



TSPNet: Translation supervised prototype network via residual learning for multimodal social relation extraction

Hankun Kang^{a,b,c,d}, Xiaoyu Li^{a,b}, Li Jin^{a,b}, Chunbo Liu^{a,b,*}, Zequn Zhang^{a,b}, Shuchao Li^{a,b}, Yanan Zhang^{a,b,c,d}

^a Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

^b Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

^c University of Chinese Academy of Sciences, Beijing 100190, China

^d School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:

Received 19 February 2022

Revised 19 July 2022

Accepted 24 July 2022

Available online 26 July 2022

2020 MSC:

00–01

99–00

Keywords:

Multimodal social relation

Knowledge triples

Few-shot scenario

Residual learning

ABSTRACT

Multimodal social relation extraction requires sufficient features fusion to identify the relation between different targets. Compared with traditional multimodal social relation extraction, there are many semantic gap issues for the few-shot scenario task, such as insufficient across-modality assistance, lacking explicit supervision, and unbalanced relations. To address the above problems, a novel Translation Supervised Prototype Network (TSPNet) is proposed, **which extracts all the features of knowledge triples, not just relation features**. First, the triple-level unimodal encoder learns textual and visual representation of knowledge triples from the entire information via two-stream encoding. Second, the triple-level multimodal extractor obtains multimodal knowledge triples by employing the residual learner to build the triple-level interaction across modalities. Finally, **the intra-triple translation supervised decoder predicts the few-shot relations based on a prototype network supervised with the intra-triple translation as an explicit constraint**. Our model achieves SOTA performance on three challenging benchmark datasets for few-shot multimodal social relation extraction, and further analysis shows that our model is effective and owns a strong generalization ability to avoid bias.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Social relations define the associations between different people and emerge in our daily lives, where the extraction task aims to identify relations between different people, helping us analyze social link connections better [1–4]. However, due to different intrinsic properties and availability of samples, some relations own rich-labeled data while others are poor during training process. In addition, there is identification ambiguity on unimodal extraction encountering insufficient textual information. Fortunately, with the advancement of social media, multimodal messages, such as text and images, can assist each other in extracting relations. It is observed that facial information for the person entity can provide auxiliary information such as age, emotion, gender, identities, and so on. As Fig. 1 shows, in case (a), unimodal extraction methods can not distinguish the relations *Son* and *Daughter* according to textual clues such as love, family, etc.

While the multimodal extraction methods can identify the true relation *Daughter* by assisting with visual clues including gender, etc. Similarly, in case (b), multimodal extraction methods can distinguish *Classmates* and *Alumni* by adding visual clues with age, provided by the facial information. Hence, the multi-dimensional information supplied by the facial features can be used to enrich unimodal textual information to boost performance.

However, there still are some critical issues with extracting relations based on the multimodal information. It is difficult to fuse information because of the gap between modalities. Under-fusion can not sufficiently use assistance between modalities, and over-fusion will discard raw unimodal information. Moreover, relations exist in unbalanced distribution, which results in identification bias between rich-labeled and poor-labeled relations. Meanwhile, there is a lack of explicit supervision to extract the core characteristics of relations in existing methods.

To address the above problems, we propose a novel framework named TSPNet (abbreviated from **T**ranslation **S**upervised **P**rototype **N**etwork via residual learning) to learn the multimodal representations of knowledge triples (consisting of head entity h , relation r , and tail entity t , i.e., $\{h, r, t\}$). Furthermore, TSPNet

* Corresponding author at: Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China.

E-mail address: liucb@aircas.ac.cn (C. Liu).

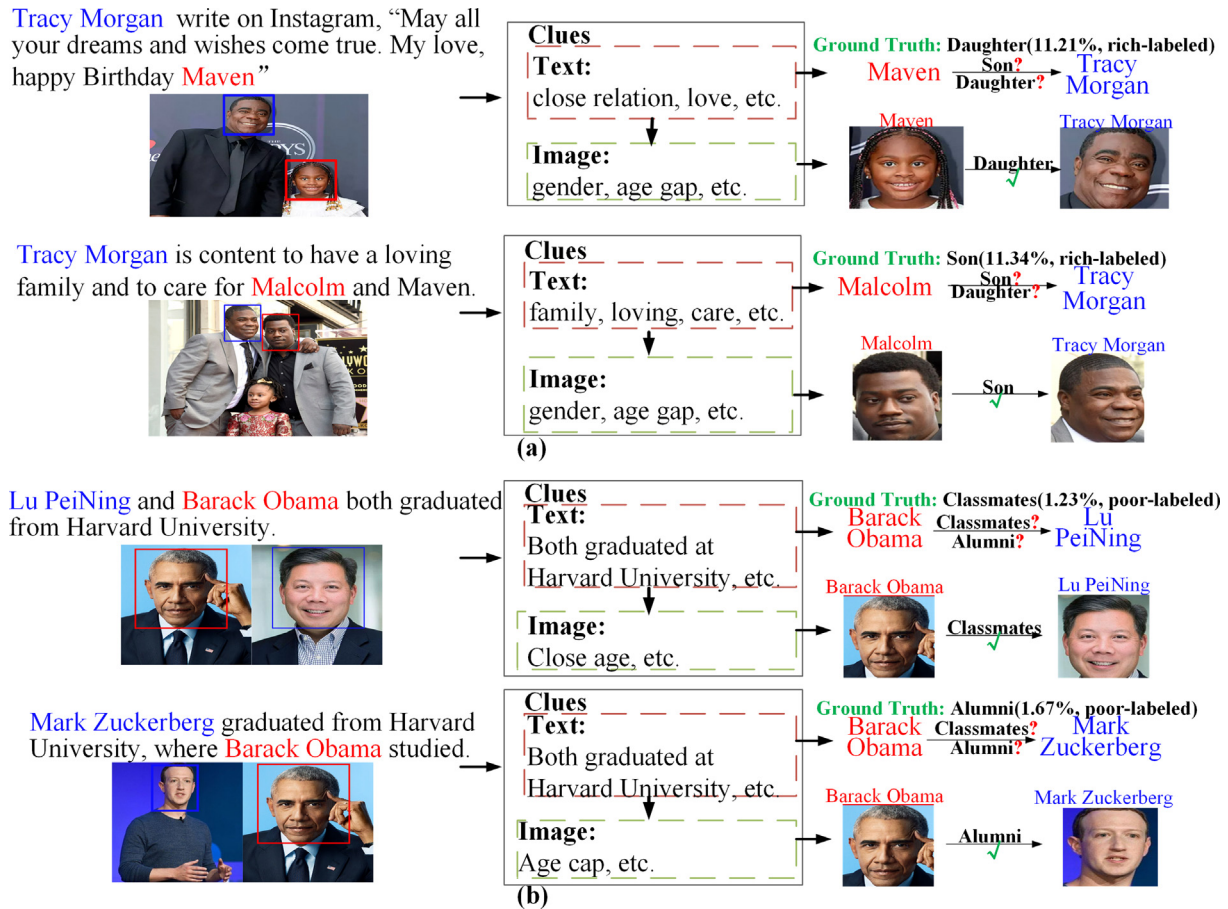


Fig. 1. Illustration for multimodal social relation extraction. The head and tail entities are highlighted by red and blue font in text and are bound by red and blue boxes in images respectively. “Clues” parts show the textual and visual clues and “Ground Truth” parts show the amounts of rich-labeled and poor-labeled relations. Data is collected from <http://openkg.cn/dataset>.

predicts the relations included in knowledge triples based on the triple-level multimodal features. Inspired by approximation theory [5–7], in order to reduce the difficulty to learn assistance and retain raw unimodal information, multi-step residual learning is employed to capture triple-level assistance. To enhance fusion, an attention variant-based convolution is designed to capture hierarchical interaction in every residual learning step. In addition, to extract unbalanced relations, existing few-shot methods [8–11] represent relations with implicit features. However, when inputs are similar, the implicit features are also similar and not essential enough, which leads to ambiguity, e.g., the directions of relations between entities are unable to be distinguished. Inspired by the translation principle in knowledge representation learning, the intra-triple translation can mine core characteristics of relations (e.g., directions). To enhance the capacity to extract relations from similar inputs, the intra-triple translation is utilized as an explicit constraint of the prototype network, improving digging essential features of relations.

In summary, the contributions of this work are as follows:

- To sufficiently employ multimodal information, we design a hierarchical attention variant to build triple-level interaction across modalities. Furthermore, to reduce the difficulty to learn assistance and retain raw unimodal information, **the interaction is learned step by step based on residual learning**. This novel idea shows potential and a new research direction for multimodal fusion.

- Considering the unbalanced distribution and explicit representations of multimodal social relations, **an intra-triple translation supervised prototype network is utilized to improve extracting core features of unbalanced relations under few-shot learning, which alleviates ambiguity and bias in relation identification**.
- The experiments show that our model achieves SOTA performance and owns a more robust generalization with weaker dependency on data. And our model alleviates ambiguity in social relations identification by extracting essential multimodal features of relations.

2. Related work

2.1. Multimodal relation extraction

Multimodal relation extraction aims to extract relations among entities (e.g., people, subjects) from online images and text as multimodal information. [12] profiled relations among people as edge semantics in a social network from multimodal information based on a multimodal graph edge variational autoencoder. [13–15] extracted relations among characters in video by employing video and text features. [16] constructed visual relation graph with text-guide information extracted by text-visual relation extractor. [17] designed a multimodal fusion graph to represent the general features in multimodal video scenes. [18] first extracted social relations from multimodal information in few-shot scenarios, which just concatenated the multimodal features without assistance.

[19] utilized image caption methods to employ visual and textual information to identify spatial relations. [20] adopted a graph network to align entities between images and text and further extract social relations between aligned entities. Existing methods also employed other modality information except for visual and textual information. [21] combined text, audio, and visual features in the same scene to extract relations between entities in the scene based on scene segmentation. [22] identified the semantic relations of unseen entities combined with textual, auditory, visual, and tonal information via constructing a knowledge graph. [23] utilized the GraphBERT model to encode the textual and molecular structure information and explore the underlying features of various modalities for biomedical relation extraction.

2.2. Few-shot relation extraction

As few-shot relation classification requires less data to train the neural network models, high costs related to data collection and labeling are eliminated [24]. Due to the different availability, the labeled data exists in long-tail distribution, which limits the effectiveness of the task. [25] used an undersampling technique to alleviate the influence of unbalanced annotations in visual relations detection. In addition, some researches formulated unbalanced relation extraction as a few-shot learning task, where [26] designed a dual graph neural network for visual relation detection that works in one-shot settings. Generally, for metric-based few-shot learning scenario, [8] learned relation prototypes as an implicit factor between entities by building a co-occurrence graph from the text. [9] adopted transformer to prototype network for enhancing generalization ability to extract data-poor relations. [10] designed distance metric between data-rich relations and data-poor relations to identify the relations with poor-labeled data. [11] addressed the uneven distribution using a convolutional siamese neural network that extracts discriminative semantic-aware features to detect relations. [27] proposed a hybrid attention-based prototype network for the problem of noisy few-shot relation extraction by designing instance-level and feature-level attention schemes. [28] adopted relational siamese networks to learn relation classifier with labeled data and further generalize the classifier to new relations. As for optimization-based few-shot relation extraction, [29] proposed a bayesian meta-learning approach to learn the prior distribution of meta relation and further extract poor relations with meta relation. [30] identified poor-labeled relations and new relations based on a combination of ideas from lifelong learning and optimization-based meta-learning.

利用更好的prototype network?

2.3. Knowledge representation learning

Both relation extraction and knowledge representation learning are the foundation task for knowledge graph. Knowledge representation learning aims to learn low-dimensional embedding containing semantic and commonsense information of knowledge triples (i.e., head entity, relation, and tail entity). Existing methods include linear models, neural models, matrix factorization models, translation models, and other models [31]. For linear models, [32–34] employed linear combination of the representations of the entities and relations to measure the probability of triples. For neural models, [35–37] computed the probability of the knowledge triples by neural networks which take the embeddings of the entities and relations as inputs. For matrix factorization models, [38,39] viewed representation learning of triples as matrix factorization operation. However, translation models adopted translation principle [40] as an optimization objective to learn the embeddings. [41] adopted the translation principle to minimize the distance between positive triples and maximize the distance between negative triples to

learn the embedding of knowledge triples, which is effective for 1-to-1 relations but has flaws when dealing with complex relations (e.g., N-to-N relations). Further, [42] used a dynamic mapping matrix to map entities to relation space to minimize translation distance for learning the embedding of complex relations. [43] projected the head entity and tail entity to the relation hyperplane and minimized translation distance in the hyperplane for learning the embeddings. Recently, Some novel methods for knowledge representation learning have been proposed. [44] utilized hierarchical relations among the types of entities to learn the embeddings. [45] formulated knowledge graph as an irregular graph to encode triples based on graph attention. [46] designed a score function by encoding regularities between relations and related entities to enhance the embedding as a constraint.

Overall, multimodal social relation extraction in few-shot scenarios is still in the exploratory stage. Intuitively, we compare the TSPNet with other existing social relation extraction methods in Table 1. The TSPNet can extract multimodal social relations with explicit representation under few-shot learning.

3. Proposed methodology

In this section, we first give the task definition and then introduce the main framework of proposed model. As shown in Fig. 2, the TSPNet consists of three main parts: (1) triple-level unimodal encoder aims to learn the textual and visual representations of knowledge triples from input text and images; (2) triple-level multimodal extractor is employed to capture the interaction between textual and visual knowledge triples; (3) intra-triple translation supervised decoder adopts a prototype network with an intra-triple translation constraint to predict multimodal social relations in few-shot scenarios.

3.1. Problem definition

Given multimodal datasets including text and images, we split the datasets into support set S_n and query set Q_n . S_n includes N social relations (e.g., {friend, employer, servant}) and every relation owns K samples. And Q_n owns Q query samples from N relations. Single sample includes one sentence with associated image from S_n and Q_n , which is assigned to a relation label y belonging to N relations. Our problem definition can be stated as follows: given the support set S_n and query set Q_n , the goal of the task is to learn a classifier that can identify the predefined relations Y of Q_n .

3.2. Triple-level unimodal encoder

Triple-level unimodal encoder aims to obtain textual and visual representations of knowledge triples from inputs. In detail, the encoder extracts textual knowledge triples $\mathbf{k}_t = \{\mathbf{h}_t, \mathbf{r}_t, \mathbf{t}_t\}$, where $\mathbf{h}_t, \mathbf{r}_t, \mathbf{t}_t \in \mathbb{R}^{d_t}$ mean textual representations of head entities, relations and tail entities, and extracts visual knowledge triples $\mathbf{k}_v = \{\mathbf{h}_v, \mathbf{r}_v, \mathbf{t}_v\}$, where $\mathbf{h}_v, \mathbf{r}_v, \mathbf{t}_v \in \mathbb{R}^{d_v}$ mean visual representations of head entities, relations and tail entities.

Table 1

Comparison between TSPNet and other existing social relation extraction methods.

Models	Multimodal	Few-shot	Explicit constraint
MEGA [20]	✓	×	×
GNN [47]	×	✓	×
SNAIL [48]	×	✓	×
SIAMESE [49]	×	✓	×
text(BERT) [18]	×	✓	×
FL-MSRE [18]	✓	✓	×
TSPNet	✓	✓	✓

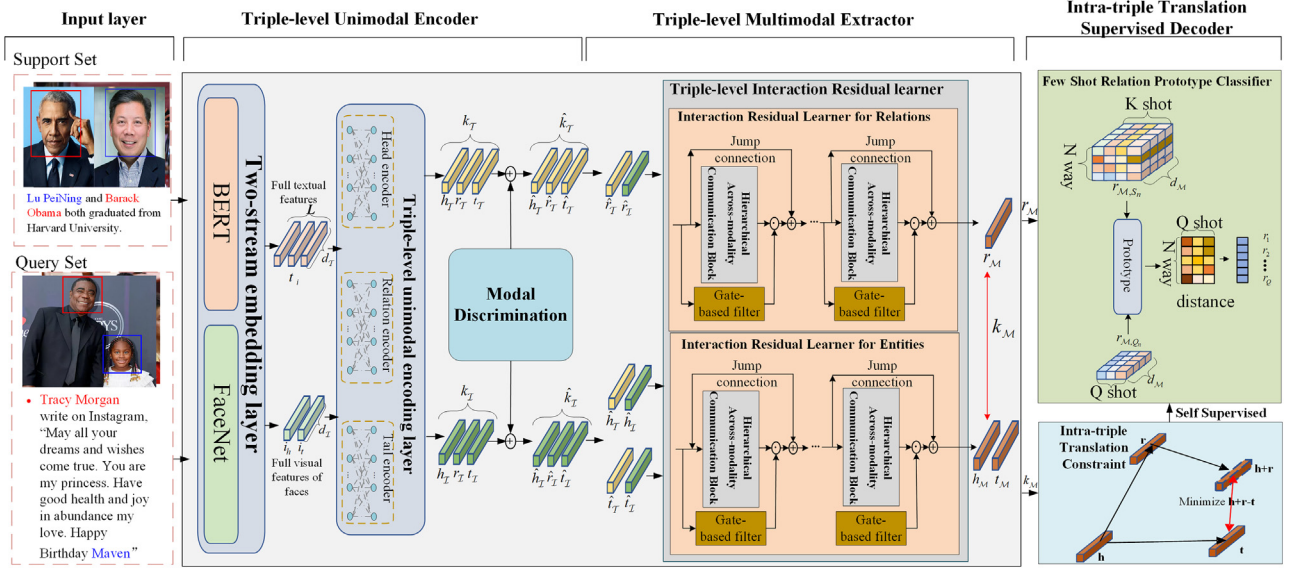


Fig. 2. The architecture of TSPNet. The model extracts the relations in query set by calculating the distance between the support and query set. The label corresponding to the shortest distance is utilized as the predefined relation.

3.2.1. Two-stream embedding layer

Two-stream embedding layer is conducted to obtain the distributed representations of inputs: (1) text encoder is employed to acquire the textual representations; (2) images encoder is utilized to earn visual features.

Text encoder employs BERT [50] to capture textual representations. To emphasize the referentiality of entity tokens, text encoder uses two learnable tokens w_h and w_t to represent head and tail entity respectively. Formally, given a sentence $S = \{w_1, w_2, \dots, w_h, \dots, w_i, \dots, w_t, \dots, w_L\}$ including head and tail entity, where L means the length of the given sentence S , w_i refers to the i th word. The BERT embeddings of text can be calculated as follows:

去除介词等，只保留动名词？

$$\mathbf{t}_i = \text{BERT}(w_i), \forall i = 1, 2, \dots, L \quad (1)$$

where $\mathbf{t}_i \in \mathbb{R}^{d_f}$ is i th the embedding for w_i .

Inspired by [51], the images encoder employs FaceNet to project images to distributed vectors for getting visual features. Given the images $f_h, f_t \in \mathbb{R}^{C \times W \times H}$, the images include head and tail entity appeared in the given sentence, where C, W and H mean the channel, width, and height of input images respectively. The visual features $\mathbf{i}_h, \mathbf{i}_t \in \mathbb{R}^{d_v}$ of the images can be calculated as follows.

$$\mathbf{i}_x = \text{FaceNet}(f_x) \quad (2)$$

where $x \in \{h, t\}$ refers to the head and tail entity.

3.2.2. Triple-level unimodal encoding layer

A triple-level unimodal encoding layer is employed to extract textual and visual knowledge triples from the full embeddings of the sentence and images respectively.

For textual knowledge triples $\mathbf{k}_{\mathcal{T}}$, fully connected layers and gelu activate function are adopted to get the triple-level textual representations as follows:

$$[\mathbf{h}_{\mathcal{T}}; \mathbf{t}_{\mathcal{T}}] = \text{gelu}(\mathbf{W}_{h,t}[\mathbf{t}_h; \mathbf{t}_t] + \mathbf{b}_{h,t}) \quad (3)$$

$$\mathbf{r}_{\mathcal{T}} = \text{gelu}(\mathbf{W}_r \sum_i \mathbf{t}_i + \mathbf{b}_r) \quad (4)$$

where gelu means gelu activate function.

As for visual knowledge triples $\mathbf{k}_{\mathcal{V}}$, by utilizing visual embeddings \mathbf{i}_h and \mathbf{i}_t , fully connected layers and relu activate function are adopted to get triple-level visual representations as follows.

$$[\mathbf{h}_{\mathcal{V}}; \mathbf{t}_{\mathcal{V}}] = \text{relu}(\mathbf{W}_{h,t}[\mathbf{i}_h; \mathbf{i}_t] + \mathbf{b}_{h,t}) \quad (5)$$

$$\mathbf{r}_{\mathcal{V}} = \text{relu}(\mathbf{W}_r \text{CONCAT}[\mathbf{i}_h; \mathbf{i}_t] + \mathbf{b}_r) \quad (6)$$

where CONCAT means concatenation operation.

3.2.3. Modality discrimination

To explicitly distinguish modalities and utilize type information, modality discrimination is designed to learn type embeddings of modalities. Especially, modality discrimination adopts a type embedding matrix of modalities. For both text and image modality, modality discrimination looks up the matrix to select related embedding and updates unimodal knowledge triples with type information of modalities as follows.

$$\hat{\mathbf{k}}_{\mathcal{X}} = \mathbf{k}_{\mathcal{X}} + \mathbf{e}_{\mathcal{X}} \quad (7)$$

where $\mathcal{X} \in \{\mathcal{T}, \mathcal{V}\}$, $\mathbf{e}_{\mathcal{X}}$ means type embeddings of modalities.

3.3. Triple-level multimodal extractor

Textual knowledge triples $\mathbf{k}_{\mathcal{T}}$ and visual knowledge triples $\mathbf{k}_{\mathcal{V}}$ are still separate and do not assist with each other. The information included in unimodal knowledge triples is insufficient and not comprehensive. Interaction between modalities is necessary to make information of different modalities assist with each other.

Triple-level multimodal extractor is designed to get multimodal knowledge triples $\mathbf{k}_{\mathcal{M}} = \{\mathbf{h}_{\mathcal{M}}, \mathbf{r}_{\mathcal{M}}, \mathbf{t}_{\mathcal{M}}\}$ with assistance of textual and visual triples. We will introduce three main parts in detail: (1) Triple-level interaction residual learner; (2) Hierarchical cross-modality communication block; (3) Gate-based dynamic filter.

3.3.1. Triple-level interaction residual learner

Triple-level interaction residual learner is employed to build the triple-level interaction between textual and visual triples. Creatively, the learner formalizes the interaction between modalities as residual learning. In every residual learning step, the learner projects textual and visual features to the Fourier space and learns

the interaction residual between textual and visual triples in the Fourier space.

Inspired by [7], the learner stacks interaction residual blocks (as Fig. 2 shows) to build the interaction residual at the loop to obtain sufficient cooperation between modalities. In addition, the input of the first residual block is initialized with raw unimodal information, and textual knowledge triples are updated in every block as follows.

$$\begin{cases} \mathbf{k}_{\mathcal{T}}^i = \hat{\mathbf{k}}_{\mathcal{T}}, & i = 0 \\ \mathbf{k}_{\mathcal{T}}^i = F^i(\mathbf{k}_{\mathcal{T}}^{i-1}; \hat{\mathbf{k}}_{\mathcal{T}}) + \mathbf{k}_{\mathcal{T}}^{i-1}, & \forall i = 1, 2, \dots, C \end{cases} \quad (8)$$

where $\mathbf{k}_{\mathcal{T}}^i$ means i th updated textual knowledge triples with interaction between modalities, F^i and C mean i th interaction residual block, the number of stacked residual blocks respectively. Fig. 3 shows the structure of the residual block and we will introduce the block in detail taking i th residual block as an example in next subsection.

Furthermore, text and images are different signals, and Fast Fourier Transform (FFT) can transform diverse signals into Fourier space [53–56]. To make textual and visual features interact effectively in a unified space, the learner adopts FFT to map the textual and visual triples to the frequency space.

$$\text{FFT} : \mathbf{x}_{\mathcal{F}}^{(j)} = \sum_{d=0}^{D-1} e^{-I \frac{2\pi}{D} dj} \mathbf{x}^{(d)}, \forall d, j = 1, 2, \dots, D-1 \quad (9)$$

$$\begin{cases} \mathbf{k}_{\mathcal{T},\mathcal{F}}^{i-1} = \text{FFT}(\mathbf{k}_{\mathcal{T}}^{i-1}), & \forall i = 1, 2, \dots, C \\ \mathbf{k}_{\mathcal{T},\mathcal{F}} = \text{FFT}(\hat{\mathbf{k}}_{\mathcal{T}}) \end{cases} \quad (10)$$

where $\mathbf{x}^{(d)}$ and $\mathbf{x}_{\mathcal{F}}^{(j)}$ mean d th element of input vector and j th element of output vectors in FFT. I and D denote imaginary unit and the dimension of the vectors. $\mathbf{k}_{\mathcal{T},\mathcal{F}}^{i-1}$ and $\mathbf{k}_{\mathcal{T},\mathcal{F}}$ refer to the fourier representations of textual triples $\mathbf{k}_{\mathcal{T}}^{i-1}$ and of visual triples $\mathbf{k}_{\mathcal{T}}$ respectively. Note that $\mathbf{k}_{\mathcal{T},\mathcal{F}}^{i-1}$ and $\mathbf{k}_{\mathcal{T},\mathcal{F}}$ indicate the residual blocks update textual triples (as the main modality) but not visual triples (as the auxiliary modality).

3.3.2. Hierarchical across-modality communication block

Local frequencies represent different information, e.g., low frequencies focus on outlines and high frequencies pay attention to details of images. Hence there are local and global assistance

between modalities. Hierarchical across-modality communication block roles as interaction residual block and designs convolution-based hierarchical attention mechanism variant to capture triple-level and hierarchical interaction across modalities. The variant applies local and global attention query, which both set textual triples as query and apply visual triples as key and value as Fig. 3 shows.

On the one hand, the residual block adopts a convolution block as a local attention query to build the local interaction, which is calculated as follows.

$$\mathbf{f}_l^i = \text{CONV}(\text{STACK}[\mathbf{k}_{\mathcal{T},\mathcal{F}}^{i-1}; \mathbf{k}_{\mathcal{T},\mathcal{F}}]) \quad (11)$$

where $\mathbf{f}_l^i = \{\mathbf{f}_{h,l}^i, \mathbf{f}_{r,l}^i, \mathbf{f}_{t,l}^i\}$ refers to local interaction between unimodal knowledge triples, **CONV** and **STACK** represent for convolution blocks and stack operation respectively.

On the other hand, the residual block uses feedforward networks as a global attention query to capture global interaction across modalities, which is calculated as follows.

$$\mathbf{f}_g^i = \text{FFN}(\text{CONCAT}[\mathbf{k}_{\mathcal{T},\mathcal{F}}^{i-1}; \mathbf{k}_{\mathcal{T},\mathcal{F}}]) \quad (12)$$

where $\mathbf{f}_g^i = \{\mathbf{f}_{h,g}^i, \mathbf{f}_{r,g}^i, \mathbf{f}_{t,g}^i\}$ denotes global interaction between unimodal knowledge triples, **FFN** and **CONCAT** mean feedforward networks and concatenation operation.

The local convolution and global feedforward network build local and global interaction between two modalities. Then, hierarchical interaction scores are calculated by summing the local and global interaction up and adopting softmax to get normalized, which are calculated as follows.

$$\mathbf{s}^{(k)} = \frac{\exp([\mathbf{f}_l^i + \mathbf{f}_g^i]^{(k)})}{\sum_{k=1}^{d_{\mathcal{F}}} \exp([\mathbf{f}_l^i + \mathbf{f}_g^i]^{(k)})} \quad (13)$$

$$\mathbf{f}^i = \tanh(\mathbf{W}(\mathbf{s} \odot \mathbf{k}_{\mathcal{T},\mathcal{F}} + \mathbf{b})) \quad (14)$$

where $\mathbf{s}^{(k)}$ denotes k th score of interaction scores $\mathbf{s} \in \mathbb{R}^{d_{\mathcal{F}}}$, $\mathbf{f}^i = \{\mathbf{f}_h^i, \mathbf{f}_r^i, \mathbf{f}_t^i\}$ represents for interaction residual learned by i th residual block and \odot means element-wise multiply.

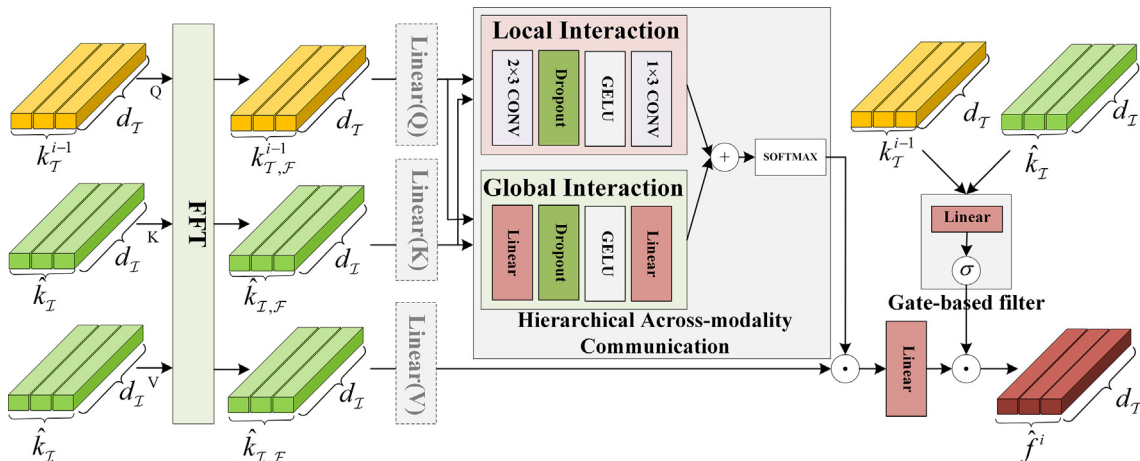


Fig. 3. The architecture of residual block for learning triple-level residual. Taking relations in triples as an example, the textual representations of relations $\mathbf{r}_{\mathcal{T}}^{i-1}$ are set as query Q and the visual representations of relations $\mathbf{r}_{\mathcal{T}}$ are set as key K and value V . According to [52], the Linear(Q), Linear(K), and Linear(V) enhance the representations of QKV in attention mechanism, which improves the capacity of the model.

3.3.3. Gate-based dynamic filter

However, the interaction is not always effective. For example, textual and visual knowledge triples maybe extract different even contrast types of relations, which reduces the overall effects. To decrease the negative impact of contradictory information, a gate-based dynamic filter is applied to remember interaction selectively in each residual block. The gate-based filter is designed by using linear layers and applies sigmoid function as follows.

$$G^i = \sigma(\mathbf{W} \cdot \text{CONCAT}[\mathbf{k}_{\mathcal{T}}^{i-1}; \hat{\mathbf{k}}_{\mathcal{V}}] + \mathbf{b}) \quad (15)$$

where G^i means gate scores for i th interaction residual. Finally, the interaction residual across modalities learned by i th residual block is calculate as follows.

$$\hat{\mathbf{f}}^i = G^i \cdot \mathbf{f}^i \quad (16)$$

where $\hat{\mathbf{f}}^i$ means filtered interaction residual between textual and visual triples built by i th residual block. The textual triples are updated by the interaction residual as follows.

$$\mathbf{k}_{\mathcal{T}}^i = \hat{\mathbf{f}}^i + \mathbf{k}_{\mathcal{T}}^{i-1} \quad (17)$$

After multi-step residual learning, the multimodal triples are calculated as follows.

$$\mathbf{k}_{\mathcal{M}} = \text{CONCAT}[\mathbf{k}_{\mathcal{T}}^C; \hat{\mathbf{k}}_{\mathcal{V}}] \quad (18)$$

where $\mathbf{k}_{\mathcal{T}}^C$ means the latest textual triples after multi-step residual learning (i.e., the textual triples after C th update in interaction residual learning).

3.4. Intra-triple translation supervised decoder

Taking the interaction-enhanced multimodal triples $\mathbf{k}_{\mathcal{M}} = \{\mathbf{h}_{\mathcal{M}}, \mathbf{r}_{\mathcal{M}}, \mathbf{t}_{\mathcal{M}}\}$ as inputs, the decoder adopts an intra-triple translation supervised prototype network to predict the relations in query set.

3.4.1. Intra-triple translation constraint

The full features are similar when the inputs are slightly different, which causes ambiguity in relation identification. As Fig. 4 shows, given inputs including two entities, the relation from LinDaiyu to XueYan is *employer*, and the relation from XueYan to

LinDaiyu is corresponding to be *servant girl*, which own similar features but contrast directions. It is difficult for the implicit extraction method to identify the two relations. In addition, the full features include all aspects of information, which is not always related to extracting relations.

Hence, it is necessary to mine essential information (e.g., directions of relations) for avoiding ambiguity. [40] found translation principle in word vector space, e.g., as Fig. 4 shows, $[\mathbf{r}_1 = \mathbf{t}(\text{XueYan}) - \mathbf{h}(\text{LinDaiyu})] \approx [\mathbf{r}_1' = \mathbf{t}(\text{FengEr}) - \mathbf{h}(\text{WangXiFeng})]$, $[\mathbf{r}_2 = \mathbf{t}(\text{LinDaiyu}) - \mathbf{h}(\text{XueYan})] \approx [\mathbf{r}_2' = \mathbf{t}(\text{WangXiFeng}) - \mathbf{h}(\text{FengEr})]$. That is, the embedding difference of two entities include features related to relations. As [41,43,42] shows, translation principle $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ can be applied as self-supervised information to mine essential characteristics of relations.

The decoder employs translation self-supervised information inside multimodal knowledge triples $\mathbf{k}_{\mathcal{M}} = \{\mathbf{h}_{\mathcal{M}}, \mathbf{r}_{\mathcal{M}}, \mathbf{t}_{\mathcal{M}}\}$ to strengthen the ability to extract essential features of $\mathbf{r}_{\mathcal{M}}$. As [41,43,42] do, the constraint utilizes $\mathbf{h}_{\mathcal{M}} + \mathbf{r}_{\mathcal{M}} \approx \mathbf{t}_{\mathcal{M}}$ as self-supervised information to build translation inside triples to obtain significant features (e.g., the direction of relation) of $\mathbf{r}_{\mathcal{M}}$ from full features. And our objective is to minimize the distance between $\mathbf{h}_{\mathcal{M}} + \mathbf{r}_{\mathcal{M}}$ and $\mathbf{t}_{\mathcal{M}}$ with mean square error (MSE) loss as follows.

$$\text{Loss}_{\text{trans}} = \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K (\mathbf{h}_{\mathcal{M}_n}^k + \mathbf{r}_{\mathcal{M}_n}^k - \mathbf{t}_{\mathcal{M}_n}^k)^2 \quad (19)$$

where N, K mean N way and K shot in prototype social relation extraction respectively. $\mathbf{k}_{\mathcal{M}_n}^k = \{\mathbf{h}_{\mathcal{M}_n}^k, \mathbf{r}_{\mathcal{M}_n}^k, \mathbf{t}_{\mathcal{M}_n}^k\}$ means the multimodal knowledge triple in k th sample which belongs to n th relation in the support set.

3.4.2. Few-shot relation prototype classifier

In prototype social relation extraction, support and query set are respectively set as seen and unseen relations to capture transfer from rich-labeled relations to poor-labeled relations by learning meta features. The classifier computes the mean of multimodal relation features $\mathbf{r}_{\mathcal{M}_n}^k \in \mathbb{R}^{d_{\mathcal{M}}}$ in support set as prototypes \mathbf{p}_n (i.e., meta relations).

$$\mathbf{p}_n = \frac{1}{K} \sum_{k=1}^K \mathbf{r}_{\mathcal{M}_n}^k, \forall n = 1, 2, \dots, N_{\text{way}}; \forall k = 1, 2, \dots, K_{\text{shot}} \quad (20)$$

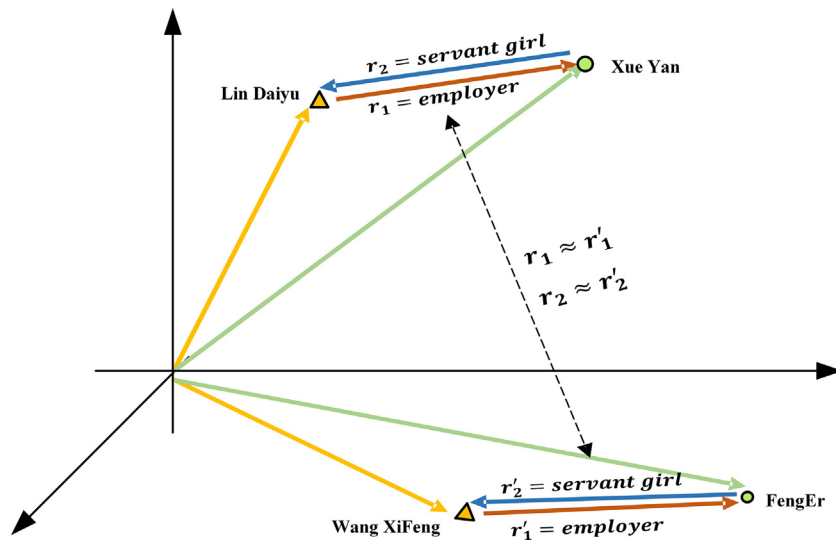


Fig. 4. Illustration of translation principle. The triangles and circles indicate head and tail entities. The brown and blue arrows refer to relations between different entities, e.g., the brown arrow from entity LinDaiyu to XueYan means $\mathbf{r}_1 = \text{employer}$ and the brown arrow from entity WangXiFeng to FengEr represents for $\mathbf{r}_1' = \text{employer}$, which satisfies $\mathbf{r}_1 \approx \mathbf{r}_1'$.

where N and K represent N classes and K samples in each class at episodes, $\mathbf{p}_n \in \mathbb{R}^{d_{\#}}$ means meta relation of n th relation in support set. Then, the classifier computes the distances between meta relations and queries $\mathbf{r}_{\#q} \in \mathbb{R}^{d_{\#}}$.

$$d(q, n) = \sum_{i=1}^{d_{\#}} (\mathbf{r}_{\#q}^{(i)} - \mathbf{p}_n^{(i)})^2, \forall q = 1, 2, \dots, Q; \forall n = 1, 2, \dots, N\text{-way} \quad (21)$$

where $Q, \mathbf{r}_{\#q}, (i)$ mean the number of the query samples, multi-modal representation of relation in q th query and i th element respectively, and $d(q, n)$ means the distance between n th meta relation and q th query. Finally, the classifier adopts *softmax* on distance matrix to compute probability $p(q, n)$, which means probability of q th query is n th relation. And the relation which has maximum probability (related to minimum distance) is the prediction as shown as follows.

$$p(q, n) = \frac{\exp(-d(q, n))}{\sum_{j=1}^N \exp(-d(q, j))}, \forall n = 1, 2, \dots, N\text{-way} \quad (22)$$

$$r_q = \arg \max_n p(q, n), \forall n = 1, 2, \dots, N\text{-way} \quad (23)$$

where r_q means relation of q th query.

3.4.3. Adjusted multi-loss

For training, we employ cross entropy as classification loss for the task.

$$Loss_{cls} = - \sum_{q=1}^Q \sum_{n=1}^N y(q, n) \log p(q, n) \quad (24)$$

where $y(q, n)$ is the true label. Combined with translation loss, the total loss is calculated as follows.

$$Loss = Loss_{cls} + \lambda \cdot Loss_{trans} \quad (25)$$

where λ means adjustment factor.

3.5. Optimization

In order to reveal a detailed data stream and training process, we show the optimization process in Algorithm 1.

Algorithm 1 Training Process of TSPNet

Input: the set of support samples, support set S ; the set of query samples, query set Q ;

Output: the relations of query set $r_q, q \in Q$

1: **initial:** Triple-level Unimodal Encoder \rightarrow TUE

2: Triple-level Multimodal Extractor \rightarrow TME

3: **repeat**

4: Sample support set and query set for current episode
 $S_n \in S, Q_n \in Q$

5: **for** $s \in S_n, q \in Q_n$ **do**

6: $[\mathbf{t}_s; \mathbf{t}_q] = \text{BERT}(s[\text{text}]; q[\text{text}])$

7: $[\mathbf{i}_s; \mathbf{i}_q] = \text{FaceNet}(s[\text{images}]; q[\text{images}])$

8: $[\mathbf{k}_{\mathcal{T},s}; \mathbf{k}_{\mathcal{T},q}] = \text{TUE}([\mathbf{t}_s; \mathbf{i}_s]), [\mathbf{k}_{\mathcal{T},q}; \mathbf{k}_{\mathcal{T},q}] = \text{TUE}([\mathbf{t}_q; \mathbf{i}_q])$

9: $\mathbf{k}_{\#s} = \text{TME}(\mathbf{k}_{\mathcal{T},s}; \mathbf{k}_{\mathcal{T},s}), \mathbf{k}_{\#q} = \text{TME}(\mathbf{k}_{\mathcal{T},q}; \mathbf{k}_{\mathcal{T},q})$

10: **for** $s \in S_n$ **do**

11: Select $\mathbf{r}_{\#s}$ from $\mathbf{k}_{\#s}$

12: **for** $\forall n = 1, 2, \dots, N\text{-way}$ **do**

13: $\mathbf{p}_n = \frac{1}{K} \sum_{k=1}^K \mathbf{r}_{\#n,s}^k, \forall k = 1, 2, \dots, K\text{-shot}$ **end for**

14: **for** $q \in Q_n$ **do**

15: Select $\mathbf{r}_{\#q}$ from $\mathbf{k}_{\#q}$

16:

$d(q, n) = (\mathbf{r}_{\#q} - \mathbf{p}_n)^2, \forall q = 1, 2, \dots, Q; \forall n = 1, 2, \dots, N\text{-way}$

17: $p(q, n) = \text{softmax}(-d(q, n))$

18: $r_q = \arg \max_n p(q, n)$ **end for**

19: $\mathcal{L} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{trans}$

20: Collect $\nabla \mathcal{L}$ via backward

21: $\mathcal{L} = \mathcal{L} - lr \cdot \nabla \mathcal{L}$

22: **until** All training episodes finished

23: **return** r_q

4. Experiments

4.1. Dataset and evaluation metrics

We evaluate our TSPNet model on FC, DRC, and OM datasets [18] which are widely used in few-shot multimodal social relation extraction. In the three datasets, the text and images are collected from the book and TV of the four Chinese masterpieces respectively. Statistics on all relations of FC, DRC, and OM datasets show in Fig. 5. Following [18], we split data in the same way 2 and use widely accepted metric *accuracy*(Acc) for evaluating our model.

4.2. Hyperparameter setting

In the embedding layer, we set the dimension of visual and textual embedding to 512 and 768 respectively. For textual triple $\mathbf{k}_{\mathcal{T}}$, we set the dimension of $\mathbf{r}_{\mathcal{T}}, \mathbf{h}_{\mathcal{T}}, \mathbf{t}_{\mathcal{T}}$ to 768. As for visual triple $\mathbf{k}_{\mathcal{V}}$, we set the dimension of $\mathbf{r}_{\mathcal{V}}, \mathbf{h}_{\mathcal{V}}, \mathbf{t}_{\mathcal{V}}$ to 1024. And we set the dimension of multimodal triple $\mathbf{r}_{\#}, \mathbf{h}_{\#}, \mathbf{t}_{\#}$ to 1792. The number of residual blocks is set to 2, and the kernel size of convolution blocks is 3. In addition, we fix the maximum length of sentence at 512 and resize the images to 160×160.

We set the learning rate to 0.1 and weight decay to 0.01. And we set the batch size to 2 and the dropout rate to 0.5. Besides, we train 20000 iterations. In the experiments, we adopt the optimizer AdamW to update the weights of our model. In addition, we train our model on two 2080 Ti and fix all seeds for absolute comparison.

4.3. Overall performance

We reimplement the following existing few-shot learning methods and all experiments adopt same hyperparameter settings for absolute comparison:

GNN [47] utilized a graph network and regarded transfer between support and query set as graph aggregation. The query can be predicted via graph reasoning.

SNAIL [48] employed a temporal convolution network to remember the past experience, and used a soft attention mechanism to verify specific information for query.

SIAMESE [49] computed the similarities between support and query set. The model predicts the query according to the similarities.

text(BERT) [18] applied BERT to extract social relations from text.

FL-MSRE [18] applied BERT and FaceNet to extract social relations via multimodal information from text and images.

Compared with GNN, SNAIL, SIAMESE, and text(BERT), our model outperforms the unimodal methods based on textual information. The methods only utilize text for social relation extraction

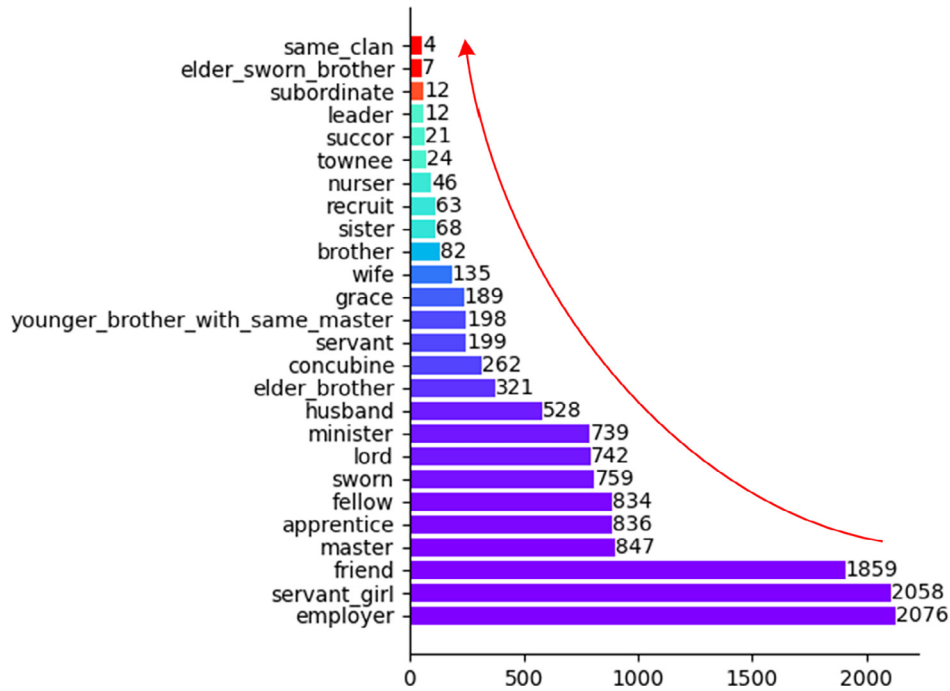


Fig. 5. The distribution of relations across three datasets. The abscissa means the number of relations. There is a serious imbalance in relations.

Table 2

Statistics on FC, DRC and OM dataset. #relation/#entity/#triple are respectively the number of relations/characters/triples/ in datasets, while #sentence/#images are respectively the number of sentences/images in datasets. **Note** that there might be some entities appearing in train, val, and test set but absolutely no relation can appear in two or more datasets.

Datasets		#relation	#entity	#triple	#sentence	#images
FC	train	14	98	4527	6485	3716
	val	5	31	560		
	test	5	46	1398		
DRC	train	3	11	319	1828	560
	val	3	39	737		
	test	3	38	773		
OM	train	5	16	79	1489	1178
	val	5	24	772		
	test	5	21	638		

task, but our model adopts images with a multimodal information enhancement. The enhancement can help alleviate the problem of insufficient information for the task.

Compared with FL-MSRE, our model also outperforms FL-MSRE. Though FL-MSRE adopts multimodal information (i.e., images and text), it simply employs a concatenation operation to fuse textual and visual features, which is weak to learn the assistance between two modalities. Our model learns the assistance to capture comprehensive multimodal information. In addition, FL-MSRE is a task-driven method to implicitly extract relations by utilizing full features, which lacks explicit supervision. However, our model adopts the translation principle as an explicit constraint, which improves the ability to mine core features of relations from full features.

Compared with text(BERT) and FL-MSRE, our model makes larger improvements in smaller-scale DRC and OM datasets than in larger-scale FC dataset. It denotes that our model owns better adaptability to extract low resource social relations.

4.4. Ablation study

To evaluate the effectiveness of translation constraint and multimodal interaction learner, we do some experiments on FC, DRC, and OM datasets without the constraint and learner respectively.

As Table 6 shows, our model drops 0.3%, 3.79% and 3.35% on FC, DRC and OM datasets without translation constraint respectively. This is because the essential features of relations become more abstract and difficult to be mined without translation constraint, which is more obvious in datasets with a smaller scale (e.g., DRC and OM).

To explore the effects of multimodal interaction learner, we remove the learner and concatenate features of two modalities. Our model drops 0.36%, 3.24%, and 2.09% on FC, DRC, and OM datasets without the learner. Because the model can not build interaction between two modalities, the multimodal features are not sufficient and comprehensive enough.

4.5. Further analysis

For further analyzing our model, we study generalization analysis across datasets, effects of multimodal fusion, impacts of the numbers of residual layers, and effects of translation constraint as following subsections.

4.5.1. Generalization analysis across datasets

Table 7 shows the evaluation on across datasets. The comparison of all results shows that our model performs better. Even the results of our model on inter-dataset exceed the results of FL-MSRE and text(BERT) on intra-dataset. We believe that our model owns a strong generalization ability to extract relations across datasets with weaker dependence on data. In addition, to discuss how the way split datasets affects model performance, we simply exchange the train and test set in the FC dataset and conduct 3way-1shot experiments for discussion as Table 8 shows. The performance of both FL-MSRE and our model dropped as the size of the train set becomes smaller after the swap, but our model still outperforms the FL-MSRE, which demonstrates our model is less dependent on data and shows stronger generalization ability.

4.5.2. Effects of multimodal fusion

Multimodal fusion learns a more accurate dividing plane to identify relations. As Fig. 6 shows, when using only unimodal information, the model incorrectly predicts the relation of query to be

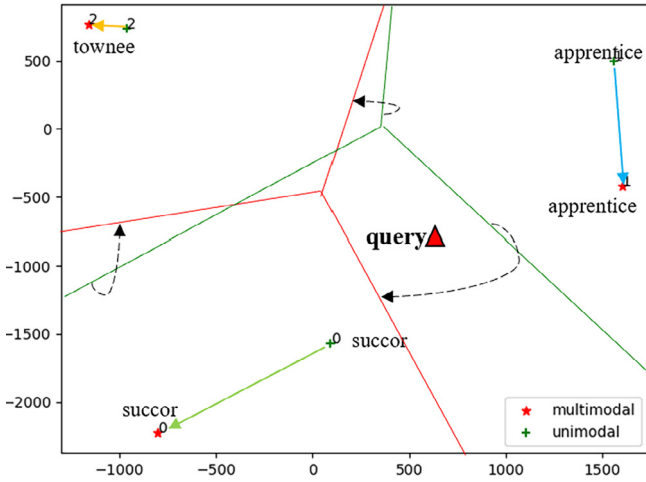


Fig. 6. Comparison between unimodal and multimodal relation extraction. The red asterisk and green plus sign denote multimodal and unimodal meta relations respectively. The red triangle means query sample. We adopt t-SNE to visualize the representations of relations.

Table 3

Overall performance compared to the state-of-the-art methods on FC test dataset. *same* and *differ* means two entities appear in same scene and different scenes respectively. The results are averaged by 8 tests. The best result of all compared methods is displayed in bold. * indicates our reimplement under the settings as same as our model.

models	scene	5way-3shot	5way-1shot	3way-3shot	3way-1shot
GNN* [47]	–	75.40 ± 0.33	70.90 ± 0.26	83.50 ± 0.31	82.80 ± 0.37
SNAIL* [48]	–	75.49 ± 0.24	72.29 ± 0.21	84.09 ± 0.38	83.03 ± 0.34
SIAMESE* [49]	–	76.17 ± 0.29	75.18 ± 0.30	87.14 ± 0.54	84.48 ± 0.28
text(BERT) [18]	–	75.03 ± 0.35	73.93 ± 0.01	86.75 ± 0.32	83.84 ± 0.37
FL-MSRE* [18]	same	80.36 ± 0.32	77.05 ± 0.34	88.46 ± 0.35	87.58 ± 0.31
	differ	78.76 ± 0.44	76.78 ± 0.32	88.55 ± 0.28	86.23 ± 0.12
ours	same	81.97 ± 0.35	79.71 ± 0.21	89.25 ± 0.47	87.85 ± 0.62
	differ	82.27 ± 0.40	79.10 ± 0.16	90.61 ± 0.48	88.17 ± 0.26

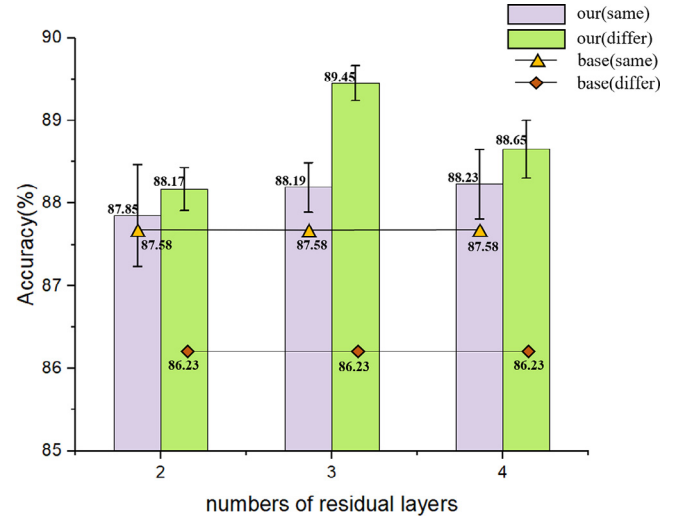


Fig. 7. The effect of different numbers of residual layers on FC dataset for 3-way 1-shot from different scenes. Because the baseline (i.e., FL-MSRE) does not adopt residual learning, the accuracy is not changed with the number of residual layers.

Table 4

Overall performance compared to the state-of-the-art methods on DRC test dataset. *same* and *differ* means two entities appear in same scene and different scenes respectively. The results are averaged by 8 tests. The best result of all compared methods is displayed in bold. * indicates our reimplement under the settings as same as our model.

models	scene	3way-3shot	3way-1shot
text (BERT) [18]	–	51.29 ± 0.16	40.00 ± 0.11
FL-MSRE* [18]	same	65.42 ± 0.48	55.35 ± 0.42
	differ	70.36 ± 0.47	57.18 ± 0.66
ours	same	72.51 ± 0.50	59.80 ± 0.60
	differ	76.80 ± 0.53	60.72 ± 0.60

Table 5

Overall performance compared to the state-of-the-art methods on OM test dataset. *same* and *differ* means two entities appear in same scene and different scenes respectively. The results are averaged by 8 tests. The best result of all compared methods is displayed in bold. * indicates our reimplement under the settings as same as our model.

models	scene	3way-3shot	3way-1shot
text(BERT)[18]	–	48.35 ± 0.43	46.28 ± 0.15
FL-MSRE* [18]	same	56.73 ± 0.66	53.25 ± 0.51
	differ	58.82 ± 0.27	55.39 ± 0.41
ours	same	63.89 ± 0.39	56.84 ± 0.84
	differ	64.48 ± 0.50	60.77 ± 0.45

Table 6

Ablation results of the different scene and 3-way 1-shot learning on FC, DRC, and OM dataset without translation constraint and multimodal interaction respectively. The results are averaged by 8 tests. The best result of all compared methods is displayed in bold.

models	FC	DRC	OM
ours	88.17 ± 0.26	60.72 ± 0.60	60.77 ± 0.45
-translation	87.87 ± 0.29	56.93 ± 0.82	57.42 ± 0.61
-interaction	87.81 ± 0.49	57.48 ± 0.65	58.68 ± 0.81

succor. However, after utilizing multimodal information via multimodal fusion, the dividing plane is adjusted and the model correctly predicts the relation to be *apprentice*.

4.5.3. Effects of the numbers of residual layers

Table 3 shows the improvement on FC dataset is less than DRC and OM datasets. It might be not adequate for FC dataset to set residual layers to 2. On the other hand, we want to explore the performance with different number of residual layers. Hence, we evaluate on FC dataset in both different and same scenes with a 3-way 1-shot episode. As shown in Fig. 7, with an increment of the number of residual layers, the performance becomes better. However, the effects get worse when our model owns too many residual layers, because the model has reached the saturation performance of residual learning [7]. Though the performance becomes worse with excessive residual layers, the performance is still better than baseline. Therefore, our model can choose a suitable number of residual layers to improve the performance and not become overfit with a slightly deeper stack (see Tables 4 and 5).

4.5.4. Effects of translation constraint

To compare the ability to extract essential features of FL-MSRE and TSPNet and analyze the effect of translation constraint, we randomly select three relations and each relation owns a hundred samples to learn the representations. Fig. 8(a) and (b) show the representations learned by FL-MSRE and our TSPNet respectively. Comparing the two figures, the representations learned by FL-MSRE have small inter-relation distances and large intra-relation distances, which easily causes the ambiguity in recognizing the relation of a query. However, our TSPNet has a larger inter-relation distance and smaller intra-relation distance. In other words, the representations learned by TSPNet are more separable, and it is easier to recognize the relation of a query without ambiguity. In addition, the representations learned by TSPNet are more concentrated, which means that TSPNet focuses more on the

Table 8

We exchange the train and test set in the FC dataset to discuss how the way to split datasets affects model performance. The “exchanged FC” column shows the new size of the train and test set.

exchanged FC	models	3way-1shot
train:1398	FL-MSRE[18]	69.02 ± 0.45
test:4527	ours	73.88 ± 0.23

essential features of the relations and is not affected by the samples. The results verify that adopting the translation principle as a explicit constraint can strengthen the ability to extract essential features of relations.

4.6. Case study

We select some typical cases to analyze the effectiveness of our model as Fig. 9 shows. In case (a), text(BERT) can identify the relation *servant* but FL-MSRE identifies the relation to *sworn* due to wrong multimodal fusion, and ours(TSPNet) can correctly extraction the relation based on suitable fusion and dynamic filter. In case (b) and (c), when inputs are similar, text(BERT) and FL-MSRE can not distinguish the directions based on implicit features. In case (d), text(BERT) identifies the relation to *servant* only based on the text, which lacks comprehensive multimodal information. Furthermore, to illustrate the identification ambiguity of text (BERT) and FL-MSRE when inputs are similar, we choose *employer* relation and *servant girl* relation, which own same entities and input contents but different directions, as Fig. 10 shows. Because the relations own same input contents, text(BERT) and FL-MSRE predict the relation of query sample Q1 to be *servant girl* but the ground truth is *employer* and predict the label of query sample Q2 to be *employer* but the ground truth is *servant girl*. However, due to adopting translation principle as a constraint to improve extracting essential features from full features, TSPNet owns a stronger ability to alleviate the identification ambiguity when inputs are similar. For example, Fig. 10 shows the different directions of meta relations extracted by our model. The directions from head entities to tail entities are consistent with the real situation. In “predictions” row, because the input contents are nearly same, the distances between queries and the representations learned by text(BERT) and FL-MSRE are close (especially text(BERT)). On the contrary, the distance matrix computed by our model are significantly in gap, which is consistent with the actual situation that input contents are almost same but the direction between entities is completely different.

Table 7

OM to DRC means that the model is trained on OM dataset but tested on DRC dataset; DRC to OM means that the model is trained on DRC dataset but tested on OM dataset. The results show the generalization ability of model on different datasets. The results are averaged by 8 tests. The best result of all compared methods is displayed in bold. * means our reimplement under the settings as same as our model.

models	scene	OM to DRC		DRC to OM	
		3way-3shot	3way-1shot	3way-3shot	3way-1shot
text(BERT)[18]	–	40.74 ± 0.18	38.38 ± 0.16	44.09 ± 0.69	37.27 ± 0.69
FL-MSRE*[18]	same	65.25 ± 0.67	51.76 ± 0.73	48.84 ± 0.75	50.31 ± 0.52
	differ	61.59 ± 0.28	50.37 ± 0.50	49.66 ± 0.32	51.45 ± 0.59
ours	same	76.08 ± 0.54	62.02 ± 0.86	58.19 ± 0.65	53.70 ± 0.84
	differ	73.19 ± 0.56	61.30 ± 0.42	56.65 ± 0.70	57.14 ± 0.61

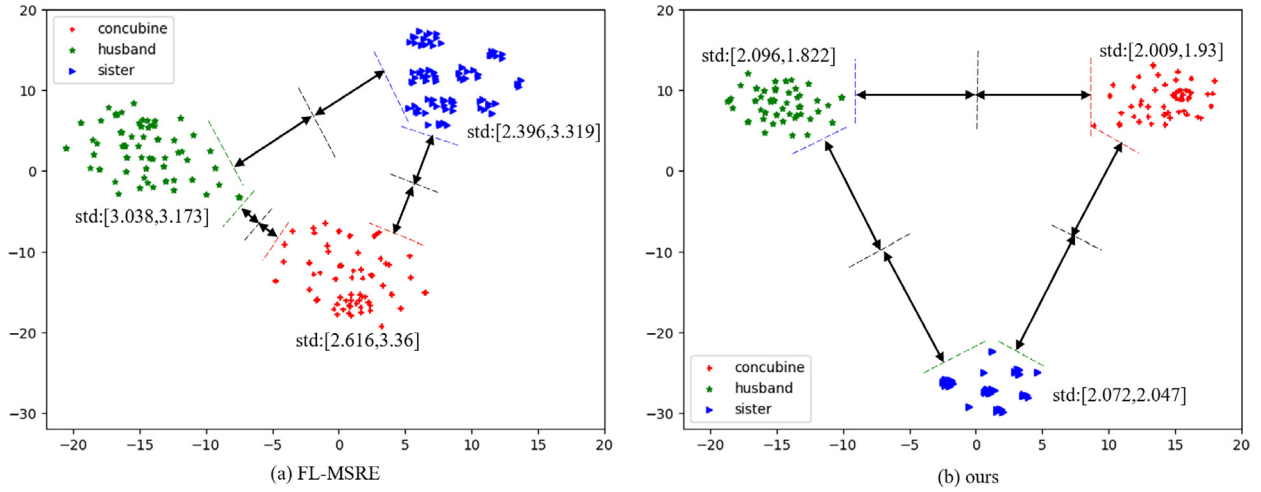


Fig. 8. Illustration of representations learned by FL-MSRE and our model. The *std* indicates the standard deviation of the relation cluster, which measures the concentration of the cluster. Two figures are set to the same scale and we adopt t-SNE to visualize the representations.



Fig. 9. The head and tail entities are highlighted in red and blue font in text, and are bounded by the red and blue boxes in images respectively. Ground Truth means the true label.

5. Conclusion

This paper proposes a novel framework named TSPNet for multimodal social relation extraction in the few-shot scenarios. Considering ambiguity to extract implicitly few-shot social relations, the TSPNet adopts intra-triple translation as an explicit constraint to obtain essential features of relations based on triple-level multimodal features. In addition, to earn triple-level multimodal features, we regard the interaction across modalities as residual learning and design a learner to capture the interaction. The

learner employs a hierarchical attention variant in each residual block. The experiments show our model achieves better performance than existing methods on few-shot multimodal social relation extraction. Further analysis suggests that our model performs more robust and has a stronger generalization ability, and ablation experiments demonstrate the effectiveness of multimodal interaction and intra-triple translation constraint. In the future, we plan to expand our model to open-main relation extraction to strengthen the ability to extract open-world relations instead of only social relations.



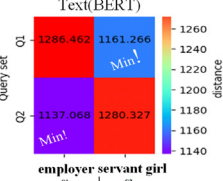
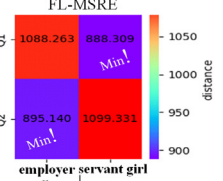
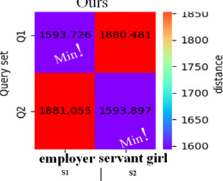
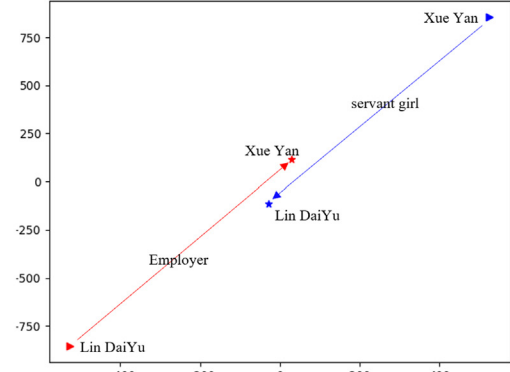
Set	Support		Query	
Input texts	Wangxifeng ordered someone to give a lotus-colored flower tent, and some brocade quilts. LinDaiyu brought only two people here: one was the nanny Wang from childhood, and the other was a ten-year-old girl XueYan .		Wang XiFeng ordered someone: “Hurry up and tell him to come quickly.” Feng Er hurriedly said: Miss Lin sent someone to invite. He response after many invitations; as soon as my grandma came in, I told him to go.	
Input images				
Ground Truth	S1: LinDaiYu → XueYan : employer	S2: XueYan → LinDaiYu : servant girl	Q1: WangXiFeng → FengEr : employer	Q2: FengEr → WangXiFeng : servant girl
Predictions	<div> <div>Text(BERT)</div>  <div>Text(BERT): Query 1: servant girl ✗ Query 2: employer ✗</div> </div> <div> <div>FL-MSRE</div>  <div>FL-MSRE: Query 1: servant girl ✗ Query 2: employer ✗</div> </div> <div> <div>Ours</div>  <div>ours: Query 1: employer ✓ Query 2: servant girl ✓</div> </div>			
Relation directions by ours				

Fig. 10. The head and tail entities are highlighted in bold font and are bounded by the red and blue boxes in input images respectively. In the “Ground Truth” row, head and tail entities are highlighted in red and blue font, which shows the directions of true relations. In the “Relation directions by ours” row, triangle and star markers mean head and tail entities. And S1 and S2 are distinguished by red and blue.

CRediT authorship contribution statement

Hankun Kang: Conceptualization, Methodology, Software, Validation, Visualization, Writing-original-draft, Formal-analysis, Data-curation, Resources. **Xiaoyu Li:** Data-curation, Writing-review-editing, Supervision, Resources, Project-administration. **Li Jin:** Investigation, Writing-review-editing, Supervision, Project-administration. **Chunbo Liu:** Investigation, Supervision, Writing-review-editing. **Zequn Zhang:** Investigation, Writing-review-editing, Supervision. **Shuchao Li:** Writing-review-editing, Investigation. **Yanan Zhang:** Writing-review-editing.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, E. Uziel, Social media recommendation based on people and tags, in: Proceedings of the international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 194–201.
- [2] Y.-M. Li, C.-W. Chen, A synthetical approach for blog recommendation: Combining trust, social relation, and semantic analysis, Expert Syst. Appl. 36 (3) (2009) 6536–6547.
- [3] Y. Jiang, H. Ma, Y. Liu, Z. Li, L. Chang, Enhancing social recommendation via two-level graph attentional networks, Neurocomputing 449 (2021) 71–84.
- [4] X. Zhao, J. Yuan, G. Li, X. Chen, Z. Li, Relationship strength estimation for online social networks with the study on facebook, Neurocomputing 95 (2012) 89–97.
- [5] E.W. Cheney, W.A. Light, A course in approximation theory, vol. 101, American Mathematical Society, 2009.
- [6] M.J.D. Powell et al., Approximation theory and methods, Cambridge University Press, 1981.
- [7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [8] Y. Cao, J. Kuang, M. Gao, A. Zhou, Y. Wen, T.-S. Chua, Learning relation prototype from unlabeled texts for long-tail relation extraction, IEEE Trans. Knowl. Data Eng. (01) (2021) 1.
- [9] W. Wen, Y. Liu, C. Ouyang, Q. Lin, T. Chung, Enhanced prototypical network for few-shot relation extraction, Inform. Process. Manage. 58 (4) (2021) 102596.

- [10] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1199–1208.
- [11] J. Yuan, H. Guo, Z. Jin, H. Jin, X. Zhang, J. Luo, One-shot learning for fine-grained relation extraction via convolutional siamese neural network, in: Proceedings of the IEEE International Conference on Big Data, 2017, pp. 2194–2199.
- [12] C. Yang, J. Zhang, H. Wang, S. Li, M. Kim, M. Walker, Y. Xiao, J. Han, Relation learning on social networks with multi-modal graph edge variational autoencoders, in: Proceedings of the International Conference on Web Search and Data Mining, 2020, pp. 699–707.
- [13] T. Xu, P. Zhou, L. Hu, X. He, Y. Hu, E. Chen, Socializing the videos: A multimodal approach for social relation recognition, *ACM Trans. Multimedia Comput., Commun., Appl.* 17 (1) (2021) 1–23.
- [14] B. Zhang, F. Yu, Y. Gao, T. Ren, G. Wu, Joint learning for relationship and interaction analysis in video with multimodal feature fusion, in: Proceedings of the ACM International Conference on Multimedia, 2021, pp. 4848–4852.
- [15] Z. Liu, W. Hou, J. Zhang, C. Cao, B. Wu, A multimodal approach for multiple-relation extraction in videos, *Multimedia Tools Appl.* 81 (4) (2022) 4909–4934.
- [16] S. Yang, G. Li, Y. Yu, Cross-modal relationship inference for grounding referring expressions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4145–4154.
- [17] C. Cao, C. Yan, F. Li, Z. Liu, Z. Wang, B. Wu, Recognizing characters and relationships from videos via spatial-temporal and multimodal cues, in: Proceedings of the IEEE International Conference on Big Knowledge, 2021, pp. 174–181.
- [18] H. Wan, M. Zhang, J. Du, Z. Huang, Y. Yang, J.Z. Pan, Fl-msre: A few-shot learning based approach to multimodal social relation extraction, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 13916–13923.
- [19] S.K. Dash, Y. Sureshchandra, Y. Mishra, P. Pakray, R. Das, A. Gelbukh, Multimodal learning based spatial relation identification, *Computación y Sistemas* 24 (3) (2020) 1327–1335.
- [20] C. Zheng, J. Feng, Z. Fu, Y. Cai, Q. Li, T. Wang, Multimodal relation extraction with efficient graph alignment, in: Proceedings of the ACM International Conference on Multimedia, 2021, pp. 5298–5306.
- [21] F. Yu, D. Wang, B. Zhang, T. Ren, Deep relationship analysis in video with multimodal feature fusion, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 4640–4644.
- [22] V. Anand, R. Ramesh, Z. Wang, Y. Feng, J. Feng, W. Lyu, T. Zhu, S. Yuan, C.-Y. Lin, Story semantic relationships from multimodal cognitions, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 4650–4654.
- [23] S. Pingali, S. Yadav, P. Dutta, S. Saha, Multimodal graph-based transformer framework for biomedical relation extraction, in: Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021, pp. 3741–3747.
- [24] B. Lv, L. Jin, X. Li, X. Sun, Z. Guo, Z. Zhang, S. Li, Dpnet: domain-aware prototypical network for interdisciplinary few-shot relation classification, *Appl. Intell.* (2022) 1–16.
- [25] L. Wang, P. Lin, J. Cheng, F. Liu, X. Ma, J. Yin, Visual relationship detection with recurrent attention and negative sampling, *Neurocomputing* 434 (2021) 55–66.
- [26] W. Wang, M. Wang, S. Wang, G. Long, L. Yao, G. Qi, Y. Chen, One-shot learning for long-tail visual relation detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12225–12232.
- [27] T. Gao, X. Han, Z. Liu, M. Sun, Hybrid attention-based prototypical networks for noisy few-shot relation classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 6407–6414.
- [28] T. Gao, X. Han, R. Xie, Z. Liu, F. Lin, L. Lin, M. Sun, Neural snowball for few-shot relation learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 7772–7779.
- [29] M. Qu, T. Gao, L.-P. Xhonneux, J. Tang, Few-shot relation extraction via bayesian meta-learning on relation graphs, in: Proceedings of the International Conference on Machine Learning, 2020, pp. 7867–7876.
- [30] A. Obamuyide, A. Vlachos, et al., Meta-learning improves lifelong relation extraction, in: Proceedings of the ACL – 4th Workshop on Representation Learning for NLP, Repl4NLP, 2019, pp. 224–229.
- [31] Y. Lin, X. Han, R. Xie, Z. Liu, M. Sun, Knowledge representation learning: A quantitative review, *arXiv preprint arXiv:1812.10901*.
- [32] R. Jenatton, N. Le Roux, A. Bordes, G. Obozinski, A latent factor model for highly multi-relational data, in: Proceedings of the International Conference on Neural Information Processing Systems, 2012, pp. 3176–3184.
- [33] A. Bordes, X. Glorot, J. Weston, Y. Bengio, A semantic matching energy function for learning with multi-relational data, *Mach. Learn.* 94 (2) (2014) 233–259.
- [34] A. Bordes, X. Glorot, J. Weston, Y. Bengio, Joint learning of words and meaning representations for open-text semantic parsing, in: Proceedings of the Artificial intelligence and statistics, 2012, pp. 127–135.
- [35] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmman, S. Sun, W. Zhang, Knowledge vault: A web-scale approach to probabilistic knowledge fusion, in: Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 601–610.
- [36] R. Socher, D. Chen, C.D. Manning, A. Ng, Reasoning with neural tensor networks for knowledge base completion, in: Proceedings of the International Conference on Neural Information Processing Systems, 2013, pp. 926–934.
- [37] Q. Liu, H. Jiang, A. Evdokimov, Z.-H. Ling, X. Zhu, S. Wei, Y. Hu, Probabilistic reasoning via deep learning: Neural association models, *arXiv preprint arXiv:1603.07704*.
- [38] M. Nickel, V. Tresp, H.-P. Kriegel, Factorizing yago: scalable machine learning for linked data, in: Proceedings of the international conference on World Wide Web, 2012, pp. 271–280.
- [39] M. Nickel, V. Tresp, H.-P. Kriegel, A three-way model for collective learning on multi-relational data, in: Proceedings of the International Conference on Machine Learning, 2011, pp. 809–816.
- [40] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the International Conference on Neural Information Processing Systems, 2013, pp. 3111–3119.
- [41] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Proceedings of the International Conference on Neural Information Processing Systems, 2013, pp. 2787–2795.
- [42] G. Ji, S. He, L. Xu, K. Liu, J. Zhao, Knowledge graph embedding via dynamic mapping matrix, in: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, 2015, pp. 687–696.
- [43] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2014, pp. 1112–1119.
- [44] Y. Xue, J. Jin, A. Song, Y. Zhang, Y. Liu, K. Wang, Relation-based multi-type aware knowledge graph embedding, *Neurocomputing* 456 (2021) 11–22.
- [45] C. Li, X. Peng, Y. Niu, S. Zhang, H. Peng, C. Zhou, J. Li, Learning graph attention-aware knowledge graph embedding, *Neurocomputing* 461 (2021) 516–529.
- [46] M. Li, Z. Sun, S. Zhang, W. Zhang, Enhancing knowledge graph embedding with relational constraints, *Neurocomputing* 429 (2021) 77–88.
- [47] V. Garcia, J. Bruna, Few-shot learning with graph neural networks, in: Proceedings of the International Conference on Learning Representations, 2018.
- [48] N. Mishra, M. Rohaninejad, X. Chen, P. Abbeel, A simple neural attentive meta-learner, in: Proceedings of the International Conference on Learning Representations, 2018.
- [49] G. Koch, R. Zemel, R. Salakhutdinov, et al., Siamese neural networks for one-shot image recognition, in: Proceedings of the International Conference on Machine Learning deep learning workshop, Vol. 2, 2015, p. 0.
- [50] J.D.M.-W.C. Kenton, L.K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
- [51] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
- [52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010.
- [53] M. Mathieu, M. Henaff, Y. LeCun, Fast training of convolutional networks through fts, in: Proceedings of the International Conference on Learning Representations, 2014.
- [54] R. Koplon, E.D. Sontag, Using fourier-neural recurrent networks to fit sequential input/output data, *Neurocomputing* 15 (3–4) (1997) 225–248.
- [55] H. Pratt, B. Williams, F. Coenen, Y. Zheng, Fcnn: Fourier convolutional neural networks, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 786–798.
- [56] J. Lee-Thorp, J. Ainslie, I. Eckstein, S. Otonan, Fnet: Mixing tokens with fourier transforms, *arXiv preprint arXiv:2105.03824*.



Hankun Kang received the B.Sc. degree from Xidian University, Xian'an, China, in 2020. He is working towards the M.Sc. degree in Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include ,few-shot learning, relation extraction and multimodal fusion.



Xiaoyu Li received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2016 and M.E. degree from Beijing University of Posts and Telecommunications in 2019. He is currently a Research Assistant Fellow at the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China. His research interests include data mining, information extraction, event logic graph and natural language processing.



Zequn Zhang received the B.Sc. degree from Peking University, Beijing, China, in 2012, and Ph.D. degree from Peking University in 2017. He is currently a Research Assistant at the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China. His research interests include information fusion, knowledge graph and natural language processing.



Li Jin received the B.S degree from Xidian University, Xi'an, China, in 2012 and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2017. He is currently an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include machine learning, knowledge graph and geographic information processing.



Shuchao Li received the B.E. degree in software engineering from North China Electric Power University, Baoding, China, in 2012, and Master's degree in computer technology from North China Electric Power University, Beijing, China, in 2017. Currently, he is a Research Assistant at the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China. His main research interests include knowledge graph, natural language processing and deep learning.



Chunbo Liu received the B.Sc. degree from Southwest University, Chong'qing, China, in 1996, M.Sc. degree from China University of Mining & Technology, Beijing, China, in 1999 and Ph.D. degree from Peking University, Beijing, China, in 2003. He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include geospatial data mining, spatiotemporal analytics and big data mining.



Yanan Zhang received the B.S. degree in communication engineering from Sichuan University, Chengdu, China, in 2016. She is currently pursuing the Ph.D. degree with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China. Her research interests include deep learning, natural language processing and question answering, especially on knowledge base question answering.