Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

A Knowledge Augmented Framework for Multimodal News Object-Entity Relation Extraction

PeiLing Li[1,2] 15073131265@163.COM and **Lin Li**[1,2,3,4*] LILIN20081@SOHU.COM
[1]*School of Computer Science, Qinghai Normal University, Xining, Qinghai, 810008, China*
[2]*The State Key Laboratory of Tibetan Intelligence, Xining, Qinghai, 810008, China*
[3]*Academy of Plateau Science and Sustainability, Qinghai Normal University, Xining,Qinghai, 810008, China*
[4]*Qinghai Sub-center, National Qinghai-Tibet Plateau Scientific Data Center, Xining, Qinghai, 810008, China*
[*]*Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Multimodal relation extraction, as an important research direction in the field of information extraction, aims to identify entities and objects from both text and images and establish cross-modal semantic associations. Current mainstream methods still face challenges in handling complex multimodal data, such as semantic alignment confusion and redundant associations, which lead to erroneous associations between irrelevant entities and objects, severely affecting system performance. To address this issue, this paper proposes a multimodal relation extraction framework that integrates knowledge graphs. This approach uses the knowledge graph as external semantic support to filter candidate entity-object pairs through structured semantic information, and leverages a multimodal alignment module to achieve precise semantic matching. Experimental results show that this method significantly outperforms existing methods on multiple benchmark datasets, especially in fine-grained relation recognition, where the F1 score increases by 4 percentage points, effectively demonstrating the framework's ability to mitigate cross-modal noise interference.

Keywords: Multimodal, Knowledge Graph, Text, Image, Relation Extraction

1. Introduction

The core challenge in multimodal relation extraction tasks arises from the noise interference during the cross-modal entity-object alignment process. When there are numerous entities and objects in the text description and the corresponding image, the number of candidate relation pairs that the system needs to handle grows combinatorially. This results in a large number of irrelevant entity-object pairs being misjudged as having semantic associations, severely affecting the accuracy of relation extraction. This challenge places higher demands on the system's semantic recognition capabilities. To address this issue, this paper proposes a multimodal relation extraction framework that integrates knowledge graphs to improve the accuracy and robustness of semantic alignment. The model deeply integrates visual-language and knowledge graph information, constructing a cross-modal representation learning system with semantic completion capabilities. The key innovations lie in: designing an adaptive knowledge graph module to supplement the missing semantic information between the image and text; establishing a knowledge-guided attention mechanism to effectively distinguish real associations from noise interference. This solution significantly improves the accuracy of cross-modal semantic alignment and provides new research ideas for multimodal relation extraction tasks.

In multimodal relation extraction tasks, the presence of multiple entities and objects in both text and images can lead to a large number of candidate relation pairs, significantly increasing computational complexity. Many of these pairs are irrelevant, especially in many-to-many scenarios like news headlines and their associated images, where incorrect alignments introduce noise and reduce extraction accuracy. Moreover, short texts often lack sufficient context, further weakening semantic understanding.

To address these challenges, this paper proposes a knowledge graph-enhanced multimodal relation extraction model. The model leverages a CLIP-based embedding layer to capture joint image-text representations, and integrates an external knowledge graph to supplement implicit semantic links between entities and objects. By enriching semantic information and guiding the alignment process, the model effectively filters out noisy pairs and improves both accuracy and robustness. The key innovation lies in the use of knowledge graphs to bridge semantic gaps in multimodal data and support more reliable relation reasoning.

2. Related Work

2.1. Unimodal Relation Extraction

Single-modality relation extraction primarily focuses on identifying and extracting semantic relationships between entities from text. Traditional text-based relation extraction methods were largely based on rules or feature engineering. However, with the advancement of deep learning techniques, end-to-end neural network-based approaches have become mainstream.

Classic Methods: Early relation extraction methods (such as those based on Support Vector Machines (SVM) and Conditional Random Fields (CRF)) relied on handcrafted features and required large amounts of annotated data for training.

Deep Learning Methods: In recent years, deep learning-based approaches have made significant progress in single-modality relation extraction. For instance, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been widely applied to relation extraction tasks in text, enabling models to automatically learn useful features from data. Zeng et al. (2014) proposed a relation extraction method based on convolutional neural networks, which could automatically extract effective features from text.

BERT and Its Variants: BERT (Bidirectional Encoder Representations from Transformers) and its variants (such as RoBERTa, DistilBERT, etc.) have become standard models in text-based relation extraction. Methods like BERT+CRF have proven effective in capturing entities and relationships in text.

2.2. Multimodal learning

Multimodal learning aims to integrate data from different modalities (such as images, text, and videos) to accomplish more complex tasks. Multimodal learning methods can be categorized into early fusion, late fusion, and intermediate fusion.

Early Fusion: This approach combines different modalities (e.g., images and text) at the feature level and trains a unified model for learning. For example, Lu et al. (2019) proposed ViLBERT, which employs a dual-stream architecture to process text and visual information separately before fusing them for multimodal learning.

Late Fusion: Here, different modalities are processed independently, and their outputs are merged at the final decision stage. For instance, Chen et al. (2019) introduced UNITER, which separately models image and text features before performing fusion during inference.

Intermediate Fusion: This method integrates multimodal information at intermediate layers of the model, allowing continuous interaction and enhancement between modalities to improve representation learning. A typical example is VisualBERT (Li et al., 2020), which embeds visual information into Transformer layers to jointly learn semantic relationships between images and text.

2.3. Multimodal Entity-Relation Extraction

Multimodal Entity-Relation Extraction (MERE) aims to extract entities from both images and text while identifying their relationships. Compared to unimodal tasks, this task involves modeling interactions between visual and textual data, presenting greater challenges.

Key Approaches in Multimodal Relation Extraction

M2ER (Multimodal Entity-Relation Extraction): Zhang et al. (2023) proposed M2ER, which jointly models entities in images and text using cross-modal self-attention mechanisms to establish relationships. By enabling cross-modal learning, it effectively improves the accuracy of entity-relation extraction.

Visual-Textual Relation Extraction: Lin et al. (2020) introduced a joint model that captures visual cues from images and linguistic patterns from text for relation extraction. The model employs visual attention mechanisms to focus on key regions in images and aligns them with textual information.

MMRE (Multimodal Relation Extraction): Wang et al. (2020) developed a method that combines visual and textual semantics to recognize relations between entities. Their model integrates visual features (CNN/ResNet) with textual embeddings (BERT/Transformer) to enhance multimodal relation understanding.

2.4. Other related studies and applications

Visual Question Answering (VQA): VQA represents a classic multimodal task, where researchers have developed various methods to answer image-related questions by integrating visual and textual information. Although VQA primarily focuses on question answering, its methodologies and techniques can be effectively applied to multimodal entity-relation extraction tasks.

Image-Text Pretraining: In recent years, numerous vision-language pretrained models (e.g., CLIP (Radford et al., 2021), BLIP, ALIGN (Jia et al., 2021)) have established robust semantic connections between images and text, emerging as pivotal technologies in multimodal learning. These pretrained models not only facilitate image-text matching but also provide unified representations for entities in both modalities, enhancing performance in entity-relation extraction tasks.

3. Model Design

Figure 1 illustrates the overall architecture of the RelFormer network. RelFormer is a multimodal relation extraction model based on the encoder-decoder paradigm. It comprises an encoder, a decoder, and an RelModule module, each of which will be elaborated upon in the subsequent sections.

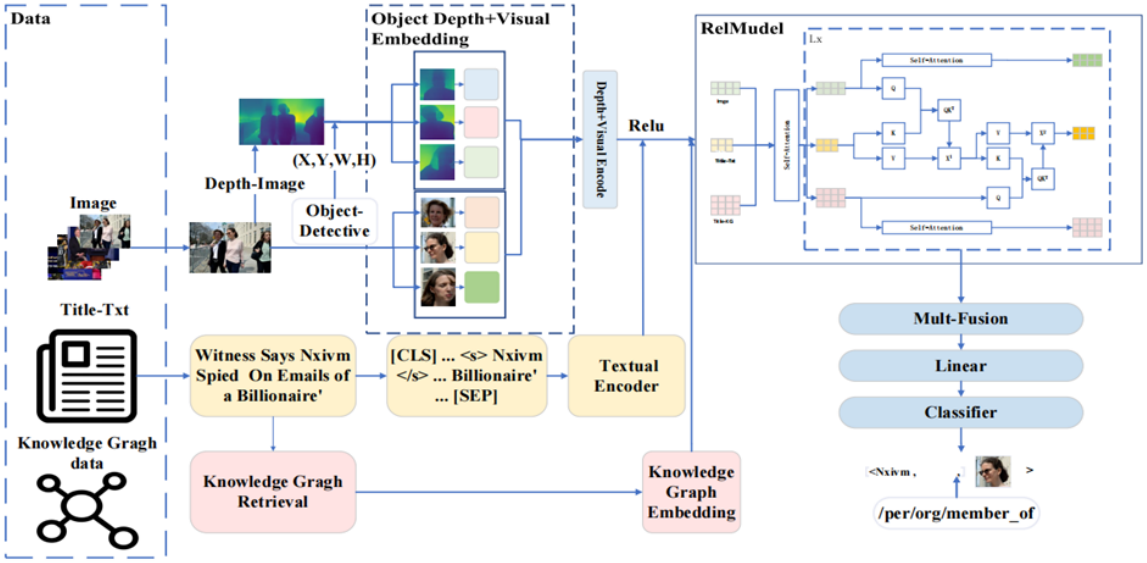

Figure 1: Overview of the proposed RelFormer

3.1. Image object recognition and depth estimation

In this study, the original image undergoes two independent processing pipelines to extract rich visual feature information. The image is first processed through the YOLO (You Only Look Once) object detection model to identify objects present in the image and obtain their positional information (e.g., bounding box coordinates). Subsequently, the detected objects and their positional information are encoded into embeddings to generate object feature representations. This step aims to leverage object detection technology to extract key object information from the image, providing structured visual input for subsequent multimodal fusion.

Concurrently, the original image is processed by a depth estimation model to generate a depth map, where each pixel value represents the distance of the corresponding object from the camera lens. Depth estimation provides the model with rich spatial information, facilitating the understanding of relative positional relationships between objects and the three-dimensional structure of the scene. To enhance the quality of the depth map, this study applies a series of post-processing operations, including denoising (e.g., Gaussian filtering), smoothing (e.g., median filtering), and normalization (mapping depth values to the [0, 1] range). These operations aim to reduce noise interference in the depth map and improve the reliability of depth information.

After obtaining the bounding box coordinates from YOLO, this study further processes the depth map using these coordinates. Based on the detected object bounding boxes, depth information from corresponding regions in the depth map is extracted and encoded into embeddings. This step not only preserves the spatial location information of objects but also tightly integrates it with depth information, thereby providing the model with more precise spatial distribution features of objects.

The object features extracted by the detection module and the depth features generated by the depth estimation module are then fused through embedding. The two types of features are combined to produce a unified visual feature representation. This fusion approach retains both the semantic

information of objects and incorporates spatial location information, thereby providing more comprehensive visual feature support for subsequent multimodal relation extraction tasks.

Through the above processing pipeline, this study achieves joint extraction and fusion of semantic and spatial location information of objects in images, establishing a solid visual feature foundation for multimodal relation extraction tasks. The introduction of depth maps not only enhances the model's ability to understand the 3D structure of scenes but also improves the accuracy of object spatial distribution by incorporating YOLO's bounding box information.

3.2. Text and knowledge graph embedding module

This study employs deep learning-based natural language processing techniques to perform multi-level feature extraction and semantic enhancement on news headline texts. Specifically, the text input is first processed through a BERT-CRF joint model for entity recognition and feature encoding. The pre-trained BERT (Bidirectional Encoder Representations from Transformers) model, with its deep bidirectional Transformer architecture, effectively captures the contextual semantic information of entities within the text; meanwhile, the Conditional Random Field (CRF) layer performs sequence optimization on the entity recognition results, ensuring the accuracy of entity boundary identification and the coherence of label prediction. This joint model can accurately recognize named entities such as person names, locations, and organization names in news headlines and encode them into low-dimensional dense vector representations, forming semantically rich entity embedding features.

To further enhance the model's semantic understanding ability, this study innovatively introduces external knowledge graphs for knowledge enhancement. The specific implementation process is as follows: first, based on the entity set identified by the BERT-CRF model and the full content of the news headline, knowledge retrieval queries are constructed; then, through API access to large-scale structured knowledge bases (such as Wikidata, DBpedia, etc.), background knowledge related to the current news topic is retrieved, including but not limited to entity attribute descriptions, relational triples between entities (e.g., ¡Beijing, is the capital of, China¿), and entity category information. These auxiliary pieces of information obtained from the knowledge graph are encoded into vectors and deeply fused with the original text features, forming knowledge-enhanced text representations.

The introduction of the knowledge graph provides critical prior knowledge support for the model. Taking the entity "Beijing" in a news headline as an example, the knowledge graph not only provides basic attribute information (such as geographical location, administrative level, etc.) but also reveals its associations with other entities (e.g., "contains landmark buildings: Tiananmen, Great Wall"). This structured knowledge can effectively compensate for the lack of contextual information in short texts, helping the model establish semantic associations between text entities and visual objects in images. For instance, when the visual object "Tiananmen" is detected in the image, the model can leverage the relational triple "Beijing–contains–Tiananmen" from the knowledge graph to more accurately infer the spatial containment relation between "Beijing" in the headline and the image content. This knowledge-guided multimodal feature fusion mechanism significantly enhances the performance of relation extraction tasks, enabling the model to maintain high recognition accuracy and robustness even in complex scenarios.

3.3. RelModule

3.3.1. SELF-ATTENTION MODULE

The encoder in this study adopts a hierarchical attention architecture. In the first 8 layers, during the intra-modal feature encoding stage, features from each modality are deeply processed through independent self-attention paths: for the textual modality, a multi-layer stacked token-level self-attention mechanism is used to gradually model long-range semantic dependencies and construct globally context-aware text representations; for the visual modality, a ViT-based spatial attention mechanism is employed, where the input image is first divided into regular grid feature regions, and spatial associations and visual semantic relationships between different regions are then established to form spatially-aware visual representations; for the knowledge modality, a two-stage processing strategy is adopted: first, a Graph Attention Network (GAT) encodes entity nodes and relation edges in the knowledge graph to capture structured information, and then a Transformer module learns high-order interaction relationships between entities, ultimately producing rich structured knowledge representations. This stage ensures that each modality obtains high-quality internal representations through sufficient intra-modal feature learning, laying a solid foundation for subsequent cross-modal interaction.As illustrated in Figure 2, the self-attention module is depicted, where the internal structure of attention is detailed to show how semantic features are captured and aggregated across modalities.The self-attention mechanisms of the text, image, and knowledge-enhancement modules are shown in Equation (1).

$$\text{SelfAttn}\,(X) = \text{Softmax}\left(\frac{XW_q(XW_k)^T}{\sqrt{d}}\right)XW_v \tag{1}$$

Among them, $X \in \{H_t, H_v, H_k\}$ they respectively represent the input features of the text, visual, and knowledge modalities, and W_q, W_k, $W_v \in R^{dxd_h}$ is the projection matrix.

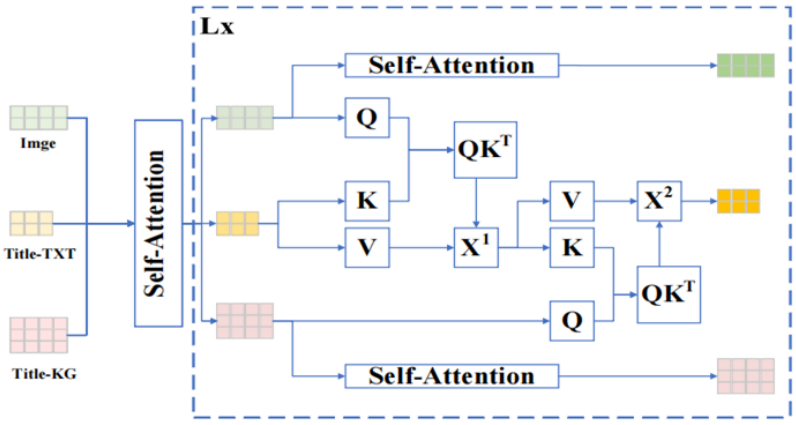

Figure 2: The structure diagram of RelModule

3.3.2. CROSS-ATTENTION MECHANISMS

In the high-level encoding stage (Layers 9–12), the system achieves deep feature fusion through an innovative bidirectional cross-modal attention mechanism. Specifically, the text-visual cross-attention module adopts a query-key-value attention architecture, where text features serve as the

query vectors and visual features as the key-value pairs. By computing cross-modal attention weights, the model achieves precise alignment between textual semantics and visual content, as shown in Equation (2). This mechanism dynamically establishes associations between concepts in the textual description (e.g., "red sports car") and corresponding visual features (such as color and shape) in the image.

$$H_t^{(l)} = LayerNorm\left(CrossAttn\left(H_v^{(l-1)}, H_t^{(l-1)}, H_t^{(l-1)}\right) + H_t^{(l-1)}\right) \tag{2}$$

Meanwhile, the text-knowledge cross-attention module operates under a similar mechanism, retrieving relevant entities from the knowledge graph based on the textual context (e.g., the concept "quantum computing" corresponding to "qubit," etc.), significantly enhancing the semantic depth of text representations, as shown in Equation (3). It is worth noting that this stage adopts a dual-path design: while performing cross-modal interactions, it also retains the self-attention processing paths of each modality, effectively preserving the feature specificity of the visual and knowledge modalities. This innovative architecture not only ensures deep semantic alignment across modalities but also avoids feature confusion.Figure 2 (right) provides a detailed illustration of the multimodal cross-attention module, where the internal structure reveals how features from different modalities are aligned and fused through attention.

$$H_k^{(l)} = LayerNorm\left(CrossAttn\left(H_t^{(l-1)}, H_k^{(l-1)}, H_k^{(l-1)}\right) + H_k^{(l-1)}\right) \tag{3}$$

3.4. Multimodal fusion

In the Multimodal Fusion module, this study performs unified integration of features from text, image, and knowledge graph modalities. Specifically, the feature representations from the three modalities are first preliminarily fused through concatenation or weighted summation. Subsequently, a Multi-Layer Perceptron (MLP) is used to further refine the fused features, as shown in Equation (4), generating a unified multimodal feature representation.

$$F' = \text{LayerNorm}\left(F + \text{MLP}\left(F\right)\right) \tag{4}$$

This multimodal fusion approach not only preserves the semantic information of text, image, and knowledge graph individually, but also incorporates cross-modal interaction information through the cross-attention mechanism, thereby providing more comprehensive and enriched feature support for the subsequent relation extraction task. The multimodal fusion module enhances the model's ability to capture semantic associations among text, image, and knowledge graph modalities, thus achieving better performance in relation extraction.

4. Experiment

4.1. Dataset

This study uses the publicly available MORE DATASET (He et al., 2023), which contains 21 different relation types and covers 20,264 multimodal relation facts annotated on 3,559 pairs of text headlines and corresponding images. In addition, the dataset includes 13,520 visual objects, with an average of 3.8 objects per image. To compensate for the sparsity of textual information and enhance semantic modeling capability, this study introduces a knowledge graph enhancement module. This

module analyzes the news text content and automatically constructs structured queries to access authoritative knowledge bases. The system retrieves relation triples and key content of the news from the knowledge base. These structured knowledge elements provide the model with rich background information.

4.2. Experimental Design and Environment

This study conducted systematic comparative experiments under the PyTorch framework to evaluate the performance of the proposed model. Several state-of-the-art multimodal models—including CNN-LSTM, BERT+Scene Graph, VisualBERT, and MoreFormer—were selected as baselines for horizontal comparison. All experiments were implemented using PyTorch and conducted on an NVIDIA A100 GPU. The models were trained with a batch size of 32, a learning rate of 2e-5, and up to 60 epochs using the AdamW optimizer, with a warmup ratio set to 0.01.

4.3. Experimental Conclusion

Experimental results reveal significant performance differences among various models in the multimodal relation extraction task. Although the traditional CNN-LSTM architecture achieved an accuracy of 60.19%, its F1-score of 34.08% and the imbalance between precision (30.10%) and recall (39.28%) indicate clear limitations in cross-modal semantic understanding when using simple feature concatenation methods. The BERT+Scene Graph model improved recall to 41.27% by incorporating scene graph structures, but its precision of 29.61% suggests that expanding coverage also introduced a considerable number of false positives.

Transformer-based models such as VisualBERT and MoreFormer demonstrated superior performance. VisualBERT achieved an accuracy of 82.84% and an F1-score of 59.66%, with relatively balanced precision (58.18%) and recall (61.22%), validating the effectiveness of a unified attention mechanism. In comparison, the proposed Re1Former model achieved the best performance across all evaluation metrics: an accuracy of 85.41%, an F1-score of 63.52%, and a highly balanced precision (64.65%) and recall (62.44%) with a margin of only 2.21 percentage points. A statistical significance test ($p < 0.01$) confirmed that Re1Former's performance improvement over existing methods is statistically significant.As shown in Table 1, these experimental results demonstrate the effectiveness of the three key innovations proposed in this study:The hierarchical attention architecture separates intra-modal feature learning from cross-modal interaction, effectively avoiding feature confusion;The knowledge graph enhancement module provides structured knowledge that significantly improves the model's ability to understand complex semantic relations;The bidirectional cross-modal attention mechanism enables more precise semantic alignment.

This study offers a novel technical solution for the multimodal relation extraction task and holds important application value in areas such as intelligent media analysis and knowledge graph construction. Future research will focus on optimizing knowledge retrieval efficiency and exploring the model's scalability in multilingual settings. Among the baseline models—CNN-LSTM, BERT+Scene Graph, VisualBERT, and MoreFormer—comparative experiments showed that Table 1 highlights the overall performance distinctions and the superiority of RelFormer.

Table 1: The overall performance of RelFormer and other state-of-the-art methods on the More+ Knowledge Graph dataset (The bold values in each column represent the best entries).

Model	Accuracy	Precision	Recall	F1-Score
CNN-LSTM	60.19	30.10	39.28	34.08
BERT+Scene Graph	61.79	29.61	41.27	34.48
VisualBERT	82.84	58.18	61.22	59.66
MoreFormer	81.72	58.74	60.35	59.53
RelFormer	**85.41**	**64.65**	**62.44**	**63.52**

5. Conclusion

This study proposes a knowledge-enhanced multimodal relation extraction model, Re1Former, which achieves effective integration of textual, visual, and knowledge graph features through a hierarchical attention mechanism. Experimental results show that the model significantly outperforms existing approaches on key metrics such as accuracy (85.41%) and F1-score (63.52%), while maintaining a good balance between precision (64.65%) and recall (62.44%). The model innovatively adopts a hierarchical attention architecture to separate intra-modal feature learning from cross-modal interaction, introduces a knowledge graph enhancement module to provide structured knowledge support, and implements a bidirectional cross-modal attention mechanism for precise semantic alignment. These components jointly construct an efficient multimodal relation extraction framework that effectively alleviates feature confusion and enhances semantic understanding. This study provides a novel solution for multimodal relation extraction tasks. Future research may explore the following directions: (1) optimizing knowledge retrieval efficiency; (2) designing dynamic knowledge selection mechanisms to improve adaptability; and (3) extending the model to multilingual scenarios.

References

Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. doi: 10.48550/arXiv.1909.11740.

L. He, H. Wang, Y. Cao, Z. Wu, J. Zhang, and X. Dai. More: A multimodal object-entity relation extraction dataset with a benchmark evaluation. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM 2023)*, pages 4564–4573, 2023. doi: 10.48550/arXiv.2312.09753.

C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. doi: 10.48550/arXiv.2102.05918.

L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2020. doi: 10.48550/arXiv.1908.03557.

J. Lin, A. Yang, Y. Zhang, J. Liu, J. Zhou, and H. Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*, 2020. doi: 10.48550/arXiv.2003.13198.

J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019. doi: 10.48550/arXiv.1908.02265.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and I. Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. doi: 10.48550/arXiv.2103.00020.

Y. Wang, B. Yu, Y. Zhang, T. Liu, H. Zhu, and L. Sun. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*, 2020. doi: 10.18653/v1/2020.coling-main.138.

D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, pages 2335–2344, 2014. doi: 10.18653/v1/C14-1220.

B. Zhang, X. Yang, G. Wang, Y. Wang, and R. Sun. M2er: Multimodal emotion recognition based on multi-party dialogue scenarios. *Applied Sciences*, 13(20):11340, 2023. doi: 10.3390/app132011340.