Understanding Social Relationship with Person-pair Relations

#229

Abstract

Social relationships understanding is to infer the social relations among people from images and videos, which has attracted increasing attention in computer vison recently. A great progress has been made since the rise of deep learning. However, they mostly focuses on the facial attributes or contextual object cues without taking into account the interaction among person pairs. Motivated by scene graph generation, we carefully analyzed the datasets and found the social relations in a still image always have high semantic relevance. For instance, if two person pair in an image are *Friends*, then the third one is always friends or at least other intimate relations but not *No Relation*. Therefore, to capture this interaction cues, we propose a novel end-to-end trainable Person-Pair Relation Network (PRN) using standard RNNs, a graph inference network that learns iteratively to improve its predictions via message passing among person pair nodes. Extensive experiments on PISC and PIPA-Relation show the superiority of our method over previous methods.

1 Introduction

Social relationships are closely related to our daily life [Barr *et al.*, 2014]. After understanding the social relationship between the person pair, we can easily explain their behavior. For machines, only when they fully understand the social relationships, can they further understand and infer the human behavior in our social life, so as to make a better response. In addition, we often leave traces that capture social relationships in many medias and we not only want the machines to be proficient at their task, but also enable them to blend in and act appropriately in different situations [Sun *et al.*, 2017]. In short, social relationship detection task is very significant in many ways. In our work, we aim to address the social relationship detection task for every picture where each picture represents a scene.

However, to solve the social relationship detection task is not so simple. For a giving picture, detecting the social relationships of all the person pair is a difficult task. The models need to be adapted to different scenes and context information to make right judgments. [Sun *et al.*, 2017] use the information of head region, body region and human attributes to predict the person-pair's social relationship separately. [Li *et al.*, 2017] make use of the pair of people in question and region proposals and allocate attention to each region to detect the social relationship of each person pair. [Wang *et al.*, 2018] takes advantage of the message propagation between person pair social relationship and the object semantic regions to solve the problem. The biggest problem of these models is that they all only detect one relationship per step which will cause that different social relationships in the same scene cannot interact with each other. Social relationships in the same scene are strongly linked but the previous models have ignored this important information.

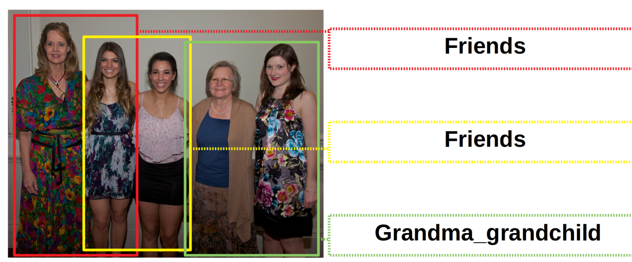


Figure 1: An example image from PIPA dataset. There are 3 person-pair social relationships in the picture: friends, friends and grandma_grandchild.

As one example in PIPA(Figure 2) where the picture denotes a scene, there are 2 "friends" relationships and 1 "grandma_grandchild" relationship in the scene. If we know only that two of these relationships are "friends, "we can infer that it is most likely a group of friends, and thus we infer that the unknown relationship is "friends" if they're similar in age. In other words, we can easily use the interaction of different relationships in the same scene to detect each social relationship. Therefore we address the issue with focusing more on the interaction between every social relationships in one scene to improve the social relationship detection performance.

However, the biggest issue of the task is that it is not as simple as people directly to judge the result. We need to design the mechanism of the interaction between social relationships and effectively model the mechanism. The other issue is how to use less effective information to model the interaction machanism and get the better result. For example, whether to introduce the information of the objects in scene has also became an important consideration for us.

To address this problem, we propose a novel end-to-end trainable Person-Pair Relation Network (PRN) which makes good use of the information of social relationship interaction in the same scene. PRN consists of 3 parts: Feature Extraction Module, Message Pooling Module and Message Padssing Module. In the Feature Extraction Module, we use a Resnet to extract features of each person and another Resnet to extract features of the person pairs in the scene. In this module, the position of each person is also taken into account. In the Message Pooling Module, we use a pooling mechanism to make all the social relationships interact with each other well and the outout of this module will be considered as the inputs of the GRU which is in the next module. In the Message Passing Module, we use the RNNs contained GRU to make the social relationship message passing proceed iteratively. As the model proceeding, this module and the Message Pooling Module will have an interaction and make the overall interaction mechanism better. We use the hidden state of the last GRU as the social relationship detection results of the scene.

We evaluate our model in classic datasets: PIPA-Relation dataset and PISC dataset. The experiment results verify the superiority of our model over previous methods. Specifically, our contributions are as follows:

- To our best knowledge, it is the first attempt to introduce the interaction of social relationships in the same scene on the social relationship detection task;
- We design a novel interaction mechanism to model the interaction of social relationships in the same scene which gets the best result in this task;
- We analyze and verify that the role played by objects in the scene is not very big which is a very novel idea in the social relationship detection task.

2 Related Work

2.1 Social Relationship Understanding

The foundation of social network is the social relationships understanding, an important multidisciplinary problem that has attracted increasing attention in computer vision recently. A much number of studies that aim to infer social relationships from images [Wang *et al.*, 2015; Li *et al.*, 2017; Wang *et al.*, 2018; 2010; Zhang *et al.*, 2015b] and videos [Ding and Yilmaz, 2010; Ramanathan *et al.*, 2013; Vinciarelli *et al.*, 2009] have been made since the rise of deep learning. For instance, motivated by psychological sudies, [Zhang *et al.*, 2015b] and [Dibeklioglu *et al.*, 2013] exploit social relationships based on facial attributs such as expression and head pose, and affective behaviour analysis. Besides, [Li *et al.*, 2017] and [Wang *et al.*, 2018] discover that contextual cues around people play a significant role in social realtionship infering. Concretely, [Li *et al.*, 2017] proposed a dual-glance model for social relationship, where the first glance makes a coarse relationship prediction for a given person pair and then the second one refines the prediction by using the objects around the pair. [Wang *et al.*, 2018] constructed a semantic-aware knowledge graph and employed Gated Graph Neural Network (GGNN) [Li *et al.*, 2015] to integrate the graph into the Graph Reasoning Model (GRM), a graph reasoning network where a proper message propagation and graph attention mechanism are introduced to explore the interaction between person pair and the contextual objects.

Unlike the aforementioned works which mainly focus on facial attributs or contextual object cues, we detaily studied the two classic datasets PISC [Li *et al.*, 2017] and PIPA-relation [Sun *et al.*, 2017] and found the social relations in a still image have high semantic relevance. Based on this discovery, we designed a novel end-to-end trainable Person-Pair Relation Network (PRN), a graph inference network to capture this semantic relevance cues via message passing among person pair nodes.

2.2 Message Passing

Graph inference is a kind of form of message passing and Conditional Random Fields (CRF) have been used extensively in this field. Johnson *et al.* used CRF to infer scene graph grounding distributions for image retrieval [Johnson *et al.*, 2015]. Yatskar *et al.* use a deep CRF model to propose situation-driven object and action prediction[Yatskar *et al.*, 2016] . Danfei Xu *et al.* use the GRU-RNNs to solve the scene graph generation problem iteratively[Xu *et al.*, 2017]. Our work is related to Graph-LSTM [Liang *et al.*, 2016] and the work of Danfei Xu *et al.*[Xu *et al.*, 2017] which formulate the message passing problem using RNN models. Danfei Xu *et al.*[Xu *et al.*, 2017] design primal graph and dual graph in their model while we just simplify the model and just use one graph to make social relationship messages to pool and achieve a better result. As what Danfei Xu *et al.*[Xu *et al.*, 2017] have done, our model iteratively refines the social relationship predictions through relationship message passing in the scene, whereas the Structural RNN model only makes one-time predictions along the temporal dimension, and thus cannot refine its past predictions[Xu *et al.*, 2017].

3 PRN Model

3.1 overview

In this section, we introduce the proposed person-pair relational network for social relationship understanding. We fomulate the relationshipes of an image as a social graph, and followed[Wang *et al.*, 2018] to extract various features. Then, These features are used to initialize the nodes on the social graph. Person pair relation network employs Gated Recurrent Unit[Cho *et al.*, 2014] to explore the interaction of relationships with each other, and we employ the attention mechanism to adaptively exploit the most relevant nodes. It integrates multiple component to learn the representations of relationshipes between peoples in an image. The proposed framework is shown in Fig. 2.

3.2 Feature Extraction Module

Given an image $\mathbf{I}$ and the bounding box of peoples, we first crop three patches for each person pair, where the first two cover each person, $\mathbf{p_1}$ and $\mathbf{p_2}$, and one for union region, $\mathbf{p_u}$,

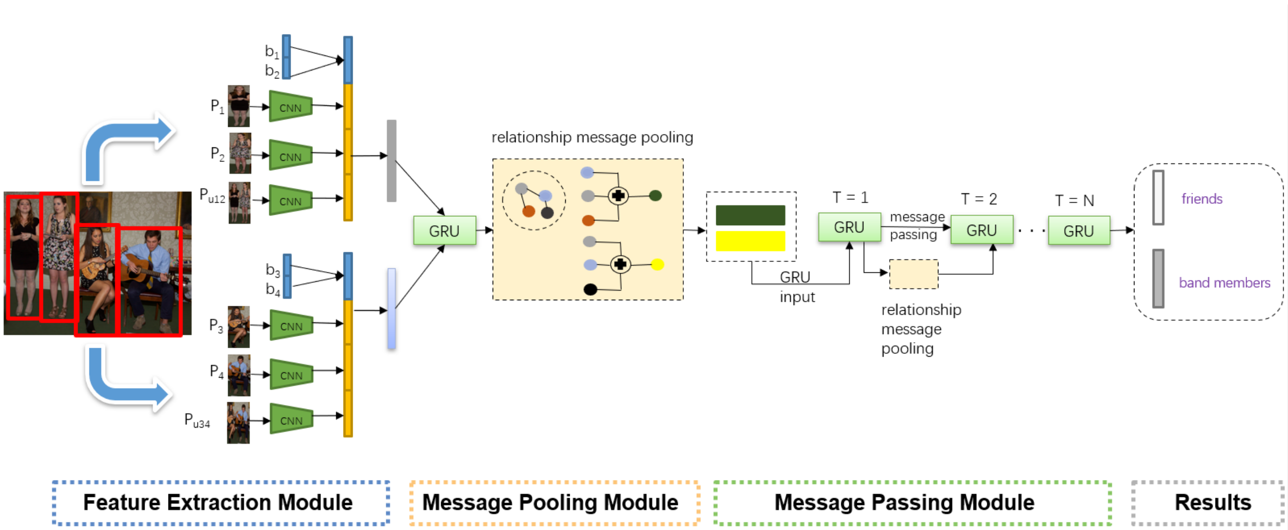


Figure 2: An illustration of our PRN. The model first extracts the features of each person and person pairs in the Feature Extraction Module. P_i denotes the ith person and P_{uij} denotes the union region of ith person and jth person. b_i denotes the position of the ith person. In the Message Pooling Module, a message pooling function computes relationship messages that are passed to the node GRU in the next iteration from the hidden states. The $\oplus$ symbol denotes a learnt weighted sum. Then in the Message Passing Module, we iteratively updates the hidden states of the GRUs. We use the hidden states of the GRUs at the last iteration step, to predict the social relationships in the scene.

that cover the both people and maintains the basic informantion for recognition. These patches are resized to 224×224 pixels and fed into three CNNs, These feature vectors from the last convolutional layer are flattened and concatenated. $\mathbf{p_1}$ and $\mathbf{p_2}$ share the same weights. In addition, geometry feature of bounding box is complementary to the visual appearance, as the relation *No Relation*, which is not easy to learn only from visual feature. We denote the geometry feature of bounding box i as $b_i^{pos} = \{x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max}, area_i\} \in \mathbf{R}^5$, where all the parameters are relative values. These features also concatenated with the CNN features for p_1, p_2 and p_u to form a single vector. Finally, both of them are fed into a fully connected layer to produce a 4096-dimension feature vector v_h.

3.3 Message Passing Module

In this module, we use the Recurrent Neural Networks(RNNs) to finish the inference of the social relationship detection task. In contrast to Zheng *et al.* [Zheng *et al.*, 2015], our model use a generic RNN module to compute the hidden states. In addition, we select the best cell Gated Recurrent Units [Wu *et al.*, 2014] for RNNs. We use the hidden states of the tth step to denote the social relationships information in the scene. So the social relationships information will be updated step by step as the GRU-based RNN going. This module can be formulated as follows:

$$\begin{aligned}
\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]), \\
\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]), \\
\hat{\boldsymbol{h}_t} &= tanh(\boldsymbol{W}[\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \\
\boldsymbol{h}_t &= (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \hat{\boldsymbol{h}_t} \\
\boldsymbol{y}_t &= \sigma(\boldsymbol{W}_o \boldsymbol{h}_t)
\end{aligned} \tag{1}$$

where σ and $tanh$ are the logistic sigmoid and hyperbolic tangent functions. In addition, $\odot$ symbol denotes the element-wise multiplication operation. $\boldsymbol{r}_t$ denotes the reset gate in t step and $\boldsymbol{z}_t$ denotes the update gate at t step. $\boldsymbol{W}_r$, $\boldsymbol{W}_Z$ and $\boldsymbol{W}_o$ denote weights of reset gate, update gate and output gate of the GRU. $\boldsymbol{W}_r$, $\boldsymbol{W}_z$ and $\boldsymbol{W}$ are trainable parameters. $\boldsymbol{h}_t$ denotes the hidden state of GRU at t step and $\boldsymbol{x}_t$ denotes the input of GRU at t step. Specially, $\boldsymbol{x}_0$ comes from the Feature Extraction Module and the hidden state will be put into the Message Pooling Module. Then the output of the Message Pooling Module will be the input of the next GRU. In this way, $\boldsymbol{x}_t$ comes from the previous Message Passing Module and the hidden state of the GRU will pass step by step which denotes the information of social relationship in the scene. At last, we use the last hidden state of the GRU to represent the social relationships and output the results.

3.4 Message pooling Module

Sec 3.2 provides a way to solve reasoning problem using RNNs. As each GRU receives multiple incoming messages, we need an aggregation function that merge information from all messages into a meaningful representation. Intuitively, the methods of standard pooling can do this, such as average pooling and max-pooling. But, it will be more effective to exploit various contextual cues and only preserve the appropriate parts when understanding social graph of an image with different. relationship. So, we utilized a message pooling function that computes the weights factors for each incoming message and aggregate the the messages using a weightes sum.

Formally, given the current GRU hidden states of nodes h_i, we takes the messages from other nodes as $m_{i,j\to i}$, and $h_{j\to i}$ as the hidden state of connected node. $m_{i,j\to i}$ is computed by function of its hidden state h_i.To be more specific, $m_{i,j\to i}$

are computed by the following message pooling functions:

$$m_{i,j\to i} = \sum_j \sigma(w^T[h_i, h_{j\to i}])h_{j\to i} \tag{2}$$

where [.] representes a the operation in the concatnation of vectors. and σ denoted a sigmoid function. w is the parameter to be learn.

3.5 Optimization

Given the predicted score $\mathbf{s}^{I,k} \in R^{|\mathcal{C}|}$ for the k-th person pair in image I, we use *softmax* to get its corresponding probability $\mathbf{p}^{I,k} \in R^{|\mathcal{C}|}$

$$p_i^{I,k} = \frac{\exp s_i^{I,k}}{\sum_{j=1}^{|\mathcal{C}|} \exp s_j^{I,k}}, i = 1, 2, \ldots, |\mathcal{C}| \tag{3}$$

where $\mathcal{C}$ donates the classes set of social relationship and $|\mathcal{C}|$ is its size. The loss function is expressed as

$$\mathcal{L} = -\frac{1}{\sum_{I\in\mathcal{I}} \mathrm{N}(I)} \sum_{I\in\mathcal{I}} \sum_{k=1}^{\mathrm{N}(I)} \sum_{i=1}^{|\mathcal{C}|} \mathrm{L}(y_i^{I,k}, p_i^{I,k}) \tag{4}$$

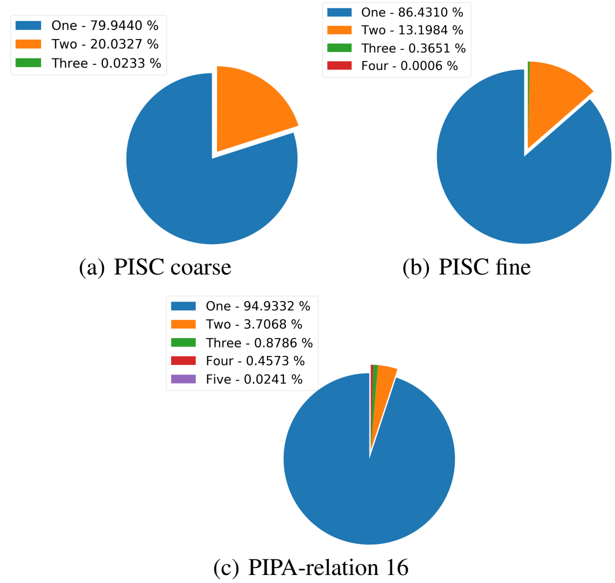
where $\mathrm{N}(I)$ returns the number of person pair in image I, $\mathrm{L}(\cdot)$ is the cross entropy loss function, $\mathcal{I}$ is the image set.

4 Experiments

4.1 Experiment Setting

Datasets. In this work, two datasets were used to evaluate our proposed method and other existing ones. The first one is the large-scale People in Social Context (PISC) [Li *et al.*, 2017] with 22,670 images and contains two-level recognition tasks: **3 Coarse-level relationship**, namely *No Relation, Intimate Relation, None-Intimate Relation* and **6 Fine-level relationship**, i.e., *Friend, Family, Couple, Professional, Commerical, No Relation*. The second one is the People in Photo Album Relation (PIPA-Relation) [Sun *et al.*, 2017], an extension verson of People in Photo Album (PIPA) [Zhang *et al.*, 2015a] with 37107 images. It also annotates 26,915 person pairs on two-level recognition tasks: **5 Social Domains** and **16 Social Relations** based on these domains. The train/val/test in PISC are 13,142/4,000/4,000 images with 14,536/25,636/15,497 person pairs on coarse level relationship, and 16,828/500/1,250 images with 55,400/1,505/3,691 person pairs on fine level relationshp, respectively. In PIPA-relation, we follow [Wang *et al.*, 2018] and focus on recognizing its 16 relationships in the experiment. The train/val/test in it are 13,729/709/5,106 person pairs.

Implementation Details. During our work, We adopt the same strategy as previous works including [Li *et al.*, 2017] and [Wang *et al.*, 2018]. First, we fine-tune the ResNet-101 model [He *et al.*, 2016] , and we set the a lower learning rate as 0.0001. For the message passing progation model, the dimension of hidden size is set as 512. The iteration time T is set as 4 and learning rate as 0.0001. Similer to [Wang *et al.*, 2018], we the fine-tuning model utilized SGD, and the message passing module is trained with ADAM.



(a) PISC coarse (b) PISC fine

(c) PIPA-relation 16

Figure 3: Social relaion categories per image on PISC and PIPA-relation.

4.2 Datasets Analysis

In this subsection, we made an analysis on PISC and PIPA-relation. Here we only used its train set and test set for statistic. As shown in Figure 3, we first calculated the social relation categories of each image, and found almost all images have only one social relation category, and the other half have two categories. For example, on PISC, approximately 79.944% of images have only one coarse category, while 20.0327% images have two coarse one.

4.3 Comparisons with State-of-the-Art Methods

We compare our proposed model with existing state-of-the-art methods on both PISC and PIPA-Relation datasets.Formally, the compared methods are as followed:

Performance on the PISC dataset

UnionCNN Following [Lu *et al.*, 2016], it generates a single CNN model to predicate relations. In this task, we also feeds the union region of person pair to s single CNN for classification.

Pair CNN[Li *et al.*, 2017] consists of two equivalent CNNs with shared weights to extracted features for image for two individuals.

Pair CNN + BBox + Union[Li *et al.*, 2017] incorporates spatial location information of two bounding box that based the previous pair CNN and Union CNN.

Dual-glance[Li *et al.*, 2017] implements coarse and fine prediction which includes three and six relationships. Dual-glance employing pair CNN + BBox + BBox + Union and utilized surrounding region proposal to refine the prediction.

GRM[Wang *et al.*, 2018] propse a graph reasoning model that unifies the frequence of co-concurrences of each relationship-object pair to facilitate social relation.

Similar to the model of GRM, we also adopt the per-class recall and mean average precision(mMP) to evaluate our

Table 1: Recall-per-class and mean average precision (mAP) evaluating our PRN model and previous methods on PISC (in %).

Methods	Coarse relationships				Fine relationships						
	Intimate	Non-Intimate	No Relation	mAP	Friends	Family	Couple	Professional	Commerical	No Relation	mAP
Union CNN [Lu *et al.*, 2016]	72.1	81.8	19.2	58.4	29.9	58.5	70.7	55.4	43.0	19.6	43.5
Pair CNN [Li *et al.*, 2017]	70.3	80.5	38.8	65.1	30.2	59.1	69.4	57.5	41.9	34.2	48.2
Pair CNN + BBox + Union [Li *et al.*, 2017]	71.1	81.2	57.9	72.2	32.5	62.1	73.9	61.4	46.0	52.1	56.9
Pair CNN + BBox + Global [Li *et al.*, 2017]	70.5	80.0	53.7	70.5	32.2	61.7	72.6	60.8	44.3	51.0	54.6
Dual-glance [Li *et al.*, 2017]	73.1	84.2	59.6	79.7	35.4	68.1	76.3	70.3	57.6	60.9	63.2
GRM [Wang *et al.*, 2018]	81.7	73.4	65.5	82.8	59.6	64.4	58.6	76.6	39.5	67.7	68.7
Ours											

Table 2: Accuracy (in %) evaluating our PRN model and previous methods on PIPA-relation.

Methods	accuracy
Two stream CNN [Zhang *et al.*, 2015a]	57.2
Dual-Glance [Li *et al.*, 2017]	59.6
GRM [Wang *et al.*, 2018]	62.3
Ours	64.7

model. The experiments data are reported in Table 1. First, both Pair CNN + BBox + Union, Pair CNN + BBox + Global, Dual-glance are incur extra Faster-RCNN[Ren *et al.*, 2015] to extract the local contextual cues(object proposal). GRM utilized the object proposal to construct a semantic-aware knowledge graph for reason about the social relationship. It is notable that both of them incur extra detection annotations that contains noisies.Specifically, our model achieves an accuracy of 75.0% and mAP of 82.1% for the coarse-level recognition. the model also takes an accuracy of 65.6% and mAP of 69.7% for the fine-level dataset, our model beating previous best model in the fine-level recognition ,but slightly lower on coarse-level than the best model before.

Performance on the PIPA-Relation dataset

On this dataset, we also compare our proposed model with the existing methods,i.e, Two stream CNN[Sun *et al.*, 2017],Dual-glance[Li *et al.*, 2017] and GRM[Wang *et al.*, 2018] that achieves the best performance before. Specifically, we directly reprint the experimental of serveral baselines from the literature. The result are presented in Table 2. Notably, our PRN significantly outperforms previous methods. Still, our model outperforms all the baselines in PIPA-relation, and beating the best of them 2.4%.

4.4 Analysis on Experimental result

In this section, we first present the comparison result of message passing mechanism and analsis the reasons behind, and the result present in Table 3. Then, we conduct a conditional experiment to investigate the effectiveness of the factor of region contexuals and the interaction between person-pair.

Table 3: The mMap and accuracy result of RCNN, our model and our model with contexual region that implements in the same way as dual-glance

Methods	PISC coarse		PISC fine	
	accuracy	mAP	accuracy	mAP
RCNN	-	63.5	-	48.4
Ours(max)	10.2	10.2	10.2	10.2
Ours(average)	10.2	10.2	10.2	10.2
Ours(atten)	10.2	10.2	10.2	10.2
Ours(atten)+region	10.2	10.2	10.2	10.2

Significance of message passing

In our framework, the core component is the introduction of message passing mechanism, and one of the key component is the message pooling functions that use learnt weights sum to aggregate hidden state of other relation nodes into message. To futher inverstigate the improvement of our approach on recognizing social relationships, we evaluate variants of our model with standard pooling methods. The first is to use average-pooling (avg. pool) instead of the learnt weighted sum to aggregate the hidden states. The second is similar to the first one, but uses max-pooling (max pool).

Analysis of contextual information

First, our model is to process the cue of contexual of relationships without contextual region that incured in Dual-glance and GRM by extra detecion annotation. From the contexual region point of view, the effect of contexual region is provide the information of scene to constraint the vector of the relationship and the information of scene is single. So, contexual regions is limited for the images with multi-class social relationship. But, the information of contexual relationships is different, and it is more appropriate for the fine relationships. For example, if the object of laptop is useful to classifier the relationship as professional that is not contribute to the relationship of friend. As tabel 3 reported, the model of RCNN is the lowest, and [Li *et al.*, 2017] trained a model with attention machanisms to exploit different contextual cues ac-

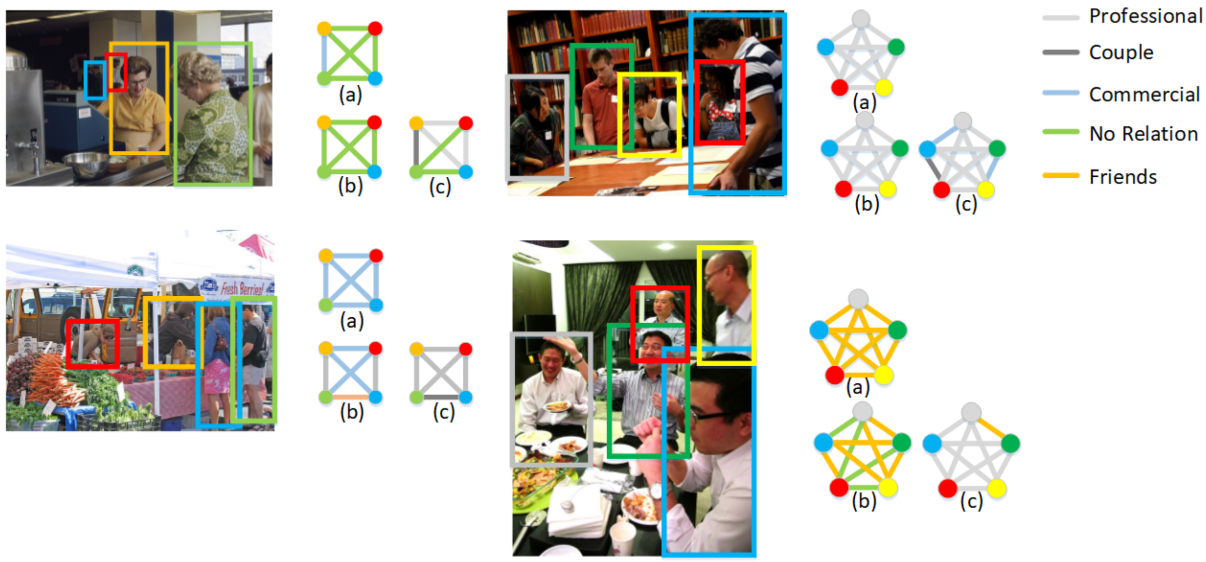


Figure 4: Comparison of social relation graphs from different sources: (a) PISC fine, (b) our PRN and (c) GRM. People with boxes of various colors on the original image correspond to the same color node on right of the image. An edge linked to two nodes represents the social relationships between them and its meaning can be found in the right legend.

cording different pairs of people. we also conduct a attention machanisms as same as Dual-glance, but the performance do not improved as reported in Table 3. One possible reason is that the contexual regions cues is coverd by contexual information of relationships cues, and the result of experiments proved the analysis before.

4.5 Case Study

Four examples in Figure 4 are shown to illustrate the ability of our PRN to infer social relationshps. We compared two predicted social relationships, one from our PRN and the another from GRM [Wang *et al.*, 2018]. We found that compared to the social relation graph in (c), the graphs in (b) in all of these examples are very similar to the graphs in (a), which means that our PRN performs better than GRM. In addition, the edges in (a) are almost identical, meaning the social relationships in a still image are always stable. More importantly, similar to (a), more than half of the edges in (b) predicted by our PRN are also consistent, which strongly suggests that the interaction cues of person pairs are very significant for social relationship understanding and our PRN can fully utilize it. Considering the first example, the true social relationship between the person with an orange box and the another with an green one is No Relation, which can be correctly predicted by our PRN while being incorrectly predicted as Couple by GRM, and the accuracy of PRN is 100% while is 33.3% in GRM.

5 Conclusion

References

[Barr *et al.*, 2014] Jeremiah R. Barr, Leonardo A. Cament, Kevin W. Bowyer, and Patrick J. Flynn. Active clustering with ensembles for social structure extraction. In *IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, March 24-26, 2014*, pages 969–976, 2014.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In Wu et al. [2014], pages 103–111.

[Dibeklioglu *et al.*, 2013] Hamdi Dibeklioglu, Albert Ali Salah, and Theo Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1497–1504, 2013.

[Ding and Yilmaz, 2010] Lei Ding and Alper Yilmaz. Learning relations among movie characters: A social network perspective. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, pages 410–423, 2010.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[Johnson *et al.*, 2015] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 3668–3678, 2015.

[Li *et al.*, 2015] Li-Jia Li, David A. Shamma, Xiangnan Kong, Sina Jafarpour, Roelof van Zwol, and Xuanhui Wang. Celebritynet: A social network constructed from large-scale online celebrity images. *TOMCCAP*, 12(1):3:1–3:22, 2015.

[Li *et al.*, 2017] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2669–2678, 2017.

[Liang *et al.*, 2016] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph LSTM. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 125–143, 2016.

[Lu *et al.*, 2016] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 852–869, 2016.

[Ramanathan *et al.*, 2013] Vignesh Ramanathan, Bangpeng Yao, and Fei-Fei Li. Social role discovery in human events. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2475–2482, 2013.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[Sun *et al.*, 2017] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 435–444, 2017.

[Vinciarelli *et al.*, 2009] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image Vision Comput.*, 27(12):1743–1759, 2009.

[Wang *et al.*, 2010] Gang Wang, Andrew C. Gallagher, Jiebo Luo, and David A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V*, pages 169–182, 2010.

[Wang *et al.*, 2015] Quan Wang, Bin Wang, and Li Guo. Knowledge base completion using embeddings and rules. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1859–1866, 2015.

[Wang *et al.*, 2018] Zhouxia Wang, Tianshui Chen, Jimmy S. J. Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 1021–1028, 2018.

[Wu *et al.*, 2014] Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi, editors. *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*. Association for Computational Linguistics, 2014.

[Xu *et al.*, 2017] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3097–3106, 2017.

[Yatskar *et al.*, 2016] Mark Yatskar, Luke S. Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5534–5542, 2016.

[Zhang *et al.*, 2015a] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir D. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4804–4813, 2015.

[Zhang *et al.*, 2015b] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3631–3639, 2015.

[Zheng *et al.*, 2015] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1529–1537, 2015.