applied sciences

MDPI

Article

Using Augmented Small Multimodal Models to Guide Large Language Models for Multimodal Relation Extraction

Wentao He [1], Hanjie Ma [1,*], Shaohua Li [1], Hui Dong [2], Haixiang Zhang [1] and Jie Feng [1]

[1] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 202130504089@mails.zstu.edu.cn (W.H.); 202120501008@mails.zstu.edu.cn (S.L.)
[2] Hangzhou Codvision Technology Co., Ltd., Hangzhou 311100, China; donghui@codvision.com
* Correspondence: mahanjie@zstu.edu.cn

Abstract: Multimodal Relation Extraction (MRE) is a core task for constructing Multimodal Knowledge images (MKGs). Most current research is based on fine-tuning small-scale single-modal image and text pre-trained models, but we find that image-text datasets from network media suffer from data scarcity, simple text data, and abstract image information, which requires a lot of external knowledge for supplementation and reasoning. We use Multimodal Relation Data augmentation (MRDA) to address the data scarcity problem in MRE, and propose a Flexible Threshold Loss (FTL) to handle the imbalanced entity pair distribution and long-tailed classes. After obtaining prompt information from the small model as a guide model, we employ a Large Language Model (LLM) as a knowledge engine to acquire common sense and reasoning abilities. Notably, both stages of our framework are flexibly replaceable, with the first stage adapting to multimodal related classification tasks for small models, and the second stage replaceable by more powerful LLMs. Through experiments, our EMRE2llm model framework achieves state-of-the-art performance on the challenging MNRE dataset, reaching an 82.95% F1 score on the test set.

Keywords: multimodal relation extraction; small multimodal guidance; multimodal relation data augmentation; flexible threshold loss; large language model

Citation: He, W.; Ma, H.; Li, S.; Dong, H.; Zhang, H.; Feng, J. Using Augmented Small Multimodal Models to Guide Large Language Models for Multimodal Relation Extraction. *Appl. Sci.* **2023**, *13*, 12208. https://doi.org/10.3390/app132212208

Academic Editor: Chilukuri K. Mohan

Received: 14 September 2023
Revised: 31 October 2023
Accepted: 7 November 2023
Published: 10 November 2023

1. Introduction

Relation Extraction (RE) aims to identify semantic relations between two entities in a given sentence, and it is one of the important tasks of Information Extraction (IE) and Knowledge Graphs (KG). KG are the underlying technologies for recommendation systems [1,2], information retrieval [3,4] and link prediction [5]. Since most of the data information on the Internet is in multimodal forms such as text and images, multimodal technologies and communities have seen rapid development recently. MKGs as an interpretable platform for multimodal data [6] have promoted the development of new technologies, such as recommendation systems on multimodal knowledge images [7], multimodal analogical reasoning [8] and multimodal product knowledge image link prediction [9]. The most critical part of constructing MKGs is ingesting unstructured image-text data. Visual data can be considered as supplementary information for multimodal knowledge graph completion (MKGC) tasks [10–12]. Technically, this is achieved through MRE techniques to acquire triplet information from multimodal data.

MRE significantly expands text-based models by taking images as additional input, since visual context helps resolve ambiguous polysemous words. As shown in Figure 1, visual information from images provides supplementary knowledge for complementing, understanding and reasoning over text. Especially for image-text datasets from Internet media, short texts and auxiliary images related to the text, pure text alone cannot offer complete and accurate information. Visual context helps resolve ambiguous textual information and provide explanatory entity information.

Text:

RT @Showbiz_IT : Kapil Sharma to team up with rival Krushna Abhishek for new show ?

Relation: /per/per/peer

Figure 1. An example of image-text multimodal data from MEGA. For an entity pair appearing in the text, the accompanying image that supplements the textual information is used to determine the relation between the entity pair.

In recent years, LLMs have undergone rapid development, demonstrating excellent capabilities in contextual reasoning, instruction following, and linguistic reasoning. These models can perform various complex language tasks in a zero-shot manner. Meanwhile, models like PICa [13], MiniGPT-4 [14], LLaMA-Adapter [15], and Prophet [16] leverage LLMs as natural language knowledge engines to accomplish advanced language abilities that smaller models cannot achieve. Inspired by these models, we employ LLMs as knowledge supplementation and reasoning engines to obtain triplet information that smaller models fail to capture. Before that, we further adapt the smaller models to fit the characteristics of media data, addressing insufficient data information and uneven relation data distribution. The enhanced smaller models are then utilized as a guiding model to obtain guiding answers for MRE, enabling the LLM engine to function smoothly. In summary, the contributions of this paper can be outlined as follows:

- We propose a data augmentation method in multimodal relation extraction (MRDA) by analyzing the characteristics of media datasets, which solves the performance bottleneck caused by insufficient data in MRE tasks.
- We improve the threshold strategy in multimodal fields and propose the Flexible Threshold Loss (FTL) to better handle the problems of uneven entity pair number distribution and long-tail classes.
- For the first time, we introduce LLMs as knowledge engines in MRE, enabling LLMs to assist MRE by implementing smaller models as guiding models and image-to-text prompting tasks.
- Experiments on the public MNRE dataset demonstrate that our model can effectively extract multimodal relations from Internet media data, and significantly outperforms current state-of-the-art models under standard supervised and low-resource settings.

2. Related Work

2.1. Multimodal Relation Extraction

In recent years, an annotated MRE dataset MNRE and baseline models have been constructed and released, demonstrating that utilizing visual information can improve Relation Extraction (RE) capabilities for social media text. The authors of MEGA [17] further corrected the MNRE dataset and proposed a dual image alignment method to capture aligned information between visual object relations and textual object relations, making better use of visual object relation information. The authors of FL-MSRE [18] proposed using few-shot learning methods to extract social relations from text and facial images.

Recently, the authors of RDS [19] separated the image-text dataset of Internet media into unimodal and multimodal subsets based on reinforcement learning, so that unimodal models perform better on unimodal subsets while multimodal models perform better on

multimodal subsets. The authors of HVPNeT [20] first introduced pyramidal features in the multimodal field, using visual representations as insertable visual prefixes to guide error-insensitive predictive decisions of textual representations. On this basis, they proposed a dynamic gated aggregation strategy to realize multi-level multi-scale visual features as fused visual prefixes. The authors of EEGA [21] were the first to propose jointly performing MNER and MRE as a joint multimodal entity-relation extraction task (JMERE). In addition to aligning visual objects and textual entities, they also aligned relations between visual objects and relations between textual entities. The authors of MKGformer [22] proposed a Transformer structure that mixes text and image modalities at multi-levels, where the prefix-guided interaction module (PGI) pre-reduces heterogeneity across modalities, and the correlation-aware fusion module (CAF) further reduces the model's sensitivity to irrelevant information between text and images.

The visual encoders of these models have evolved from early CNNs (e.g., Fast-RCNN [23], Resnet [24]) to later Vision Transformers (e.g., clip [25]), while most textual encoders are Transformer [26] models (e.g., bert [27], roberta [28]). These works have achieved great success. However, they fine-tune pre-trained unimodal models, while the multimodal parts are trained directly on MRE data. That is, modality alignment is secondary to MRE. Moreover, our analysis of image-text media data shows that visual information is more abundant yet noisier than text. This is also evidenced in ALBEF [29], which indicates that larger visual models are needed to fully utilize visual information. Therefore, we argue plain small models are insufficient for proper MRE due to dataset scale limitations and model capacity constraints. Previous models faied to achieve good modality alignment. To address this, we propose the MRDA and FTL methods, resolving data volume and relation category distribution problems.

2.2. Vision and Language Pre-Training

Extention of pre-training and fine-tuning schemes to the joint visio-linguistic domain has produced VLP models, such as Vilbert [30] and VinVL [31]. The pre-training tasks of these models mainly include masked language modeling (MLM), masked region classification (MRC) and image-text matching (ITM). During pre-training, multimodal fusion and alignment of encoded image and text features is achieved, followed by fine-tuning on downstream visio-linguistic tasks.

In recent years, ViLT [32] authors improved inference speed by lightening the visual encoder, but at the cost of performance drop. ALBEF [28] authors unified contrastive learning and fusion learning for multimodality and used momentum distillation to alleviate noise issues in large-scale multimodal data. VLMO [33] introduced the Mixture-of-Modality-Experts (MoME) Transformer with modality-specific expert pools and a shared cross-attention layer in each block. BLIP [34] combined encoder and decoder to form a unified multimodal understanding and generation model, further improving performance via data augmentation and data cleansing. However, most of these models are pre-trained on image captioning or VQA datasets. Applying current visual language models to MRE tasks may not lead to good performance.

2.3. Multimodal Tasks Using LLM

The developing large language models (LLMs) have shown powerful general knowledge understanding and logical reasoning abilities. An increasing number of works leverages LLMs for vision-language tasks, notably by training visual encoders and other extra structures to adapt LLMs. Recently, LLaVA [35] authors constructed image-text-related instruction tuning datasets to extend LLMs to multimodal domains. mPLUG-Owl [36] authors adopted a modular training idea to transfer LLMs into a large multimodal model, and proposed a comprehensive test suite OwlEval for visually grounded instructions.

InstructBLIP [37] authors proposed a visual language instruction tuning framework to enable general models to solve various vision tasks through a unified natural language interface. Although these models excel in image captioning and VQA, they do not perform

well on specialized downstream tasks like MRE, since the datasets are mostly for image captioning or VQA. Also, instruction tuning requires additional MRE instruction datasets. Therefore, we propose that in the field of MRE, prompts for large model few-shot learning can be constructed through the guidance of small models so as to enable large models as assisting knowledge engines to complete the complex task of MRE.

3. Methodology

This section introduces our EMRE2llm framework, which consists of two stages: small model guidance and LLM knowledge enhancement. The first stage trains a small model tailored for MRE to obtain guiding information. The training model is based on pure Transformer [26] cross-attention mechanisms to achieve multimodal learning for MRE. The second stage transforms the guiding information from the first stage into prompts to address modality disconnection and task disconnection issues, and finally realizes LLM knowledge enhancement (GPT-3 [38] is used here). The overall EMRE2llm framework is shown in Figure 2.

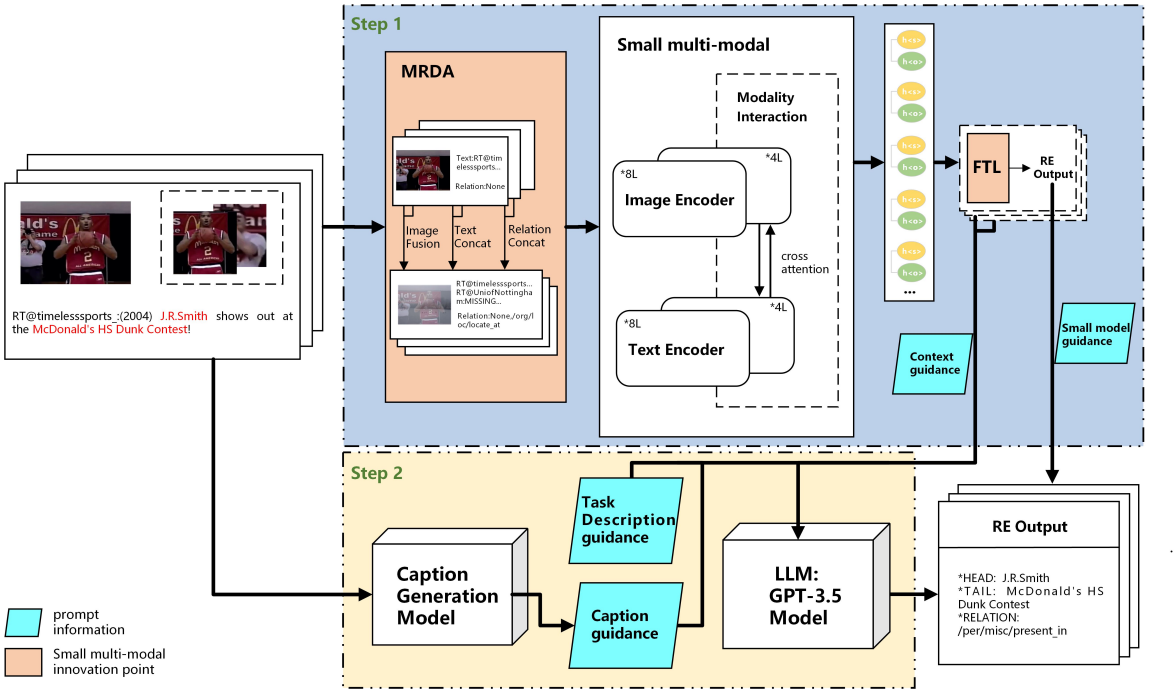

Figure 2. The EMRE2llm framework mainly consists of two stages: small model guidance and LLM knowledge enhancement. In the small model guidance stage, we enhance the small model through multimodal relation data augmentation (MRDA) and flexible threshold loss (FTL). In the LLM knowledge enhancement stage, first, the context guidance from small model outputs and small model guidance information are obtained; then, caption guidance information is generated by processing images with a caption generation model; finally, the task description guidance is combined to be fed into the LLM. The LLM knowledge enhancement stage can process data that the small model stage fails to handle, which requires additional knowledge bases for supplementation and further reasoning.

3.1. Small Multi-Model Guidance

LLMs suffer from severe hallucination or overgeneration issues [39], so the usage of large models alone cannot achieve good performance on subtasks. Small model guidance not only provides relation category confidence scores for the LLM to refer to, but also compensates for the LLM's uncertainties. Both visual and textual modules use pure Transformer models [40], with cross-attention for modality interaction. Here, we choose

MKGformer [22] as the baseline. For the MRE task, we propose MRDA and FTL to further enhance the small model, aiming to address insufficient datasets and imbalanced relations.

The small model uses Transformer-based architecture for both visual and textual modules. Specifically, the Image Encoder and Text Encoder use the encoder architectures from BERT [27] and the first K layers of CLIP [25], respectively, while the Modality Interaction uses their last L-K layers containing cross-attention interactions. Based on the Transformer architecture with L stacked layers, each layer contains multi-head self-attention (MHA), feed forward network (FFN), layer normalization (LN) and residual connections. We adopt the M-Encoder architecture from MKGformer [22] to implement the Modality Interaction part, reducing modality heterogeneity by incorporating the prefix-guided interaction module (PGI) into self-attention. In the self-attention layer, visual queries and textual keys/values are operated to aggregate textual information into the visual Transformer side. MHA performs the attention function in parallel over N_h heads, where each head is separately parameterized by W_q , W_k , $W_v \in R^{d \times d_h}$ to project inputs to queries, keys, and values. Given the input sequence vectors $x \in R^{n \times d}$, the attention heads $head^{M_t}$ and $head^{M_v}$ for text and vision are formulated as

$$head^{M_t} = Attn(x^t W_q^t, x^t W_k^t, x^t W_v^t), \tag{1}$$

$$head^{M_v} = Attn(x^v W_q^v, [x^v W_k^v, x^t W_k^t], [x^v W_v^v, x^t W_v^t]). \tag{2}$$

3.1.1. Multimodal Relation Data Augmentation

Data augmentation is a necessity for improving data efficiency and addressing insufficient data volume. Due to the image-text coupling constraints in multimodality, previous works like ViLT [32] only used RandAugment [41] for image augmentation, and CLIP [25] just randomly resized cropping.

These multimodal data augmentations are far from sufficient. Additionally, using traditional image augmentation techniques such as randomly adjusting cropping, flipping, and changing colors can cause mismatch between text and image information, leading to the inability of text to keep up with changes in image information. Therefore, we adopt a method similar to MixGen [42], augmenting data via image fusion and text concatenation without breaking image-text semantics. Image fusion through overlapping retains positional and color information of the original images, while text concatenation can seamlessly fuse textual information without changing semantics. For multimodal relational data, relation category fusion is performed by text concatenation as well, i.e., the relation information also doubles correspondingly for the generated image-text pairs with doubled information volume, as shown in Figure 3.

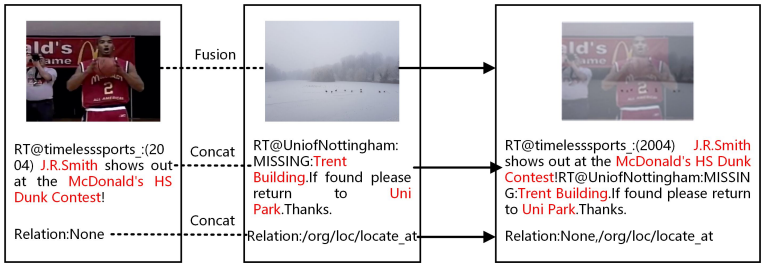

Figure 3. Our proposed Multimodal Relation Data Augmentation (MRDA) generates new image-text data by fusing images and concatenating texts and relation sequences for any image-text pair.

For an image-text dataset with N pairs, images and texts are denoted as I and T, respectively; then, for any two given image-text pairs (I_i, T_i) and (I_j, T_j) where $i, j \in \{1, ..., N\}$ and $i \neq j$, the formulation for generating a new sample (I_k, T_k) is

$$I_k = \lambda \cdot I_i + (1 - \lambda) \cdot I_j, \tag{3}$$

$$T_k = concat(T_i, T_j), \tag{4}$$

$$R_k = concat(R_i, R_j). \tag{5}$$

3.1.2. Flexible Threshold Loss

MRE is essentially a multi-label classification problem with images as additional information, where the output probability range $P(r|e_s, e_o)$ is [0, 1]. Thresholding is required to convert probabilities into relation labels. In the field of MRE, most methods use a fixed global threshold, i.e., binary cross-entropy loss. However, using the same global threshold for different entity pairs cannot classify well, and different relations have unequal data volumes, so global thresholding cannot enable fair learning. Especially for image-text datasets from Internet media, global thresholding fails to handle imbalanced entity pair number distributions and long-tail classes.

Our FTL draws inspiration from ATLOP's [43] clever threshold class T_H setting, where the positive–negative split for each entity pair (e_s, e_o) is determined by the special threshold class T_H. The logit for (e_s, e_o) greater than the logit for T_H is predicted as positive, otherwise it is predicted as negative. Of course, this threshold class T_H is learned in the same way as other classes.

Specifically, we divide the labels for an entity pair $T = (e_s, e_o)$ into two subsets: the positive subset P_T and the negative subset N_T. If T has a positive relation, it must be in P_T, otherwise P_T is empty. The negative subset N_T contains non-positive relations, $N_T = R \backslash P_T$. Unlike ATLOP's [43] unified computation, we compute the probabilities $P(r_i|(e_s, e_o))$ and $P(r_{TH}|(e_s, e_o))$ separately for each relation r_i and the threshold relation r_{TH} of the entity pair. To address the long-tail issue, we modify ATLOP's [43] loss by incorporating focal loss [44] to balance the positive logits. The formulas are as follows:

$$P(r_i|(e_s, e_o)) = \frac{exp(logit_{r_i}^{(s,o)})}{exp(logit_{r_i}^{(s,o)}) + exp(logit_{TH}^{(s,o)})}, \tag{6}$$

$$P(r_{TH}|(e_s, e_o)) = \frac{exp(logit_{r_{TH}}^{(s,o)})}{\sum_{r_j \in N_T \cup \{TH\}} exp(logit_{r_j}^{(s,o)})}, \tag{7}$$

$$L_{RE} = \sum_{r_i \in P_T} (1 - P(r_i|(e_s, e_o)))^{\gamma} log(P(r_i|(e_s, e_o))) + P(r_{TH}|(e_s, e_o)). \tag{8}$$

In the above loss function, γ is a hyperparameter that aims to pay more attention to classes with low confidence. That is, if $P(r_i|(e_s, e_o))$ is low, the loss contribution from the relevant class will be higher, which enables better optimization for long-tail classes.

3.2. Large Language Models for Knowledge Augmentation

LLMs [45–48] have shown impressive capabilities in contextual learning. Although the performance of LLMs on downstream subtasks is still far below the supervised baselines, LLMs possess powerful external knowledge bases and reasoning abilities that small models lack. We filter out the non-deterministic results obtained by the small model, which require reasoning and assistance from external knowledge bases. That is, we further obtain knowledge-enhanced results through LLMs [16].

3.2.1. Introduction to the GPT-3.5

GPT-3 [38], one of the most outstanding LMMs, is an autoregressive language model pretrained on massive datasets with powerful zero-shot capabilities. In contextual reasoning, few-shot learning formulates new downstream tasks as a text sequence generation task on top of the frozen GPT-3 model. The GPT-3.5 series is trained using instruction learning and reinforcement learning from human feedback (RLHF) to guide the model,

which connects a few examples of a task with the input as a prompt, and does not require parameter updates. However, it is noteworthy that such effective few-shot learning requires an enormous model as support. We utilize the few-shot ability of LLMs in our second stage. Although the results may not be as good as continuing to train the LLMs on vertical domains and task-specific tuning, this increases the flexibility and convenience of our framework, with the focus being on ways to elicit the potential capabilities of LLMs under insufficient resource conditions.

In the few-shot inference process of LLMs, for a frozen GPT-3.5 model with input sample x and target y, the generation condition is composed of the context examples $E_k = \{e_1, e_2, ...e_n\}$ and the prompt information pro, denoted as $p(x, E_k, pro)$. For a y_i at some time step from the model output $Y_k = \{y_1, y_2, ..., y_n\}$, we have

$$y_i = \underset{y_i^{'}}{^{argmax}} \, p(y_i^{'}|p, y_{<i}). \tag{9}$$

Here, we chose the two most advanced GPT-3.5 models: text-davinci-003 and GPT-3.5-turbo. Although GPT-3.5-turbo is an improvement over text-davinci-003, optimized for chat applications, according to our experiments, GPT-3.5-turbo is more unstable compared to text-davinci-003 for the MRE task. This is reflected in issues like ambiguous answers and redundant information.

3.2.2. Prompt Construction

The prompt input consists of a task description prompt, a caption prompt, a context prompt and a small model prompt, as shown in Figure 4. The task description prompt introduces the MRE task and reasoning ideas, instructing the LLM to first judge entity category information, and then infer relations based on the caption and small model prompts. The caption prompt is generated by describing the image multiple times using BLIP2 [49] and concatenating as direct image information conversion. The small model prompt provides relation category and confidence scores, serving as auxiliary information for GPT-3.5 to determine relations. Of course, not all data are used as prompt information, because GPT-3.5 is an autoregressive language model for text generation. Its direct application to MRE tasks introduces discrepancies, leading to poor performance on simple instances. We select difficult samples with low confidence scores from the small model as inputs to GPT-3.5 for knowledge enhancement.

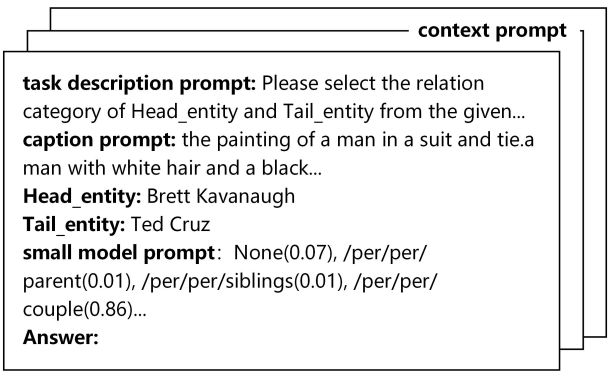

Figure 4. The guidance information fed into the LLMs consists of task description guidance, caption guidance, context guidance, and small model guidance to guide GPT-3.5 in predicting answers for uncertain data.

4. Experiments

Next, we introduce specific details of the experimental settings, including the training stage of the augmented small model and the invocation stage of GPT-3.5. Then, we evaluate the effectiveness of the whole model on the MNRE dataset, and finally conduct comprehensive ablation experiments.

4.1. Datasets

We evaluate on MNRE, currently the primary dataset for MRE. The text and image posts are crawled from Twitter, containing 15 k samples and 9 k images with 23 relation categories. The multimodal relation extraction MNRE dataset consists of the multimodal entity extraction datasets Twitter15 [50] and Twitter17 [51], as well as real data crawled from Twitter. The topics of these data include music, sports, and social events, and can be divided into 23 types of relational information, as shown in Figure 5.

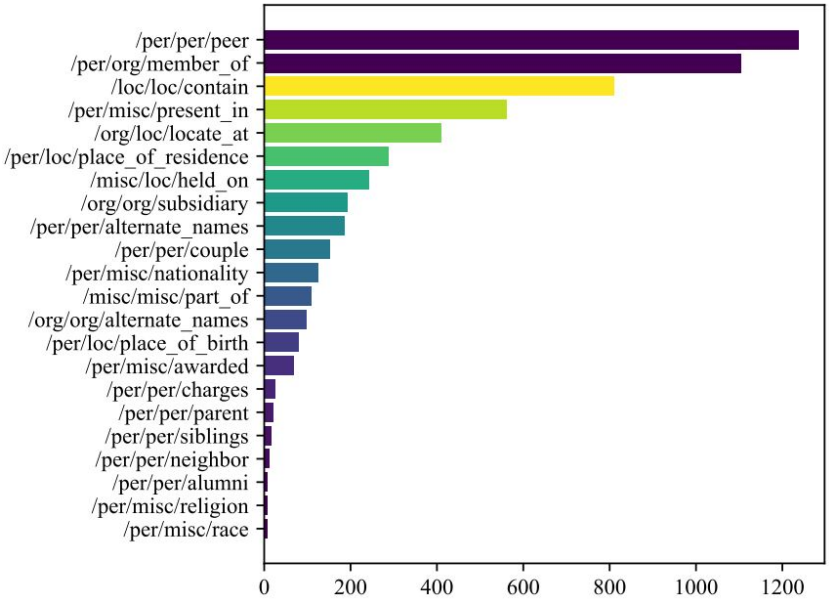

Figure 5. The relation categories and quantity per category in the MNRE dataset.

The statistics on the number of sentences, entities, and images in the multimodal dataset are shown in Table 1.

Table 1. The statistics on sentences, entities, etc., in the MNRE dataset.

Dataset	Sentence	Entity	Image
MNRE	9201	30,970	9201

We use MKGformer [22] as the small MRE model to generate small model guidance. To improve the small model's performance on insufficient and imbalanced data, we incorporate MRDA and FTL. Data augmentation happens during small model training. Since the image dataset is obtained via object subimages from RCNN [52] and visual grounding, the augmentation corresponds to random overlapping images, concatenating texts and relations for the one-to-many image-text case within the same batch. FTL addresses problems of fixed threshold and class imbalance by building on ATLOP's [43] idea of using a threshold class TH for splitting.

For LLM knowledge enhancement, we use GPT-3.5 text-davinci-003 and GPT-3.5-turbo with sampling temperature set to zero to remove randomness. For prompt data construction, we use $CAPTION\_N = 3$ and $EXAMPLES\_N = 10$ by default. The caption model is obtained by further fine-tuning BLIP2_OPT on COCO. Nucleus Sampling is used for decoding. The random context examples are extracted from the train set, similar to the prediction process, but requiring additional variables to mark whether the train predictions are correct to enable uniform sampling. We note that when the descriptive prompt is good enough to compensate for task disconnection, random context examples provide little gain.

4.2. Overall Performance

We compare our EMRE2llm model against several baselines and recent models for MRE. These models can be categorized into text-only models and image-text models, to demonstrate the superiority of our model.

First, text-only traditional models are used to demonstrate the indispensability of visual information in multimodality. We select two representative text RE models: PCNN [53] and MTB [54], a BERT-based RE pre-training model.

Second, we select previous SOTA models on image-text datasets. UMT [55] uses multimodal Transformers with an auxiliary entity span detection module. BERT + SG [17] connects BERT-derived text representations with scene image features. VisualBERT [56] is a single-stream pretrained visual language model. MEGA [17] employs dual image alignment to capture correlations between textual and visual objects.

Finally, we conduct a comparison with the latest SOTA models. EEGA [21] first proposes the joint extraction task JMERE of multimodal entities and relations, representing text and images as two images and aligning visual and textual objects. HVPNeT [20] proposes a dynamic gated aggregation strategy to realize multi-level multi-scale visual features as fused visual prefixes. MKGformer [22] uses pure Transformer architecture with PGI and CAF cross-attention techniques to jointly learn textual and visual information.

Table 2 shows the results of our EMRE2llm model and other models on the MNRE dataset, also contrasting our enhanced small model EMRE2llm_nollm. The EMRE2llm_nollm model is our enhanced small model in the first stage. Its output is processed simply based on the confidence scores, which means the relation labels are directly assigned according to the predicted probability values for each category.

Table 2. Performance comparison of various baseline models on the MNRE dataset. Here, MKGformer* is our re-implemented result. Even compared to the 81.95 F1 of MKGformer in the original paper, our EMRE2llmt has noticeable gains. The best scores are bolded.

Modality	Methods	MNRE		
		Precision	Recall	F1
Text	PCNN [53]	62.85	49.69	55.49
	MTB [54]	64.46	57.81	60.86
Text + Imag	UMT [55]	62.93	63.88	63.46
	BERT + SG [17]	62.95	62.95	62.80
	VisualBERT [56]	57.15	59.48	58.30
	UMGF [2]	77.51	76.01	76.75
	MEGA [17]	64.51	68.44	66.41
	EEGA [21]	78.27	78.91	78.59
	HVPNeT [20]	83.64	80.78	81.85
	MKGformer* [22]	80.83	79.06	79.94
	EMRE2llm_nollm(ours)	84.34	80.78	82.52
	EMRE2llmt(ours)	**85.06**	**80.94**	**82.95**

In order to better utilize the prompt information from small models and guide GPT-3.5, we adopt confidence screening strategies and context guiding strategies. Specifically, the confidence screening strategy divides the confidence scores of the relations predicted by the small model into intervals using the γ hyperparameter to obtain data with uncertain small model correctness. This ensures the effectiveness of the small model on data without excessive external knowledge bases while avoiding the hallucination problem of large models as much as possible, and dealing more with data with diffuse factors. For the context guiding strategy, we select context example data consisting of part of the training

data, with specific implementation methods including manually selected fixed numbers and random sampling. Both help guide GPT-3.5's understanding of the relation extraction task to some extent.

4.3. Ablation Studies

We conduct ablation studies on our model using the default settings in Section 4.2 to further demonstrate the efficacy of each module. This includes the impact of data augmentation parameters for MRDA, the thresholding and long-tail enhancement effects of FTL, and the selection of prompting ratios for GPT-3.5.

4.3.1. Enhancement Strategies for MRDA

We adopt random concatenation within the same batch and test the impact of different augmentation degrees. As shown in Table 3, augmenting half of the data works best, i.e., preserving equal original data for training. Since prediction uses raw test data, augmenting half for training retains real data characteristics while providing augmentation; hence, it is the most suitable.

Table 3. Experiments of applying MRDA to MKGformer-based models on the MNRE dataset. By comparing different augmentation degrees, the better parameter setting is obtained. The best scores are bolded.

Method = MKGformer_MRDA (batch = 64)	**MNRE_F1**	
	Test	**Val**
num = 16	80.54	76.92
num = 32	**81.32**	**77.18**
num = 64	81.10	76.42

4.3.2. The Efficacy of FTL

In Table 4, we compare vanilla MKGformer, MKGformer with TH thresholding, and our proposed FTL on the MNRE dataset, validating the superiority of FTL.

Table 4. Comparison of the phenotype of MKGformer on the MNRE dataset without data augmentation, with TH threshold data augmentation, and with our proposed FTL. The best scores are bolded.

Method	**MNRE_F1**	
	Test	**Val**
MKGformer_MRDA	81.32	77.18
MKGformer_MRDA + TH_threshold loss	81.00	76.91
MKGformer_MRDA + FTL	**82.36**	**77.51**

4.3.3. LLM Knowledge Enhancement

To enable fair comparison, we implement our proposed MRDA and FTL on top of MKGformer, and compare the benefits of LLM knowledge enhancement under the same dataset. As shown in Table 5, by comparing different uncertainty data ratio hyperparameters, the good stability of small models on significantly correct data is verified, while the fact that large models complement the limitations of small models under the assistance of external knowledge bases is also confirmed. As the table shows, it can be seen that the uncertainty and instability of large models lead to relatively lower scores on MRE tasks without the assistance of large models. The uncertain data of small models under test confidence scores below 0.47 need further assistance from large models.

Table 5. Demonstration of the effect of LLM knowledge enhancement and the appropriateness of confidence scores. The best scores are bolded. We divide the confidence score intervals of the relations predicted by the small model using the γ hyperparameter and use grid search to obtain the optimal γ hyperparameter. Below, we only show some representative parameter values for illustration.

Method	acc	micro_f1
EMRE2llm_nollm	91.51	82.52
$\gamma < 0.4$	**91.64**	82.68
$\gamma < 0.47$	91.57	**82.79**
$\gamma < 0.5$	91.45	82.31
$\gamma < 0.7$	90.33	80.36
only llm	85.56	73.60

5. Conclusions

In this paper, we introduce a new MRE framework EMRE2llm. Specifically, the first stage further enhances the ability of small models to deal with data scarcity, fixed thresholds, and long-tail distributions, achieving state-of-the-art performance of small models for MRE. The second stage better leverages GPT-3.5 to resolve examples that small models cannot solve, which often require common sense reasoning and external knowledge base capabilities, through the guiding information generated by small models as well as other prompt information like image-to-text. We conduct extensive ablation experiments on the MNRE dataset, a major MRE dataset, to demonstrate the effectiveness of each module, and show that EMRE2llm outperforms all existing state-of-the-art methods.

Moreover, our framework can be extended to other classification tasks handled by small models, as the two stages have no strong coupling and can be flexibly replaced iteratively. For example, in the key phrase extraction task, we can use a well-designed small model (such as the common BERT-CRF) to output confidence scores for each phrase category, which can serve as our first stage. The second stage can then directly use our method of prompting GPT-3.5, such as the design of Prompts. The advantage of such a framework is that it can better handle samples that require external knowledge bases and complex reasoning. Of course, our second stage utilizes the few-shot ability of LLMs. If we need the LLM to be more effective in the vertical domain of this task, we need to continue training the LLM, but this would cause a loss of the flexibility of our framework.

Author Contributions: Methodology, W.H.; Formal analysis, W.H. and S.L.; Writing—original draft, W.H.; Writing—review & editing, W.H.; Visualization, H.M., S.L., H.D., H.Z. and J.F.; Supervision, H.M., H.D., H.Z. and J.F.; Project administration, H.M.; Funding acquisition, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found from the public MNRE dataset at: https://github.com/thecharm/MNRE, accessed on 13 September 2023. Follow-up work and the source code are available on reasonable request from the corresponding author.

Conflicts of Interest: Hui Dong was employed by the company Hangzhou Codvision Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Cui, X.; Qu, X.; Li, D.; Yang, Y.; Li, Y.; Zhang, X. MKGCN: Multi-Modal Knowledge Graph Convolutional Network for Music Recommender Systems. *Electronics* **2023**, *12*, 2688. [CrossRef]
2. Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; Zhou, G. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; pp. 14347–14355.

3. Sha, Y.; Feng, Y.; He, M.; Liu, S.; Ji, Y. Retrieval-Augmented Knowledge Graph Reasoning for Commonsense Question Answering. *Mathematics* **2023**, *11*, 3269. [CrossRef]
4. Yang, Z. Biomedical information retrieval incorporating knowledge graph for explainable precision medicine. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 25–30 July 2020; pp. 2486–2486.
5. Rossi, A.; Barbosa, D.; Firmani, D.; Matinata, A.; Merialdo, P. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Trans. Knowl. Discov. Data (TKDD)* **2021**, *15*, 1–49. [CrossRef]
6. Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; Rosenblum, D.S. MMKG: Multi-modal knowledge graphs. In Proceedings of the the Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, 2–6 June 2019; pp. 459–474.
7. Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; Zheng, K. Multi-modal knowledge graphs for recommender systems. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 19–23 October 2020; pp. 1405–1414.
8. Zhang, N.; Li, L.; Chen, X.; Liang, X.; Deng, S.; Chen, H. Multimodal analogical reasoning over knowledge graphs. *arXiv* **2022**, arXiv:2210.00312.
9. Wang, Z.; Li, L.; Li, Q.; Zeng, D. Multimodal data enhanced representation learning for knowledge graphs. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
10. Chen, X.; Zhang, N.; Xie, X.; Deng, S.; Yao, Y.; Tan, C.; Huang, F.; Si, L.; Chen, H. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In Proceedings of the ACM Web Conference 2022, New York, NY, USA, 25–29 April 2022; pp. 2778-2788.
11. Chen, Y.; Ge, X.; Yang, S.; Hu, L.; Li, J.; Zhang, J. A Survey on Multimodal Knowledge Graphs: Construction, Completion and Applications. *Mathematics* **2023**, *11*, 1815. [CrossRef]
12. Li, S.; Zhang, H.; Ma, H.; Feng, J.; Jiang, M. CSIT: Channel Spatial Integrated Transformer for human pose estimation. *IET Image Process.* **2023, *12*, 3244.** . [CrossRef]
13. Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; Wang, L. An empirical study of gpt-3 for few-shot knowledge-based vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 3081–3089.
14. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
15. Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; Qiao, Y. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv* **2023**, arXiv:2303.16199.
16. Shao, Z.; Yu, Z.; Wang, M.; Yu, J. Prompting large language models with answer heuristics for knowledge-based visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14974–14983.
17. Zheng, C.; Feng, J.; Fu, Z.; Cai, Y.; Li, Q.; Wang, T. Multimodal relation extraction with efficient graph alignment. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 20–24 October 2021; pp. 5298–5306.
18. Wan, H.; Zhang, M.; Du, J.; Huang, Z.; Yang, Y.; Pan, J.Z. FL-MSRE: A few-shot learning based approach to multimodal social relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; pp. 13916–13923.
19. Xu, B.; Huang, S.; Du, M.; Wang, H.; Song, H.; Sha, C.; Xiao, Y. Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 1855–1864.
20. Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; Chen, H. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *arXiv* **2022**, arXiv:2205.03521.
21. Yuan, L.; Cai, Y.; Wang, J.; Li, Q. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 11051–11059.
22. Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; Chen, H. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 904–915.
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef]
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
26. Ranaldi, L.; Pucci, G. Knowing knowledge: Epistemological study of knowledge in transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]
27. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

28. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
29. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
30. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **2019**, .. [CrossRef]
31. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl: Making visual representations matter in vision-language models. *arXiv* **2021**, arXiv:2101.00529.
32. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 5583–5594.
33. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Piao, S.; Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32897–32912.
34. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
35. Liu, H.; Li, C.; Wu, Q.; Lee, Y. Visual instruction tuning. *arXiv* **2023**, arXiv:2304.08485.
36. Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y. mplug-owl: Modularization empowers large language models with multimodality. *arXiv* **2023**, arXiv:2304.14178.
37. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; Hoi, S. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv* **2023**, arXiv:2305.06500.
38. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process.* **2020**, *33*, 1877–1901.
39. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
40. Xu, P.; Zhu, X.; Clifton, D.; Intelligence, M. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef] [PubMed]
41. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 702–703.
42. Hao, X.; Zhu, Y.; Appalaraju, S.; Zhang, A.; Zhang, W.; Li, B.; Li, M. Mixgen: A new multi-modal data augmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Vancouver, BC, Canada, 17–24 June 2023; pp. 379–389.
43. Zhou, W.; Huang, K.; Ma, T.; Huang, J. Document-level relation extraction with adaptive thresholding and localized context pooling. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; pp. 14612–14620.
44. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
45. Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv* **2022**, arXiv:2201.11990.
46. Du, N.; Huang, Y.; Dai, A.M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A.W.; Firat, O. Glam: Efficient scaling of language models with mixture-of-experts. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 5547–5569.
47. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
48. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
49. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv* **2023**, arXiv:2301.12597.
50. Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; Ji, H. Visual attention model for name tagging in multimodal social media. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 1990–1999.
51. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive co-attention network for named entity recognition in tweets. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
52. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
53. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762.

54. Soares, L.B.; FitzGerald, N.; Ling, J.; Kwiatkowski, T. Matching the blanks: Distributional similarity for relation learning. *arXiv* **2019**, arXiv:1906.03158.
55. Yu, J.; Jiang, J.; Yang, L.; Xia, R. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
56. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; Chang, K. Visualbert: A simple and performant baseline for vision and language. *arXiv* **2019**, arXiv:1908.03557.