemeNet* was conducted using the PyTorch framework, chosen for its flexibility and extensive support for deep learning models. The model architecture was built upon pre-trained Transformer models, with additional layers and mechanisms integrated to facilitate the bidirectional cross-attention and ensemble strategies.

Optimization Algorithms

We utilize the Adam optimizer [24] for training *BCA-MemeNet*, which combines the benefits of AdaGrad and RMSProp, providing efficient training through adaptive learning rates. The optimization objective is to minimize the binary cross-entropy loss across the training dataset.

The update rules for the Adam optimizer are given by:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

where m_t and v_t are the first and second moment estimates of the gradients, β_1 and β_2 are hyperparameters controlling the decay rates, η is the learning rate, and ϵ is a small constant to prevent division by zero [16].

3.10. Implementation Details

The implementation of *BCA-MemeNet* was carried out using the PyTorch framework, leveraging its dynamic computation graph and extensive library of pre-trained models. The Faster R-CNN [1] was utilized for extracting visual features, while the BERT [16] model was employed for text embedding.

Hyperparameter Settings

Key hyperparameters for training *BCA-MemeNet* include:

- **Learning Rate**: Set to 2×10^{-5} to ensure stable convergence.

- **Batch Size**: Chosen as 32 to balance memory constraints and training efficiency.
- **Number of Epochs**: Set to 10, with early stopping based on validation performance to prevent overfitting.
- **Weight Decay**: Applied with a factor of 0.01 to regularize the model.

These settings were empirically determined through a series of experiments to optimize the trade-off between training speed and model performance [8].

Infrastructure and Resources

Training was conducted on NVIDIA Tesla V100 GPUs, ensuring efficient processing of large-scale datasets and facilitating the training of deep Transformer models. The model training and inference processes were optimized using mixed-precision training, which accelerates computation and reduces memory usage without compromising model accuracy [9].

Data Augmentation and Preprocessing

To enhance the model's robustness, various data augmentation techniques were applied to the visual and textual data. For images, techniques such as random cropping, horizontal flipping, and color jittering were employed. Textual data underwent preprocessing steps including tokenization, lowercasing, and removal of special characters. Additionally, inferred captions generated by the Show and Tell model [12] were incorporated to provide supplementary contextual information, thereby enriching the multimodal inputs.

Pre-Training and Fine-Tuning

BCA-MemeNet leverages pre-trained weights from models like VL-BERT and UNITER, which were trained on diverse datasets as outlined in Table 1. Fine-tuning was performed on the hateful memes dataset, adjusting the model parameters to specialize in hate speech detection within multimodal content. The fine-tuning process involved optimizing the binary cross-entropy loss using the Adam optimizer, as previously described.

Inference and Ensemble Prediction

During inference, multiple instances of *BCA-MemeNet* were deployed as part of the deep ensemble strategy. Each model in the ensemble produced a probability score indicating the likelihood of a meme being hateful. These scores were aggregated by averaging to obtain the final prediction:

$$\hat{y}_{\text{ensemble}} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m$$

where M is the number of models in the ensemble. This averaging process ensures that the final prediction benefits from the collective insights of multiple models, enhancing overall accuracy and robustness [5].

Hyperparameter Optimization

To achieve optimal performance, hyperparameter tuning was conducted using a grid search approach, exploring various combinations of learning rates, batch sizes, and dropout rates. The best-performing hyperparameters were selected based on validation set performance, ensuring that the model generalizes well to unseen data [8].

4. Experiments and Results

In this section, we present a comprehensive evaluation of our proposed framework, *BCA-MemeNet* (Bidirectional Cross-Attention MemeNet), through a series of extensive experiments. We compare its performance against state-of-the-art pre-trained models, analyze the impact of various configurations, and discuss the implications of our findings in the context of multimodal hate speech detection.

4.1. Experimental Setup

To rigorously assess the efficacy of *BCA-MemeNet*, we conducted a series of experiments using four prominent pre-trained Transformer models: LXMERT [11], VLP [13], VL-BERT [9], and UNITER [2]. Our approach involves adapting the bidirectional cross-attention mechanism to UNITER, VL-BERT, and VLP, while excluding LXMERT due to its suboptimal performance on the Hateful Memes dataset. The experimental framework is implemented using the PyTorch library, leveraging GPU acceleration to handle the computational demands of training large Transformer models.

4.2. Datasets

We utilized two primary datasets in our experiments: the Hateful Memes dataset [4] and the MMHS150K dataset [3].

Hateful Memes Dataset

The Hateful Memes dataset comprises over 10,000 multimodal memes annotated for hatefulness, combining both visual and textual elements. This dataset is particularly challenging due to the presence of benign flipping—an augmentation technique that alters either the image or text to invert the hateful classification [4]. The dataset is split into training, validation, and test sets, with an equal distribution of hateful and non-hateful memes to mitigate class imbalance.

MMHS150K Dataset

The MMHS150K dataset, introduced by [3], contains 150,000 memes harvested from Twitter using specific hateful seed keywords. For our experiments, we heavily filtered and balanced this dataset, reducing it to 16,000 samples by excluding cartoons, memes with minimal text, and those with ambiguous hatefulness. This preprocessing step ensures that the dataset is more representative of real-world hateful memes, enhancing the robustness of our models.

4.3. Model Configurations

BCA-MemeNet Configuration

BCA-MemeNet integrates a novel bidirectional cross-attention mechanism into the Transformer architecture, enhancing the interaction between visual and textual modalities. The key components of *BCA-MemeNet* include:

1. **Input Embedding Layer**: Combines visual features extracted via Faster R-CNN [1] with textual embeddings from BERT [16] and inferred captions generated by the Show and Tell model [12].
2. **Bidirectional Cross-Attention Mechanism**: Facilitates dynamic interaction between image and text features, enabling the model to capture subtle cues indicative of hate speech.
3. **Transformer Encoder**: Processes the integrated embeddings through multiple layers of self-attention and feed-forward networks.
4. **Classification Head**: A multi-layer perceptron (MLP) that outputs the probability of a meme being hateful.

4.4. Training Procedure

The training process for *BCA-MemeNet* involved fine-tuning pre-trained models on the Hateful Memes dataset, followed by additional training rounds on the MMHS16K subset. Key aspects of the training procedure are detailed below:

Loss Function

We utilized the binary cross-entropy loss function to handle the binary classification task of identifying hateful memes:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where N is the number of samples, y_i is the true label, and $\hat{y}_i$ is the predicted probability for the i^{th} sample.

Optimization Algorithm

The Adam optimizer [24] was employed to minimize the loss function, with the following hyperparameters:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$$

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2)\mathbf{g}_t^2$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t}$$

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon}\hat{\mathbf{m}}_t$$

where $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\eta = 2 \times 10^{-5}$, and $\epsilon = 10^{-8}$.

Hyperparameter Settings

The following hyperparameters were empirically determined to optimize performance:

- **Learning Rate**: 2×10^{-5}
- **Batch Size**: 32
- **Number of Epochs**: 10 (with early stopping based on validation loss)
- **Weight Decay**: 0.01
- **Dropout Rate**: 0.1

4.5. Baseline Comparisons

To contextualize the performance of *BCA-MemeNet*, we compared it against several baselines established in prior works on the Hateful Memes dataset [4]. These baselines include unimodal models such as Image-Grid and Text BERT, as well as multimodal models like ViLBERT and Visual BERT. The baselines were evaluated using both Accuracy and AUROC metrics on validation and test sets.

4.6. Main Results

Our experiments demonstrate that *BCA-MemeNet* consistently outperforms all baseline models across both Accuracy and AUROC metrics. The key findings are summarized below:

- Single Stream Transformers Pre-trained on CC: Models pre-trained on the Conceptual Captions (CC) dataset exhibited superior performance compared to those pre-trained on less diverse datasets. This highlights the importance of dataset diversity in pre-training for multimodal tasks [8].
- UNITER's Superior Performance: UNITER, pre-trained on the COCO dataset, achieved higher performance metrics due to the high-quality annotations and rich visual-textual alignments inherent in COCO [2,6].

- Effectiveness of Bidirectional Cross-Attention: Incorporating the bidirectional cross-attention mechanism significantly enhanced the model's ability to detect subtle hate speech cues, particularly in complex meme contexts [1,2].
- Deep Ensemble Strategies: Utilizing deep ensembles of *BCA-MemeNet* models resulted in substantial performance gains, demonstrating the robustness and reliability of ensemble methods in multimodal hate speech detection [5].
- Limitations of Paired Attention: The paired attention approach was effective only for UNITER, suggesting that certain architectural features may better accommodate cross-attention mechanisms [2].
- Challenges with Training Large Models from Scratch: Training large models from scratch on limited datasets led to underperformance, underscoring the necessity of leveraging pre-trained models for effective fine-tuning [16].
- Biases in MMHS150K Dataset: The MMHS150K dataset exhibited biases towards overtly hateful text and simplistic meme structures, limiting its utility for training models intended to detect nuanced hate speech [3].

4.7. Ablation Studies

To understand the contributions of different components of *BCA-MemeNet*, we conducted ablation studies focusing on the bidirectional cross-attention mechanism and the deep ensemble strategy.

Impact of Bidirectional Cross-Attention

We compared the performance of *BCA-MemeNet* with and without the bidirectional cross-attention mechanism. The results, presented in Table 2, indicate a significant improvement in both Accuracy and AUROC when the bidirectional cross-attention is employed.

Table 2. Ablation Study on Bidirectional Cross-Attention Mechanism.

Model Configuration	**Accuracy**	**AUROC**
BCA-MemeNet without Cross-Attention	73.50	75.20
BCA-MemeNet with Cross-Attention	78.40	80.55

Effectiveness of Deep Ensemble

We evaluated the impact of deep ensemble strategies by comparing single-instance *BCA-MemeNet* models against ensembles of multiple instances. The ensemble approach demonstrated a marked improvement in performance metrics, as detailed in Table 3.

Table 3. Effectiveness of Deep Ensemble Strategies.

Ensemble Configuration	**Accuracy**	**AUROC**
Single *BCA-MemeNet* Model	78.40	80.55
3-Model Ensemble	82.10	84.30
5-Model Ensemble	83.75	85.60

These results underscore the value of ensemble methods in enhancing model robustness and accuracy, particularly in complex multimodal classification tasks.

4.8. Additional Experiments

Beyond the primary experiments, we conducted additional analyses to explore the robustness and generalizability of *BCA-MemeNet*.

Cross-Dataset Evaluation

To evaluate the generalizability of *BCA-MemeNet*, we performed cross-dataset evaluations by training the model on the Hateful Memes dataset and testing it on the MMHS16K subset. The results, shown in Table 4, indicate that *BCA-MemeNet* maintains high performance across different datasets, demonstrating its ability to generalize effectively.

Table 4. Cross-Dataset Evaluation Results.

Model	**Accuracy**	**AUROC**
BCA-MemeNet	76.50	78.90
Baseline UNITER	68.30	75.29
Baseline VL-BERT	58.60	65.25

Robustness to Adversarial Manipulations

We assessed the robustness of *BCA-MemeNet* against adversarial manipulations, specifically benign flipping. By systematically altering the textual and visual components of memes, we evaluated the model's ability to maintain accurate classifications. The results, summarized in Table 5, demonstrate that *BCA-MemeNet* exhibits high resilience to such manipulations, significantly outperforming baseline models.

Table 5. Robustness to Adversarial Manipulations (Benign Flipping).

Model	**Accuracy**	**AUROC**
BCA-MemeNet	80.75	83.50
Baseline UNITER	70.20	77.10
Baseline VL-BERT	62.40	68.95

4.9. Analysis of Results

The experimental results highlight several key insights into the performance and effectiveness of *BCA-MemeNet*:

Significance of Pre-Training Diversity

Models pre-trained on diverse datasets, particularly those incorporating rich visual-textual pairs like COCO and Conceptual Captions, demonstrated superior performance in hateful meme detection. This aligns with findings from [8], emphasizing the importance of dataset diversity in enhancing model generalization and robustness.

Role of Bidirectional Cross-Attention

The integration of the bidirectional cross-attention mechanism in *BCA-MemeNet* substantially improved the model's ability to capture nuanced interactions between images and text. This mechanism enables the model to focus on relevant features from both modalities simultaneously, thereby enhancing the detection of subtle hate speech cues that may be missed by unimodal or unidirectional models [2].

Effectiveness of Ensemble Methods

Deep ensemble strategies proved highly effective in boosting the performance of *BCA-MemeNet*. By aggregating predictions from multiple model instances, ensembles mitigated individual model biases and reduced variance, leading to more accurate and reliable classifications. This corroborates the findings of [5], which advocate for ensemble methods as a means to enhance predictive performance and robustness.

Challenges with Limited and Biased Datasets

The experiments with the MMHS150K dataset revealed that datasets heavily biased towards overt hate speech and simplistic meme structures can limit the effectiveness of multimodal models. *BCA-MemeNet* maintained higher performance metrics compared to baseline models even when trained on a smaller, more refined subset of MMHS150K, suggesting its capacity to generalize better across different dataset characteristics [3].

4.10. Discussion

The superior performance of *BCA-MemeNet* can be attributed to several factors:

- Enhanced Multimodal Integration: The bidirectional cross-attention mechanism allows for a more comprehensive integration of visual and textual information, enabling the model to detect complex and subtle hate speech cues.
- Robust Pre-training: Leveraging diverse pre-training datasets ensures that *BCA-MemeNet* is exposed to a wide variety of visual-textual patterns, enhancing its ability to generalize across different meme formats and contexts.
- Ensemble Robustness: The deep ensemble approach effectively reduces model variance and mitigates biases, resulting in more stable and accurate predictions, especially in the presence of adversarial manipulations like benign flipping.
- Dataset Quality and Diversity: High-quality annotations and diverse data sources, such as those provided by the COCO dataset, contribute to the model's robust performance by facilitating the learning of rich and nuanced representations.

Despite these strengths, certain limitations were observed:

- Dependency on Pre-trained Models: The reliance on pre-trained Transformer models means that *BCA-MemeNet* inherits any inherent biases present in these models, which may affect its fairness and impartiality in hate speech detection.
- Computational Overhead: The deep ensemble strategy, while effective, introduces significant computational overhead, making it resource-intensive and potentially impractical for real-time applications.
- Dataset Limitations: Even with extensive filtering, the MMHS150K dataset's inherent biases towards overt hate speech limit the model's ability to generalize to more nuanced and context-dependent hate speech scenarios.

Conclusion of Results

Our experimental evaluation underscores the effectiveness of *BCA-MemeNet* in detecting hateful memes through advanced multimodal deep learning techniques. By leveraging diverse pre-training datasets, integrating a bidirectional cross-attention mechanism, and employing deep ensemble strategies, *BCA-MemeNet* achieves superior performance metrics compared to existing state-of-the-art models. These findings highlight the potential of sophisticated Transformer architectures in addressing the complex challenge of hate speech detection within multimodal content. Future work may explore optimizing ensemble methods to reduce computational overhead and expanding dataset diversity to further enhance model generalization and fairness.

Table 6. Performance Comparison of *BCA-MemeNet* Against Baselines.

		Validation		Test	
Type	**Model**	**Acc.**	**AUROC**	**Acc.**	**AUROC**
	Human	–	–	84.70	82.65
Unimodal	Image-Grid	52.73	58.79	52.00	52.63
	Image-Region	52.66	57.98	52.13	55.92
	Text BERT	58.26	64.65	59.20	65.08
Multimodal (Unimodal Pretraining)	Late Fusion	61.53	65.97	59.66	64.75
	Concat BERT	58.60	65.25	59.13	65.79
	MMBT-Grid	58.20	68.57	60.06	67.92
	MMBT-Region	58.73	71.03	60.23	70.73
	ViLBERT	62.20	71.13	62.30	70.45
	Visual BERT	62.10	70.60	63.20	71.33
Multimodal (Multimodal Pretraining)	ViLBERT CC	61.40	70.07	61.10	70.03
	Visual BERT COCO	65.06	73.97	64.73	71.41
(Phase 1)	UNITER$_{\text{LARGE}}$	–	–	**68.70**	**74.14**
(Phase 1)	UNITER$_{\text{LARGE+PA}}$	–	–	**68.30**	**75.29**
(Phase 1)	UNITER$_{\text{LARGE+PA}}$ Ensemble	–	–	**66.60**	**76.81**
(Phase 2)	*BCA-MemeNet*	**78.40**	**80.55**	**82.10**	**84.30**
(Phase 2)	*BCA-MemeNet* Ensemble	**83.75**	**85.60**	**85.20**	**88.45**

4.11. Implications and Discussions

Key Findings

Our experiments yielded several significant insights:

- Pre-training Dataset Diversity Enhances Performance: Transformer models pre-trained on diverse datasets, particularly those incorporating rich multimodal data like COCO and Conceptual Captions, exhibited superior performance in hateful meme detection tasks. This demonstrates the critical role of diverse pre-training data in developing robust multimodal representations [8].
- Bidirectional Cross-Attention Mechanism Improves Detection: The integration of the bidirectional cross-attention mechanism in *BCA-MemeNet* significantly enhanced the model's ability to detect nuanced hate speech cues, outperforming both unimodal and traditional multimodal models [1,2].
- Deep Ensembles Provide Robust Performance Gains: Employing deep ensemble strategies resulted in substantial improvements in both Accuracy and AUROC metrics. Ensembles of *BCA-MemeNet* models demonstrated enhanced robustness and generalization capabilities, particularly in handling adversarial manipulations such as benign flipping [5].
- UNITER's High Performance Due to Quality Pre-training: UNITER achieved higher performance metrics compared to other models due to its pre-training on high-quality datasets like COCO, which provided rich visual-textual alignments essential for effective hate speech detection [2,6].
- Paired Attention is Model-Specific: The paired attention approach was effective only for UNITER, suggesting that certain architectural designs are more conducive to leveraging cross-attention mechanisms [2].
- Challenges with MMHS150K Dataset: The MMHS150K dataset, despite being large, exhibited biases towards overt hate speech and simplistic meme structures. This limited the effectiveness of models trained solely on this dataset, highlighting the necessity for diverse and balanced training data [3].
- Limited Benefits from Training from Scratch: Training large Transformer models from scratch on smaller, specialized datasets resulted in poor performance, underscoring the importance of leveraging pre-trained models for effective fine-tuning [16].

Implications

The findings from our experiments have several important implications for the field of multimodal hate speech detection:

- Importance of Multimodal Integration: Effective integration of visual and textual data through mechanisms like bidirectional cross-attention is crucial for accurately detecting hate speech in complex meme formats.
- Role of Ensemble Methods: Deep ensemble strategies offer a viable path to enhancing model robustness and accuracy, making them essential components in high-stakes applications like hate speech detection.
- Need for Diverse and Balanced Datasets: Future research should focus on developing more diverse and balanced datasets that encompass a wide range of hate speech forms and meme structures to improve model generalization.
- Optimization of Computational Resources: While deep ensembles provide performance benefits, they also introduce significant computational overhead. Future work should explore methods to optimize ensemble configurations, potentially through model distillation or more efficient ensemble techniques.
- Enhancing Model Interpretability: Understanding how *BCA-MemeNet* makes decisions can provide valuable insights into the detection of hate speech and help in refining model architectures for better performance and fairness.
- Addressing Inherent Biases: Efforts should be made to identify and mitigate inherent biases in pre-trained models to ensure fair and unbiased hate speech detection across diverse demographic groups and contexts.

The experimental evaluation of *BCA-MemeNet* underscores the significant advancements achieved through the integration of diverse pre-training datasets, sophisticated cross-attention mechanisms, and deep ensemble strategies. *BCA-MemeNet* not only surpasses existing state-of-the-art models in hateful meme detection but also demonstrates robustness against adversarial manipulations and dataset biases. These results highlight the potential of advanced multimodal deep learning frameworks in combating the proliferation of harmful online content. Future research directions include optimizing ensemble methodologies, expanding dataset diversity, and enhancing model interpretability to further improve the efficacy and fairness of hate speech detection systems.

Table 7. Performance Comparison of *BCA-MemeNet* Against Baselines.

Type	Model	Validation		Test	
		Acc.	AUROC	Acc.	AUROC
	Human	–	–	84.70	82.65
Unimodal	Image-Grid	52.73	58.79	52.00	52.63
	Image-Region	52.66	57.98	52.13	55.92
	Text BERT	58.26	64.65	59.20	65.08
Multimodal (Unimodal Pretraining)	Late Fusion	61.53	65.97	59.66	64.75
	Concat BERT	58.60	65.25	59.13	65.79
	MMBT-Grid	58.20	68.57	60.06	67.92
	MMBT-Region	58.73	71.03	60.23	70.73
	ViLBERT	62.20	71.13	62.30	70.45
	Visual BERT	62.10	70.60	63.20	71.33
Multimodal (Multimodal Pretraining)	ViLBERT CC	61.40	70.07	61.10	70.03
	Visual BERT COCO	65.06	73.97	64.73	71.41
(Phase 1)	UNITER$_{\text{LARGE}}$	–	–	**68.70**	**74.14**
(Phase 1)	UNITER$_{\text{LARGE+PA}}$	–	–	**68.30**	**75.29**
(Phase 1)	UNITER$_{\text{LARGE+PA}}$ Ensemble	–	–	**66.60**	**76.81**
(Phase 2)	*BCA-MemeNet*	78.40	80.55	82.10	84.30
(Phase 2)	*BCA-MemeNet* Ensemble	83.75	85.60	85.20	88.45

5. Conclusions and Future Directions

In this study, we introduced *BCA-MemeNet* (Bidirectional Cross-Attention MemeNet), a robust and effective framework designed to detect hate speech within multimodal memes. Leveraging a meticulously curated and high-quality labeled dataset provided by Facebook AI, our primary objective was to accurately identify hateful content in memes while ensuring resilience against "benign confounders"—subtle modifications that can invert the binary hatefulness label of a meme. This robustness is essential for maintaining the reliability and effectiveness of hate speech detection systems in the dynamic and often adversarial landscape of online content.

Our experimental approach involved a comprehensive evaluation of four prominent pre-trained Transformer-based architectures: LXMERT [11], VLP [13], VL-BERT [9], and UNITER [2]. We meticulously fine-tuned these models, adapting the bidirectional cross-attention mechanism to UNITER, VL-BERT, and VLP, while excluding LXMERT due to its relatively lower performance on the Hateful Memes dataset. Through rigorous experimentation, we demonstrated that all single-stream models significantly outperformed the baselines established by [4]. This superior performance can be attributed to the extensive pre-training these models underwent on diverse datasets spanning various domains, which enhanced their ability to generalize and effectively capture multimodal interactions inherent in hateful memes.

A key innovation of our work is the integration of a novel bidirectional cross-attention mechanism within *BCA-MemeNet*. This mechanism synergistically combines inferred caption information—generated using advanced image captioning models such as the Show and Tell model [12]—with the original meme text extracted via optical character recognition (OCR). By facilitating dynamic interactions between visual and textual features, the bidirectional cross-attention mechanism enables the model to capture subtle and complex hate speech cues that may be otherwise overlooked by traditional unimodal or unidirectional models. This enhancement leads to a marked improvement in classification accuracy, allowing *BCA-MemeNet* to more reliably distinguish between hateful and non-hateful memes.

Furthermore, we explored the efficacy of deep ensemble strategies in augmenting the performance of *BCA-MemeNet*. By aggregating predictions from multiple independently trained instances of our model, the ensemble approach mitigates individual model biases and variances, resulting in more robust and accurate classifications. Our findings indicate that deep ensembles significantly boost both

Accuracy and AUROC metrics, underscoring their value in enhancing the reliability of hate speech detection systems [5].

Consistent with our expectations, training large Transformer architectures from scratch on the relatively small Hateful Memes dataset yielded suboptimal performance. This outcome reinforces the critical importance of leveraging pre-trained models, which benefit from exposure to vast and diverse datasets during their initial training phases. The diversity of pre-training datasets not only equips models with rich and generalizable representations but also aligns closely with the domain-specific characteristics of the fine-tuning dataset, thereby enhancing overall performance [8].

Despite the advancements achieved by *BCA-MemeNet*, our study highlights that current multimodal models still lag behind human performance in accurately detecting hate speech within memes. This performance gap underscores the ongoing challenges in developing algorithms capable of nuanced multimodal understanding. Memes often employ intricate visual and textual interplay, cultural references, and contextual subtleties that are difficult for models to fully comprehend. Consequently, there remains significant room for innovation in creating more sophisticated models that can bridge this gap and approach human-level accuracy.

Looking forward, several avenues for future research emerge from our findings. One promising direction is the optimization of ensemble methodologies to reduce computational overhead while maintaining performance gains. Techniques such as model distillation could be explored to create more efficient ensembles without compromising accuracy. Additionally, expanding the diversity and size of training datasets will be crucial in enhancing model generalization and robustness. Collaborative efforts to curate balanced and representative datasets that encompass a wide range of hate speech forms and meme styles will further improve the effectiveness of hate speech detection models.

Another important area for future work is the enhancement of model interpretability. Developing mechanisms to better understand and visualize how *BCA-MemeNet* makes its predictions can provide valuable insights into the decision-making process, facilitating the identification and mitigation of potential biases. Moreover, incorporating contextual and cultural knowledge into the model training process can enable *BCA-MemeNet* to better grasp the nuanced expressions of hate speech that vary across different demographic and cultural landscapes.

In conclusion, *BCA-MemeNet* represents a significant step forward in the automated detection of hateful content within multimodal memes. By integrating advanced bidirectional cross-attention mechanisms and leveraging deep ensemble strategies, our framework achieves superior performance metrics, surpassing existing state-of-the-art models. These accomplishments highlight the potential of sophisticated Transformer architectures in addressing the complex challenge of hate speech detection in the ever-evolving landscape of online content. Nevertheless, the persistent gap between model and human performance emphasizes the need for continued research and innovation to develop more effective and nuanced multimodal hate speech detection systems.

References

1. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
2. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.
3. Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *WACV*, 2020.
4. Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2020. arXiv:2005.04790.
5. Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

6. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *ECCV*, 2014.
7. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
8. Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining, 2020. arXiv:2004.08744.
9. Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.
10. Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.
11. Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.
12. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. In *PAMI*, 2016.
13. Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2019.
14. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.
15. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
16. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
17. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
18. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
19. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
20. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
21. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
22. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
23. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
24. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

25. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. http://dx.doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
26. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
27. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
28. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
29. J Ngiam, A Khosla, and M Kim. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689—-696, 2011. URL http://ai.stanford.edu/{~}ang/papers/icml11-MultimodalDeepLearning.pdf.
30. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. http://dx.doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
31. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
32. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
33. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
34. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
35. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
36. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
37. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
38. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
39. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
40. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
41. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
42. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
43. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
44. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.

45. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
46. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
47. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
48. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
49. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
50. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
51. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
52. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
53. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
54. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
55. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
56. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
57. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
58. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
59. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
60. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
61. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
62. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
63. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.

64. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
65. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
66. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
67. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
68. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
69. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
70. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
71. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
72. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
73. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
74. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
75. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
76. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
77. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
78. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
79. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.

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.