Graph-Based Social Relation Inference with Free-Form Attention

Anonymous CVPR submission

Paper ID 8731

Abstract

Social relation inference intrinsically requires high-level semantic representation learning. In order to accurately infer relations of persons in images, one needs to not only understand scenes and objects in images, but also adaptively attend to important clues for relations. Existing studies classify social relations using attention on detected objects, which constrains the scope of attention and may miss vital clues such as human interactions. In contrast, we propose a free-form attention mechanism, where the attention can be adaptively on scenes, objects and human interactions, for social relation inference. We develop a transformer-style network, which consists of three modules, i.e.*, a feature extraction module, a novel graph-based query module and a transformer reasoning module. Specifically, in the proposed Graph-based Relation Inference Transformer (*i.e.*, GRIT), we design a graph-based query module to generate informative relation queries, which fuses local person features and global context. Moreover, in the transformer reasoning module, we fully take advantage of global self-attention units in the transformer-style network for classifying social relations. To the best of our knowledge, the proposed GRIT is the first for inferring social relations with free-form attention. GRIT is trained in an end-to-end manner and consistently outperforms other methods on two benchmark datasets under different backbone settings. All the codes and experimental results are publicly available on github*[1].

1. Introduction

Social relations, which are fundamental to the daily life of human beings [13], characterize the connections among two or more individuals. In light of [2], common social relations include family, couple, friends, colleagues, professional, etc. Nowadays, billions of people share images in social media platforms such as Facebook and Twitter. There has been a dramatically increasing interest in understanding social relations among persons in still images due to the broad computer vision applications including social recommendations based on images [28], group behavior analysis [10], and image caption generation [4].

The problem of social relation inference is challenging and complicated because it requires high-level semantic understanding of images. In order to accurately infer social relations of persons in images, one needs to not only identify the category of scenes and background objects in images, but also have free-form attention on the scenes, objects and human interactions, *e.g.*, hugging and handshaking. Free-form attention does not require a pre-processing step of object detection. Besides, free-form attention has two advantages. The first one is attention on scenes, objects or human interactions in a global view for different relations within an image. The second one is differential attention not just on objects but also on scenes and human interactions for the same social relation in different images.

We note that there are two studies for social relation inference with attention. In [14], a dual-glance model was developed in which the first glance obtains feature of person pair and the second glance attends to region proposals of detected objects in images. In [26], a knowledge graph with persons and objects was proposed to infer social relations. The graph model learns to classify relations with weights on contextual objects. Both methods rely on object detection algorithms as a pre-processing step and only attend to region proposals of detected objects. Thus we call the attention mechanism in [14,26] as constrained attention.

To illustrate the importance of free-form attention and its difference from constrained attention in this task, we visualize four examples of the attention in relation inference by the dual-glance model in [14] and our GRIT model in Figure 1. It is clear that free-form attention helps infer social relations because scenes, objects or human interactions may contribute to the inference of relations in different ways for various scenarios. By contrast, the constrained attention mechanism in the models of [14,26] was heavily restricted by the performance of object detection models. For example, the number of object category in backbone models for object detection based on COCO dataset is 80, which might

[1] https://github.com/IFBigData/GRIT

(a) Method in [14] attends to a person and wrongly predicts friend. True label is no relation.

(b) Method in [14] attends to a cup in hands and a tie, and correctly predicts friend. True label is friend.

(c) Method in [14] attends to a bottle of water and wrongly predicts no relation. True label is friend.

(d) Method in [14] attends to a person and a cup on the table, and wrongly predicts family. True label is friend.

(e) GRIT attends to the background (*e.g.*, candle), and correctly predicts no relation. True label is no relation.

(f) GRIT attends to human interactions and correctly predicts friends. True label is friend.

(g) GRIT attends to faces, dishes and neighbeurs, and correctly predicts friend. True label is friend.

(h) GRIT attends to faces, chairs and standing persons, and correctly predicts friend. True label is friend.

Figure 1. Comparisons of attention in social relation inference between the dual-glance model [14] and our GRIT. The proposed GRIT achieves free-form attention in various scenarios. Figure 1e and Figure 1f show free-form attention on the background and human interactions for different relations within an image. Figure 1g and Figure 1h show differential attention in a free-form manner for the same social relation. Attention is shown in terms of heatmap. Best view in color.

miss plenty of less frequent but important objects or clues affecting social relation inference.

Motivated by solving the problem of social relation inference with free-form attention, we propose a transformer-style network. We note that transformer networks are powerful in global self-attention, and thus may help achieve free-form attention in this task. In [3, 6, 16], various transformers were proposed to tackle computer vision tasks with outstanding performance, especially for the tasks requiring attention in a global view.

In this paper, we propose Graph-based Relation Inference with Transformer (GRIT), which is trained in an end-to-end manner. GRIT consists of a Feature Extraction Module (FEM), a novel Graph-based Query Module (GQM) and a Transformer Reasoning Module (TRM). The proposed GQM is a graph neural network that is designed to fuse local person features and global context to generate informative relation queries for TRM. The relation queries may help to capture logical constraints among different types of social relations. Inspired by [23], we design TRM with an encoder-decoder transformer for classifying social relations, which contains powerful self-attention units. The combination of TRM and GQM achieves free-form attention in GRIT. To the best of our knowledge, we are the first to tackle the social relation inference problem with free-form attention. GRIT in this study achieves new state-of-the-art results on two benchmark datasets for social relation inference.

We summarize our contributions in this work as follows.

- We develop a novel method which is named GRIT by taking advantage of global self-attention units and graph representation learning for social relation inference. The proposed GRIT achieves free-form attention, and does not require a pre-processing step of object detection in classifying social relations.
- We design a graph-based query module in the method, which is experimentally verified to be effective for social relation inference. The graph-based query module enables GRIT to classify the relations of all person pairs in an image within a single pass.
- The proposed GRIT significantly outperforms previous attention-based methods, *e.g.*, 6.3% absolute improvement in fine relation recognition on PISC dataset. Moreover, GRIT consistently surpasses other methods on standard datasets under different backbone settings.

2. Related Work

To assess our contributions in the classification of social relations, we consider three streams of studies: social relation inference, graph neural networks and transformer.

2.1. Social Relation Inference

Recently, there has been a recent surge of research interest in social relation inference based on images. Based on [14, 30], for the problem of social relation inference, the input of a model is images with annotated bounding boxes of all persons, and the model is required to predict the social relation of each person pair. There are two benchmark datasets for this task. Specifically, one dataset is named as People In Photo Albums (PIPA) in [30] and the other is named as People in Social Context (PISC) in [14].

The pioneering work on social relation inference dates back to 2010 [24], where the authors developed a model to characterize the interaction between multi-person actions, facial appearances and identities. The model recognized the family relationships, such as husband-wife, siblings and grandparent-child, etc. Later, Zhang et al. [31] developed a deep neural network to learn social relation traits from rich facial attributes, such as expression, gender and age.

With PIPA and PISC, several interesting works move forward along the research line of social relation understanding [8, 15, 20]. For instance, Sun et al. [20] presented a model with semantic attributes to classify social relations. Goel et al. [8] trained a model in an end-to-end manner with multi-task loss. Recently, Li et al. [15] proposed a social graph to restrict logical connections of persons, which achieved state-of-the-art results in social relation inference. We find that these models may get a less representative and discriminative feature for subtle but important clues of different relations due to the lack of attention mechanism.

We note that there are two studies for social relation inference with constrained attention. Wang et al. [26] constructed a knowledge graph of persons and objects by object detection algorithms. The knowledge graph learned the weight mapping from detected objects to social relations. Li et al. [14] developed a dual-glance method attending to region proposals of detected objects in images. Both methods with attention mechanism must adopt object detection algorithms as a pre-processing step, which limits the robustness of the models. It is urgent and essential to solve the problem of social relation inference with free-form attention.

2.2. Graph Neural Networks (GNNs)

Inspired by the success of convolutional networks in the computer vision domain, GNNs [18, 22, 27] were proposed to redefine the notation of convolution for graph structured data [12]. Most recently, GNNs have been adopted to social relation reasoning [15, 26, 29]. For instance, Wang et. al. [26] applied gated GNN with a graph-attention mechanism on knowledge graphs to facilitate social relationship recognition. Zhang et al. [29] designed a person-object graph and a person-pose graph, and conducted social relation reasoning on these two graphs by GNN. Li et al. [15] proposed a graph relation reasoning network to infer social relations by building a graph for each image, where the nodes represent the persons and the edges represent the relations. In this paper, inspired by the design of GNN in [12], we propose a graph-based query module to extract relation queries with logical constraints among different types of social relations from an image.

2.3. Transformer

Transformer was first introduced by Vaswani et al. [23] as a new building block with self-attention mechanism in natural language processing for machine translation. The attention mechanism in transformers usually refers to a neural network that aggregates information from the entire input, e.g., the whole sentence in machine translation.

Transformer was originally designed for the sequence-to-sequence [21] tasks, and then transferred to other domains [5, 25]. In [23], stacked encoders and decoders were designed to capture global dependencies regardless of token distance in a sequence. In [6], vision transformer was developed by stacking encoders and splitting images into patches. The challenge for vision transformer is its intensive computation complexity due to image size. Recently, in [16], Swin Transformer was proposed for solving the problem of computation complexity, as well as generalizing the structure of transformer. We note that there are several studies on human object interactions with transformer [11, 17, 32]. But those methods cannot be directly adopted to solve the problem of social relation inference because social relations characterize the connections between persons and also social relation inference requires more high-level understanding on scenes, objects and human interactions. Thus, in this work, we design a transformer-style network to attack the challenge of free-form attention in social relation inference.

3. Method

The overall architecture is presented in Figure 2. The framework of GRIT consists of an FEM, a GQM and a TRM. In GRIT, an image is first passed to FEM to extract RoI features of all persons and the global image feature. Then the person features and global image feature are fed into GQM. Specifically, the GQM is designed to fuse local person features and global context to generate informative relation queries for TRM. The relation queries may help to capture logical constraints among different types of social relations. Inside GQM, a stack of graph convolutional layers updates edges and nodes in an alternative manner. Based on relation queries and the flattened image feature, TRM infers the relations of all person pairs in an image within a single pass. The combination of GQM and TRM can adaptively attend to important clues for social relation recognition in a free-form manner. GRIT is trained in an end-to-end manner. In the following, details of the three modules will be introduced separately.

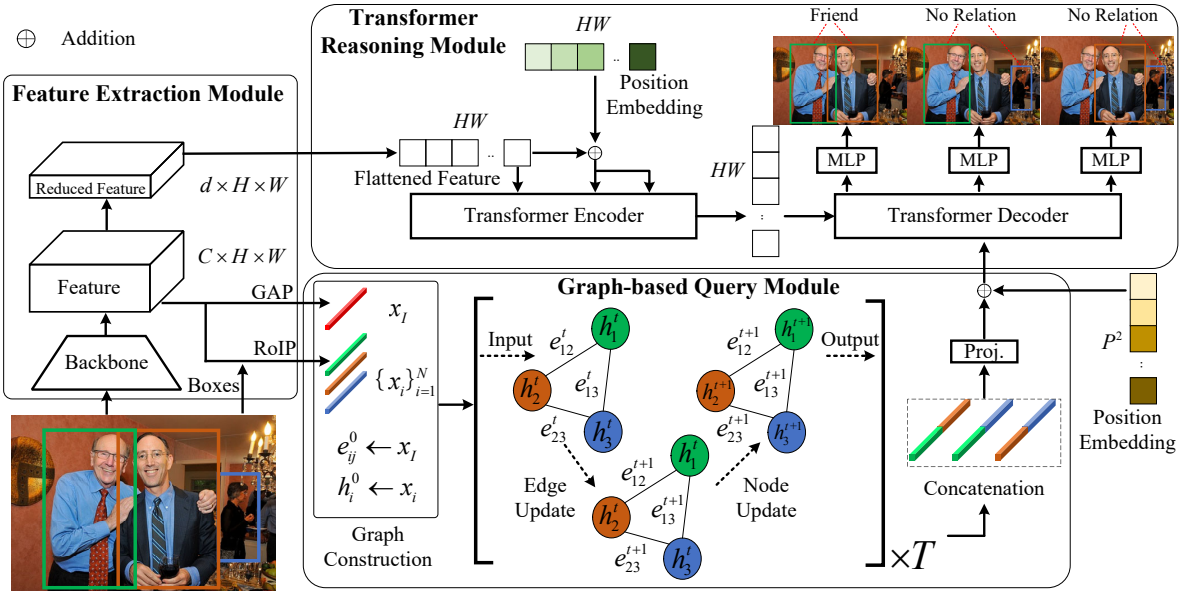


Figure 2. The proposed end-to-end network (*i.e.*, GRIT) for social relation inference. Given an image with person bounding boxes as input, the Feature Extraction Module extracts a set of features, including a dimension-reduced image feature map, a global image feature x_I (red) and RoI-pooled box features $\{x_i\}_{i=1}^N$. In the Graph-based Query Module, a complete graph is constructed with the node and edge features initialized from x_i and x_I, respectively. The edge and node features are updated alternatively in each layer of the Graph-based Query Module. After T rounds of updates, we concatenate pairs of node features to generate the relation queries. The relation queries and the reduced image features are fed into the Transformer Reasoning Module to produce relation predictions. *RoIP* and *GAP* refer to RoI pooling and global average pooling, respectively. The "Proj." means a linear layer for dimension reduction.

3.1. Feature Extraction Module (FEM)

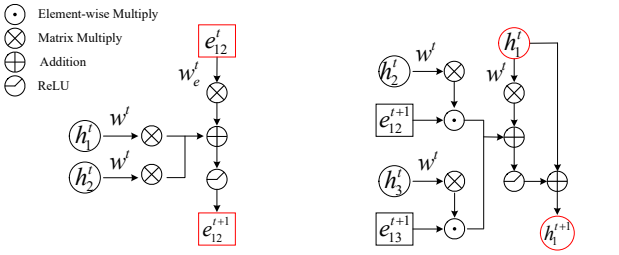
Given an image with its person bounding boxes, we use a backbone model to extract features for GRIT. Our FEM extracts two types of features, *i.e.*, the features of persons and the global image feature.

Starting from the initial image $I \in \mathbb{R}^{3\times H_0\times W_0}$, a backbone generates a feature map $f_I \in \mathbb{R}^{C\times H\times W}$. Typically, $H = H_0/32$, $W = W_0/32$, and the value of C depends on the backbone model. For instance, $C = 1024$ if the backbone is Swin Transformer Base model [16] and $C = 2048$ if the backbone is Resnet101 [9]. We then extract the feature representations of each person and the whole image from the feature map f_I using RoI pooling (RoIP) and global average pooling (GAP), respectively. Specifically, given feature map f_I with N bounding boxes $b_1, b_2, ..., b_N$ for N persons in image I, we obtain the feature representations of the i-th person in the image, denoted as x_i as follows:

$$x_i = RoIP(f_I, b_i) \in \mathbb{R}^C, \tag{1}$$

and the feature representation of the whole image x_I is extracted by applying a GAP layer on the f_I:

$$x_I = GAP(f_I) \in \mathbb{R}^C. \tag{2}$$



(a) Edge Update $e_{12}^t \rightarrow e_{12}^{t+1}$ (b) Node Update $h_1^t \rightarrow h_1^{t+1}$

Figure 3. Illustration of edge and node updates in GQM. The red boxes emphasize the units in $e_{12}^t \rightarrow e_{12}^{t+1}$ and $h_1^t \rightarrow h_1^{t+1}$.

3.2. Graph-based Query Module (GQM)

The GQM is designed to fuse local person features x_i and global image feature x_I to generate relation queries for all person pairs in the image, which are used as inputs to the transformer's decoder in TRM. In essence, the design of GQM is inspired by [15, 26, 29]. In the following, we introduce the details of graph construction and iteration.

Graph Construction We build a graph for each image to generate relation queries. In the graph, each person in an image is modeled as a node, and each pair of persons in the image has a message-passing edge, *i.e.*, there is a com-

plete graph for each image. Denote $\mathcal{G} = (\mathcal{V}, \xi)$ as the complete graph with node set $\mathcal{V}$ and edge set ξ in an image. For the i-th node $v_i \in \mathcal{V}$ in $\mathcal{G}$, we set its initial node feature representation as $h_i^0 = x_i \in \mathbb{R}^C$. Correspondingly, each edge has a feature representation, and we set the initial edge feature representation between node i and node j as $e_{ij}^0 = x_I \in \mathbb{R}^C$.

Graph Iteration The GQM is a stack of graph convolutional layers, and there is an edge update and a node update in each layer, as illustrated in Figure 3. In total, the edge and node feature representations are updated iteratively for T times in GQM. Specifically, at $(t+1)$-th layer the edge and node representations are updated as follows:

$$e_{ij}^{t+1} = \sigma(W^t h_i^t + W^t h_j^t + W_e^t e_{ij}^t), \quad (3)$$

$$h_i^{t+1} = h_i^t + \sigma(W^t h_i^t + \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} e_{ij}^{t+1} \odot W^t h_j^t), \quad (4)$$

where $\mathcal{N}_i$ is the set of neighbors for node i, $W^t, W_e^t \in \mathbb{R}^{C \times C}, t = 0, 1, ..., T-1$ are the learnable parameters at each layer, σ is the ReLU function, and $\odot$ is the Hadamard point-wise multiplication operator. Intuitively, through iterative updates on edge and node representations, the local personal feature and global image feature are fused to form better feature representations.

Finally, we obtain relation query for each person pair by concatenating the node feature of the two persons from the last layer. Namely, we have

$$q_{ij} = < h_i^T, h_j^T >,$$

where $<>$ means concatenating feature vectors of two persons.

We note that GQM in Equations (3)-(4) is an anisotropic variant of GCN [7]. Similar to Residual GatedGCN [1], GQM has residual connections on the node feature update, and explicitly maintains edge feature at each layer.

Although there are some pioneering studies that apply GCN in social relation inference [26,29], most of them treat the problem as pairwise social relation recognition, *i.e.*, they predict social relations of person pairs independently. This clearly limits the power of GCN. Most recently, Li et al. [15] proposed a GCN-based model called GR^{2}N to restrict logical connections of persons, which can simultaneously reason all relations for each image and achieved the state-of-the-art results in social relation inference. Although GQM and GR^{2}N [15] both are variants of GCN, our GQM is more scalable since the number of parameters of GQM is independent of the number of social relation categories in contrast to GR^{2}N.

3.3. Transformer Reasoning Module (TRM)

The TRM is designed based on conventional encoder-decoder transformer [23]. Intuitively, attention layers inside the transformer allow for automatically identifying important clues in images to help social relation inference.

Transformer Encoder We first apply a 1×1 convolution layer to reduce the dimension of the feature map f_I from C to d, and then flatten f_I into a sequence of visual tokens $\{f_i \in \mathbb{R}^d, i = 1, 2, \ldots, HW\}$. Each encoder layer has a standard architecture and consists of a multi-head self-attention module and a feed-forward network. Following traditional settings in other transformer models, we add a learnable positional embedding to the input of the encoder.

Transformer Decoder The inputs to the decoder are the relation queries among all person pairs from GQM and a learnable positional embedding. As different images may have different number of persons, we pad the relation queries to the maximal number of person pairs in the dataset. Denote the maximal number of persons in the dataset as P, the relation queries are pad to have length of P^2. In addition, we add learnable positional embedding to the relation queries. By design, the positional embedding helps the decoder to be aware of different person pairs.

4. Experiments

In this section, we conduct extensive experiments based on PIPA and PISC datasets. We first present the description of datasets and the implementation details. Then we evaluate the performance of GRIT through comparisons with benchmarks and ablation study. Finally, we visualize the attention map of sample images from GRIT and prior works to show the effectiveness of free-form attention.

4.1. Datasets

We conduct experiments on two social relation datasets, *i.e.*, the PIPA dataset [30] and the PISC dataset [14]. The PIPA dataset partitions social life into five social domains (coarse) and sixteen social relations (fine). The *accuracy* over all classes is typically used to evaluate all methods on PIPA dataset. The PISC dataset has a hierarchy of three coarse-level relations (intimate, non-intimate, no relation) and six fine-level relations (friend, family, couple, professional, commercial, and no relation). We follow the standard train/val/test split in [14]. The *per-class recalls* and *mean Average Precision (mAP)* are used to evaluate all methods on PISC dataset.

4.2. Implementation Details

We train GRIT with Adam Optimizer in an end-to-end manner and set the learning rate of backbone and the rest of network to 10^{-5} and 10^{-4}, respectively. The GRIT network is trained for 10 epochs with a batch size of 16. The number of layers in GQM is set to be 2, *i.e.*, $T = 2$. The transformer has 3 encoder and 3 decoder layers, each with 8 self-attention heads. The dropout in TRM is set as 0.2.

We adopt different backbones in GRIT. We report the results with ImageNet pretrained Resnet101 model as backbone and call the model as GRIT-R101. Following [15], the input images are resized to 448×448 for GRIT-R101. In addition, we also report results with ImageNet pretrained Swin Transformer Base model (*i.e.*, Swin-B in [16]) as backbone, denoted as SW224. The input image are resized to 224×224 and the corresponding models are called GRIT-SW224. More results based on different backbones are reported in supplementary materials.

4.3. Comparisons with Benchmarks

In this subsection, we compare GRIT with the following existing methods. For fair comparisons, we report the best results in experiments for Tables 2 and 3 following the routine in this research field.

Discussions on Benchmarks We present the details of prior methods as follows.

Pair-CNN [14] A backbone with RoI pooling is used to extract features of two persons, which are concatenated and fed into an MLP for classification.

Dual-Glance [14] The first glance focuses on the pair of persons. The second glance extracts the information of objects in the context to refine the prediction.

SRG-GN [8] Scene and human attribute context features are extracted by five CNNs.

GRM [26] A weighted graph is constructed to represent the persons and objects existing in an image, and then a gated graph network is applied to predict social relations.

MGR [29] Two GNNs are applied to learn features in the person-pose graph and the person-object graph.

GR^{2}N [15] A GNN is designed to model all relationships in a graph which can provide strong logical constraints among different types of social relations.

For fair comparisons, we also report the results in Pair-CNN, Dual-Glance, GRM and GR^{2}N by substituting their backbone with SW224. For ease of presentation, the corresponding models are called Pair-SW224, Dual-Glance-SW224, GRM-SW224 and GR^{2}N-SW224, respectively. We note that the source codes of SRG-GN and MGR are not available and their results cannot be reproduced. Besides, as their performance are inferior to GR^{2}N, we do not include the paper results of SRG-GN and MGR in Tables 2 and 3. Interested readers can refer to their papers [8, 29] for performance comparison.

All of the prior methods, except GR^{2}N, treat the problem as pairwise social relation recognition, *i.e.*, they predict social relations of person pairs separately. In contrast, GR^{2}N and our GRIT consider the social relations among all persons in one image jointly. Besides, GR^{2}N lacks attention mechanism by design. Dual-Glance and GRM use constrained attention to assist in social relation inference. Our GRIT adopts a graph-based transformer-style network to achieve free-form attention.

Table 1. Comparisons between GRIT and other attention-based methods on PIPA and PISC datasets with backbones *Resnet101* and *VGG*. We note that our GRIT only adopts *Resnet101* as backbone. We report accuracy on PIPA and mAP on PISC (in %).

Dataset	Method	Coarse (Domain)	Fine (Relation)
PIPA	Dual-Glance	65.2	59.6
	GRM	-	62.3
	GRIT (ours)	**73.4** (+8.2)	**66.7** (+4.4)
PISC	Dual-Glance	79.7	63.2
	GRM	82.8	68.7
	GRIT (ours)	**84.6** (+1.8)	**75.0** (+6.3)

Table 2. Comparisons of the accuracy (in %) between GRIT and other state-of-the-art methods on PIPA dataset.

Backbone	Method	Coarse (Domain)	Fine (Relation)
R101 /VGG	Pair-CNN	65.9	58.0
	Dual-Glance	65.2	59.6
	GRM	-	62.3
	GR^{2}N	72.3	64.3
	GRIT (ours)	**73.4** (+1.1)	**66.7** (+2.4)
SW224	Pair-CNN	77.6	70.7
	Dual-Glance	79.1	70.5
	GRM	-	69.9
	GR^{2}N	78.8	69.2
	GRIT (ours)	**80.4** (+1.3)	**71.5** (+0.8)

Experimental Comparison In the comparisons, we address the following questions with experimental results shown in Tables 1, 2 and 3.

Q1: How does GRIT perform by comparing all the attention-based methods?

Q2: By using Resnet101 or VGG [19] as backbones, how do GRIT and other methods perform?

Q3: By using SW224 as backbone, how do GRIT and other methods perform?

For Q1, we note that Dual-Glance, GRM and GRIT all use attention mechanism. The results of these methods are presented in Table 1. We observe that GRIT significantly outperforms other two attention-based methods. For instance, GRIT-R101 achieves 6.3% absolute improvement in fine relation recognition on PISC dataset compared to GRM. These results validate the effectiveness of free-form attention in GRIT.

For Q2, by following the backbone setting of GR^{2}N, we only show the results of GRIT with Resnet101 in Tables 2 and 3. We observe that GRIT-R101 outperforms all other

Table 3. Comparisons of the per-class recall for each relation and the mAP over all relations (in %) between our GRIT and other state-of-the-art methods on PISC dataset. "Back.": Backbone, Int: Intimate, Non: Non-Intimate, NoR: No Relation, Fri: Friend, Fam: Family, Cou: Couple, Pro: Professional, Com: Commercial.

Back.	Method	Coarse relationships				Fine relationships						
		Int	Non	NoR	mAP	Fri	Fam	Cou	Pro	Com	NoR	mAP
R101 /VGG	Pair-CNN	70.3	80.5	38.8	65.1	30.2	59.1	69.4	57.5	41.9	34.2	48.2
	Dual-Glance	73.1	84.2	59.6	79.7	34.4	68.1	76.3	70.3	57.6	60.9	63.2
	GRM	81.7	73.4	65.5	82.8	59.6	64.4	58.6	76.6	39.5	67.7	68.7
	GR^{2}N	81.6	74.3	70.8	83.1	60.8	65.9	84.8	73.0	51.7	70.4	72.7
	GRIT (ours)	83.6	74.0	70.1	**84.6** (+1.5)	66.3	65.4	52.0	79.4	33.3	79.6	**75.0** (+2.3)
SW224	Pair-CNN	82.2	74.2	49.0	75.9	71.8	69.0	53.1	84.7	29.1	51.9	69.6
	Dual-Glance	84.4	76.8	53.1	78.9	69.5	68.4	59.8	87.9	33.6	56.1	70.3
	GRM	81.0	75.4	56.0	78.2	68.8	69.2	60.5	84.6	34.5	56.0	70.9
	GR^{2}N	85.8	78.9	67.5	84.2	70.0	67.5	62.9	88.6	25.1	82.4	75.3
	GRIT (ours)	84.7	76.8	70.8	**85.6** (+1.4)	70.6	71.9	55.5	87.7	25.7	78.3	**78.3** (+3.0)

Table 4. Ablation study of GRIT's module on PISC dataset. We report the mean and standard deviation of mAP (in %) among 3 random runs on PISC dataset.

Settings	Coarse	Fine
Pair-SW224	84.1 ± 0.1	74.5 ± 0.3
GRIT-SW224 w/o GQM	84.6 ± 0.2	75.9 ± 0.6
GRIT-SW224 w/o TRM	84.4 ± 0.1	76.4 ± 0.4
GRIT-SW224	$\mathbf{85.4 \pm 0.1}$	$\mathbf{77.6 \pm 0.3}$

Table 5. Ablation study of transformer module with different number of encoder and decoder on PISC dataset. We report the mean and standard deviation of mAP (in %) among 3 random runs on PISC dataset.

Settings	Coarse	Fine
Enc1 & Dec1	85.0 ± 0.2	78.0 ± 0.3
Enc3 & Dec3 (Default)	85.4 ± 0.1	77.6 ± 0.3
Enc6 & Dec6	85.0 ± 0.1	77.2 ± 0.1

methods on PIPA dataset and improves the current state-of-the-art method, *i.e.*, GR^{2}N, by 1.1% absolute improvement for coarse relation recognition and 2.4% absolute improvement for fine relation recognition, respectively. Consistent observations can be found from Table 3. We observe that GRIT-R101 achieves an mAP of 84.6% for the coarse-level recognition and 75.0% for the fine-level recognition, outperforming GR^{2}N by 1.5% and 2.3% in absolute improvements, respectively. This result demonstrates the superiority of GRIT with Resnet101 as backbone.

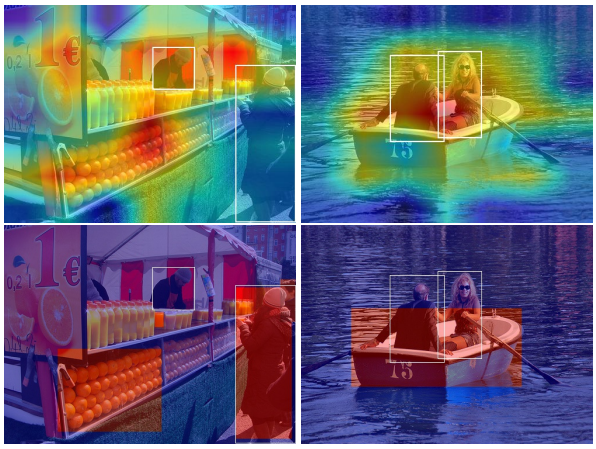
For Q3, we observe that GRIT-SW224 consistently outperforms all previous methods and establishes new state-of-the-art results on both datasets. In particular, GRIT-SW224 achieves 1.4% and 3.0% absolute improvements in PISC coarse and fine relationships recognition respectively as compared to previous best method. In contrast, GR^{2}N-SW224 outperforms other previous methods on PISC dataset, but its performance is inferior to other methods on PIPA dataset. The results in Tables 2 and 3 based on SW224 as backbone further validate the advantage of our proposed network architecture.

By combining the results from Q2 and Q3, we observe that most of the methods have improvements if we replace their backbone with SW224. For instance, GR^{2}N-SW224 achieves an mAP of 75.3% in PISC fine relationships recognition, which is 2.6% higher in absolute value as compared to GR^{2}N. It suggests that Swin Transformer can extract more representative features than Resnet model.

Overall, we observe that GRIT consistently surpasses other methods on PIPA and PISC datasets under different backbone settings.

4.4. Ablation Study

Module Ablation The modules of TRM and GQM are the core of our GRIT model, which work together to perform free-form attention. To verify the effectiveness of these two modules, we conduct a module ablation experiment. By default, we use SW224 as the backbone. We remove GQM (GRIT-SW224 w/o GQM) and TRM (GRIT-SW224 w/o TRM) from GRIT-SW224 respectively. We also remove GQM and TRM simultaneously (Pair-SW224). The experiment results are shown in Table 4. Instead of reporting only the best mAP, we report the mean and standard deviation of mAP among 3 random runs, which can reduce the influence of randomness. We observe that the mAP in both levels of PISC dataset drops significantly after removing GQM or



(a) Commercial. (b) Couple.

(c) Friend. (d) Sport team members.

Figure 4. Visualization of sample images with attention heatmap. Images in (a) and (b) are from PISC dataset. Images in (c) and (d) are from PIPA dataset. Images in the first and second rows are results from GRIT and GRM, respectively. Best view in color.

TRM. The performance degradation demonstrates that the proposed GRIT needs TRM and GQM to work together, and thus can effectively perform social relation inference.

Structure Ablation To further explore the influence of structure in GRIT, we design an experiment that uses different numbers of encoder and decoder layers in TRM. As shown in Table 5, we observe that the changes in mAP on both coarse and fine levels on PISC dataset are very minor. These results suggest that GRIT-SW224 is insensitive to the depth of TRM. By default, we set the number of layers in both encoder and decoder to be 3.

4.5. Qualitative Evaluation

In this subsection, we visualize the attention results from GRIT and GRM to demonstrate the difference between free-form and constrained attention, and show the effectiveness of the free-form attention mechanism.

(a) True label is no relation. The predicted label is Professional. (b) True label is family. Predicted label is friend.

Figure 5. Visualization of sample images with attention heatmap where GRIT makes wrong predictions. Best view in color.

Sample images with attention heatmap are shown in Figure 4. In Figure 4a, one customer is buying juice in front of a booth. GRIT understands the context by paying attentions on the juice, the oranges and the signboard in the booth. However, limited by the power of object detection algorithms in the pre-processing step, GRM only has attention weights on the oranges and the customer. In Figure 4b, a couple were rowing a boat on the lake. Although both GRIT and GRM notice about the boat, GRIT has free-form attention on extra information such as the persons and the water. In Figure 4c, two friends are looking at the computer. GRIT infers their relation by attending on the computer, while GRM ignores this vital information. In Figure 4d, two sport team members are playing soccer on a field. GRM attends to the soccer only, while GRIT has free-form attention on the soccer, the field and human interactions.

We also present two failure cases of GRIT in Figure 5. In Figure 5a, we observe that our method fails to make correct predictions when the persons in the images are too small or being occluded by other objects. When there exists overlapping of persons, it is difficult for GRIT to attend and make correct predictions, as shown in Figure 5b.

5. Conclusions

In this paper, we propose GRIT for social relation inference. GRIT consists of an FEM, a GQM and a TRM. We design a transformer-based network to achieve free-form attention of images in a global view for classifying social relations. We develop an iterative update strategy for a graph to generate informative relation queries in GQM. Extensive experiments clearly demonstrate that the proposed GRIT is superior than the prior methods, and establishes new state-of-the-art results on PIPA and PISC datasets. Ablation study and qualitative evaluation show the effectiveness of the proposed modules. The query design in this study might inspire new directions of investigations on new transformer for computer vision and natural language processing. Future directions lie in free-form attention networks for image caption generation, behaviour analysis, etc.

References

[1] Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *ICLR*, 2018. 5

[2] David G Bromley and Bruce C Busching. Understanding the structure of contractual and covenantal social relations: Implications for the sociology of religion. *SA. Sociological analysis*, pages 15S–32S, 1988. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 2

[4] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, pages 2422–2431, 2015. 1

[5] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *NeurIPS*, 1(2):3, 2021. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 3

[7] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *TSP*, 12:50–500. 5

[8] Arushi Goel, Keng Teck Ma, and Cheston Tan. An end-to-end network for generating social relationship graphs. In *CVPR*, pages 11186–11195, 2019. 3, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[10] Minh Hoai and Andrew Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *CVPR*, pages 875–882, 2014. 1

[11] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, pages 74–83, 2021. 3

[12] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3

[13] Shinobu Kitayama and Hazel Rose Markus. The pursuit of happiness and the realization of sympathy: Cultural patterns of self, social relations, and well-being. *Culture and subjective well-being*, 1:113–161, 2000. 1

[14] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *ICCV*, pages 2650–2659, 2017. 1, 2, 3, 5, 6

[15] Wanhua Li, Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Graph-based social relation reasoning. In *ECCV*, pages 18–34. Springer, 2020. 3, 4, 5, 6

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021. 2, 3, 4, 6

[17] Debaditya Roy and Basura Fernando. Action anticipation using pairwise human-object interactions and transformers. *IEEE Transactions on Image Processing*, 30:8116–8129, 2021. 3

[18] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 3

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[20] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *CVPR*, pages 3481–3490, 2017. 3

[21] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. 3

[22] Kiran K Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. Attention-based graph neural network for semi-supervised learning. *ICLR*, 2018. 3

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 3, 5

[24] Gang Wang, Andrew Gallagher, Jiebo Luo, and David Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, pages 169–182. Springer, 2010. 3

[25] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021. 3

[26] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *IJCAI*, pages 1021–1028, 2018. 1, 3, 4, 5, 6

[27] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *KDD*, pages 793–803, 2019. 3

[28] Jing Zhang, Ying Yang, Li Zhuo, Qi Tian, and Xi Liang. Personalized recommendation of social images by constructing a user interest tree with deep features and tag trees. *IEEE Transactions on Multimedia*, 21(11):2762–2775, 2019. 1

[29] Meng Zhang, Xinchen Liu, Wu Liu, Anfu Zhou, Huadong Ma, and Tao Mei. Multi-granularity reasoning for social relation recognition from images. In *ICME*, pages 1618–1623. IEEE, 2019. 3, 4, 5, 6

[30] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, pages 4804–4813, 2015. 3, 5

[31] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *ICCV*, pages 3631–3639, 2015. 3

[32] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, pages 11825–11834, 2021. 3