Article

Not peer-reviewed version

An Integrated Multimodal System for Identifying Offensive Memes

Ella Cohen , Jaden Levy * , Jannat Roy

Posted Date: 15 November 2024

doi: 10.20944/preprints202411.1131.v1

Keywords: Multimodal Detection; Hate Speech; Memes; Ensemble Learning

Article

An Integrated Multimodal System for Identifying Offensive Memes

Ella Cohen, Jaden Levy * and Jannat Roy

Brandeis University
* Correspondence: jlevy@brandeis.edu

Abstract: The manifestation of online hate speech has increasingly adopted a multimodal format, prominently featuring *memes* that combine both visual imagery and textual content. Developing automated systems capable of accurately identifying such hateful material is crucial for reducing its negative impact on society. Detecting hate speech within memes presents a complex and unresolved challenge, as it necessitates the simultaneous analysis of images and text, thereby requiring sophisticated multimodal reasoning and a unified understanding of both visual and linguistic elements. In this study, we aim to push the boundaries of current research by introducing an integrated multimodal system, MemeGuardNet, designed specifically for the identification of offensive memes. Our MemeGuardNet enhances the efficacy of existing multimodal detection techniques beyond mere fine-tuning. Notably, we demonstrate the benefits of augmenting contrastive examples through upsampling to promote multimodal integration and employ ensemble learning techniques grounded in cross-validation to bolster the system's robustness. Additionally, we conduct a thorough examination of instances where the model fails to classify correctly, leading us to propose several hypothesis-driven data augmentations and evaluate their impact on overall performance. These insights are intended to guide and inspire future investigations in this domain. Our most effective model configuration utilizes an ensemble architectures and achieves an Area Under the Receiver Operating Characteristic curve (AUROC) score of 80.53. Furthermore, we introduce the following performance metric formula to quantify the model's effectiveness. This metric provides a comprehensive measure of the model's discriminative ability across all classification thresholds.

Keywords: Multimodal Detection, Hate Speech, Memes, Ensemble Learning

1. Introduction

The proliferation of social media platforms such as Twitter, Facebook, and Instagram has significantly transformed the way individuals communicate and interact online. However, this transformation has also given rise to a pressing societal issue: online abuse. A substantial portion of internet users have either been subjected to or have witnessed various forms of abusive behavior in digital spaces. According to recent studies, approximately 41% of American adults have experienced online harassment [7]. This alarming statistic underscores the pervasive nature of online abuse and highlights the urgent need for effective mitigation strategies.

Among the different manifestations of online abuse, hate speech stands out due to its severe implications for both individuals and communities. Hate speech encompasses any communication that denigrates or discriminates against individuals or groups based on attributes such as race, gender, religion, sexual orientation, or physical appearance [18]. In recent years, there has been a notable surge in the prevalence of hate speech, exacerbated by the anonymity and reach provided by digital platforms. This increase poses significant challenges for maintaining social harmony and ensuring the well-being of targeted populations.

A particularly insidious form of hate speech that has gained traction is the use of *memes*. Traditionally, memes are image macros shared on social media for entertainment purposes, often employing humor or satire to engage audiences. However, their adaptability and virality have also made them effective tools for disseminating hateful content. *Hateful memes* leverage both visual and textual elements to convey derogatory messages, often targeting specific communities or individuals by perpetuating stereotypes or expressing overt hostility. For instance, such memes may use offensive

 doi:10.20944/preprints202411.1131.v1

imagery alongside discriminatory language to reinforce prejudiced views, thereby contributing to the normalization of racism [35] and sexism [6].

The dual-modality of memes—combining images and text—complicates the detection of hateful content. Traditional text-based or image-based analysis methods fall short in accurately interpreting the nuanced interplay between visual and linguistic cues present in memes. This challenge is further compounded by the creative and often subtle ways in which hateful messages are embedded within seemingly benign content. As a result, there is a critical need for advanced systems capable of performing multimodal reasoning to effectively identify and mitigate the spread of hateful memes.

Addressing this challenge requires the development of sophisticated automated detection systems that can seamlessly integrate and analyze both visual and textual information. The complexity of multimodal hate speech detection lies in the necessity for these systems to understand context, sarcasm, and the implicit meanings conveyed through the combination of images and text. Consequently, enhancing multimodal reasoning and joint visual-linguistic understanding is paramount for improving the accuracy and reliability of hate speech detection mechanisms.

In response to this pressing issue, Facebook initiated the *Hateful Memes Challenge* [15] as part of the NeurIPS 2020 competition track. The primary objective of this challenge is to spur innovation in multimodal reasoning and to foster the creation of robust systems capable of accurately detecting hateful memes. The challenge frames the task as a binary classification problem, wherein each meme is categorized as either *hateful* or *not hateful*. To facilitate this, the organizers introduced the Hateful Memes (HM) dataset [15], comprising over $10,000$ memes annotated for hate speech content.

A key feature of the HM dataset is the inclusion of *benign confounders*, also known as *contrastive* or *counterfactual* examples, for a subset of hateful memes. These confounders are non-hateful memes that are semantically similar to hateful ones but lack the hateful content, thereby serving to minimize accidental biases in system classifications. This design choice emphasizes the necessity for models to engage in genuine multimodal reasoning rather than relying on superficial or unimodal cues. Empirical evaluations have demonstrated a significant performance gap between unimodal and multimodal systems on this task [15], with the latter still lagging behind human performance. This gap underscores the critical need for advancements in multimodal understanding to enhance the efficacy of automated hate speech detection systems.

Building upon the insights from the Hateful Memes Challenge, our research explores various early-fusion multimodal architectures to tackle the complexities of hateful meme detection. Early-fusion models integrate visual and textual data at initial stages, allowing for more comprehensive feature interactions compared to late-fusion models that combine modalities at later stages. Specifically, we investigate architectures such as LXMERT [34], UNITER [3], and Oscar [20] for their potential in capturing the intricate relationships between image and text components in memes.

To enhance the performance of these models beyond conventional fine-tuning techniques, we introduce several innovative strategies. One such strategy involves the upsampling of contrastive examples, which aims to bolster the model's ability to discern between hateful and benign content by emphasizing multimodal interactions. Additionally, we employ ensemble learning methods based on cross-validation to improve the robustness and generalizability of our detection system. Ensemble methods aggregate the predictions from multiple models, thereby mitigating individual model biases and enhancing overall performance. Our most effective system, **MemeGuardNet**, is an ensemble of UNITER-based architectures that achieves an Area Under the Receiver Operating Characteristic curve (AUROC) score of 80.53. The AUROC metric is defined as:

$$\text{AUROC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{1}$$

where $\text{TPR}(t)$ denotes the true positive rate and $\text{FPR}(t)$ represents the false positive rate at a given threshold t. This metric provides a comprehensive evaluation of the model's ability to distinguish between hateful and non-hateful memes across all classification thresholds.

 doi:10.20944/preprints202411.1131.v1

Furthermore, we conduct an in-depth error analysis to understand the patterns and causes of model misclassifications. This analysis informs the development of hypothesis-driven model augmentations, such as multi-task learning approaches, which aim to refine the model's understanding of multimodal interactions and improve its discriminatory capabilities. By systematically enhancing MemeGuardNet, we contribute valuable insights and methodologies that can inform and inspire future research endeavors in the domain of multimodal hate speech detection.

In summary, this study addresses the critical challenge of detecting hateful memes by leveraging advanced multimodal architectures and innovative training strategies. Through the development of MemeGuardNet, we demonstrate significant improvements in the accuracy and robustness of hate speech detection systems, thereby contributing to the broader efforts of mitigating online abuse and fostering a safer digital environment.

2. Literature Review

2.1. Multimodal Representation Learning

In recent years, multimodal representation learning has emerged as a pivotal area of research, driven by the limitations of unimodal models in handling tasks that require the integration of multiple data types. Traditional unimodal approaches, which process either visual or textual data in isolation, often fall short in capturing the complex interactions between different modalities. This inadequacy is particularly evident in tasks such as Visual Question Answering (VQA) [1,12] and Visual Reasoning [33], where understanding the synergy between image and text is crucial for accurate performance.

Multimodal learning seeks to bridge this gap by developing models that can effectively combine and interpret information from diverse sources. The primary challenge lies in designing architectures that can seamlessly integrate visual and linguistic data, enabling the model to reason about both modalities simultaneously. Early efforts in this domain adopted either late-fusion (LF) or early-fusion (EF) strategies to combine the two modalities. Late-fusion methods [14,15] typically involve processing each modality independently using unimodal encoders and then merging their representations at a later stage, often through simple concatenation or other straightforward combination techniques. While this approach benefits from modularity and simplicity, it may fail to capture intricate interactions between modalities that are essential for nuanced understanding.

In contrast, early-fusion methods aim to integrate visual and textual information at earlier stages of the model, allowing for more profound and comprehensive feature interactions. Prominent examples of early-fusion architectures include MMBT [14], VisualBERT [19], and ViLBERT [23]. These models employ sophisticated mechanisms such as joint embedding spaces and cross-attention layers to facilitate the seamless fusion of visual and linguistic data. By doing so, they enhance the model's ability to perform tasks that require a deep understanding of both modalities.

Recent advancements in multimodal representation learning have been propelled by models like UNITER (UNiversal Image-TExt Representation) [3], LXMERT (Learning Cross-Modality Encoder Representations from Transformers) [34], and Oscar (Object-Semantics Aligned Pre-training) [20]. These models leverage large-scale pretraining on diverse V, L, and V+L tasks, such as image captioning, visual grounding, and VQA, to develop robust and versatile representations. UNITER, for instance, unifies various pretraining tasks to learn a comprehensive multimodal embedding, while LXMERT employs a dual-stream architecture with separate transformers for visual and textual inputs that interact through cross-modal attention mechanisms. Oscar introduces object-tags as anchor points to align visual and textual modalities more effectively, resulting in enhanced performance on downstream tasks.

Oscar, being the most recent among these, has demonstrated superior performance over its predecessors on several benchmarks, including VQA [1], GQA [12], and NLVR2 [33]. It achieves state-of-the-art results by effectively aligning object semantics with textual information, thereby facilitating more accurate and contextually relevant responses. The continual evolution of these early-fusion

 doi:10.20944/preprints202411.1131.v1

models underscores the importance of sophisticated multimodal integration techniques in advancing the field of multimodal representation learning.

Beyond these architectures, recent research has explored various enhancements to improve multimodal understanding. Techniques such as attention mechanisms, graph-based representations, and contrastive learning objectives have been employed to refine the interaction between modalities. Additionally, the incorporation of external knowledge bases and the use of advanced pretraining strategies have further augmented the capabilities of multimodal models. These innovations collectively contribute to the development of more intelligent systems capable of nuanced and context-aware interpretations of complex, multimodal data.

2.2. Multimodal Hate Speech Detection

The detection of hate speech has predominantly been explored within the realm of text-based analysis, leveraging linguistic features to identify offensive or discriminatory language. However, the advent of multimodal content, particularly memes, has introduced new challenges and opportunities for hate speech detection. Mishra et al. [25] provide a comprehensive overview of emerging trends, resources, and challenges in the domain of online abusive language detection, highlighting the shift towards incorporating visual information alongside textual content. Despite this growing interest, research in the vision and multimodal domains remains comparatively sparse, largely due to the limited availability of annotated datasets that encompass both images and text.

One notable contribution to this field is the MMHS150K dataset introduced by Gomez et al. [9], which comprises multimodal tweets annotated for hate speech. The authors developed and evaluated three multimodal models—Feature Concatenation Model (FCM), Spatial Concatenation Model (SCM), and Textual Kernels Model (TKM)—aimed at integrating visual and textual features for hate speech detection. However, their findings revealed that these multimodal approaches did not significantly outperform unimodal counterparts, suggesting that effective multimodal integration remains a challenging task.

Further investigations by Hosseinmardi et al. [11] focused on cyberbullying and cyberaggression in Instagram posts and comments. They evaluated the performance of classifiers such as Naive Bayes and linear Support Vector Machines (SVM) using a variety of features, including word n-grams, image categories, and metadata like the number of followers and likes. Their results indicated that while multimodal features could enhance classification accuracy, the improvement was modest, and more sophisticated feature engineering was necessary to achieve substantial gains.

In the context of the *Hateful Memes Challenge* [15], the organizers provided several baseline models to evaluate the effectiveness of multimodal hate speech detection systems. These baselines encompassed both unimodal and multimodal architectures pretrained using diverse methodologies. The experimental results demonstrated that multimodal approaches consistently outperformed unimodal systems, reinforcing the necessity for models to process image and text signals jointly to accurately detect hateful content. However, even the best-performing multimodal baselines exhibited performance gaps when compared to human-level accuracy, underscoring the complexity of the task and the need for further advancements in multimodal reasoning.

Among the baseline models, ViLBERT [23] and VisualBERT [19] emerged as the top performers. ViLBERT extends the BERT architecture into a multimodal, two-stream model that employs separate transformers for processing visual and textual inputs, which interact through co-attentional transformer layers to learn joint representations. VisualBERT, on the other hand, adopts a single-stream architecture that integrates visual and textual information within a unified self-attention mechanism, allowing for more cohesive feature interactions. These models were pretrained on large-scale datasets such as Conceptual Captions (CC) [31] and Common Objects in Context (COCO) [21], which provided diverse and rich multimodal data for learning robust representations.

Despite the progress made by these models, challenges remain in effectively capturing the nuanced interplay between image and text in hateful memes. The subtlety and creativity with which

 doi:10.20944/preprints202411.1131.v1

hate speech is embedded within memes often require a deeper contextual and cultural understanding, which current models may not fully possess. Additionally, the presence of benign confounders in datasets like HM introduces further complexity, as models must discern the underlying hateful intent despite superficial similarities to non-hateful content.

Recent studies have explored various strategies to enhance multimodal hate speech detection. Approaches such as attention-based fusion, where the model selectively focuses on relevant parts of the image and text, and the use of adversarial training to improve robustness against deceptive or misleading content, have shown promise. Moreover, incorporating external knowledge sources and leveraging transfer learning techniques from related domains are avenues that researchers are actively investigating to bolster model performance.

In summary, while significant strides have been made in multimodal hate speech detection, particularly through the development of advanced multimodal architectures and the creation of specialized datasets, the field continues to face substantial challenges. The complexity of effectively integrating visual and textual information, coupled with the dynamic and context-dependent nature of hate speech, necessitates ongoing research and innovation. Our work builds upon these foundational studies by introducing MemeGuardNet, a novel multimodal framework designed to enhance the detection of hateful memes through improved multimodal reasoning and robust feature integration.

3. MemeGuardNet

In light of the superior performance exhibited by early-fusion architectures in the detection of hateful memes [15], we introduce our novel multimodal framework, **MemeGuardNet**. MemeGuardNet is designed to leverage the strengths of existing transformer-based models while introducing enhancements tailored specifically for the nuanced task of hateful meme detection. Unlike traditional models that process visual and textual data independently or with basic integration techniques, MemeGuardNet employs an advanced early-fusion strategy that facilitates deeper interaction between modalities, thereby enabling more sophisticated multimodal reasoning.

3.1. Architecture of MemeGuardNet

MemeGuardNet is built upon a transformer-based architecture that integrates both visual and textual information from memes in a cohesive manner. The architecture comprises three primary components:

1. **Visual Encoder**: Utilizes a pretrained Faster R-CNN [29] to extract rich visual features from meme images. These features include object detection results, bounding box coordinates, and semantic labels, which are crucial for understanding the context and content of the image.
2. **Textual Encoder**: Employs a pretrained BERT tokenizer [5] to process the textual content of memes. The tokenizer converts raw text into token embeddings, capturing linguistic nuances and contextual information.
3. **Multimodal Fusion Module**: Integrates the visual and textual features using a sophisticated attention mechanism that enables the model to focus on relevant parts of both modalities simultaneously. This module is pivotal in facilitating the joint understanding required for accurate hate speech detection.

The fusion module operates by aligning the visual and textual embeddings in a shared latent space, allowing for seamless interaction and mutual reinforcement of features. This alignment is achieved through cross-attention layers that dynamically adjust the influence of each modality based on the context, thereby enhancing the model's ability to discern subtle cues indicative of hateful content.

3.2. Pretraining and Fine-Tuning Strategies

MemeGuardNet undergoes a two-phase training process: pretraining and fine-tuning.

 doi:10.20944/preprints202411.1131.v1

3.2.1. Pretraining

During the pretraining phase, MemeGuardNet is exposed to large-scale multimodal datasets such as COCO [21], Visual Genome [17], and SBU Captions [26]. The objectives during pretraining include:

- **Masked Language Modeling (MLM)**: Randomly masks tokens in the textual input and trains the model to predict them based on the surrounding context and visual features.
- **Masked Region Prediction (MRP)**: Masks regions in the image and requires the model to predict the masked visual features using both the visible image regions and the accompanying text.
- **Image-Text Matching (ITM)**: Determines whether a given image and text pair are semantically aligned, promoting a deeper understanding of the relationship between modalities.
- **Word-Region Alignment (WRA)**: Aligns specific words in the text with corresponding regions in the image, enhancing the model's ability to associate linguistic and visual elements accurately.

These pretraining tasks equip MemeGuardNet with a robust foundation for multimodal comprehension, enabling it to capture intricate patterns and relationships within the data.

3.2.2. Fine-Tuning

Following pretraining, MemeGuardNet undergoes fine-tuning on the Hateful Memes (HM) dataset [15] to adapt the model specifically for the task of hateful meme detection. The fine-tuning process involves optimizing the model parameters using a binary classification objective, where each meme is classified as either *hateful* or *not hateful*. The loss function employed during this phase is the Binary Cross-Entropy (BCE) loss, defined as:

$$\mathcal{L}_{BCE}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log f_{\boldsymbol{\theta}}(x_i) + (1-y_i)\log(1-f_{\boldsymbol{\theta}}(x_i))] \tag{2}$$

where:

- N is the total number of memes in the training set.
- y_i is the ground truth label for meme i.
- $f_{\boldsymbol{\theta}}(x_i)$ is the predicted probability that meme i is hateful.
- $\boldsymbol{\theta}$ represents the model parameters.

To address class imbalance in the HM dataset, where hateful memes constitute a minority, we incorporate a weighted loss function that assigns higher importance to the hateful class:

$$\mathcal{L}_{HW}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\left[\alpha_{pos} y_i \log f_{\boldsymbol{\theta}}(x_i) + \alpha_{neg}(1-y_i)\log(1-f_{\boldsymbol{\theta}}(x_i))\right] \tag{3}$$

Here, α_{pos} and α_{neg} are weights assigned to the positive (hateful) and negative (not hateful) classes, respectively, ensuring that the model pays more attention to the underrepresented class.

3.3. Enhancements in MemeGuardNet

To further bolster the performance of MemeGuardNet, we introduce several key enhancements:

3.3.1. Confounder Upsampling (CFU)

The HM dataset includes benign confounders—non-hateful memes that are semantically similar to hateful ones—to mitigate the risk of models exploiting unimodal cues. However, our initial experiments revealed that MemeGuardNet struggled to effectively utilize these confounders, particularly those involving textual alterations. To address this, we implement a Confounder Upsampling (CFU) strategy, where we increase the frequency of confounder instances during training. This approach ensures that the model is exposed to a balanced mix of hateful and benign examples, fostering better multimodal reasoning.

 doi:10.20944/preprints202411.1131.v1

3.3.2. Cross-Validation Ensemble (CVE)

Given the limited size of the HM dataset, overfitting is a significant concern. To enhance generalization, we adopt a Cross-Validation Ensemble (CVE) technique. Specifically, we train multiple instances of MemeGuardNet on different subsets of the training data and aggregate their predictions through a weighted average. The weights for each model in the ensemble are optimized using an evolutionary algorithm (EA) to maximize the Area Under the Receiver Operating Characteristic curve (AUROC) on the validation set. This ensemble approach not only stabilizes predictions but also leverages the diversity of individual models to improve overall performance.

3.3.3. Incorporation of Fine-Grained Object Tags

Through qualitative analysis, we identified that MemeGuardNet occasionally misclassifies memes targeting specific communities based on attributes such as religion, gender, and race. To mitigate this, we augment the model with fine-grained object tags derived from YOLO9000 [28], an advanced object detection model. By integrating a subset of 97 relevant object classes, such as "Nigerian", "Muslimah", "revolver", and "amputee", MemeGuardNet gains additional semantic context that aids in identifying targeted groups within the meme imagery. The inclusion of these tags is formalized as follows:

$$\mathcal{L}_{MemeGuardNet} = \mathcal{L}_{HW} + \gamma \mathcal{L}_{MR} + \beta \mathcal{L}_{YOLO} \tag{4}$$

where γ and β are hyperparameters controlling the influence of margin ranking loss ($\mathcal{L}_{MR}$) and YOLO-based loss ($\mathcal{L}_{YOLO}$), respectively.

3.4. Training Procedure

The training of MemeGuardNet is conducted in multiple stages to ensure optimal performance:

1. **Initial Fine-Tuning**: MemeGuardNet is first fine-tuned on the HM dataset using the BCE loss with class weighting to address class imbalance.
2. **Confounder Upsampling**: The training data is augmented by upsampling benign confounders, particularly text confounders, to enhance the model's ability to perform multimodal reasoning.
3. **Ensemble Training**: Multiple instances of MemeGuardNet are trained using different cross-validation folds. An evolutionary algorithm optimizes the ensemble weights based on AUROC performance on the validation set.
4. **Incorporation of YOLO Tags**: The model is further fine-tuned with the inclusion of fine-grained object tags to improve the detection of targeted groups within memes.

Throughout the training process, we employ techniques such as learning rate scheduling, dropout regularization, and gradient clipping to prevent overfitting and ensure stable convergence. The final model, MemeGuardNet, represents an ensemble of optimally weighted MemeGuardNet instances, each contributing to a robust and accurate detection system.

4. Feature Engineering on MemeGuardNet

In this section, we delineate the comprehensive methodology underpinning our proposed framework, MemeGuardNet, for the detection of hateful memes. Our approach encompasses several key components: image feature extraction, advanced training strategies, ensemble optimization, and the integration of fine-grained object information. Each component is meticulously designed to address the challenges inherent in multimodal hate speech detection, ensuring that MemeGuardNet operates with enhanced accuracy and robustness.

4.1. Image Feature Extraction and Base Model Configuration

MemeGuardNet operates by ingesting both textual and visual data from memes. To facilitate this, we employ a two-pronged feature extraction process:

 doi:10.20944/preprints202411.1131.v1

- **Visual Features**: We utilize a pretrained Faster R-CNN [29] model to extract high-dimensional visual features from meme images. Specifically, the model generates bounding box coordinates, object classifications, and associated feature vectors for each detected object within an image. To maintain consistency with MemeGuardNet's pretrained parameters, we source these features directly from the original model checkpoints [1], ensuring that the visual representations align seamlessly with the model's expectations.
- **Textual Features**: The textual content of memes is tokenized using the standard pretrained BERT tokenizer [5], converting raw text into a sequence of token embeddings. These embeddings capture both syntactic and semantic information, providing a rich representation of the meme's linguistic elements.

With these features in hand, MemeGuardNet is fine-tuned on the HM dataset using a binary classification objective to distinguish between hateful and non-hateful memes. The primary optimization target is the BCE loss, defined as:

$$\mathcal{L}_{BCE}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log f_{\boldsymbol{\theta}}(x_i) + (1-y_i)\log(1-f_{\boldsymbol{\theta}}(x_i))] \tag{5}$$

where $\boldsymbol{\theta}$ represents the model parameters, x_i denotes the input features for meme i, and y_i is the corresponding ground truth label.

Empirical evaluations revealed that MemeGuardNet, based on the UNITER architecture [3], consistently outperformed other baseline models such as LXMERT [34] and Oscar [20] in both base and large configurations (refer to Table 2 for detailed results). This superior performance is attributed to UNITER's comprehensive pretraining tasks, which encompass masked language modeling, masked region prediction, cross-modality matching, and image question answering. Additionally, larger model variants exhibited diminished generalization capabilities due to overfitting, underscoring the importance of selecting appropriately sized models for this task. Consequently, we adopt the UNITER-base variant as the foundational architecture for MemeGuardNet, ensuring a balance between performance and generalizability.

4.2. Enhancing Multimodal Reasoning Through Confounder Upsampling and Loss Re-weighting

A pivotal aspect of the HM dataset is the inclusion of benign confounders—non-hateful memes that are semantically akin to their hateful counterparts. These confounders are instrumental in preventing models from relying solely on unimodal cues, thereby necessitating genuine multimodal reasoning. However, initial experiments indicated that MemeGuardNet's performance on benign confounders, particularly those involving textual modifications, was suboptimal.

4.2.1. Confounder Upsampling (CFU)

To mitigate this issue, we implement a Confounder Upsampling (CFU) strategy, wherein confounder instances are disproportionately represented during training. Specifically, text confounders are upsampled within each training batch, ensuring that MemeGuardNet receives a balanced exposure to both hateful and benign examples. This adjustment encourages the model to focus on the interplay between visual and textual modalities, enhancing its ability to discern subtle differences indicative of hate speech.

Mathematically, the upsampling process can be formalized as:

$$\text{Batch}_{\text{aug}} = \text{Batch}_{\text{original}} \cup \lambda \cdot \text{Confounder}_{\text{upsampled}} \tag{6}$$

[1] https://github.com/MILVLG/bottom-up-attention.pytorch

 doi:10.20944/preprints202411.1131.v1

where λ represents the upsampling factor, determining the extent to which confounders are emphasized within each batch.

4.2.2. Loss Re-weighting (HW)

The HM dataset exhibits an imbalanced distribution, with 36% of memes labeled as hateful and 64% as non-hateful. To address this imbalance and ensure that MemeGuardNet adequately learns to identify hateful content, we introduce a Loss Re-weighting (HW) mechanism. This involves assigning higher weights to the hateful class within the BCE loss function, thereby amplifying its influence during the optimization process.

The re-weighted loss function is defined as:

$$\mathcal{L}_{HW}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\left[\alpha_{pos} y_i \log f_{\boldsymbol{\theta}}(x_i) + \alpha_{neg}(1-y_i)\log(1-f_{\boldsymbol{\theta}}(x_i))\right] \tag{7}$$

where α_{pos} and α_{neg} are the weights assigned to the positive (hateful) and negative (non-hateful) classes, respectively, with the constraint $\alpha_{pos} > \alpha_{neg}$ and $\alpha_{pos} + \alpha_{neg} = 1$. This weighting scheme ensures that misclassifications of hateful memes incur a higher penalty, thereby driving the model to prioritize accurate detection of hate speech.

4.3. Cross-Validation Ensemble Optimization

Given the relatively limited size of the HM dataset, MemeGuardNet is susceptible to overfitting when trained on small subsets of data. To enhance generalization and stabilize predictions, we employ a Cross-Validation Ensemble (CVE) strategy. This involves training multiple instances of MemeGuardNet on distinct subsets of the training data and aggregating their predictions to form a more robust final output.

4.3.1. Ensemble Construction

The ensemble is constructed by partitioning the HM training set into K cross-validation folds. For each fold $k \in \{1, \ldots, K\}$, MemeGuardNet is trained on $K-1$ folds and validated on the remaining fold. This process is repeated K times, resulting in K distinct model instances, each trained on different data distributions. The predictions from these models are then combined using a weighted average, where the weights are optimized to maximize the ensemble's performance on a validation set.

$$\hat{y} = \sum_{k=1}^{K} \alpha_k f_{\boldsymbol{\theta}_k}(x) \tag{8}$$

where $\hat{y}$ is the final predicted probability, $f_{\boldsymbol{\theta}_k}(x)$ is the prediction from the k-th model, and α_k is the corresponding weight.

4.3.2. Weight Optimization via Evolutionary Algorithm (EA)

To determine the optimal weights $\{\alpha_k\}_{k=1}^{K}$ for the ensemble, we employ an Evolutionary Algorithm (EA). The EA iteratively searches for the weight configuration that maximizes the AUROC score on the validation set. The optimization process is guided by the following objective function:

$$\max_{\{\alpha_k\}} \ \text{AUROC}\left(\sum_{k=1}^{K} \alpha_k f_{\boldsymbol{\theta}_k}(x)\right) \tag{9}$$

Subject to:

$$\sum_{k=1}^{K} \alpha_k = 1 \quad \text{and} \quad \alpha_k \geq 0 \quad \forall k \tag{10}$$

The EA operates through a population-based search, evolving candidate weight vectors over multiple generations through selection, crossover, and mutation operations. This stochastic optimization

 doi:10.20944/preprints202411.1131.v1

technique is well-suited for navigating the complex, non-convex landscape of weight assignments, ensuring that the ensemble leverages the strengths of individual models while mitigating their weaknesses.

4.4. Integration of Fine-Grained YOLO9000 Object Tags

Through extensive error analysis, we identified that MemeGuardNet occasionally misclassifies memes that target specific communities or individuals based on attributes such as religion, gender, and race. To enhance the model's ability to recognize and interpret these targeted attributes, we integrate fine-grained object tags derived from YOLO9000 [28], a state-of-the-art object detection model capable of identifying a diverse set of 9000 classes.

4.4.1. Selection of Relevant Object Classes

Not all YOLO9000 classes are pertinent to the HM dataset. Therefore, we curate a subset of 97 object classes that are most relevant to the context of hateful memes. Examples of these classes include "Nigerian", "Muslimah", "revolver", and "amputee". This selective approach ensures that the additional information provided to MemeGuardNet is both meaningful and contextually appropriate.

4.4.2. Incorporation into MemeGuardNet

For each meme, YOLO9000 processes the image to generate a set of detected object tags. These tags are then filtered to retain only the selected subset of 97 classes. The resulting object tags are integrated into MemeGuardNet as additional input features, alongside the original visual and textual data. Formally, the input to MemeGuardNet is augmented as follows:

$$x_i = \{\text{Text}_i, \text{YOLOTags}_i, \text{VisualFeatures}_i\} \tag{11}$$

This multi-faceted input enables MemeGuardNet to leverage fine-grained object information, enhancing its ability to identify and interpret the targeted groups within memes. The integration of YOLO9000 tags is achieved through an additional embedding layer that maps the object tags into the shared latent space, facilitating their seamless incorporation into the multimodal fusion process.

4.4.3. Impact on Model Performance

The inclusion of fine-grained object tags has a twofold impact on MemeGuardNet:

1. **Enhanced Contextual Understanding**: By explicitly identifying objects and attributes relevant to hate speech, the model gains a deeper contextual understanding of the meme's content, enabling more accurate classification.
2. **Improved Target Group Identification**: The object tags assist in pinpointing specific groups or individuals being targeted, thereby refining the model's ability to detect nuanced expressions of hate that might otherwise be overlooked.

Empirical evaluations demonstrate that the incorporation of YOLO9000 tags leads to a measurable improvement in MemeGuardNet's AUROC score, validating the effectiveness of this enhancement.

4.5. Comprehensive Training Pipeline

The training pipeline of MemeGuardNet integrates all the aforementioned components in a cohesive manner:

1. **Data Preprocessing**: Extract visual features using Faster R-CNN and tokenize textual content using BERT tokenizer. Additionally, generate fine-grained object tags using YOLO9000.
2. **Model Initialization**: Initialize MemeGuardNet with pretrained weights from UNITER-base, ensuring a strong starting point for multimodal understanding.
3. **Pretraining Phase**: Fine-tune MemeGuardNet on large-scale multimodal datasets using MLM, MRP, ITM, and WRA objectives to establish robust multimodal representations.

 doi:10.20944/preprints202411.1131.v1

4. **Fine-Tuning Phase**: Adapt MemeGuardNet to the HM dataset using the weighted BCE loss, incorporating CFU and HW strategies to address confounders and class imbalance.
5. **Ensemble Training**: Train multiple MemeGuardNet instances across different cross-validation folds and optimize ensemble weights using EA to enhance generalization.
6. **Integration of YOLO Tags**: Fine-tune the ensemble with the inclusion of YOLO9000 object tags, further refining the model's ability to detect targeted hate speech.
7. **Evaluation**: Assess the final ensemble's performance on the unseen test set using AUROC and other relevant metrics, ensuring that MemeGuardNet meets the desired detection standards.

This comprehensive pipeline ensures that MemeGuardNet is meticulously trained to handle the intricacies of multimodal hate speech detection, leveraging advanced feature extraction, sophisticated training strategies, and robust ensemble optimization to achieve superior performance.

4.6. Mathematical Formulation of the Ensemble Prediction

To formalize the ensemble prediction mechanism, let us define the prediction function for MemeGuardNet as follows:

$$\hat{y} = \sum_{k=1}^{K} \alpha_k f_{\boldsymbol{\theta}_k}(x) \tag{12}$$

where:

- K is the number of models in the ensemble.
- $f_{\boldsymbol{\theta}_k}(x)$ is the predicted probability from the k-th MemeGuardNet model for input x.
- α_k is the optimized weight for the k-th model, constrained by $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k \geq 0$.

The ensemble prediction $\hat{y}$ represents a weighted aggregation of individual model predictions, optimized to maximize the overall AUROC score on the validation set. This formulation ensures that each model contributes proportionally to its performance, thereby enhancing the ensemble's discriminative capability.

4.7. Optimization Objective

The primary objective during training is to minimize the combined loss function, which incorporates both classification and ranking objectives:

$$\mathcal{L}_{Total}(\boldsymbol{\theta}) = \mathcal{L}_{HW}(\boldsymbol{\theta}) + \gamma \mathcal{L}_{MR}(\boldsymbol{\theta}) + \beta \mathcal{L}_{YOLO}(\boldsymbol{\theta}) \tag{13}$$

where:

- $\mathcal{L}_{HW}$ is the weighted BCE loss addressing class imbalance.
- $\mathcal{L}_{MR}$ is the margin ranking loss encouraging correct ordering of hateful and non-hateful predictions.
- $\mathcal{L}_{YOLO}$ is the loss associated with the integration of YOLO9000 object tags.
- γ and β are hyperparameters controlling the relative importance of each loss component.

By jointly optimizing these loss components, MemeGuardNet is trained to not only accurately classify memes but also to prioritize the correct identification of hateful content through enhanced multimodal reasoning and contextual understanding.

4.8. Implementation Details

MemeGuardNet is implemented using the PyTorch framework, leveraging the MMF (Multimodal Framework) library [32] for streamlined model training and evaluation. The model is trained on NVIDIA Tesla V100 GPUs with a batch size of 32 and a learning rate of 2×10^{-5}. Training is conducted for a maximum of 10 epochs, with early stopping based on validation AUROC to prevent overfitting.

 doi:10.20944/preprints202411.1131.v1

4.9. Evaluation Metrics

To comprehensively assess MemeGuardNet's performance, we employ the following evaluation metrics:

- **Area Under the Receiver Operating Characteristic curve (AUROC)**: Measures the model's ability to distinguish between hateful and non-hateful memes across all classification thresholds.
- **F1 Score**: Balances precision and recall, providing a single metric that accounts for both false positives and false negatives.
- **Precision and Recall**: Evaluate the model's accuracy in identifying hateful memes and its ability to capture all relevant instances, respectively.

These metrics collectively offer a nuanced understanding of MemeGuardNet's effectiveness, ensuring that the model not only achieves high discriminative performance but also maintains a balanced approach to precision and recall.

5. Experimental Setup and Performance Evaluation

In this section, we detail the experimental framework, including implementation specifics and hyperparameter configurations for our top-performing models. Subsequently, we present and analyze the performance metrics of various model iterations, highlighting the efficacy of our proposed strategies in enhancing hateful meme detection.

5.1. Dataset Overview

Our experiments leverage the *Hateful Memes* (HM) dataset [15], curated by Facebook AI, which serves as the benchmark for the *Hateful Memes Challenge*. The HM dataset comprises over 10,000 memes, each combining visual imagery with textual content, meticulously labeled as either *hateful* or *non-hateful*. The classification adheres to the stringent definition provided by Kiela et al. [15], characterizing hatefulness as "a direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, religion, caste, sex, etc." For comprehensive details, readers are directed to Kiela et al. [15].

In addition to the primary labeled set, the dataset includes an ancillary collection of 2,000 unlabeled memes designated for testing purposes, segmented into two distinct phases: Phase 1 (*seen* test set) and Phase 2 (*unseen* and final test set). This bifurcation ensures a rigorous evaluation of model generalizability and robustness across varying data distributions.

A unique feature of the HM dataset is the incorporation of *benign confounders*. These are non-hateful memes derived from hateful counterparts through minimal alterations in either image or text, resulting in label inversions. Specifically, there are two categories of confounders:

1. **Image Confounders**: Memes sharing identical images but differing in textual content.
2. **Text Confounders**: Memes with identical text but distinct images.

Overall, the dataset encompasses five meme types:

- *Multimodal Hate*: Memes combining image and text to convey hate.
- *Unimodal Hate*: Memes where either image or text alone conveys hate.
- *Benign Image Confounders*: Memes with hateful text and benign images.
- *Benign Text Confounders*: Memes with hateful images and benign text.
- *Random Non-Hateful Examples*: Memes devoid of hate speech, serving as control instances.

The distribution of these meme types across various dataset splits is detailed in Table 1.

 doi:10.20944/preprints202411.1131.v1

Table 1. Distribution of Meme Categories in the Hateful Memes Dataset

Dataset Split	Multimodal Hateful	Unimodal Hateful	Benign Confounders	Random Benign	Dynamic Adversarial Confounders	Total
Train	1,300	1,750	3,200	2,250	–	8,500
Dev-seen	200	50	200	50	–	500
Test-seen	400	100	400	100	–	1,000
Dev-unseen	200	–	200	–	140	540
Test-unseen	729	–	597	–	674	2,000

The HM dataset's meticulous construction ensures that multimodal approaches are indispensable for optimal performance. The presence of benign confounders challenges models to engage in genuine multimodal reasoning rather than relying on unimodal heuristics, thereby enhancing the robustness of hate speech detection mechanisms.

5.2. Implementation Details and Hyperparameter Configuration

Our primary model, **MemeGuardNet**, is instantiated using the *UNITER-base* architecture [3] as the foundational framework. The following hyperparameters and training configurations were employed across all experiments to ensure consistency and reproducibility:

- **Batch Size**: Set to 16, with gradient accumulation over 2 steps to achieve an effective batch size of 32. This adjustment accommodates memory constraints without compromising training stability.
- **Learning Rate**: Initialized at 3×10^{-5}, optimized using the Adam optimizer [16] with a weight decay parameter of 1×10^{-3}.
- **Learning Rate Scheduler**: Utilized a cosine learning rate scheduler with 500 warmup steps to facilitate smooth convergence and prevent abrupt learning rate changes.
- **Loss Function**: Employed Binary Cross-Entropy (BCE) loss, further refined with class weighting to address the inherent class imbalance in the HM dataset. Specifically, the loss for the hateful class was scaled by a factor of 1.8 to emphasize minority class learning.
- **Training Epochs**: Configured to run for a maximum of 30 epochs, with early stopping triggered if there was no improvement in the Area Under the Receiver Operating Characteristic curve (AUROC) for 5 consecutive epochs. This strategy prevents overfitting and ensures efficient utilization of computational resources.
- **Sequence Lengths**: Capped the maximum text length at 60 tokens and limited image bounding box features to 100 per meme to balance computational efficiency with representational richness.
- **Confounder Upsampling**: Implemented a confounder upsampling (CFU) strategy, where text confounders were upsampled by a factor of 3. This technique increases the likelihood of confounder instances being sampled during training, thereby enhancing the model's capacity for multimodal reasoning.
- **Cross-Validation Folds**: Explored varying numbers of cross-validation (CV) folds, ultimately determining that a 15-fold cross-validation setting yielded the best performance metrics.
- **Ensemble Weight Optimization**: Leveraged an evolutionary algorithm (EA) with tournament selection, Gaussian noise mutation, and uniform crossover strategies [8,24] to optimize ensemble weights. The EA was configured with a population size of 512 individuals over 100 generations, selecting the optimal weight set based on the highest AUROC score achieved on the development set (*dev-seen*).

These hyperparameter settings were meticulously chosen based on preliminary experiments and literature benchmarks to strike an optimal balance between model complexity and generalization capability.

 doi:10.20944/preprints202411.1131.v1

5.3. Performance Evaluation

Table 2 encapsulates the AUROC scores achieved by our various model configurations on the development set (*dev-seen*) and the Phase 1 and Phase 2 test sets of the challenge. It is noteworthy that the number of submissions for Phase 1 and Phase 2 was restricted to one per day and three in total, respectively, resulting in some models not being evaluated across all test phases.

Table 2. AUROC Performance of Various Model Configurations on Development and Test Sets

Model	AUROC		
	Development	Phase 1	Phase 2
ViLBERT CC	70.07	70.03	–
VisualBERT COCO	73.97	71.41	–
MemeGuardNet	78.04	74.73	–
LXMERT	72.33	–	–
Oscar	72.00	–	–
$\text{MemeGuardNet}_{\text{CV10}}$	79.81	–	–
$\text{MemeGuardNet}_{\text{CV10 + CFU}}$	79.64	–	–
$\text{MemeGuardNet}_{\text{CV10 + CFU + HW}}$	80.01	78.60	–
$\text{MemeGuardNet}_{\text{CV15 + CFU + HW}}$	80.65	79.06	–
$\text{MemeGuardNet}_{\text{CV30 + CFU + HW}}$	81.36	78.98	–
$\text{MemeGuardNet}_{\text{CV15 + CFU + HW + MRL}}$	80.44	78.14	–
$\text{MemeGuardNet}_{\text{CV15 + CFU + HW + YOLO}}$	80.67	78.21	–
$\text{MemeGuardNet}_{\text{ENSEMBLE 1}}$	81.71	79.13	80.33
$\text{MemeGuardNet}_{\text{ENSEMBLE 2}}$	81.76	79.10	80.40
$\text{MemeGuardNet}_{\text{FINAL}}$	77.39	79.07	**80.53**

Interpretation of Results

The AUROC scores presented in Table 2 demonstrate the progressive enhancements in performance achieved through iterative model refinements and strategic training methodologies. Notably, *ViLBERT CC* and *VisualBERT COCO* serve as baseline multimodal models, attaining AUROC scores of 70.07 and 73.97 on the development set, respectively. In contrast, our foundational model, **MemeGuardNet**, significantly surpasses these baselines with an AUROC of 78.04, underscoring the efficacy of our initial fine-tuning approach on the HM dataset.

Delving deeper, the introduction of cross-validation techniques (*CV10*) further augments performance, yielding an AUROC of 79.81 on the development set. The subsequent application of confounder upsampling (CFU) and hateful class loss re-weighting (HW) enhances this score to 80.01, indicating the benefits of addressing data imbalance and emphasizing multimodal reasoning through targeted training strategies.

Expanding the cross-validation framework to 15 folds (*CV15 + CFU + HW*) yields a marginal improvement to 80.65, while an extensive 30-fold cross-validation (*CV30 + CFU + HW*) achieves the peak development set performance of 81.36. However, it is noteworthy that increasing the number of folds beyond 15 does not translate to proportional gains on the Phase 1 test set, suggesting potential diminishing returns and the risk of overfitting with excessively granular cross-validation.

Further enhancements involving margin ranking loss (MRL) and the integration of YOLO9000 object tags (*YOLO*) demonstrate mixed outcomes. While *MRL* variants slightly reduce performance on the development set to 80.44, the incorporation of YOLO9000 tags modestly improves the AUROC to 80.67. However, both modifications do not consistently enhance performance across all test phases, potentially due to the introduction of noise and misalignment between object tags and hate speech indicators.

 doi:10.20944/preprints202411.1131.v1

The most promising results are observed with our ensemble models. *ENSEMBLE 1*, comprising an EA-optimized aggregation of various MemeGuardNet instances, achieves an impressive AUROC of 81.71 on the development set and 79.13 on Phase 1. *ENSEMBLE 2*, which integrates additional MemeGuardNet variants trained with varying seeds and batch sizes, marginally improves these scores to 81.76 and 79.10, respectively. Finally, our flagship model, *FINAL*, which incorporates cross-validation training with part of the development set included in training folds, culminates in an AUROC of 80.53 on Phase 2, securing the fourth position on the Phase 2 leaderboard.

These results collectively affirm that our strategic enhancements—particularly cross-validation ensembles and targeted training adjustments—substantially bolster MemeGuardNet's capability to discern and classify hateful memes with high accuracy and reliability.

6. Error Analysis and Model Refinements

To gain deeper insights into MemeGuardNet's performance, we conducted a comprehensive error analysis, focusing on misclassification patterns within the *dev-seen* set. This analysis not only elucidates the model's strengths and weaknesses but also guides subsequent refinements aimed at mitigating identified shortcomings. Additionally, we explore supplementary model variants designed to incorporate complementary information and address overfitting tendencies.

6.1. Misclassification Patterns

Our investigation centered on the misclassifications made by $\textbf{MemeGuardNet}_{\textbf{CV15 + CFU + HW}}$, one of our top-performing models. Through qualitative examination, we identified the following key patterns:

- **False Negatives**: These are instances where hateful memes were incorrectly classified as non-hateful. Analysis revealed that these memes often possess a genuinely multimodal nature, where the hateful intent is only discernible through the synergistic interplay of image and text. The model struggled particularly with memes targeting specific groups such as Muslims, wheelchair users, and members of extremist organizations like the Ku Klux Klan. Additionally, it failed to recognize subtle references to real-life individuals (e.g., Anne Frank) and symbolic representations, indicating a gap in contextual and cultural understanding.
- **False Positives**: These instances involved non-hateful memes being erroneously classified as hateful. Many of these memes exhibited characteristics that could be construed as hateful under different contexts, suggesting that the model might be overly sensitive to certain visual or textual cues without sufficient contextual grounding. This tendency highlights the challenge of distinguishing between benign and potentially harmful content based solely on surface-level features.

Although the model demonstrates robust performance overall, these misclassification trends underscore the necessity for enhanced contextual reasoning and finer-grained feature extraction to accurately interpret nuanced expressions of hate.

6.2. Incorporating Target Group Information

Recognizing that accurately identifying the target groups within memes could significantly improve classification performance, we explored methods to embed this information into MemeGuardNet. The following approaches were undertaken:

Social Bias Frames Integration

We experimented with integrating the *Social Bias Inference Corpus* (SBIC) [30] to provide MemeGuardNet with structured annotations related to offensive language, intent, and targeted groups. The SBIC dataset comprises approximately $150,000$ annotations across $44,761$ social media posts, encompassing various bias-related attributes.

 doi:10.20944/preprints202411.1131.v1

- **Multi-Task Learning (MTL)**: Initially, we implemented a multi-task learning framework where MemeGuardNet was trained to perform both hateful meme classification and SBIC-based tasks, such as offensiveness detection and targeted group prediction. This dual-objective training aimed to enrich the model's understanding of nuanced language and context. However, empirical results indicated a decline in performance on the *dev-seen* set compared to single-task fine-tuning, suggesting potential conflicts between task objectives or insufficient alignment of dataset domains.
- **Label Generation via RoBERTa**: To circumvent the limitations observed with MTL, we employed a two-step approach. First, we fine-tuned a RoBERTa model [22] exclusively on the SBIC target group classification task. Subsequently, this fine-tuned model was utilized to generate target group labels for each meme based solely on textual input. These generated labels were then incorporated into MemeGuardNet as additional features, akin to the YOLO9000 object tags described in Section 4.4. Despite this augmentation, performance improvements were not realized, potentially due to the introduction of noisy labels and the inherent challenge of inferring target groups from text without corresponding visual context.

Emotion Detection Augmentation

Inspired by Rajamanickam et al. [27], who demonstrated that incorporating emotion detection can enhance abusive language detection, we explored integrating emotional context into MemeGuardNet. Utilizing the *GoEmotions* dataset [4], which annotates Reddit comments with 27 emotion categories, we designed a multi-task learning variant as follows:

- **Multi-Task Learning Setup**: MemeGuardNet was trained concurrently on the primary task of hateful meme classification and an auxiliary emotion detection task. The model was configured to predict both the binary hate label and the associated emotion categories for each meme.
- **Training Dynamics**: Batches were sampled with varying ratios of primary to auxiliary tasks, and distinct classification heads were employed for each task to prevent interference. Despite these efforts, the integration of emotion detection did not yield significant performance gains on the *dev-seen* set, suggesting that emotion annotations may not directly correlate with the nuanced nature of hateful content in memes.

6.3. Mitigating Overfitting on Textual Features

Observing that MemeGuardNet exhibited a propensity to over-rely on textual features—resulting in strong performance on image confounders but weaker on text confounders—we sought strategies to balance the model's attention across modalities. Specifically, we addressed the overfitting of textual features through architectural modifications:

- **Independent Attention Mechanisms**: We restructured the internal attention layers of MemeGuardNet's transformer architecture to incorporate four distinct attention blocks: *text-to-text*, *text-to-image*, *image-to-image*, and *image-to-text*. This division allowed us to apply varying dropout rates tailored to each modality interaction, thereby reducing the model's tendency to overemphasize text-based cues.
- **Dropout Rate Adjustments**: Recognizing the imbalance in modality influence, we assigned higher dropout rates to the *text-to-text* attention blocks to discourage excessive reliance on textual features. Conversely, lower dropout rates were applied to *image-to-text* and *text-to-image* blocks to maintain sufficient interaction between visual and textual modalities.
- **Consistent Dropout Across Attention Heads**: To ensure uniformity and prevent discrepancies across multiple attention heads, dropout rates were consistently applied across all heads within each attention block. This approach maintained structural integrity while enforcing modality-specific regularization.
- **Empirical Outcomes**: Despite these architectural enhancements, the adjusted attention mechanisms did not translate into improved performance on the *dev-seen* set. This outcome suggests

 doi:10.20944/preprints202411.1131.v1

that overfitting may stem from more complex interdependencies between modalities or that additional regularization techniques are required to effectively balance feature utilization.

Our exploratory efforts to augment MemeGuardNet with target group information and mitigate overfitting on textual features, though conceptually promising, did not yield the anticipated enhancements in performance. These findings highlight the intricate challenges of multimodal hate speech detection, where the interplay between image and text necessitates sophisticated and contextually aware modeling strategies. Future research may benefit from more integrated approaches, such as leveraging external knowledge bases or employing advanced regularization techniques, to further refine MemeGuardNet's multimodal reasoning capabilities.

7. Conclusions and Future Directions

In this study, we introduced **MemeGuardNet**, a sophisticated multimodal framework designed for the detection of hateful memes. Our approach achieved a commendable fourth-place ranking in Phase 2 of the 2020 *Hateful Memes Challenge* organized by Facebook, underscoring its effectiveness in addressing the complexities of multimodal hate speech detection. MemeGuardNet integrates a suite of advanced techniques that collectively enhance its reasoning and classification capabilities. Key among these techniques are the strategic utilization of truly multimodal memes (confounders) during training, the optimization of an ensemble of models through cross-validation and evolutionary algorithms for precise weight tuning, the implementation of loss re-weighting to address class imbalances, and the meticulous alignment of image features within the appropriate feature space by employing identical checkpoints of the object detector backbone as utilized in the model's pretraining phase.

Our comprehensive error analysis, based on model misclassifications, provided valuable insights into the strengths and limitations of MemeGuardNet. This analysis revealed specific areas where the model excels, as well as scenarios where it falters, thereby informing targeted refinements and future research directions. Additionally, we conducted a series of supplementary experiments aimed at further enhancing model performance. Although some of these experiments did not yield the desired improvements, they highlighted critical challenges and potential avenues for future exploration.

7.1. Implications of MemeGuardNet

The successful deployment of MemeGuardNet demonstrates the viability of leveraging multimodal data for hate speech detection in digital environments. By effectively integrating visual and textual information, MemeGuardNet surpasses traditional unimodal models, offering a more nuanced understanding of meme content. This capability is particularly crucial given the pervasive and often subtle nature of hate speech in memes, where the interplay between image and text can obscure malicious intent from simplistic analysis.

Furthermore, MemeGuardNet's robust performance in the competitive challenge setting indicates its potential applicability in real-world scenarios. Social media platforms, content moderation services, and online communities could benefit from deploying such advanced detection systems to proactively identify and mitigate the spread of hateful content. This proactive stance is essential for fostering healthier online environments and safeguarding marginalized communities from targeted abuse.

7.2. Limitations and Challenges

Despite its strengths, MemeGuardNet is not without limitations. One primary challenge lies in the inherent ambiguity and contextual dependency of hate speech in memes. The model occasionally struggles with accurately interpreting the nuanced interplay between image and text, especially in cases involving sarcasm, irony, or culturally specific references. Additionally, the reliance on predefined object detectors and tokenizers may constrain the model's ability to generalize to novel or evolving forms of hateful content.

Another significant limitation pertains to the dataset's representation. While the HM dataset is comprehensive, it may not encompass the full spectrum of hateful expressions found across diverse

 doi:10.20944/preprints202411.1131.v1

cultures and languages. This limitation underscores the necessity for continual dataset expansion and diversification to enhance the model's adaptability and effectiveness across varied contexts.

7.3. Future Work

Building on the foundation laid by MemeGuardNet, several promising directions emerge for future research:

7.3.1. Enhanced Multimodal Reasoning

To address the challenges of nuanced interpretation, future iterations of MemeGuardNet could incorporate more sophisticated multimodal reasoning mechanisms. Integrating external knowledge bases and context-aware embeddings can provide the model with a deeper understanding of cultural and contextual cues, thereby improving its ability to discern subtle hate speech elements. Additionally, leveraging advancements in transformer architectures, such as ERNIE-ViL [37], which incorporates structured knowledge from scene graphs [13], could further refine the model's contextual comprehension and semantic alignment between image and text modalities.

7.3.2. Dynamic and Continual Learning

Given the evolving nature of hate speech, implementing dynamic and continual learning paradigms within MemeGuardNet can enhance its adaptability to new forms of hateful content. Techniques such as online learning, where the model incrementally updates its parameters in response to new data, can ensure that MemeGuardNet remains current and effective against emerging hate speech trends. Moreover, incorporating mechanisms for detecting and mitigating concept drift—where the statistical properties of the target variable change over time—can sustain the model's performance in dynamic environments.

7.3.3. Fine-Tuning of Image Extractors

While MemeGuardNet employs pretrained image extractors to obtain visual features, fine-tuning these extractors during the training phase could significantly improve image understanding specific to hate speech detection. By adapting the image feature extractor to the nuances of hateful content, the model can achieve a more refined and task-specific visual representation, thereby enhancing overall classification accuracy.

7.3.4. Incorporation of User and Contextual Metadata

Integrating user and contextual metadata, such as the source of the meme, user demographics, and temporal information, can provide additional layers of context that inform hate speech detection. This metadata can help the model discern intent and contextual relevance, thereby reducing false positives and improving the precision of hate speech identification.

7.3.5. Expanding and Diversifying Training Datasets

To bolster MemeGuardNet's generalizability, future work should focus on expanding and diversifying training datasets. Incorporating memes from a wider array of cultures, languages, and social contexts can equip the model with a more comprehensive understanding of global hate speech dynamics. Additionally, augmenting datasets with more granular annotations related to sentiment, intent, and target groups can facilitate more nuanced model training and evaluation.

7.3.6. Ethical Considerations and Bias Mitigation

As with any automated content moderation tool, ethical considerations and bias mitigation are paramount. Future research should explore strategies to minimize inherent biases in MemeGuardNet, ensuring that the model does not disproportionately target specific groups or perpetuate existing stereotypes. Techniques such as fairness-aware training, bias detection and correction mechanisms, and

 doi:10.20944/preprints202411.1131.v1

transparent model auditing can contribute to the ethical deployment of MemeGuardNet in real-world applications.

References

1. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
2. Prithvi Bhattacharya. 2019. Social degeneration through social media: A study of the adverse impact of 'memes'. In *2019 Sixth HCT Information Technology Trends (ITT)*, pages 44–46.
3. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.
4. Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
6. Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. Old jokes, new media–online sexism and constructions of gender in internet memes. *Feminism & Psychology*, 28(1):109–127.
7. Maeve Duggan. 2017. Men, women experience and view online harassment differently. *Pew Research Center*.
8. A. E. Eiben and James E. Smith. 2015. *Introduction to Evolutionary Computing*, 2nd edition. Springer Publishing Company, Incorporated.
9. Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1470–1478.
10. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and D. Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.
11. Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.
12. Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.
13. Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
14. Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.
15. Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.
16. Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
17. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.
18. Zachary Laub. 2019. Hate speech on social media: Global comparisons. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.
19. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

20. Xiujun Li, Xi Yin, C. Li, X. Hu, Pengchuan Zhang, Lei Zhang, Longguang Wang, H. Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.
21. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
22. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
23. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.
24. B. Miller and D. Goldberg. 1995. Genetic algorithms, tournament selection, and the effects of noise. *Complex Syst.*, 9.
25. Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.
26. Vicente Ordonez, G. Kulkarni, and T. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *NIPS*.
27. Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. *arXiv preprint arXiv:2005.14028*.
28. Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271.
29. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
30. Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
31. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.
32. Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.
33. Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.
34. Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
35. Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of internet memes. *Computers in Human Behavior*, 63:424 – 432.
36. Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
37. Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*.
38. Yuke Zhu, O. Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.
39. 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.
40. 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.

41. 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.
42. 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).
43. 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.
44. 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.
45. 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)*.
46. 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.
47. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
48. 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).
49. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. 10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
50. 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/.
51. 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.
52. 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.
53. 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.
54. 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. 10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
55. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
56. 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.
57. 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.

58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
73. 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.
74. 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.
75. 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.
76. 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. 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
77. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.

78. 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.
79. 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.
80. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.

97. 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.
98. 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.
99. 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.
100. 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.
101. 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.
102. 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.
103. 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.