FL-MSRE: A Few-Shot Learning based Approach to Multimodal Social Relation Extraction

Hai Wan,[1] **Manrong Zhang,** [1] **Jianfeng Du,** [2,4*] **Ziling Huang,** [1] **Yufei Yang,** [1] **Jeff Z. Pan** [3]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, P.R.China
[2] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou 510006, P.R.China
[3] School of Informatics, The University of Edinburgh, Edinburgh, UK
[4] Pazhou Lab, Guangzhou 510330, P.R.China
wanhai@mail.sysu.edu.cn, jfdu@gdufs.edu.cn, {zhangmr7, huangzling, yangyf35}@mail3.sysu.edu.cn, jeff.z.pan@abdn.ac.uk

Abstract

Social relation extraction (SRE for short), which aims to infer the social relation between two people in daily life, has been demonstrated to be of great value in reality. Existing methods for SRE consider extracting social relation only from unimodal information such as text or image, ignoring the high coupling of multimodal information. Moreover, previous studies overlook the serious unbalance distribution on social relations. To address these issues, this paper proposes FL-MSRE, a few-shot learning based approach to extracting social relations from both texts and face images. Considering the lack of multimodal social relation datasets, this paper also presents three multimodal datasets annotated from four classical masterpieces and corresponding TV series. Inspired by the success of BERT, we propose a strong BERT based baseline to extract social relation from text only. FL-MSRE is empirically shown to outperform the baseline significantly. This demonstrates that using face images benefits text-based SRE. Further experiments also show that using two faces from different images achieves similar performance as from the same image. This means that FL-MSRE is suitable for a wide range of SRE applications where the faces of two people can only be collected from different images.[1]

Introduction

Social relations are specific relations in human daily life. They define the associations between two people in either the physical or the virtual world. *Social relation extraction* (*SRE*) aims to infer the social relation between two people from texts, personal albums, or films, *etc.*. It has been demonstrated to be of great value in reality. For example, it can capture social connections and enable machines better understand human behaviors.

In the past years SRE has drawn increasing research interests. (Fairclough 2003) extracts social relations from the text of microblogs. (Du et al. 2019a) extracts social relations from four classical masterpieces of Chinese literature. (Du et al. 2019b) mines social relations from the text of microposts. (Zhang et al. 2015) proposes to learn face representations to capture facial attributes and performs pairwise face reasoning for relation prediction. All these studies address SRE only from unimodal information and ignore the high coupling in multimodal information.

To explain why SRE requires multimodal information, we provide three examples in Figure 1. Consider Example (a). From the mention of "the Obamas" in text we can easily infer that Malia and Sasha are family members. From the faces of Malia and Sasha, we may find that they are girls with similar ages. Thus, by utilizing both face information and text information we can conclude that Malia and Sasha are sisters. Consider Example (b) and Example (c). The word "hug" in both two texts indicates a close relation between two people. Taking the information of faces into account, we may find that Barack is a man, Malia is a girl and Michelle is a woman; moreover, there is a significant age gap between Barack and Malia, while Barack and Michelle have similar ages. Hence Barack is more likely to be the father of Malia, whereas he is more likely to be the husband of Michelle.

Motivated by these examples, we focus on multimodal SRE from both texts and images. In this line it is crucial to clarify the following two questions:

- Can introducing face image information into a text-based model improve the performance for SRE?
- Can facial features extracted from different images achieve similar performance as from the same image?

Answering these questions calls for multimodal datasets having both texts and face images. Since existing SRE datasets (Zhang et al. 2015; Sun, Schiele, and Fritz 2017) only contain unimodal information, we construct three datasets from four classical masterpieces and corresponding TV series. These datasets reflect a complete set of social relations in a certain period of Chinese society, but most social relations therein only contain a few instances, in accord with the relation imbalance characteristics (Li et al. 2017).

Existing methods for SRE overlook the serious unbalance distribution on real-life social relations. To cope with the unbalance, we resort to few-shot learning techniques (Koch, Zemel, and Salakhutdinov 2015; Vinyals et al. 2016;

*Corresponding author

[1]The code and datasets are available at https://github.com/sysulic/FL-MSRE.

Figure 1: Representative examples collected from the Web, where (a) comes from https://en.wikipedia.org/wikiFamily_of_Barack_Obama, (b) from https://www.cbsnews.com/pictures/malia-and-sasha-obama/4/ and (c) from https://time.com/barack-michelle-obama-love-story-photos/.The head entities are labeled with blue bounding boxes in images and highlighted with blue texts, whereas the tail entities are labeled with red bounding boxes in images and highlighted with red texts.

Snell, Swersky, and Zemel 2017) and propose a **F**ew-shot **L**earning based approach to **M**ultimodal **S**ocial **R**elation **E**xtraction (FL-MSRE for short). Taking a sentence and a list of images as input, FL-MSRE generates an integrated representation for head/tail entities via a multimodal encoder and adapts prototypical network (Snell, Swersky, and Zemel 2017) for few-shot learning. The multimodal encoder consists of a sentence encoder, a face encoder and a cross-modality encoder. The sentence encoder extracts textual features by a pre-trained language model. The face encoder generates facial features by extracting bounding boxes and corresponding visual features from a pre-trained face recognition model. The cross-modality encoder integrates textual and facial features to the multimodal representation.

By comparing FL-MSRE with a strong baseline for SRE which predicts social relations from text by the same pre-trained language model, we show that SRE with multimodal information outperforms SRE with text only; moreover, FL-MSRE with faces from different images achieves similar performance as using faces from the same image.

Our contributions can be summarized as follows:

- We present multimodal social relation datasets, which can facilitate future research on multimodal SRE.
- To leverage both texts and face images, we propose a novel approach FL-MSRE for SRE.
- Extensive experiments demonstrate that FL-MSRE is effective in improving the performance of SRE.

Related Work

Social Relation Extraction

Previous studies on SRE focus on unimodal information only. (Du et al. 2019a) integrates text relational extraction into logic-based abduction to predict social relation between people in four classical masterpieces of Chinese literature. (Du et al. 2019b) proposes a kernel-based learning algorithm to mine social relations between people from the text of microposts. Alternatively, by merely using visual information, existing studies demonstrate that facial appearance, attributes and landmarks play an important role for social relation prediction. As a pioneer study for SRE from images, (Zhang et al. 2015) constructs a dataset for 16 social relations and proposes a deep model with bridging layers. Subsequently, (Li et al. 2017) defines a social relation hierarchy with 3 coarse levels and 6 fine levels by following the theory of relational models (Fiske 1992); meanwhile, it presents another dataset on SRE named People in Social Context (PISC).

As a whole, the above studies solely explore either visual information or textual information for SRE. Different from these studies, we propose to utilize both texts and face images for SRE and propose an approach to multimodal SRE accordingly. Moreover, existing studies for image-based SRE only address the prediction of social relations from a single image. To widen the range of SRE applications, we also demonstrate that our approach still works well for social relation prediction between two people whose faces are collected from different images.

Few-Shot Learning

Few-shot learning aims to learn classification models from few instances. Meta-learning and metric-learning constitute two main categories of few-shot learning, where the former focuses on extracting meta-knowledge that can adapt to new tasks and the latter focuses on learning distance distributions among classes. Among existing meta-learning methods, (Ravi and Larochelle 2017) exploits the long short-term memory (LSTM) network for self-training from an episode, whereas (Finn, Abbeel, and Levine 2017) proposes the Model-Agnostic Meta-Learning (MAML) approach to learning initialization parameters so as to adapt a trained model to new data. Among existing metric-learning methods, (Vinyals et al. 2016) proposes matching networks and applies an attention mechanism over a learned embedding of examples in the support set to classify instances in the query set, whereas (Koch, Zemel, and Salakhutdinov 2015) employs a unique structure to naturally rank similarity between inputs to learn Siamese neural networks (Koch, Zemel, and Salakhutdinov 2015). Prototypical network (Snell, Swersky, and Zemel 2017) assumes that there exists a prototype for every class and learns a representation function to calculate

the prototypes for classes of supporting instances. Our work adapts prototypical network for few-shot learning because of its excellent performance.

Multimodal Learning

Multimodal methods extract relevant information from different modalities and combine them collaboratively. In terms of visual question answering (Das et al. 2018), image captioning (Chen et al. 2015) and visual reasoning (Zellers et al. 2019), state-of-the-art methods have demonstrated that the performance of these tasks can be improved by multimodal information. Pre-trained models are considered to contain abundant syntax and semantic information to understand plain texts, benefiting the process of text information. The first pre-trained model for visual-linguistic is VideoBERT (Sun et al. 2019), where images cropped from videos are encoded into feature vectors and assigned into different clusters which can be treated as visual tokens. ViLBERT (Lu et al. 2019) extends BERT (Devlin et al. 2019) to a multimodal two-stream model, which processes both visual and textual inputs in separate streams through a co-attention Transformer layer. LXMERT (Tan and Bansal 2019) is also a multimodal model built upon the Transformer model. It consists of an object relationship encoder, a language encoder and a cross-modality encoder. In addition, VL-BERT (Su et al. 2020) is a single-stream network which extends the Transformer model to cope with visual inputs.

Multimodal Social Relation Datasets

The supplement of text descriptions to existing benchmark datasets for image-based SRE (Zhang et al. 2015; Li et al. 2017; Xia, Shao, and Fu 2011) is hard or even impossible, since it requires manual drafting of the text descriptions. Thus we attempt to enhance benchmark datasets for text-based SRE rather than for image-based SRE. We notice that there exist TV series for presenting the four classical masterpieces of Chinese literature that have corresponding benchmark datasets for text-based SRE (Du et al. 2019a). It allows us to collect images from the TV series to supplement image information into the text-based benchmarks.

To be specific, we enhance four text-based benchmark datasets (Du et al. 2019a), which are respectively collected from four e-books for the four Chinese masterpieces. In these original datasets, every e-book has been separated into sentences and all sentences mentioning at least two people are picked out, but whether a sentence entails a certain social relation is determined by distant supervision; *i.e.*, every sentence mentioning two people is assumed to contribute to the entailment of the relation between the two people.

We first confirm whether a sentence actually entails a social relation between two mentioning people. To this end, we recruit several annotators familiar with the four Chinese masterpieces to mark whether a sentence can be recognized as evidence of a relation. A sentence is finally marked as evidence only when more than half of annotators support it. Following (Du et al. 2019a), two people and their social relations are expressed as $\langle h, t, r \rangle$, where h is the head entity, r is the social relation and t is the tail entity. We remove the relations that only occur in one triple. Since there may exist hierarchical relations between two people, such as family and brother, we only keep the most specific relation to make sure that the relation between two people is unique.

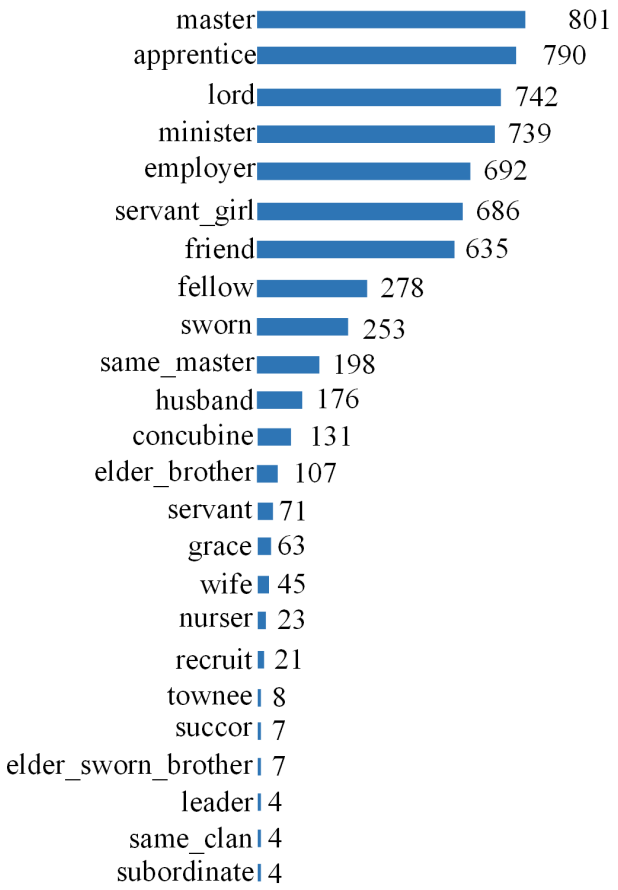

Figure 2: # of sentences for every social relation in FC-TF.

We then collect images from the TV adaptations of the four Chinese masterpieces. Initially, we use FFmpeg[2] to erase subtitles to prevent the information leakage of text messages and delete duplicate images. Afterwards, we apply FaceNet (Schroff, Kalenichenko, and Philbin 2015) for face recognition and retain images containing at least two people. Every person in the image is labeled with the character name and the bounding box of the face region. We collect 10 face images from different angles for every person and adapt a pre-trained FaceNet model[3] to recognize faces in each image and to generate the annotation. Then we invite annotators who are familiar with the TV series to check the annotations in all images. Lastly, we pair sentences and images for every triple with character names of the people.

At last, since some enhanced datasets have insufficiently many relations, we reorganize the four enhanced datasets into three new datasets, *i.e.*, *Dream of the Red Chamber*

[2]http://ffmpeg.org/

[3]https://github.com/tbmoon/facenet

(*DRC-TF*), *Outlaws of the Marsh* (*OM-TF*) and *the Four Classic* (*FC-TF*), where *TF* denotes that the dataset contains both texts and face images. For these datasets, FC-TF contains 24 social relations, DRC-TF contains 15 social relations and OM-TF contains 9 social relations. Every social relation is supported by multiple triples; meanwhile, every triple is supported by multiple pairs of face images about the two entities as well as by multiple sentences mentioning both entities. The number of sentences for each social relation in FC-TF can be seen in Figure 2. It exhibits a serious unbalance distribution on social relations.

The Proposed Approach FL-MSRE

Problem Formulation

In the task of multimodal social relation extraction, we intend to classify the social relation between two people based on text and image inputs. A text input is a sentence containing two people whose relation needs to be predicted. An image input consists of a head entity, a tail entity and two possibly identical images, where one image contains the face of the head entity and the other contains the face of the tail entity. Every entity e consists of two parts, *i.e.*, a bounding box $b_e = (x_1, y_1, x_2, y_2)$ and a character name c_e, where x_1, y_1, x_2 and y_2 represent the four coordinates of the two corners of the entity face in the image.

We follow the N way K shot setting, which has been widely adopted in recent research on few-shot learning (Han et al. 2018; Gao et al. 2019). Suppose the input dataset is represented by a set of tuples (s, h, t, g_h, g_t, r), where s is a sentence, h is a head entity, t is a tail entity, g_h is an image contains the face of h, g_t is an image contains the face of t, and r is the social relation between h and t. Basically, the N way K shot setting is to learn a representation function for a tuple such that, when given a random support set of NK tuples on N different social relations (where each relation corresponds to K tuples) as well as a random query set of N tuples on the same N social relations (where each relation corresponds to one tuple), a prediction model trained on the support set achieves as high accuracy as possible in predicting the social relations in the query set.

Overview

Our proposed approach FL-MSRE adapts prototypical network (Snell, Swersky, and Zemel 2017) which is built upon a multimodal encoder for SRE. The architecture of the multimodal encoder is shown in Figure 3. The multimodal encoder consists of three parts, *i.e.*, sentence encoder, face encoder and cross-modality encoder. Sentence encoder extracts textual features using a pre-trained language model BERT (Devlin et al. 2019). Face encoder generates facial features by extracting bounding boxes and corresponding visual features from a pre-trained face recognition model FaceNet (Schroff, Kalenichenko, and Philbin 2015). Cross-modality encoder integrates textual features and facial features and outputs the final multimodal representation.

Multimodal Encoder

Sentence Encoder In the sentence encoder, each sentence s will be encoded into a vector. We use the pre-trained language model BERT as the language encoder. Given a sentence s consisting of m tokens $w_1, w_2, \ldots, w_m$, in which the names of characters have been masked. We construct a token sequence $s' =$ "[CLS], $w_1, w_2, \cdots, w_m$,[SEP]", where [CLS] and [SEP] are special tokens introduced in BERT. BERT takes the token sequence s' as input, and outputs a sequence of 768-dimensional vectors $V_{[CLS]}, V_{w_1}, V_{w_2}, \ldots, V_{w_m}, V_{[SEP]}$. By transforming the first vector $V_{[CLS]}$ through a fully-connected layer, we get the final vector B_s for sentence s, formally defined as

$$B_s = \tanh(W_1 V_{[CLS]} + b_1) \quad (1)$$

where $W_1 \in \mathbb{R}^{768\times768}$ and $b_1 \in \mathbb{R}^{768}$ are trainable.

Face Encoder This encoder is built upon FaceNet (Schroff, Kalenichenko, and Philbin 2015), using GoogLeNet style inception network (Szegedy et al. 2015) as the backbone. More specifically, we remove the last fully-connected layer of the inception network and extract facial features from the last average pooling layer. For the given two images g_h and g_t containing the faces of the head entity h and the tail entity t respectively, we can get two vectors f_{h,g_h} and f_{t,g_t} in $\mathbb{R}^{1792}$ by applying FaceNet, defined as

$$f_{h,g_h} = \phi(\text{crop}(g_h, b_h)) \quad (2)$$
$$f_{t,g_t} = \phi(\text{crop}(g_t, b_t)) \quad (3)$$

where $\text{crop}(g_e, b_e)$ (where $e \in \{h, t\}$) denotes the image cropped from g_e to the bounding box b_e and resized to 160×160 pixels, and ϕ denotes the modified inception network.

To obtain high-level features, a fully-connected layer and a batch normalization layer are added to get final facial features of the two entities. More precisely, two vectors v_{h,g_h} and v_{t,g_t}, respectively representing final facial features of h and t, are refined from f_{h,g_h} and f_{t,g_t} by

$$v_{h,g_h} = \text{BatchNorm}(W_2 f_{h,g_h} + b_2) \quad (4)$$
$$v_{t,g_t} = \text{BatchNorm}(W_3 f_{t,g_t} + b_3) \quad (5)$$

where $W_2, W_3 \in \mathbb{R}^{512\times1792}$ and $b_2, b_3 \in \mathbb{R}^{512}$ are trainable.

Finally, we apply L2 normalization to v_{h,g_h} and v_{t,g_t} and concatenate them to achieve the final vector F_{h,g_h,t,g_t} of facial features, which is defined as

$$F_{h,g_h,t,g_t} = [\text{L2Norm}(v_{h,g_h}); \text{L2Norm}(v_{t,g_t})] \quad (6)$$

where $[;]$ indicates the concatenation of two vectors.

Cross-modality Encoder This encoder intends to integrate textual features and facial features into the multimodal representation. We simply concatenate the vector of textual features and the vector of facial features and then apply a layer normalization layer to avoid vanishing or exploding of gradients. For an input tuple (s, h, g_h, t, g_t), the multimodal representation for it is formally defined as

$$L_{s,h,g_h,t,g_t} = \text{LayerNorm}[B_s; F_{h,g_h,t,g_t}] \quad (7)$$

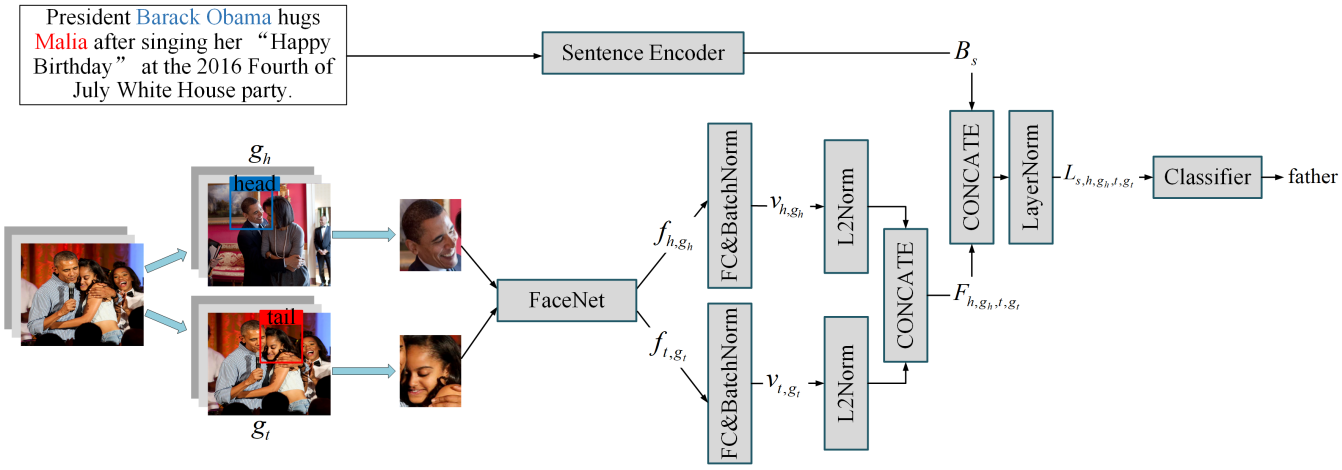

Figure 3: The architecture of the multimodal encoder in FL-MSRE. Given a tuple (s, h, t, g_h, g_t) where s is a sentence, h is a head entity, t is a tail entity, g_h is an image contains the face of h, and g_t is an image contains the face of t, the sentence encoder extracts a vector B_s of textual feature from a pre-trained language model, while the the face encoder first applies FaceNet to extract a vector f_{h,g_h} of facial features for h and a vector f_{t,g_t} of facial features for t, then applies a fully-connected layer and a normalization layer to refine f_{h,g_h} and f_{t,g_t} respectively to v_{h,g_h} and v_{t,g_t}, and finally concatenates v_{h,g_h} and v_{t,g_t} to yield a final vector F_{h,g_h,t,g_t} of facial features. Afterwards, the cross-modality encoder concatenates the two vectors B_s and F_{h,g_h,t,g_t} to yield the multimodal representation L_{s,h,g_h,t,g_t} of the tuple.

Prototypical Network

Given a support set $S =$

$$\begin{aligned}&\{(s_{11}, h_{11}, t_{11}, g_{h,11}, g_{t,11}, r_1), \quad \ldots,\\ &(s_{1K}, h_{1K}, t_{1K}, g_{h,1K}, g_{t,1K}, r_1),\\ &\ldots\ldots,\\ &(s_{N1}, h_{N1}, t_{N1}, g_{h,N1}, g_{t,N1}, r_N), \quad \ldots,\\ &(s_{NK}, h_{NK}, t_{NK}, g_{h,NK}, g_{t,NK}, r_N)\}\end{aligned}$$

of tuples in the N way K shot setting, prototypical network assumes that there exists a prototype for each of the N social relations in S. It computes the prototype for social relation r_m based on the multimodal representations of K tuples $(s_{m1}, h_{m1}, t_{m1}, g_{h,m1}, g_{t,m1}, r_m)$, $\ldots$, $(s_{mK}, h_{mK}, t_{mK}, g_{h,mK}, g_{t,mK}, r_m)$ in S. More precisely, the prototype representation $P_m(S)$ for r_m is defined as

$$P_m(S) = \frac{1}{K}\sum_{i=1}^{K} L_{s_{mi},h_{mi},g_{h,mi},t_{mi},g_{t,mi}} \quad (8)$$

To predict a social relation among the N ways, prototypical network calculates the Euclidean distance d between a query tuple $q = (s, h, t, g_h, g_t)$ and each prototype $P_m(S)$ and applies softmax over the distance vector to generate a probability distribution on social relations. Formally, the probability on r_m, written $\Pr(r_m \mid q)$, is defined as

$$\Pr(r_m \mid q) = \frac{\exp(-d(L_{s,h,g_h,t,g_t}, P_m(S)))}{\sum_{i=1}^{N}\exp(-d(L_{s,h,g_h,t,g_t}, P_i(S)))} \quad (9)$$

Experiments

The Baseline Approach

To verify the effectiveness of our approach, we employ a strong baseline to extract social relations from text based on BERT (Devlin et al. 2019), a pre-trained language model which has been proved to be very promising in a number of tasks for natural language processing. This baseline approach is similar to ours except that the multimodal encoder in our approach is replaced with a BERT encoder which generates a 768-dimensional vector for every tuple (s, h, t), where s is a sentence, h is a head entity and t is a tail entity. The BERT encoder is also fine-tuned with prototypical network to achieve the best performance for SRE. By Proto (BERT) we denote this baseline throughout our experiments.

Datasets	#rel	#char	#triple	#sen	#img
DRC-TF	9	47	59	1828	560
OM-TF	15	43	54	1489	1178
FC-TF	24	121	166	6485	3716

Table 1: The statistics on three datasets. #rel/#char/#triple are respectively the number of relations/characters/triples/ in text inputs, while #sen/#img are respectively the number of sentences/images in image inputs.

Image Sampling

To answer the two questions raised in Introduction, we consider the following two methods for image sampling:

1. **the same image**: this method extracts pairwise bounding boxes for the faces of the head entity and the tail entity from the same image, and we denote this method by FL-MSRE (same) throughout our experiments;
2. **different images**: this method extracts bounding boxes for the faces of the head entity and the tail entity from different images, and we denote this method by FL-MSRE

Methods	FC-TF				DRC-TF		OM-TF	
	5 way 1 shot	5 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot
Proto (BERT)	73.93±0.01	75.03±0.35	83.84±0.37	86.75±0.32	40.00±0.11	51.29±0.16	46.28±0.15	48.35±0.43
FL-MSRE (same)	**79.05±0.10**	79.59±0.14	87.57±0.41	**90.24±0.32**	58.44±0.21	71.96±0.49	54.16±0.22	61.67±0.67
FL-MSRE (different)	78.61±0.48	**79.95±0.23**	**87.64±0.15**	89.98±0.17	**62.03±0.24**	**77.29±0.54**	**54.53±0.68**	**62.30±0.39**

Table 2: Comparison results on FC-TF, DRC-TF and OM-TF datasets. The results under each setting has been averaged by 10 random runs and recorded as percentage. The best result of all compared methods is displayed in bold.

Methods	DRC-TF (trained on OM-TF)		OM-TF (trained on DRC-TF)	
	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot
Proto (BERT)	38.38±0.16	40.74±0.18	37.27±0.69	44.09±0.69
FL-MSRE (same)	**50.69±0.81**	63.83±0.42	**55.82±0.76**	**58.50±0.45**
FL-MSRE (different)	50.34±0.15	**64.59±0.56**	52.76±0.91	55.43±1.25

Table 3: Comparison results when training on one dataset and testing on another dataset. The results under each setting has been averaged by 10 random runs and recorded as percentage. The best result of all compared methods is displayed in bold.

(different) throughout our experiments.

These two methods have their own advantages in the task of SRE. On one hand, considering faces in the same image may indicate a more accurate relation between the two entities since a person will show different facial expressions when getting along with different people. On the other hand, considering faces in different images widens the range of SRE applications especially in cases where the two considering people never appear in the same image; moreover, essential attributes of people that are useful for SRE such as age and gender may still be preserved in different images.

By considering that random sample of images may impact the stability of performance in SRE, we combine every image among n randomly selected images with the sentence to predict social relations and treat the majority relation as the result. In our experiments we set n to 5.

Experiment Setting

Dataset Analysis and Splits We compare our approach with the baseline on the aforementioned three datasets. The statistics of these datasets are reported in Table 1. Following the traditional way of few-shot learning (Han et al. 2018; Gao et al. 2019), we use disjoint sets of social relations for training, validation and test. For DRC-TF, we respectively use 3, 3 and 3 relations for training, validation and test. For OW-TF, we respectively use 5, 5 and 5 relations for training, validation and test. For FC-TF, we respectively use 14, 5 and 5 relations for training, validation and test. Since people may have the same surname if they are family members, to avoid this information leakage for SRE, we replace the name of the head entity with *#head#* and the name of the tail entity with *$tail$*.

Implementation Details We apply FaceNet (Schroff, Kalenichenko, and Philbin 2015) pre-trained on VGGFace2 (Cao et al. 2018) for the implementation of ϕ in Equation (2) and (3). For the sentence encoder, we fine-tune the parameters of the pre-trained BERT model and learn the remaining parameters from scratch. In all of our experiments, we use the Adam optimizer (Kingma and Ba 2015) and tune hyper-parameters to the best values according to the accuracy on the validation set. To be specific, the batch size is set to 8 and the weight delay is set to 0.1. The learning rate is initialized as 2×10^{-5} for the baseline and initialized as 2×10^{-6} for our approach.

Result Analysis

We consider four different settings of few-shot learning in our experiments: 3 way 1 shot, 3 way 3 shot, 5 way 1 shot and 5 way 3 shot. We use prediction accuracy as the evaluation metric in all experiments. We compare the performance of FL-MSRE with two different methods for image sampling with that of the baseline on all the three datasets. To alleviate the impact of random factors in experimental results, we repeat experiments ten times in the test stage under each setting and calculate the mean and variance of the accuracies. The results are reported in Table 2.

For FC-TF, FL-MSRE with either method for image sampling achieves an absolute improvement of 3.2% to 5.2% over the baseline Proto (BERT). Moreover, FL-MSRE performs more stably than the baseline when the number of training instances decreases. For the other two smaller datasets OM-TF and DRC-TF, FL-MSRE with either method for image sampling achieves a more significant improvement over the baseline. In particular, since most of the social relations in DRC-CF are highly relevant to character attributes, FL-MSRE achieves the most significant improvement over the baseline.

Cross-Dataset Analysis To further investigate the effectiveness of our approach, we conduct cross-dataset analysis by training on one dataset and testing on another one. The results are reported in Table 3. As can be seen, FL-MSRE with either method for image sampling still significantly outperforms the baseline.

Answering the Two Questions To answer the two questions raised in Introduction, detailed comparison results are reported in Table 4 and Table 5 to quantify the differences

Methods	FC-TF				DRC-TF		OM-TF	
	5 way 1 shot	5 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot
FL-MSRE (same)	5.12±0.04	4.56±0.12	3.74±0.16	3.49±0.17	18.44±0.13	20.67±0.08	7.88±0.37	13.31±0.12
FL-MSRE (different)	4.67±0.10	4.92±0.06	3.80±0.22	3.22±0.25	22.03±0.19	26.00±0.21	8.24±0.29	13.94±0.40

Table 4: The performance improvements beyond the baseline Proto (BERT).

Method	FC-TF				DRC-TF		OM-TF	
	5 way 1 shot	5 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot	3 way 1 shot	3 way 3 shot
FL-MSRE (different)	-0.44±0.10	0.36±0.12	0.63±0.15	-0.26±0.22	3.60±0.22	5.33±0.39	0.36±0.66	0.63±0.25

Table 5: The performance difference (former−latter) between FL-MSRE (different) and FL-MSRE (same).

Figure 4: Two test instances that are incorrectly predicted by the baseline Proto (BERT) but correctly predicted by FL-MSRE. The head entities are labeled with blue bounding boxes in images and highlighted with blue texts, whereas the tail entities are labeled with red bounding boxes in images and highlighted with red texts.

between the performances of different approaches. Specifically, we calculate the mean and the variance of the differences between two performance results. Based on these results, we answer these two questions in the following.

FL-MSRE is succeeded in integrating text information and image information into a representation model and draws significant improvements over a pure text-based representation model on all the three datasets, showing the effectiveness in utilizing face images for SRE.

After replacing the input of pairwise faces in the same image with two faces from different images, FL-MSRE achieves comparable performance on FC-TF and better performance on both DRC-TF and OM-TF. These results show that using faces from different images does have advantages. A probable reason is that this treatment enforces the learnt representation model to focus more on fixed attributes of human faces, where these attributes contribute to more accurate predictions on social relations.

Case Study

To better clarify the advantage of the face information, we select two examples in Figure 4 and compare our approach with the baseline. As for the first example, the baseline incorrectly predicts the social relation between the two people as “sworn brother” (simply “sworn”), while FL-MSRE with either method for image sampling predicts the correct relation “wife”. Looking into the sentence, we find that the wrong prediction may be due to the word “brother” mentioned between the two entities. By considering the faces of the two entities, the head entity is revealed as a woman and the tail entity as a man, thus FL-MSRE is able to remove the “sworn brother” relation since only men can become sworn brothers. As for the second example, the relation between the two entities is wrongly predicted as “wife” by the baseline. This prediction is impossible by considering faces because the two entities are both men. To summarize, by embedding face information into a model for social relation prediction, impossible social relations can be removed and thus the robustness of SRE is improved.

Conclusion

In this paper, we have proposed FL-MSRE, a new approach leveraging both text information and face information for social relation extraction. To clarify its effectiveness, we present a strong baseline and compare its performance with the performance of two variants of FL-MSRE on three datasets. Experimental results demonstrate that FL-MSRE consistently outperforms the baseline on all the datasets. They also show that the performance of FL-MSRE with faces collected from different images is comparable to or even better than that with faces collected from the same image. This implies a special value of FL-MSRE, *i.e.*, it is also good at predicting the social relation between people who never appear in the same image. Future work will explore social relation extraction for more application scenarios.

Acknowledgments

This work is supported by the National Key R&D Program of China (No.2018YFC0830600), the National Natural Science Foundation of China (No. 61876204 and 61976232), Guangdong Province Natural Science Foundation (No. 2017A070706010 (softscience) and 2018A030313086), Guangdong Province Science and Technology Plan projects (2017B010110011), and Guangzhou Science and Technology Project (No. 201804010496).

References

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *FG*, 67–74.

Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR,abs/1504.00325* .

Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; and Batra, D. 2018. Embodied Question Answering. In *CVPR*, 1–10.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Du, J.; Pan, J. Z.; Wang, S.; Qi, K.; Shen, Y.; and Deng, Y. 2019a. Validation of Growing Knowledge Graphs by Abductive Text Evidences. In *AAAI*, 2784–2791.

Du, Y.; Su, F.; Yang, A.; Li, X.; and Fan, Y. 2019b. Extracting Deep Personae Social Relations in Microblog Posts. *IEEE Access* 8: 5488–5501.

Fairclough, N. 2003. *Analysing Discourse: Textual Analysis for Social Research*. Psychology Press.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, 1126–1135.

Fiske, A. 1992. The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations. *Psychological review* 99: 689–723.

Gao, T.; Han, X.; Zhu, H.; Liu, Z.; Li, P.; Sun, M.; and Zhou, J. 2019. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In *EMNLP-IJCNLP*, 6251–6256.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *EMNLP*, 4803–4809.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese Neural Networks for One-shot Image Recognition. In *ICML deep learning workshop*, volume 2.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2017. Dual-Glance Model for Deciphering Social Relationships. In *ICCV*, 2650–2659.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NIPS*, 13–23.

Ravi, S.; and Larochelle, H. 2017. Optimization as a Model for Few-Shot Learning. In *ICLR*, 4077–4087.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*, 815–823.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In *NIPS*, 4077–4087.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *ICCV*, 7463–7472.

Sun, Q.; Schiele, B.; and Fritz, M. 2017. A Domain Based Approach to Social Relation Recognition. In *CVPR*, 3481–3490.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going Deeper with Convolutions. In *CVPR*, 1–9.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*, 5099–5110.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *NIPS*, 3630–3638.

Xia, S.; Shao, M.; and Fu, Y. 2011. Kinship verification through transfer learning. In *IJCAI*, 2539–2544.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*, 6720–6731.

Zhang, Z.; Luo, P.; Loy, C.-C.; and Tang, X. 2015. Learning Social Relation Traits from Face Images. In *ICCV*, 3631–3639.