Detecting Hateful Meme

Hunjun Shin, Dhruv Agarwal, Wonhee Lee

Abstract— **Hateful meme classification poses a unique challenge due to the intricate interplay between visual and textual elements, necessitating a nuanced understanding of both modalities. This study presents a comprehensive multimodal approach that synergizes image and text analysis to enhance hateful meme detection. A ResNet-based model processes visual features, while DistilBERT interprets textual content, augmented by Visual Question Answering (VQA) and image captioning to extract additional contextual information. These multimodal data streams are embedded into a unified feature space using CLIP embeddings, enabling seamless integration of representations. A Graph Attention Network (GAT) is employed to model relationships between embeddings derived from images, text, VQA, and captions, capturing complex interdependencies critical for accurate classification. Experimental results highlight that this graph-based framework significantly improves classification accuracy and AUROC scores, outperforming traditional unimodal and basic multimodal approaches. By leveraging graph-based relational learning, the proposed method offers a robust solution for nuanced hateful meme detection, advancing the development of effective content moderation tools. The code is available at *https://github.com/Dhruv-2020EE30592/EECE7205-Project-HateMeme*.**

I. Introduction

Meme content, a pervasive form of communication on social media, combines visual and textual elements to convey messages that are often humorous, satirical, or provocative. While memes are typically benign, they can also propagate harmful ideologies, hate speech, and discriminatory narratives. Detecting hateful memes presents a significant challenge due to their multimodal nature, where the interplay between images and text can create implicit, context-dependent meanings. Unlike traditional content moderation tasks that focus solely on text or image analysis, hateful meme classification requires an integrated approach to capture subtle and nuanced relationships between visual and textual elements.

Conventional unimodal methods fall short in effectively addressing this problem, as they lack the capacity to analyze the intricate connections between modalities. Simple multimodal approaches, while integrating visual and textual features, often fail to fully leverage contextual cues or model complex interdependencies. Furthermore, the multimodal nature of hateful memes often involves implicit connotations, requiring not only feature extraction but also contextual reasoning to understand the intent behind the content.

To address these challenges, this study proposes a novel multimodal approach for hateful meme classification that combines advanced feature extraction and graph-based relational learning. Our method integrates image and text analysis using a ResNet-based model and DistilBERT, respectively, and enriches contextual understanding through Visual Question Answering (VQA) and image captioning. These modalities are unified in a shared feature space using CLIP [6] embeddings, ensuring seamless integration of visual and textual representations. To model the relationships between these diverse features, we construct a graph structure where embeddings act as interconnected nodes and apply a Graph Attention Network (GAT) to capture nuanced dependencies and context.

Method	Accuracy (%)	AUROC (%)
ResNet	61.58	59.83
DistillBERT	77.46	82.92

TABLE I
ACCURACY AND AUROC OF UNIMODAL MODELS IN DETECTING HATEFUL MEMES.

This innovative framework bridges the gap between modalities, offering a more holistic analysis of hateful memes. Experimental results demonstrate that our approach significantly improves classification performance compared to traditional methods, paving the way for more effective and reliable tools for moderating harmful content in online environments.

Through these innovations, ISSUES achieves state-of-the-art performance, demonstrating its effectiveness in capturing the complex relationships inherent in multimodal hateful memes.

II. Approaches

A. *mappIng memeS to wordS for mUltimodal mEme claSsification (ISSUES) [1]*

The ISSUES (Mapping Memes to Words for Multimodal Meme Classification) model is a novel approach designed to address the challenges of hateful meme classification by leveraging advanced multimodal techniques. ISSUES builds upon the pre-trained CLIP vision-language model and incorporates a textual inversion technique to create a multimodal representation in the textual embedding space. This is achieved by mapping an image into a pseudo-word token that resides within the CLIP token embedding space, allowing the integration of both textual and visual information.

The model enhances meme classification by focusing on three core aspects:

- **Textual Inversion:** This process enriches the textual representation of memes by generating pseudo-word tokens that encapsulate visual features, enabling the

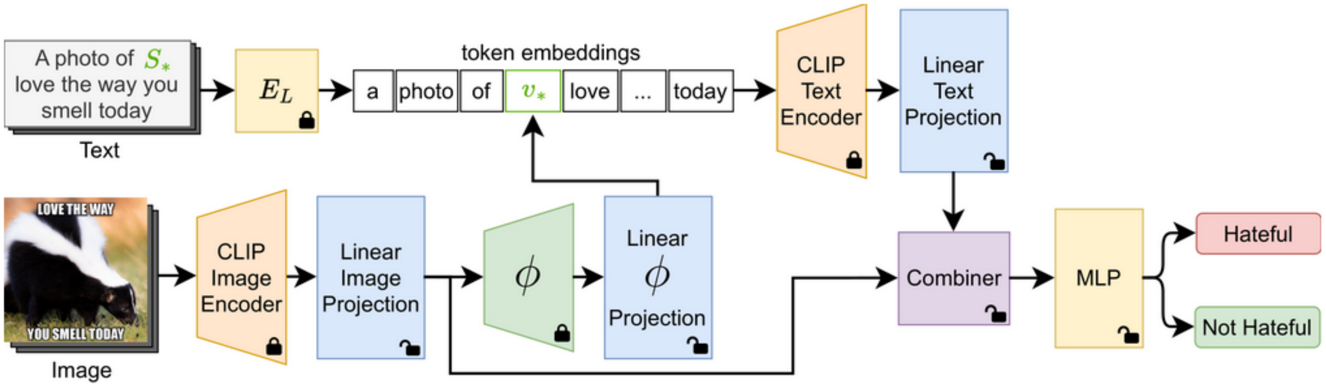

Fig. 1. Overview of ISSUES.

creation of a multimodal embedding that integrates textual and visual data effectively.

- **Embedding Space Adaptation:** Linear projections are trained to disentangle and adapt the image and text features within the shared multimodal latent space to better suit the specific task of meme classification.
- **Multimodal Fusion:** ISSUES employs a Combiner network to fuse the adapted textual and visual embeddings, enhancing its capacity to model the nuanced semantic interactions between the modalities.

B. Retrieval-Guided Contrastive Learning (RGCL) [4]

Retrieval-Guided Contrastive Learning (RGCL) is an advanced framework designed to enhance hateful meme detection by creating a hatefulness-aware embedding space for joint vision-language representations. RGCL addresses a critical issue in existing systems, where subtle differences in memes—particularly confounders—are not adequately distinguished due to proximity in the embedding space of similar text or image content.

- **Dynamic Example Retrieval:**
  - **Pseudo-Gold Positive Examples:** These are same-class examples with high similarity scores, which help align semantically similar memes in the embedding space.
  - **Hard Negative Examples:** Opposite-class examples with high similarity scores are used to enhance the model's ability to distinguish between subtle confounder memes.
  - **In-Batch Negative Examples:** Randomly selected samples from the same batch with different labels introduce diversity in gradient signals.
- **Contrastive Objective:** RGCL optimizes a joint loss combining a contrastive loss and cross-entropy loss. This dual approach ensures improved classification performance by balancing embedding alignment and classification accuracy.
- **Dense Retrieval Integration:** Dense retrieval techniques are utilized to dynamically identify pseudo-gold positive and hard negative examples during training, ensuring that the model adapts effectively to evolving patterns in hateful memes.
- **K-Nearest Neighbor (KNN) Inference:** RGCL-trained embeddings allow for the use of a retrieval-based KNN majority voting classifier, enabling efficient system updates without retraining. This is particularly beneficial in real-world applications where online hate evolves rapidly.

RGCL's innovative methodology addresses critical gaps in multimodal hateful meme detection, offering a scalable, efficient, and accurate solution adaptable to real-world content moderation challenges.

C. Visual Question Answering (VQA) & Image Captioning

Visual Question Answering (VQA) [2]: We utilize the BLIP model to generate answers from images in response to carefully crafted questions, guided by insights from "Rethinking Visual Question Answering" (Chen et al., 2023). This approach leverages the alignment between visual context and targeted questions to enhance the model's ability to extract relevant information from images. By tailoring questions to the specific characteristics of the visual content, we enable the BLIP model to achieve a more nuanced understanding of the interplay between image and text elements. This refined question-answering framework significantly improves the model's capacity for accurate interpretation and classification of multimodal relationships, making it particularly effective in tasks that require precise contextual reasoning. The approach not only boosts performance but also underscores the importance of question design in maximizing the potential of visual question-answering systems.

Image Captioning: The BLIP model is also employed for image captioning, a process that generates descriptive text to encapsulate the visual content of an image. This involves identifying and articulating key elements such as objects, actions, contexts, and settings depicted in the image. By translating visual data into meaningful textual descriptions, image captioning facilitates a deeper understanding of the image's content and context. These generated captions play a crucial role in refining the analysis of image-text relationships, enabling a more seamless integration of visual

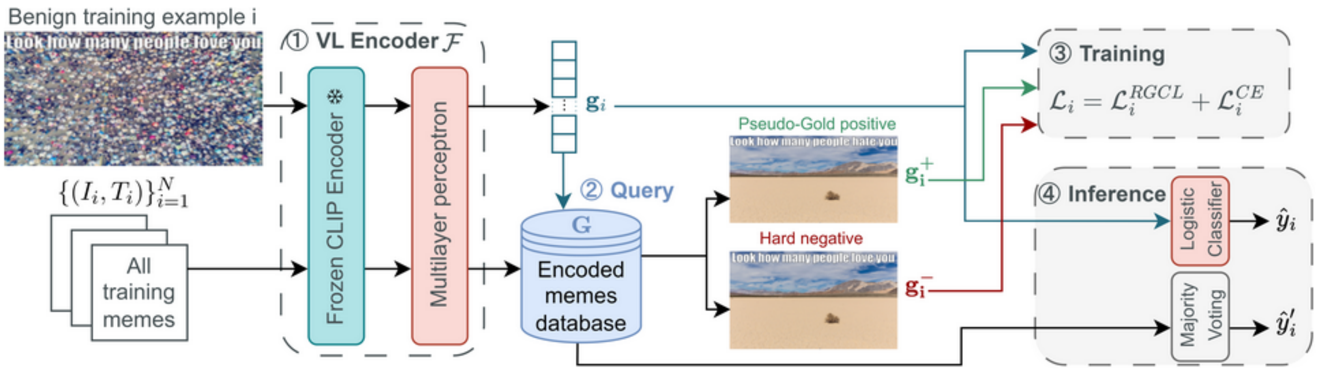

Fig. 2. Overview of RGCL.

and textual modalities. This not only supports tasks like classification but also enhances the performance of broader multimodal applications by providing rich, context-aware descriptions that bridge the gap between visual and linguistic representations.

D. Graph Attention Networks (GAT) [5]

Our methodology constructs a graph-based representation of multimodal data using embeddings derived from CLIP. In this representation, images, text, VQA question-answer pairs, and captions are treated as individual nodes. Two graph structures were explored to establish edges between these nodes:

1) **Threshold-Based Graphs:** Edges are formed between nodes if their cosine similarity exceeds a predefined threshold (0.7). This approach emphasizes strong relationships while reducing noise by pruning weaker connections.
2) **Fully Connected Graphs:** All nodes are connected, regardless of similarity. This configuration ensures that even weak but potentially informative relationships are preserved, supporting richer feature aggregation.

Experimental results highlighted that fully connected graphs outperformed threshold-based graphs, as retaining weaker connections allowed the model to capture more holistic relationships within the data.

For feature aggregation, we employ a Graph Attention Network (GAT), which computes attention scores to determine the relative importance of neighboring nodes. These scores enable weighted aggregation of features, allowing the model to focus on the most relevant connections. The aggregated features are then processed through one of two pooling strategies to generate a graph-level representation:

1) **Global Mean Pooling:** This method averages features across all nodes, treating them equally. It was found to be more effective in our setup, particularly with fully connected graphs, where all nodes provide meaningful information.
2) **Global Attention Pooling:** This approach assigns attention-based weights to nodes for feature aggregation. However, it underperformed in our experiments, probably because of the relatively uniform importance of nodes in the multimodal data.

The graph-level representation, derived through pooling, is passed through fully connected layers for binary classification, predicting whether a meme is harmful or not.

To evaluate the robustness of CLIP embeddings, we also experimented with alternative embeddings, including ResNet for images and DistilBERT for text. All embeddings were mapped to a uniform 512-dimensional space to ensure compatibility as graph node features, enabling seamless integration of multimodal data.

Our findings emphasize the importance of aligning graph structures and pooling strategies with the multimodal nature of the task. Fully connected graphs combined with global mean pooling emerged as the most effective configuration, leveraging both strong and weak relationships while maintaining computational simplicity. This framework effectively captures the diverse elements of memes, facilitating robust classification performance.

III. Experimental Results

A. Datasets

For our study, we used the HarMeme data set, which is made up of real-world COVID-19-related memes. Each meme in the dataset is annotated with one of three labels: very harmful, partially harmful, and harmless, based on the perceived degree of harm conveyed through the content. To streamline our analysis and consistent with previous work such as [5], the first two labels (very harmful and partially harmful) were merged into a single hateful class, resulting in a binary classification task. The HarMeme dataset consists of 3,013 memes in the training set and 354 memes in the test set.

This data set provides valuable information on the nuances of harmful online communication, making it particularly suitable for tasks like multi-modal hateful meme classification. Unlike synthetic datasets such as HMC, HarMeme features real-world examples ”in the wild,” offering a more realistic perspective on harmful meme detection.

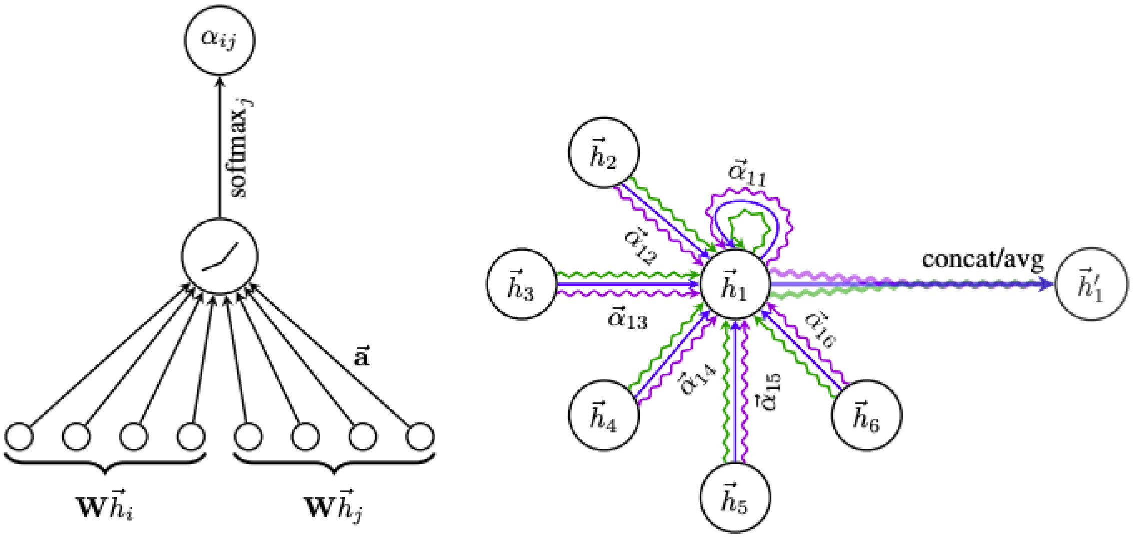

Fig. 3. Overview of GAT

B. Quantitative Results

To assess the performance of our proposed approaches for hateful meme detection, we conducted comprehensive experiments using Accuracy and the Area Under the Receiver Operating Characteristic Curve (AUROC) as evaluation metrics. These metrics were chosen to capture both the overall classification correctness and the model's ability to distinguish between hateful and non-hateful memes across varying decision thresholds. We evaluated a combination of range of methods, including caption generation, visual question answering, CLIP, and Graph Attention Networks. This section presents a detailed analysis of the results, highlighting the comparative strengths of these approaches and providing insights into the interplay between textual and visual modalities in hateful content detection.

Method	Caption	VQA	Acc. (%)	AUROC (%)
ResNet			61.58	59.83
DistillBERT			77.46	82.92
	✓		80.79	82.40
		✓	80.79	82.19
	✓	✓	79.66	80.43

TABLE II
PERFORMANCE COMPARISON OF VARIOUS METHODS FOR HATEFUL MEME DETECTION USING ACCURACY (ACC.) AND AUROC AS EVALUATION METRICS.

Unimodal Approach: ResNet alone demonstrates limited performance, achieving an accuracy of 61.58% and an AUROC of 59.83%, indicating its inability to effectively capture the nuanced relationships between visual and textual features. On the other hand, DistillBERT significantly outperforms ResNet when used alone, achieving 77.46% accuracy and 82.92% AUROC, showcasing the strength of text-based analysis in this task.

Integrating captions and visual question answering (VQA) further enhances DistillBERT's performance, with captions alone increasing accuracy to 80.79%, though there is a slight dip in AUROC to 82.40%. Similarly, adding VQA achieves comparable accuracy (80.79%) with a slight decrease in AUROC (82.19%). However, combining both captions and VQA leads to a marginal decline in performance, with accuracy dropping to 79.66% and AUROC to 80.43%. This suggests that while captions and VQA individually contribute to performance gains, their simultaneous integration may introduce redundancy or noise that slightly affects the model's efficacy.

Graph Attention Networks: Graph Attention Networks (GAT) leverage attention mechanisms to dynamically assign importance to graph edges, enabling effective learning of relationships in complex data. In this study, graph construction methods and pooling strategies were evaluated to assess their alignment with GAT principles and their performance in hateful meme detection.

Method	Acc. (%)	AUROC (%)
Cosine Similarity + Mean Pooling	81.41	85.85
Cosine Similarity + Attention Pooling	80.28	82.74
Fully Connected + Mean Pooling	81.36	88.95
Fully Connected + Attention Pooling	79.66	88.05

TABLE III
COMPARISON OF GRAPH CONSTRUCTION METHODS (COSINE SIMILARITY VS. FULLY CONNECTED) AND POOLING STRATEGIES (MEAN POOLING VS. ATTENTION POOLING) BASED ON ACCURACY AND AUROC.

Cosine Similarity with Mean Pooling achieved an accuracy of 81.41% and an AUROC of 85.85%, highlighting the efficacy of static edge relationships when combined with a straightforward pooling method. However, when paired with Attention Pooling, the performance dropped to 80.28% accuracy and 82.74% AUROC. This suggests that Attention Pooling may not effectively leverage the static graph structure or is less robust in this setup compared to Mean Pooling, which offers a simpler and more reliable aggregation strategy.

Fully Connected graphs, which inherently model dense relationships, performed consistently well. Mean Pooling achieved an accuracy of 81.36% and an AUROC of 88.95%, slightly outperforming Cosine Similarity. Attention Pooling in Fully Connected graphs maintained a high AUROC of 88.05%, though accuracy slightly decreased to 79.66%. These results suggest that while attention-based pooling can capture subtle feature interactions, the simplicity and robustness of Mean Pooling often provide more reliable aggregation.

These findings indicate the potential for incorporating GAT-inspired mechanisms into graph-based hateful meme detection frameworks. By introducing dynamic edge weighting and end-to-end optimization, future models could further enhance the ability to capture multimodal relationships and improve overall performance.

GAT-Enhanced Multimodal Detection: This study evaluates the performance of combining unimodal models (ResNet + DistillBERT) and a native multimodal approach (CLIP) for hateful meme detection, with both enhanced using Graph Attention Networks (GAT) to improve multimodal interaction. The analysis investigates how captions and visual question answering (VQA) contribute to the performance of these methods.

Method	Caption	VQA	Acc. (%)	AUROC (%)
ResNet + DistillBERT			75.14	78.60
	✓		75.14	80.07
		✓	77.97	78.30
	✓	✓	76.65	79.69
CLIP			83.33	89.28
	✓		83.62	90.96
		✓	84.75	90.94
	✓	✓	81.36	88.95

TABLE IV
PERFORMANCE EVALUATION OF COMBINED RESNET AND DISTILLBERT MODELS AND THE CLIP MODEL FOR HATEFUL MEME DETECTION

The ResNet + DistillBERT combination demonstrates moderate performance, with a baseline accuracy of 75.14% and AUROC of 78.60%. Adding captions improves AUROC to 80.07% without affecting accuracy, while incorporating VQA alone raises accuracy to 77.97% but slightly reduces AUROC to 78.30%. The inclusion of both captions and VQA provides a balance, achieving 76.65% accuracy and 79.69% AUROC. These results suggest that GAT can enhance feature integration in unimodal combinations, but the limitations of the underlying models constrain its ability to fully exploit multimodal relationships.

In contrast, CLIP, as a multimodal model, consistently outperforms the unimodal combination. Without additional inputs, it achieves 83.33% accuracy and 89.28% AUROC. Incorporating captions boosts AUROC to 90.96% and accuracy to 83.62%, while VQA integration achieves the highest accuracy of 84.75% and AUROC of 90.94%. However, the simultaneous inclusion of captions and VQA results in a slight drop to 81.36% accuracy and 88.95% AUROC, possibly due to redundancy or conflicting signals in the inputs. GAT effectively leverages CLIP's multimodal design, making it highly efficient in capturing cross-modal relationships.

These findings highlight the importance of multimodal frameworks like CLIP in hateful meme detection. While GAT provides notable improvements in both unimodal and multimodal approaches, its synergy with inherently multimodal models like CLIP offers a more robust solution for leveraging textual and visual modalities.

ISSUES & RGCL: ISSUES achieves an accuracy of 81.64% and an AUROC of 92.83%, demonstrating its strong ability to identify hateful memes. Similarly, RGCL achieves robust results, with an accuracy of 84.75% and an AUROC of 91.32%. These results indicate that both methods independently perform well and are effective at leveraging multimodal information.

Method	Acc. (%)	AUROC (%)
ISSUES	81.64	92.83
RGCL	84.75	91.32
ISSUES + RGCL	66.38	64.08

TABLE V
COMPARISON OF ISSUES, RGCL, AND THEIR COMBINATION FOR HATEFUL MEME DETECTION.

However, when the two methods are combined (ISSUES + RGCL), the performance drops significantly to an accuracy of 66.38% and an AUROC of 64.08%. This sharp decline suggests potential issues in the integration process. It is possible that the implementation of the combination was not executed effectively, leading to suboptimal interaction between the two methods.

IV. CONCLUSION

In this study, we tackled the challenging task of hateful meme detection by exploring diverse approaches that leverage both textual and visual modalities. Through methods such as text-only and image-only models, caption generation, visual question answering, and advanced frameworks like the CLIP model and Graph Attention Networks, we demonstrated the multifaceted nature of this problem.

Our findings highlight the importance of integrating textual and visual cues for effective hateful content detection, as neither modality alone suffices to capture the nuanced interplay between text and imagery. Advanced models like CLIP and Graph Attention Networks showed significant potential by leveraging cross-modal interactions, underscoring the promise of multimodal learning in addressing complex social challenges

Performance in multimodal tasks is largely determined by how well the embedding space is optimized. Inspired by the sampling techniques and projection layers in RGCL and ISSUE, which adjusted embedding spaces based on CLIP, we applied similar ideas to the GAT model. Specifically, we propose exploring approaches where embeddings generated by ISSUE or RGCL, finalized just before classification, are used as node inputs for GAT. This methodology, which effectively adjusts the embedding space for multimodal representation, remains a promising avenue for future work.

As hateful content continues to evolve, future work could focus on refining these models, incorporating larger datasets, and exploring more sophisticated multimodal architectures. By advancing these efforts, we aim to contribute to a safer and more inclusive digital space.

REFERENCES

[1] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. D. Bimbo, “Mapping Memes to Words for Multimodal Hateful Meme Classification,” Oct. 2023, arXiv:2310.08368. [Online]. Available: http://arxiv.org/abs/2310.08368

[2] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, and J. Jiang, “Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection,” Aug. 2023, arXiv:2308.08088. [Online]. Available: http://arxiv.org/abs/2308.08088

[3] S. Sharma, A. Kulkarni, T. Suresh, H. Mathur, P. Nakov, M. S. Akhtar, and T. Chakraborty, “Characterizing the Entities in Harmful Memes: Who is the Hero, the Villain, the Victim?” Apr. 2023, arXiv:2301.11219. [Online]. Available: http://arxiv.org/abs/2301.11219

[4] J. Mei, J. Chen, W. Lin, B. Byrne, and M. Tomalin, “Improving Hateful Meme Detection through Retrieval-Guided Contrastive Learning,” Oct. 2024, arXiv:2311.08110. [Online]. Available: http://arxiv.org/abs/2311.08110

[5] P. Velivckovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” 2017, arXiv:1710.10903v3. [Online]. Available: https://arxiv.org/abs/1710.10903v3

[6] J. Chen, J. Mei, W. Lin *et al.*, “Rethinking visual question answering,” *arXiv preprint arXiv:2308.08088*, 2023, accessed November 2024. [Online]. Available: https://arxiv.org/pdf/2308.08088

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” Feb. 2021, arXiv:2103.00020 [cs]. [Online]. Available: http://arxiv.org/abs/2103.00020