A Hierarchical Network for Multimodal Document-Level Relation Extraction

Lingxing Kong[1], Jiuliang Wang[1], Zheng Ma[1,2*], Qifeng Zhou[1,3], Jianbing Zhang[1,3†], Liang He[1], Jiajun Chen[1]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Institute for AI Industry Research (AIR), Tsinghua University
[3] School of Artificial Intelligence, Nanjing University, China
{konglingxing, wangjlcs20, maz, zhouqf, heliang}@smail.nju.edu.cn
{zjb, chenjj}@nju.edu.cn

Abstract

Document-level relation extraction aims to extract entity relations that span across multiple sentences. This task faces two critical issues: long dependency and mention selection. Prior works address the above problems from the textual perspective, however, it is hard to handle these problems solely based on text information. In this paper, we leverage video information to provide additional evidence for understanding long dependencies and offer a wider perspective for identifying relevant mentions, thus giving rise to a new task named Multimodal Document-level Relation Extraction (MDocRE). To tackle this new task, we construct a human-annotated dataset including documents and relevant videos, which, to the best of our knowledge, is the first document-level relation extraction dataset equipped with video clips. We also propose a hierarchical framework to learn interactions between different dependency levels and a textual-guided transformer architecture that incorporates both textual and video modalities. In addition, we utilize a mention gate module to address the mention-selection problem in both modalities. Experiments on our proposed dataset show that 1) incorporating video information greatly improves model performance; 2) our hierarchical framework has state-of-the-art results compared with both unimodal and multimodal baselines; 3) through collaborating with video information, our model better solves the long-dependency and mention-selection problems.

Introduction

Relation Extraction (RE) is a crucial task in Natural Language Processing (NLP). Traditionally, RE works have focused on extracting relations between entities in a single sentence. However, natural language often expresses complex relations over multiple sentences, necessitating the extension of RE from sentence-level to document-level for real-world applications.

One challenge in Document-level Relation Extraction (DocRE) is to capture the long dependencies between entities that are described separately and distantly in text. Most existing DocRE works utilize a document graph to organize long dependencies between entities (Quirk and Poon 2017; Peng et al. 2017; Verga, Strubell, and McCallum 2018). These works typically use a graph neural network to learn and reason for relation prediction. Recent works (Wang et al. 2019; Xu et al. 2021; Zhou et al. 2021; Tan et al. 2022) leverage Pre-trained Language Models (PrLMs) such as BERT to capture long-dependency features between entities. The transformer structure of these models has been shown to excel in dealing with long sequential data. However, existing works only explore a single textual modality for DocRE. Rather than only relying on the textual modality, Figure 1 illustrates how incorporating additional information from corresponding videos can be beneficial in addressing the long-dependency challenge in real-world scenarios. Specifically, multiple frames capture the interactions between **Donald Trump** and **Hillary Clinton**, providing evidence of their tense relationship. So, to accurately predict the target relation **Competitor**, it is necessary to effectively organize evidence from both modalities.

Another challenge in DocRE is the mention-selection problem. Due to the abundance of information in a document, an entity may appear multiple times, resulting in multiple mentions. Previous approaches (Wang et al. 2019; Xu et al. 2021; Zhou et al. 2021; Tan et al. 2022) have not distinguished between entity mentions, instead, they have simply averaged all mentions to represent the entity. However, in reality, not all mentions may be relevant to the target relation. As shown in Figure 1, only the first two mentions of **Hillary Clinton** are associated with the relational triple (**Donald Trump**, **Competitor**, **Hillary Clinton**), whereas the third mention in the last sentence "On Friday, Hillary returned to Detroit on her private plane." should be disregarded when extracting this target triple. The difficulty in distinguishing mentions is exacerbated by the complexity of the written language, which may include rhetorical devices such as irony, puns, and hyperboles. In such instances, accurately assessing the importance of a mention based solely on the textual modality becomes challenging. To overcome these hurdles, incorporating corresponding video information can prove invaluable. Visual descriptions captured in different frames offer diverse perspectives of an entity, thereby aiding in more accurately measuring its significance.

To overcome these challenges, we present an innovative extension of the DocRE task that incorporates an additional video modality, giving rise to a new task called Multimodal Document-level Relation Extraction (MDocRE). To support

*Internship at AIR, Tsinghua University
†Corresponding Author.

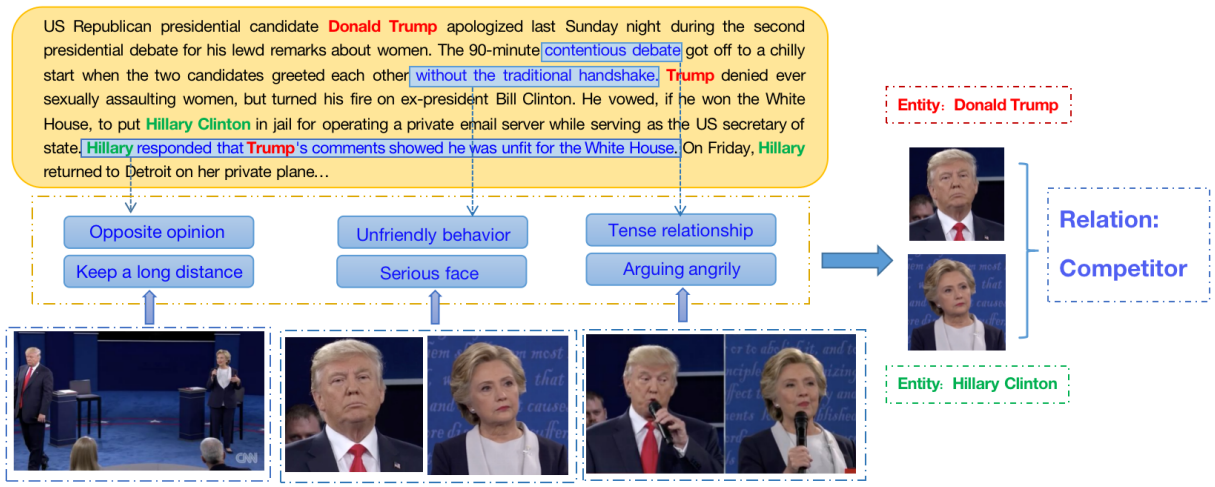

Figure 1: An example of combining the text and video modalities for document-level relation extraction. Different evidence can be extracted from different modalities to help predict the target relation between "Donald Trump" and "Hillary Clinton".

.

this novel task, we construct a human-annotated dataset with VOA news scripts and videos. Our approach to addressing this task is based on a hierarchical network that adeptly captures and fuses multimodal features at two distinct levels. At the local level, we align intra-frame visual structures with textual phrases, while at the global level, we align inter-frame visual bindings with textual semantics. During the multimodal fusion process, we propose a textual-guided transformer that leverages textual features to guide the learning of the multimodal embeddings. Furthermore, we address the mention-selection issue by proposing a mention gate module, which evaluates the correlation between entity mentions from a multimodal perspective. Experimental results show that our approach outperforms all unimodal and multimodal state-of-the-art baselines on the constructed dataset.

Related Work

One promising category of DocRE research is based on document graphs, which was first introduced by Quirk and Poon. Their document graph is constructed by adding semantic-dependency edges between entity nodes. In subsequent studies (Guo, Zhang, and Lu 2020; Nan et al. 2020; Christopoulou, Miwa, and Ananiadou 2019; Zeng et al. 2020; Wang et al. 2020), graph neural networks have been further applied to model the document graph. Additionally, some recent works directly utilize pre-trained language models to model cross-sentence dependencies (Xu et al. 2021; Wang et al. 2019; Zhou et al. 2021; Tan et al. 2022). Specifically, Zhou et al. propose adaptive thresholding and local context pooling to solve the multi-label and multi-entity problems in DocRE. Tan et al. propose adaptive focal loss to tackle the class imbalance problem. However, unlike previous efforts in other NLP tasks that have integrated multimodal information (Zhao et al. 2022; Xing et al. 2023; Zhao et al. 2023), the existing studies on DocRE focus solely on unique modal information without extending the extraction scope to the multimodal area.

Methodology

The task of MDocRE requires the model to extract the relation between two entities from both the input document text and the corresponding video, where each entity may have multiple mentions in both modalities. To be more specific, let $X = [x_1, x_2, ..., x_n]$ denote the input document text, and the subject and object textual entities are denoted as $E_s = \{E_i\}_{i=1}^{n}$ and $E_o = \{E_j\}_{j=1}^{m}$ respectively, where E_i and E_j represent textual mentions. Similarly, let $V = [v_1, v_2, ..., v_m]$ denote the input video, and the subject and object visual entities are denoted as $I_s = \{I_i\}_{i=1}^{n'}$ and $I_o = \{I_j\}_{j=1}^{m'}$ respectively, where I_i and I_j represent visual mentions. During the testing process, given (X, V) and multimodal information (E_s, E_o, I_s, I_o) for a certain entity pair, the model needs to predict the relation label r. It is worth noting that each (X, V) contains multiple entity pairs, and the model needs to predict the relation labels for all of them.

To address the MDocRE task, we propose a hierarchical framework that consists of a local encoder and a global encoder to capture features at different levels. In both encoders, the textual modality is used as the primary cue during the fusion of multimodal information. Specifically, each encoder contains a multimodal fusion layer, within which we propose a textual-guided transformer to align the visual information with the textual information. Moreover, a mention gate module is applied to solve the mention-selection problem in both modalities. Figure 2 shows the architecture of our entire model.

Local Encoder

The local encoder is responsible for capturing and fusing the local features, which include the intra-frame features from the visual modality and the phrasal features from the textual modality.

Textual Modality To extract phrasal features from the textual input, we utilize the lower layers of the pre-trained lan-

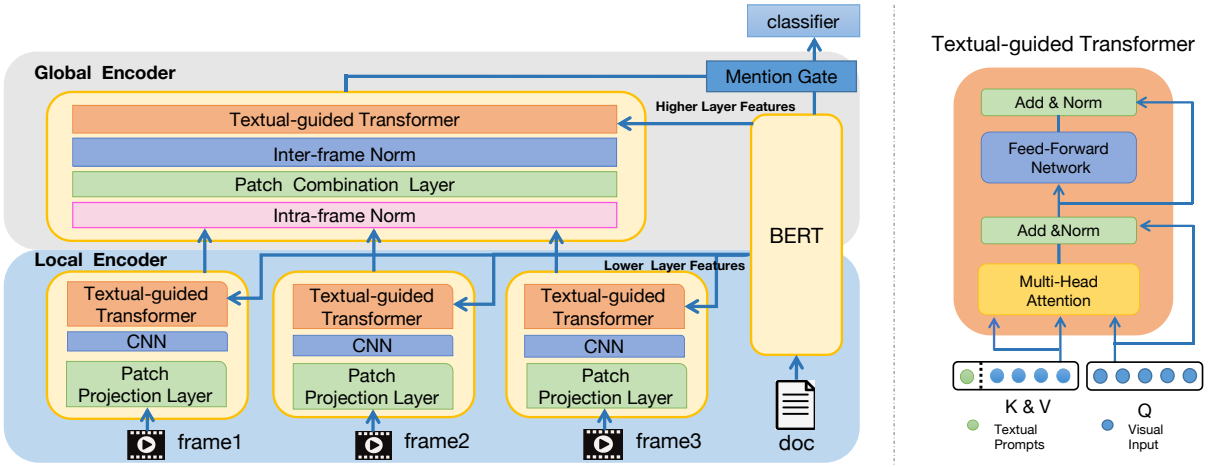

Figure 2: The framework of our approach. Our model utilizes a hierarchical structure to fuse the long textual and visual features.

guage model BERT. Studies have shown that BERT encodes a hierarchy of linguistic features, where the lower layers capture phrase-level features, the middle layers capture syntactic features, and the higher layers capture semantic features (Jawahar, Sagot, and Seddah 2019). So, we use the first two layers of BERT ($\text{BERT}_{\text{lower}}$) to generate the phrasal embedding matrix $\mathbf{L} \in \mathbb{R}^{n \times d}$ as follows:

$$\mathbf{L} = \text{Average}(\text{BERT}_{\text{lower}}(X))$$

Visual Modality In the local encoder, we process each frame v_m of the video input separately. First, we divide one frame into k 16*16 patches, which we project into a linear sequence denoted as $P_m = [p_1, p_2, ..., p_k]$. Then, we apply a convolution layer (CNN) followed by a normalization layer (LN) to generate the intra-frame features $\mathbf{P_m} \in \mathbb{R}^{k \times d}$ as follows:

$$\mathbf{P_m} = \text{LN}(\text{CNN}(P_m) + pos)$$

where $pos \in \mathbb{R}^{k \times d}$ is the positional embedding.

Multimodal Fusion After collecting the phrasal features $\mathbf{L}$ and intra-frame features $\mathbf{P_m}$, the main challenge is how to align and fuse these multimodal features. In most current vision-language research, multimodal fusion is typically accomplished through the use of either a one-stream or two-stream transformer architecture (Chen et al. 2023). However, for the task of MDocRE, we place greater emphasis on the textual modality and have developed a textual-guided transformer architecture. This is because the textual modality emphasizes logical expression and therefore neighboring sentences contain abundant continuous dependencies that are critical for reasoning a target relation. In contrast, the visual modality provides a more intuitive expression and thus the scenes in neighboring frames may be changeable. Inspired by Chen et al., we propose a textual-guided transformer architecture, as shown in Figure 2, which uses textual features as prompts to guide the fusion process. Specifically, we prepend these textual prompts to the visual features. In contrast to Chen et al., who used visual prompts, we leverage textual clues to achieve more accurate and robust video understanding.

At the local-level fusion, we first project the phrasal features $\mathbf{L}$ into the same embedding space as the intra-frame features $\mathbf{P_m}$ using a set of linear transformations $\mathbf{W}_t^1 \in \mathbb{R}^{d \times 2 \times d}$. This yields the textual prompts $\phi_k^1, \phi_v^1 \in \mathbb{R}^{n \times d}$, given by:

$$\{\phi_k^1, \phi_v^1\} = \mathbf{L}\mathbf{W}_t^1$$

Next, we calculate the attention scores by computing the query ($\mathbf{Q}_1$) based on the intra-frame embedding $\mathbf{P_m}$, the key ($\mathbf{K}_1$) based on $[\phi_k^1; \mathbf{P_m}]$, and the value ($\mathbf{V}_1$) based on $[\phi_v^1; \mathbf{P_m}]$. The fused embedding $\mathbf{G_m}$ is obtained by applying the softmax function to the scaled dot-product of $\mathbf{Q}_1$ and $\mathbf{K}_1$, and then multiplying the result by $\mathbf{V_1}$. We implement this as follows:

$$\mathbf{G_m} = \text{softmax}(\frac{\mathbf{Q}_1\mathbf{K}_1^T}{\sqrt{d}})\mathbf{V_1}$$

$$\mathbf{Q}_1 = \mathbf{P_m}\mathbf{W}_Q^1; \mathbf{K}_1 = [\phi_k^1; \mathbf{P_m}]\mathbf{W}_K^1; \mathbf{V}_1 = [\phi_v^1; \mathbf{P_m}]\mathbf{W}_V^1$$

Global Encoder

The global encoder aims to extract inter-frame features from the visual modality and fuse them with the semantic features from the textual modality.

Textual Modality To generate the semantic features, we use the 11$^{\text{th}}$ and 12$^{\text{th}}$ layer of BERT. As mentioned, these higher layers of BERT ($\text{BERT}_{\text{higher}}$), are shown to focus on semantic information in probing tasks (Jawahar, Sagot, and Seddah 2019). The resulting semantic embedding matrix $\mathbf{H}$ is obtained by:

$$\mathbf{H} = \text{Average}(\text{BERT}_{\text{higher}}(X))$$

Visual Modality We use a frame organization layer to reorganize the visual feature $\mathbf{G_m}$ generated by the local encoder. Recall that $\mathbf{G_m} = [\mathbf{g_1}, \mathbf{g_2}, ..., \mathbf{g_k}]$ is a sequence of patch embeddings. We concatenate (Concat) and normalize (LN) the patch embeddings to obtain the frame embedding $\mathbf{F_m}$ for the m^{th} frame:

$$\mathbf{F_m} = \text{LN}(\text{Concat}(\mathbf{G_m}))$$

Then, the frame embeddings are projected onto the video sequence by concatenating them and adding a frame-level positional embedding $pos' \in \mathbb{R}^{m \times d}$:

$$\mathbf{F} = \mathrm{LN}([\mathbf{F_1}, \mathbf{F_2}, ..., \mathbf{F_m}] + pos')$$

Multimodal Fusion We use another textual-guided transformer to fuse the global-level multimodal features. The textual prompts are generated based on the semantic embedding $\mathbf{H}$:

$$\{\phi_k^2, \phi_v^2\} = \mathbf{H}\mathbf{W}_t^2$$

Based on the generated prompts and inter-frame embedding $\mathbf{F}$, the fused embedding $\mathbf{M}$ is achieved through the following implementation:

$$\mathbf{M} = \mathrm{softmax}(\frac{\mathbf{Q}_2\mathbf{K}_2^T}{\sqrt{d}})\mathbf{V}_2$$
$$\mathbf{Q}_2 = \mathbf{F}\mathbf{W}_Q^2; \mathbf{K}_2 = [\phi_k^2; \mathbf{F}]\mathbf{W}_K^2; \mathbf{V}_2 = [\phi_v^2; \mathbf{F}]\mathbf{W}_V^2$$

The effectiveness of our proposed textual-guided transformer architecture has been extensively verified through experimental comparisons, and we will discuss the details in the experiment section. After global-level multimodal fusion, we further combine the aligned visual feature $\mathbf{M}$ with the textual feature $\mathbf{H}$ to generate the entity pair representation ($\mathbf{Z}_s$, $\mathbf{Z}_o$):

$$\mathbf{Z}_s = tanh(\mathbf{W}_h^1\mathbf{H}_s + \mathbf{W}_m^1\mathbf{M}_s) \quad (1)$$

$$\mathbf{Z}_o = tanh(\mathbf{W}_h^2\mathbf{H}_o + \mathbf{W}_m^2\mathbf{M}_o) \quad (2)$$

where $\mathbf{H}_s, \mathbf{M}_s \in \mathbb{R}^{d \times 1}$ are the subject representations in the two modalities, while $\mathbf{H}_o, \mathbf{M}_o \in \mathbb{R}^{d \times 1}$ are the object representations in the two modalities. $\mathbf{W}_h^1$, $\mathbf{W}_m^1$, $\mathbf{W}_h^2$, $\mathbf{W}_m^2 \in \mathbb{R}^{d \times d}$ are the learnable weights. Then the multimodal representations $\mathbf{Z}_s$ and $\mathbf{Z}_o$ are utilized to predict the relation type r:

$$P(r|E_s, E_o) = \mathrm{softmax}(\mathbf{Z}_s^T\mathbf{W}_z\mathbf{Z}_o + b)$$

where $\mathbf{W}_z \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}$ are the learnable weights and bias.

Mention Gate Module

To address the issue of mention selection, we propose a mention gate module that amplifies highly correlated mentions. The mention gate module aims to assess the mention-relation correlations from both intra-modality and inter-modality perspectives. Specifically, given the i$^{\text{th}}$ mention embedding $\mathbf{m_i}$ of a textual or visual entity, we compute the intra-modality attention $\mathbf{A}_i \in \mathbb{R}^{n \times 1}$ by computing the attention score between $\mathbf{m_i}$ and all n mentions from the same modality. Similarly, the inter-modality attention $\mathbf{B}_i \in \mathbb{R}^{m \times 1}$ is calculated between $\mathbf{m_i}$ and all m mentions from the other modality. We then reframe the task of evaluating mention-relation correlations as evaluating mention-mentions correlations. The mention weight is determined based on the gate output $\boldsymbol{\alpha}$, which is calculated as follows:

$$c_i = \mathbf{W_a}\mathbf{A}_i + \mathbf{W_b}\mathbf{B}_i$$
$$\mathbf{c} = [..., c_i, ..., c_{n+m}]$$
$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{c})$$

Where $\mathbf{W_a} \in \mathbb{R}^{1 \times n}$ and $\mathbf{W_b} \in \mathbb{R}^{1 \times m}$ are learnable weights. Once we obtain the mention weight, we calculate the entity representations $\mathbf{H}_s$, $\mathbf{H}_o$ in Eq. (1), Eq. (2) for the textual modality as follows:

$$\mathbf{H}_s = \sum_{i=1}^{s}(\alpha_i \cdot \mathbf{H}_i)$$
$$\mathbf{H}_o = \sum_{j=1}^{o}(\alpha_j \cdot \mathbf{H}_j)$$

Where s and o are the number of subject mentions and object mentions, respectively, $\mathbf{H}_i$ and $\mathbf{H}_j$ represent the subject and object mention embeddings in $\mathbf{H}$, and α_i, α_j are mention weights. Similarly, for the visual modality, we compute the entity representations $\mathbf{M}_s$, $\mathbf{M}_o$ in Eq. (1), Eq. (2) as follows:

$$\mathbf{M}_s = \sum_{k=1}^{s'}(\alpha_k \cdot \mathbf{M}_k)$$
$$\mathbf{M}_o = \sum_{l=1}^{o'}(\alpha_l \cdot \mathbf{M}_l)$$

Experiments

Dataset

We collect data from a free English learning website, which deliberately offers both VOA news scripts and videos to assist English learners in comprehending the content. These news reports serve as suitable materials for studying the MDocRE task for several reasons. Firstly, they provide an appropriate length of content (a few hundred words of document and several minutes of video). Secondly, both modalities cover the same topic, allowing for the meaningful supplementation of information between the two modalities. Lastly, most entities have multiple mentions within the context, facilitating the study of the mention selection issue. Moreover, the news scripts from this website are available in both Chinese and English. We employ 20 trained annotators and generate the sample labels in both languages to support future bilingual research.

We compare the statistics of our dataset with those of existing datasets in Table 1. It is evident that our dataset introduces a much more complex scenario compared to existing ones. In terms of the textual modality, our dataset poses significant challenges due to its larger number of words and sentences compared to existing datasets like DocRED. Moving to the visual modality, we believe our dataset is pioneering in introducing long videos, with an average length of 210 seconds, to the field of information extraction. This sets it apart from existing datasets like MNRE(Zheng et al. 2021), which primarily focuses on relation extraction from single image and short text in social media. Furthermore, our dataset is highly suitable for studying the mention selection issue, given its inclusion of abundant entity mentions.

Implementation Details

We split the constructed dataset into a training set with 2300 samples, a development set with 343 samples, and a test-

Statistics	MNRE	DocRED	Our Dataset
#Word	172k	1,002k	1,372k
#Sentence	14,796	40,276	70,703
#Doc.	-	5,053	3,043
#Video	-	-	3,043
#Image	10k	-	15,975k
#Relation	31	96	21
Avg. Mention	1.0	1.3	2.1
#Modality	2	1	2
#Language	1	1	2

Table 1: Statistics of our dataset compared to existing datasets. Avg. Mention equals #Mention divided by #Entity.

ing set with 400 samples. The hyper-parameters for all models are tuned on the development set. We set the number of textual-guided transformer layers L_{N1} in the Global Encoder and L_{N2} in the Local Encoder to 1 and 2, respectively. We set the number of heads N to 12. The maximum sequence length for textual and visual inputs is set to 512 and 128, respectively. During training, we use a batch size of 4, a learning rate of 1e-5, and a dropout rate of 0.2. Our training process comprises two stages. Initially, the textual encoder is frozen to enable rapid alignment of the visual modules. Then, in the second stage, the textual encoder begins learning all weights. Following Yao et al., we use F1 and Ign F1 scores as our evaluation metrics. We make our resources available (https://github.com/acddca/MDocRE).

Main Results

Baselines We conduct a comprehensive comparison of our MDocRE model with various baselines. First, we compare our model with unimodal DocRE baselines, including graph-based models **LSR** (Nan et al. 2020) and **AGGCN** (Guo, Zhang, and Lu 2020), and transformer-based models **ATLOP** (Zhou et al. 2021) and **BERT-E** (Zhou et al. 2021), to observe the performance changes resulting from the introduction of an extra video modality. Next, we choose the recent video-language pre-trained model **ClipBERT** (Lei et al. 2021) as one of our multimodal baselines and fine-tune it for the MDocRE task. Additionally, we consider different combinations of state-of-the-art video encoders, text encoders, and fusion architectures as our other multimodal baselines. Specifically, we utilize a pre-trained model **VideoMAE** (Tong et al. 2022) and **ViViT** (Arnab et al. 2021) to obtain visual features alternatively, **BERT** (Zhou et al. 2021) and **BERT-E** (Zhou et al. 2021) to obtain document-level textual features alternatively, and one-stream architecture **Os** (Li et al. 2019) and two-stream architecture **Ts** (Lu et al. 2019) to fuse the multimodal features alternatively. We adapt all the models above to the proposed MDocRE Dataset based on their open implementation.

Results and Analysis As shown in Table 2, our model outperforms all the state-of-the-art baselines, both unimodal and multimodal, on the proposed dataset. Compared with the unimodal baselines, our model outperforms both the graph-based methods and the transformer-based methods. This indicates that the long dependency between entities cannot be well-handled by considering only a single modality but requires coupling information from multiple modalities. Additionally, as shown in Table 3, the performance of unimodal models on MDocRE deteriorates compared to their performance on DocRED, which shows that the multiple-mention setting in MDocRE introduces new challenges for these models. Moreover, both the one-stream and two-stream multimodal baselines obtain inferior results compared to ours, which further verifies the effectiveness of our proposed hierarchical fusion framework and textual-guided transformer architecture.

Surprisingly, we observe that some unimodal baselines, such as ATLOP and BERT-E, outperform several multimodal baselines on our dataset. This phenomenon could be attributed to the introduction of noise rather than helpful information when long visual sequences are fed into inappropriate fusion architectures. To further validate that introducing video modality can indeed enhance model performance, we conduct supplementary experiments by combining our proposed multimodal fusion architecture with the top

Models	Dev		Test	
	F1	IgnF1	F1	IgnF1
Unimodal Models				
ATLOP	38.63	35.92	36.89	33.15
AGGCN	18.12	17.88	17.51	17.01
LSR	36.30	34.39	34.08	32.10
BERT-E	38.12	35.67	36.56	33.29
Multimodal Models				
ClipBERT	23.91	22.75	20.88	19.66
VideoMAE+BERT+Os	31.88	30.47	26.62	24.94
VideoMAE+BERT+Ts	37.73	35.83	34.26	31.67
VideoMAE+BERT-E+Os	32.28	30.63	27.46	25.49
VideoMAE+BERT-E+Ts	37.94	36.10	34.18	31.31
ViViT+BERT+Os	37.08	35.24	34.31	31.40
ViViT+BERT+Ts	38.87	36.12	36.45	33.20
ViViT+BERT-E+Os	38.95	36.67	35.96	32.77
ViViT+BERT-E+Ts	38.35	36.11	37.45	34.46
Ours	**42.37**	**40.13**	**40.54**	**37.22**

Table 2: Model performance on the proposed dataset.

Unimodal Models	DocRED		Our Dataset	
	F1	IgnF1	F1	IgnF1
Sequence-based Models				
CNN*	43.45	41.58	16.52	15.66
BiLSTM*	50.94	48.87	32.39	31.44
Graph-based Models				
AGGCN	52.47	46.29	18.12	17.88
LSR	59.00	52.43	36.30	34.39
Transformer-based Models				
BERT-E	56.31	54.29	38.12	35.67
ATLOP	61.46	59.40	38.63	35.92

Table 3: Comparison of unimodal models on DocRED and our dataset. We report both F1 and Ign F1 scores on the development sets. Results with $*$ are based on the open implementation in (Yao et al. 2019).

Unimodal Models	**F1**	**Ign F1**
BERT-E	38.12	35.67
ATLOP	38.63	35.92
Multimodal Models	**F1**	**Ign F1**
BERT-E (Multimodal)	40.35	37.97
ATLOP (Multimodal)	40.29	37.63

Table 4: Comparison of top unimodal models with their multimodal versions on our dataset.

Ablation of HF	**F1**	**Ign F1**
Non-hierarchical Tg	↓ 3.72	↓ 3.19
Non-hierarchical Ts	↓ 5.55	↓ 5.14
Non-hierarchical Os	↓ 19.00	↓ 17.61

Table 5: Ablation of the hierarchical fusion framework.

unimodal baselines for generating more sophisticated joint-modal embeddings. Specifically, we keep the local context pooling technique in ATLOP and logsumexp pooling in BERT-E, respectively, when generating the textual entity embedding. Then, for introducing the visual information, we apply the hierarchical framework and textual-guided transformer architecture in our proposed approach to generate the joint-modal embeddings. To ensure a fair comparison, in the combined approaches, we remove the proposed mention gate module, and the joint-modal embeddings are directly sent to the relation classifier for predicting the target relations. The results, shown in Table 4, demonstrate performance improvements compared with the original unimodal ATLOP and BERT-E, thereby proving the advantages of introducing an extra video modality for long-dependency relation extraction. These results also highlight the challenges of the MDocRE task and the importance of selecting an appropriate fusion architecture.

Ablation Study

To analyze the effectiveness of the proposed approach, we conduct ablation studies on its essential components.

Ablation Setup

Hierarchical Fusion Framework We replace the proposed *Hierarchical Fusion (HF)* framework with a standard *Non-hierarchical* transformer block. We implement non-hierarchical *Textual-guided (Tg)*, *One-stream (Os)* (Li et al. 2019), and *Two-stream (Ts)* (Lu et al. 2019) transformer architectures for the replacement.

Textual-guided Transformer Directly removing the *Textual-guided (Tg)* transformer architecture resulted in a significant performance drop or even rendered the model untrainable. Thus, we replace the original *Tg* architecture with either a *Visual-guided (Vg)* transformer architecture or a standard *Two-stream (Ts)* transformer architecture in which the two modalities are treated equally.

Ablation of Tg	**F1**	**Ign F1**
Vg	↓ 11.55	↓ 10.78
Hierarchical Ts	↓ 4.02	↓ 4.02

Table 6: Ablation of the textual-guided transformer module.

Ablation of MG	**F1**	**Ign F1**
-MG	↓ 2.02	↓ 2.16
-Textual MG	↓ 1.39	↓ 1.30
-Visual MG	↓ 1.11	↓ 1.90

Table 7: Ablation of the mention gate module.

Mention Gate We alternatively remove all the *Mention Gate (MG)* modules, only the textual *MG* module, and only the visual *MG* module to observe the performance changes.

Ablation Analysis

Hierarchical vs Non-hierarchical Removing the *HF* framework results in a notable performance drop. As shown in Table 5, all the non-hierarchical *Tg*, *Os*, and *Ts* variations get inferior results than the original hierarchical design. This indicates the importance of learning multi-granularity dependencies for the MDocRE task. Moreover, we observe that the non-hierarchical *Tg* outperforms the non-hierarchical *Os* and *Ts* by 15.28 and 1.83 F1 score, respectively. This also underscores the advantages of our proposed textual-guided transformer, which captures contextual information from the text and guides the fusion process at a certain granularity level.

Visual vs Textual The textual modality plays a more critical role in the MDocRE task. As shown in Table 6, compared to the original *Tg* design, the overall performance of *Vg* and *Ts* decreases by 11.55 and 4.02 F1 score, respectively. This verifies the effectiveness of using the textual modality as guidance during the fusion process. Moreover, *Vg* performs even worse than *Ts*, which suggests that using the visual modality as guidance instead will hurt the model's performance. We believe this is because the visual modality focuses more on intuitive expressions and provides less logical information, which is crucial for capturing long dependencies compared to its textual counterpart.

Multiple Mentions Selection The mention-selection issue is essential in the MDocRE task. As shown in Table 7, removing all the *MG* modules results in a performance drop of 2.02 F1 score. This demonstrates the importance of differentiating mentions when generating an entity representation. Additionally, removing either the textual *MG* module or the visual *MG* module results in a similar performance drop. This suggests that handling the mention-selection issue is equally crucial for both modalities.

Visualization

To enhance the interpretability of the proposed mention gate module, we generate visualizations of the calculated mention weights in various samples, as shown in Figure 3. When

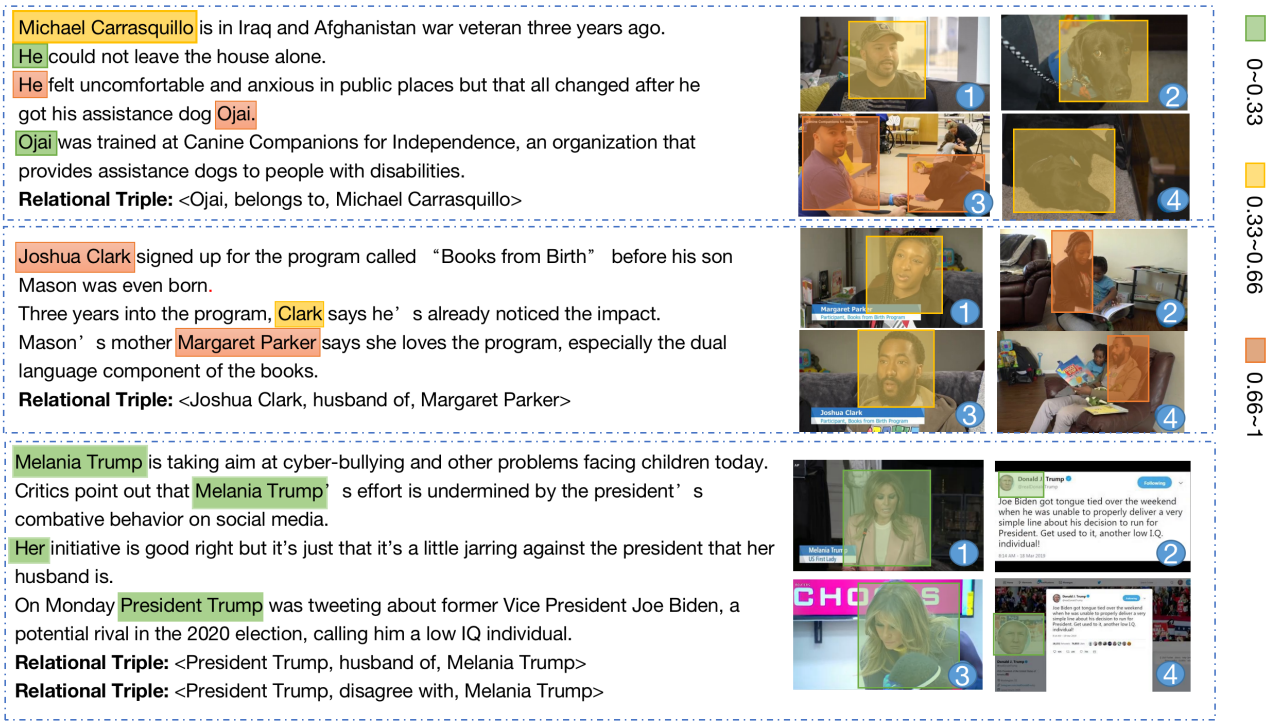

Figure 3: Visualization of mention weights. In both the textual and visual modalities, the red, yellow, and green rectangles represent different mention weights.

testing on these samples, all baseline models listed in Table 2 fail to extract any relational triple. On the other hand, our model successfully extracts the triples from the first two samples but fails to extract all the triples in the third sample.

Filtering Out Related Mentions Our approach can effectively focus on the mentions that exhibit high correlations with the target relational triple in both modalities. In the visual modality, our approach assigns more importance to the common scenes of the entity pair. For example, in the first sample, the visual mention of **Michael Carrasquillo** is given a higher weight in the third frame where he plays with **Ojai**, as compared to the first frame where he appears alone. Similarly, in the textual modality, our approach assigns a higher weight to the mentions that occur within sentences providing direct evidence or descriptions of the target relation. In the second sample, the mention of **Mason** as the common son provides direct evidence for inferring the **husband of** relation between **Joshua Clark** and **Margaret Parker**. In response to this evidence, our approach attributes greater weights to the mentions of **Joshua Clark** and **Margaret Parker** when co-occurring with **Mason** in sentences.

Advantages of Our Approach Our approach outperforms existing models that treat all mentions equally. For instance, our approach successfully extracts the target relations from the first two samples, while other state-of-the-art baselines fail to do so. This highlights the significance of eliminating interference from extraneous mentions.

Limitations of Our Approach In scenarios involving overlapping relations, such as the third sample, our approach encounters limitations. When tested on the third sample, where the same entity pair is involved in multiple relational triples, our model only extracts the (**President Trump**, **husband of**, **Melania Trump**) triple, while the (**President Trump**, **disagree with**, **Melania Trump**) triple remains undetected. Moreover, we observe that nearly identical weights are assigned to mentions in both modalities. This suggests that our approach faces challenges in distinguishing between mentions within such overlapping relation scenarios. In future work, we plan to explore more solutions for handling document-level overlapping relations.

Conclusion

In this study, we introduce the MDocRE task, which utilizes both video and document text information for relation extraction. To support this task, we construct the first human-annotated MDocRE dataset and propose a hierarchical framework with a textual-guided transformer to capture multimodal features at different levels. Moreover, we employ a mention gate module that selects the most relevant mentions for generating entity representations. Experimental results demonstrate that our proposed approach outperforms both unimodal and multimodal state-of-the-art baselines.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by NSFC (No. 62176115) and National Key R&D Program of China (2022ZD0160501).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6836–6846.

Chen, F.-L.; Zhang, D.-Z.; Han, M.-L.; Chen, X.-Y.; Shi, J.; Xu, S.; and Xu, B. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1): 38–56.

Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022. Good Visual Guidance Makes A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. arXiv:2205.03521.

Christopoulou, F.; Miwa, M.; and Ananiadou, S. 2019. Connecting the Dots: Document-level Neural Relation Extraction with Edge-oriented Graphs. arXiv:1909.00228.

Guo, Z.; Zhang, Y.; and Lu, W. 2020. Attention Guided Graph Convolutional Networks for Relation Extraction. arXiv:1906.07510.

Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Nan, G.; Guo, Z.; Sekulić, I.; and Lu, W. 2020. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction. arXiv:2005.06312.

Peng, N.; Poon, H.; Quirk, C.; Toutanova, K.; and Yih, W.-t. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5: 101–115.

Quirk, C.; and Poon, H. 2017. Distant Supervision for Relation Extraction beyond the Sentence Boundary. arXiv:1609.04873.

Tan, Q.; He, R.; Bing, L.; and Ng, H. T. 2022. Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. arXiv:2203.10900.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Verga, P.; Strubell, E.; and McCallum, A. 2018. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. arXiv:1802.10569.

Wang, D.; Hu, W.; Cao, E.; and Sun, W. 2020. Global-to-Local Neural Networks for Document-Level Relation Extraction. arXiv:2009.10359.

Wang, H.; Focke, C.; Sylvester, R.; Mishra, N.; and Wang, W. 2019. Fine-tune Bert for DocRED with Two-step Process. arXiv:1909.11898.

Xing, S.; Zhao, F.; Wu, Z.; Li, C.; Zhang, J.; and Dai, X. 2023. DRIN: Dynamic Relation Interactive Network for Multimodal Entity Linking. In El-Saddik, A.; Mei, T.; Cucchiara, R.; Bertini, M.; Vallejo, D. P. T.; Atrey, P. K.; and Hossain, M. S., eds., *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 3599–3608. ACM.

Xu, B.; Wang, Q.; Lyu, Y.; Zhu, Y.; and Mao, Z. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14149–14157.

Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; and Sun, M. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. arXiv:1906.06127.

Zeng, S.; Xu, R.; Chang, B.; and Li, L. 2020. Double Graph Based Reasoning for Document-level Relation Extraction. arXiv:2009.13752.

Zhao, F.; Li, C.; Wu, Z.; Ouyang, Y.; Zhang, J.; and Dai, X. 2023. M2DF: Multi-grained Multi-curriculum Denoising Framework for Multimodal Aspect-based Sentiment Analysis. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 9057–9070. Association for Computational Linguistics.

Zhao, F.; Li, C.; Wu, Z.; Xing, S.; and Dai, X. 2022. Learning from Different text-image Pairs: A Relation-enhanced Graph Convolutional Network for Multimodal NER. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, 3983–3992. ACM.

Zheng, C.; Wu, Z.; Feng, J.; Fu, Z.; and Cai, Y. 2021. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Zhou, W.; Huang, K.; Ma, T.; and Huang, J. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of*

the AAAI conference on artificial intelligence, volume 35, 14612–14620.