

An End-to-End Network for Generating Social Relationship Graphs

Arushi Goel¹, Keng Teck Ma^{1,2}, and Cheston Tan²

¹A*STAR Artificial Intelligence Initiative, Singapore, ²Institute for Infocomm Research, A*STAR, Singapore
 goela@ihpc.a-star.edu.sg, ma.keng.teck@scei.a-star.edu.sg, cheston-tan@i2r.a-star.edu.sg

Abstract

Socially-intelligent agents are of growing interest in artificial intelligence. To this end, we need systems that can understand social relationships in diverse social contexts. Inferring the social context in a given visual scene not only involves recognizing objects, but also demands a more in-depth understanding of the relationships and attributes of the people involved. To achieve this, one computational approach for representing human relationships and attributes is to use an explicit knowledge graph, which allows for high-level reasoning. We introduce a novel end-to-end-trainable neural network that is capable of generating a Social Relationship Graph – a structured, unified representation of social relationships and attributes – from a given input image. Our Social Relationship Graph Generation Network (SRG-GN) is the first to use memory cells like Gated Recurrent Units (GRUs) to iteratively update the social relationship states in a graph using scene and attribute context. The neural network exploits the recurrent connections among the GRUs to implement message passing between nodes and edges in the graph, and results in significant improvement over previous methods for social relationship recognition.

1. Introduction

The understanding of human relationships in computer vision research is in its nascent stage. In comparison, significant efforts have been made by social psychologists and other researchers to study social relationships in humans [8, 12]. The pioneering work of Sun *et al.* [24] proposes a social relationship framework based on Bugental’s Social Domain Theory [3] to classify social relationships and domains. In this paper, we take a step further in understanding social relationships from images by generating a *Social Relationship Graph (SRG)*, as illustrated in Figure 1.

In recent computer vision research, predicting relationships of the “subject-predicate-object” kind have gained major research attention. These can be used for multiple high-level tasks like image retrieval, image captioning, and

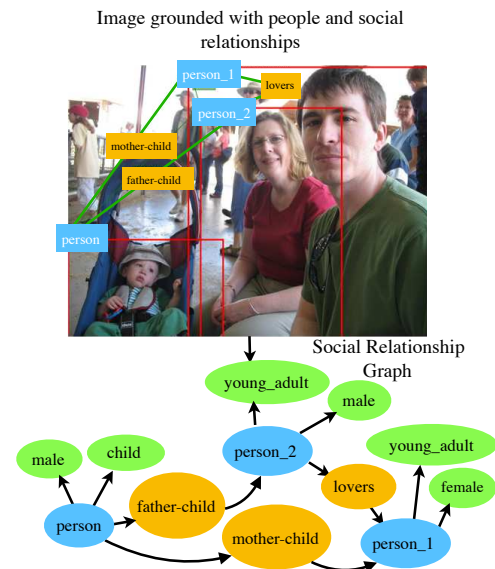


Figure 1: For a given scene, our network generates a structured representation – a Social Relationship Graph. Graph representations have shown good results on a variety of high-level vision tasks, e.g. image retrieval and visual Q&A.

visual question answering [10, 25, 2]. The recent work for the generation of scene graphs using an end-to-end model [28, 13, 29] gives the best results on the Visual Genome Dataset [11]. Since such graphs are human-interpretable, we propose to build a Social Relationship Graph, which encodes relationship and attribute information and captures the rich semantic structure of a scene.

The task of understanding human relationships is a challenging problem given the wide variations that humans pose in their environments. There is unobservable, latent information in images which we as humans find easy to interpret. For developing human-level understanding in such situations, computational models are based on the theories of social and cognitive psychology [23]. Based on the social psychology theories of Bugental [3], we focus on human attributes and environments for social relationships.

Scene and global contextual cues have the best results for social relationships [12]. Furthermore, the activity that peo-

ple are partaking in provides crucial features for social relationship classification [24]. In social psychology research [3], it has been shown that appearance cues such as age, gender and clothing are useful in understanding social relationships. We thus use scene context, activity and appearance features for social relationship graph inference.

We formulate our problem as graph inference that encodes the interactions between nodes and edges in a graph. Our problem is more challenging than scene graph generation [28, 13, 29] as our work requires understanding of high-level semantic features (*e.g. social context*) and low-level visual features (*e.g. spatial arrangement of objects*).

We devise a novel end-to-end model for predicting social relationships using a *Social Relationship Graph Generation Network (SRG-GN)* that combines inputs from a *Multi-Network Convolutional Neural Network (MN-CNN)* to iteratively update the hidden states of the nodes (persons) and edges (relationships) in a *Social Relationship Graph Inference Network (SRG-IN)* by passing messages between two types of Gated Recurrent Units (GRUs) [5].

The *Rship GRUs* (edges) have the scene and activity features as the input, while the *PPair GRUs* (nodes) have the human attribute features as input. The hidden state for each edge gets updated by combining the updated node state and updated edge state. Thus, the relationship (edge) state gets updated by the fine-grained attribute features of the adjacent nodes and the scene and activity context from nearby edges.

The main contributions of this paper are: 1) a novel structured representation (Social Relationship Graph) for social understanding in visual scenes; 2) a novel end-to-end-trainable neural network architecture using GRUs and semantic attributes for graph-generation; 3) new state-of-the-art results for social relationship recognition on the PIPA-relation [24] and PISC [12] datasets. This is the first architecture that builds on social relationships and attributes using memory cells, and our results demonstrate the importance of message passing and scene context.

2. Related Work

2.1. Social Relationship Recognition

The area of social relationships is of growing interest to the community, as social chatbots and personal assistants need to understand social interactions. Many researchers have tried to understand social relationships, roles and interactions. Zhang *et al.* [30] have studied interpersonal relationships using facial expressions with a Siamese-like architecture. There are studies on Kinship recognition [20] and Kinship verification [6]. Wang *et al.* [26] studies family relationships in personal image collections. Jinna *et al.* [16] introduced a video dataset for coarse-grained social relationships between humans. Li *et al.* [12] predicts social relationships in images using an Attentive-RCNN model

for 6-relationship categorization. Ramanathan *et al.* [19] recognize social roles played by people in various events. Chakraborty *et al.* [4] classify photos into classes such as ‘couple, family, group, or crowd’. Sun *et al.* [24] predict social relationships for fine-grained relationships between humans in everyday images. Many of the above-mentioned works have used physical appearance or cues like activity, proximity, emotion, expression, context etc. Our work differs by combining the essential attribute features with memory cells providing a richer framework for our problem.

2.2. Graph-Based Representations

There is a lot of recent interest in using structured graph representations for visual grounding of images. Knowledge graphs are being widely used for object detection and image classification [7, 17]. Johnson *et al.* [10] introduced ground-truth annotated scene graphs for the task of image retrieval using object relationships and attributes. Since then, the task of generating scene graphs directly from images by using intrinsic graph properties and surrounding context has gained attention [28, 13, 29, 9]. The use of vision and language modules together has also been explored by researchers for identifying relationships between objects [15]. Graph Neural Networks [21, 14] have gained a lot of attention for learning structured knowledge in graphs. Wang *et al.* [27] exploit the knowledge of surrounding objects with graph neural networks for updating social relationships. In our work, we present a novel framework for generating graphs, focusing on social relationships and attributes of people involved in a scene, unlike the focus on pair-wise relationships or object relationships in existing work.

3. Model Definition

In this section, we provide an overview of our method for generating Social Relationship Graphs from images using our Social Relationship Graph Generation Network (SRG-GN). The framework in Figure 2 gives a more detailed description of our two modules: A Multi-Network Convolutional Neural Network (MN-CNN) module for Attribute and Relationship representations followed by a Social Relationship Graph Inference Network (SRG-IN) module for generating a structured graph representation. The model is trained end-to-end to predict relationships, domains and attributes as part of a scene in the form of a structured semantic directed graph representation.

3.1. Multi-Network Convolutional Neural Network (MN-CNN) for Relationships and Attributes

We have an input image \mathbf{I} and a set of bounding box annotations B_i for the people in image \mathbf{I} where $i = 1, 2, \dots, N$. These annotations are cropped for a single-body image of a person, I_i and resized into 227×227 pixels. For every annotated relationship between two people, we define a ‘‘context

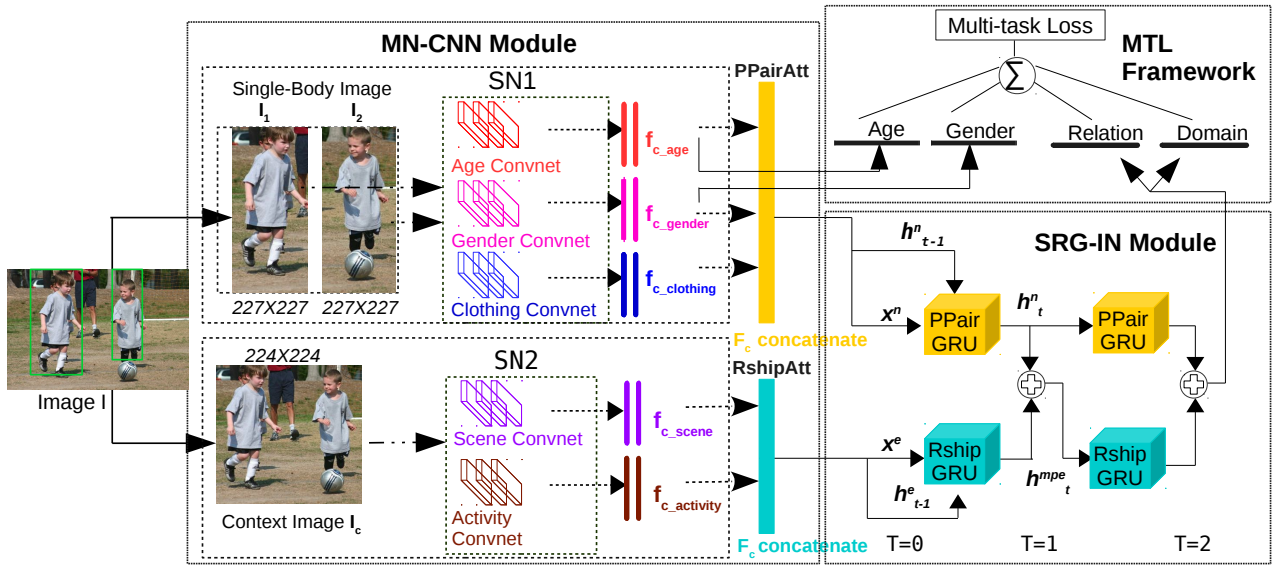


Figure 2: SRG-GN: Our proposed end-to-end network for Social Relationship Graph generation. We take the single body images, I_1 and I_2 , and the “context image” (smallest image that contains both single-body images), I_c as input to the SN1 and SN2 sub-modules of the MN-CNN module and fine-tune the fully-connected layers of all the attributes. These fully-connected layers are concatenated and fed as input to the SRG-IN module and the hidden edge state gets iteratively updated by mean-pooling the edge (relation) and node (person/attribute) hidden states. The final updated edge state is used for predicting social relationships in the given image. For the multi-task learning framework, age and gender attributes from the fully-connected layers of the MN-CNN module also contribute to the joint optimization of the individual cross-entropy losses. The symbol \sum denotes summation and \oplus denotes mean-pooling.

image” (smallest image that contains both single-body images) I_c , resized into 224×224 pixels.

The MN-CNN module has two sub-modules (SN1 and SN2) with the inputs I_i and I_c respectively. I_i is passed through the sub-module, SN1, which is an Attribute ConvNet architecture with 5 conv layers and 2 fully-connected layers (fc6 and fc7), each for the 3 attributes – age, gender and clothing. The weights for these 3 ConvNet layers are the pre-trained weights as discussed later in Section 4.3. We fine-tune the fully-connected layers for each attribute and then the features from the fc7 layers are concatenated into a single feature vector, which we assign to $PPairAtt$.

$$PPairAtt = [f_{c_{age}|4096d}, f_{c_{gender}|4096d}, f_{c_{clothing}|4096d}] \quad (1)$$

The sub-module SN2 is a network of pairwise-relationship ConvNet architectures. There are two VGG-16 architectures [22] to compute activity and scene features from the context images of people. Activity has an important correlation to identifying relationships between people, say, two people “marrying” are more likely to be lovers. Scene context information can also be leveraged for improving the model efficacy to predict relationships. As humans too, we understand images by looking at the whole image scene and not only the objects under consideration. This gives more coarse-grained information to comprehend the given task. We fine-tune the fully-connected layers for both of these

sub-architectures, then concatenate the fc7 layers to form a high-dimensional vector, which we assign to $RshipAtt$.

$$RshipAtt = [f_{c_{activity}|1024d}, f_{c_{scene}|4096d}] \quad (2)$$

3.2. Social Relationship Graph Inference Network (SRG-IN)

We formulate the task of classifying social relationships between people in the form of a social graph inference problem, where we predict the relationships in an image by considering relationship triplets $\langle person1, relation, person2 \rangle$. Consider a pair of people in the given image I with some social relationship between them. In our network, each relationship in an image gets information from its nearby nodes (person attributes) and also its nearby edges (relationships). This is achieved by using Gated Recurrent Units (GRUs) to aggregate messages from the adjacent nodes and relationships and iteratively update those messages to improve the predicted edge states (relationships) between the given nodes (persons). Thus, we are able to exploit the information in the scene context and the individual attributes to improve the relationships in the Social Relationship Graph.

3.2.1 Inference using GRUs and Message Passing Scheme

Mathematically, we formulate our inference task as a probability function: given an input image I , bounding box values

B_i and x as the representation of the SRG:

$$x = \{x_i^{age}, x_i^{gender}, x_{i \rightarrow j}^{relation}, I | i = 1, 2, \dots, N, j = 1, 2, \dots, N\} \quad (3)$$

where x_i^{age} and x_i^{gender} are the age and gender attributes of the person and $x_{i \rightarrow j}^{relation}$ is the social relationship between the persons i and j , and N is the total number of people in an image. We have to find an optimal value of x ,

$$x^* = \operatorname{argmax}_x Pr(x|I, B_i) \quad (4)$$

where,

$$Pr(x|I, B_i) = \prod_{i=1}^N \prod_{j=1}^N Pr(x_i^{age}, x_i^{gender}, x_{i \rightarrow j}^{relation} | I, B_i) \quad (5)$$

We perform this inference using an end-to-end network of Social Relationship Graph Generation where the *MN-CNN module* provides the initial inputs for the nodes and the edges in the *SRG-IN module*.

Gated Recurrent Units (GRUs) are the most reliable and lightweight RNN memory units. The GRUs operate using a reset gate and an update gate and have the ability to keep memory from previous activations allowing them to remember features for a long time. Let us briefly revisit the functioning of a single GRU cell. The reset gate r is defined as

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (6)$$

where σ is the sigmoid function, W_r is the learnable weight matrix, h_{t-1} is the previous hidden state, x_t is the input to the GRU cell and $[\cdot]$ denotes concatenation. The update gate z is given by

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (7)$$

The actual activation in the memory unit is given by

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (8)$$

where,

$$\tilde{h}_t = \tanh(Wx_t + U(r_t * h_{t-1})) \quad (9)$$

W and U are weight matrices that are learned and $*$ is the element-wise multiplication. As empirically evaluated [5], the reset gate r sits between the previous activation and the next candidate activation to forget the previous state, and the update gate z decides how much of the candidate activation to use in updating the cell state.

Our network has two sets of GRUs (Relationship (*Rship*) and Person-Pair (*PPair*)). The initial state of the GRUs can be set to zero or some random vector, and the input to the unit is a sequence of features or symbols. To compute activations from the *PPair* GRU, we take the feature vector, *PPairAtt*, from the SN1 sub-module of the MN-CNN module as the initial state and input to the *PPair* GRU. We concatenate the features from the two nodes (persons) with a

relationship and take this integrated message as input. To compute activations from the *Rship* GRU, we take the feature vector, *RshipAtt*, from the SN2 sub-module of the MN-CNN as the initial state and input to the *Rship* GRU. When the state of the *PPair* GRU is updated, we update the state of the *Rship* GRU by including the node state information into the edge state information to provide context to the edges from its adjacent nodes.

Each of the two GRUs receives incoming messages and we concatenate these messages using a standard pooling operation, mean pooling. Mean pooling aggregates messages in a more meaningful representation as shown in Section 5.2. The *PPair* GRU receives $[f_i, f_j]$ as input, x_n where, f_i and f_j are the attribute features (*PPairAtt*) of the nodes i and j respectively and $[\cdot]$ denotes concatenation. The previous node state h_{t-1}^n is also initialized using $[f_i, f_j]$ and updates the node state to h_t^n using x_n as input. The *Rship* GRU receives $f_{i \rightarrow j}$ as input, x_e where, $f_{i \rightarrow j}$ are the relationship features from the MN-CNN module. The previous edge state h_{t-1}^e is initialized using $f_{i \rightarrow j}$ and the edge state is updated to the "mean-pooled" edge state, h_t^{mpe} , given by:

$$h_t^{mpe} = \frac{h_t^e + h_t^n}{2} \quad (10)$$

This includes the semantic node information into the edge context for updating the edge state with meaningful information from the adjacent nodes and edges. In the next iteration of the GRU, the input to the GRUs are messages from the previous time step. The updated edge representations are used to predict the relationships between nodes.

3.3. Multi-Task Learning (MTL) Framework

In Multi-Task Learning, we simultaneously learn multiple tasks with some shared layers except for one task-specific layer. This can be achieved if the same dataset has multiple labels for learning. For our problem, we have four task labels (age, gender, domain and relationship) that can be learned using the same network. We jointly optimize the loss function by combining the individual loss functions for all these four tasks. We learn the domain labels together with the relationship labels, so that the network can share some relevant information between these two tasks to improve the overall loss function. For instance, the "Reciprocity Domain" refers to relationships that have a reciprocal nature, such as, "friends", "siblings" and "classmates". The output from the *Rship* GRUs are used to predict the domain and relationship labels, whereas the f_{cage} and the $f_{cgender}$ feature vectors from the MN-CNN module are used to predict the age and gender attribute labels respectively using a cross-entropy loss function. We only consider age and gender attribute predictions because the dataset is limited to only these two attributes. Figure 2 shows how we incorporate the MTL framework in our SRG-GN model.

4. Empirical Evaluation

4.1. Dataset Preparation

The PIPA-relation dataset [24] has 16 fine-grained relationship categories¹. We extend their dataset to a PIPA-relation graph dataset. We expand the ground-truth annotations for faces in PIPA into full human body annotations by following the body proportion measurements; 3 x face width and 6 x face height. This gives us ground-truth annotations for single-body images. The context images are cropped from the full images using bounding box values of the people with relationship annotations. We construct our PIPA-relation graph dataset using two attributes (age and gender) from the attribute annotations published on the PIPA dataset [18]. The train/val/test set has 6289 images with 13,672 relationships and 16,145 attributes, 270 images with 706 relationships and 753 attributes, 2649 images with 5075 relationships and 6655 attributes.

We further validate the performance of our model on the large-scale People in Social Context (PISC) dataset released by Li *et al.* [12]. The PISC dataset has 22,670 images where the person pairs are annotated for 3 coarse-grained relationships (*intimate, not-intimate and no relation*) and 6 fine-grained relationships (*commercial, couple, family, friends, professional and no-relation*). The train/val/test set consist of 16,828 images with 55,400 relationship instances, 500 images and 1,505 instances, 1,250 images and 3,961 instances, respectively.

4.2. Baselines

Comparison models for PIPA-relation dataset: Our baselines are the two end-to-end models trained on the PIPA-relation dataset by Sun *et al.* [24] and the end-to-end model for Scene Graph Generation by Xu *et al.* [28] as below:

Double-Stream (DS) CaffeNet: Trained from scratch on the entire dataset using a two stream network for each single body of a person to predict relationships between them.

Finetuned model from pre-trained on Imagenet: Uses fixed weights of the conv layers from the Imagenet pre-trained weights and fine-tuned the fully-connected layers on the PIPA-relation dataset.

Primal-Dual graph model: Trained the primal-dual graph model [28] on the PIPA-relation graph dataset.

Comparison models for PISC dataset: We compare our models with the models proposed by Li *et al.* [12]. An overview of the baseline models by [12] is given below:

Pair-CNN+BBox: Two CNNs for each cropped person image with geometry bounding box features.

¹*father-child, mother-child, grandpa-grandchild, grandma-grandchild, friends, siblings, classmates, lovers/spouses, presenter-audience, teacher-student, trainer-trainee, leader-subordinate, band members, dance team members, sport team members and colleagues*

| MODEL | Accuracy |
|------------------------------------|---------------|
| Double-Stream CaffeNet | 34.40% |
| Primal-Dual model (Our trained) | 44.91% |
| Fine-tuned pre-trained on Imagenet | 46.20% |
| Our MN-CNN module only | 49.75% |
| Our SRG-GN without Scene | 51.79% |
| Our SRG-GN (final model) | 53.56% |

Table 1: Accuracy for the task of Social Relationship Recognition (SRRec on PIPA-relation graph dataset). Chance-level accuracy is 6.25% (1 in 16).

Pair-CNN+BBox+Union: Pair-CNN+BBox with a single CNN for union region of interest features.

Pair-CNN+BBox+Global: Pair-CNN+BBox with the whole image as context.

Pair-CNN+BBox+Scene: Pair-CNN+BBox with scene features as context.

Dual-Glance: Combines Pair-CNN+BBox+Union with attention from contextual information to refine predictions.

4.3. Implementation Details

The pre-trained weights for age, gender, clothing and activity models are publicly available [24]. The pre-trained weights for the Scene ConvNet architecture are from the models published by Zhou *et al.* [31]. We freeze the weights for all the layers and only fine-tune the fully-connected layers of the MN-CNN module, and the GRUs. The output of both the GRUs have a dimension of 512. A softmax layer computes the final scores for age and gender attributes, domains and relationship labels. In case of PISC dataset, we only get scores for domain and relationships as there are no labels for attributes. We sum all the losses and jointly optimize the total weighted loss, as part of the MTL framework. A learning rate of 10^{-6} and 2 time-steps for the GRU are used to train the model. To prevent over-fitting, methods like early-stopping, dropout and regularization are employed. Our model is implemented using Tensorflow [1].

4.4. Results

We evaluate the performance of our model on the PIPA-relation graph dataset and the PISC dataset. The PIPA-relation graph dataset additionally has 6 age labels (*infant, child, young adult, middle age, senior and unknown*) and 2 gender labels (*male and female*).

4.4.1 Quantitative Results

We evaluate our model for two setups:

Social Relationship Recognition (SRRec): To evaluate this, we only consider the triplet predictions of *person-relationship-person* and calculate the accuracy score for social relationship recognition.

Social Relationship Graph Generation (SRGGen): We consider two triplet predictions (*person-relationship-*

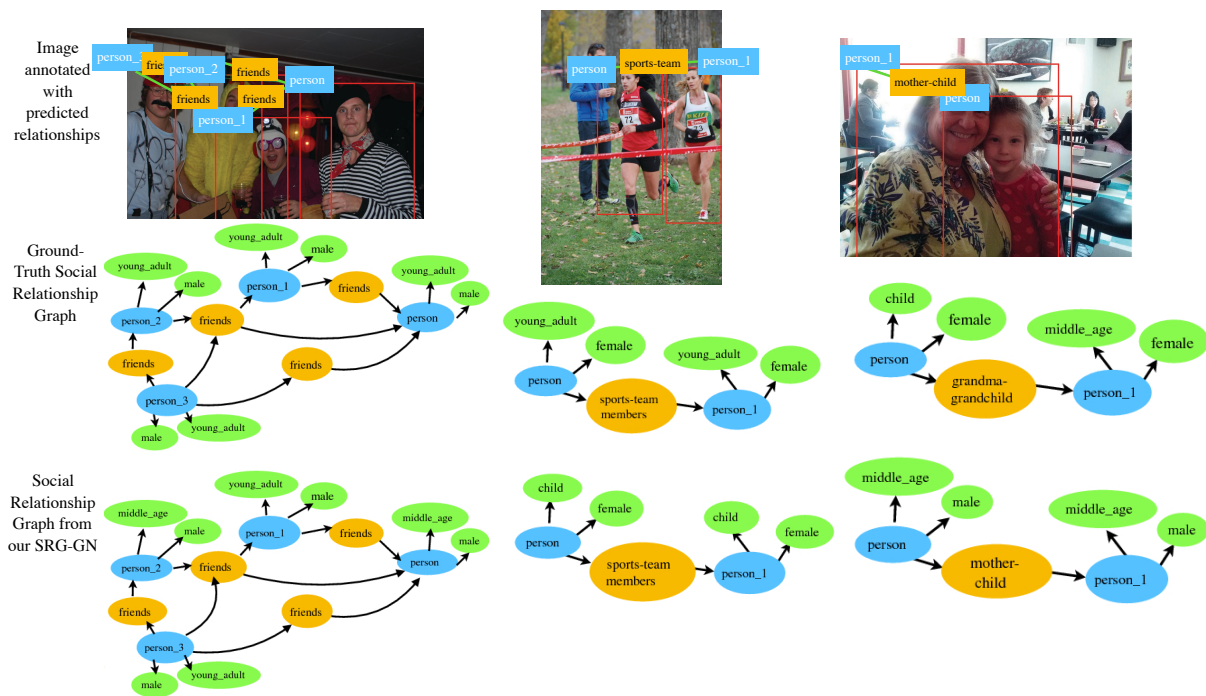


Figure 3: Example Social Relationship Graph generation results from our final model on PIPA-relation graph dataset, and comparison with ground-truth social relationship graphs. Each person (blue ovals) has related age and gender attributes (green ovals) with social relationships between each pair of persons (orange ovals). We only visualize relationships that have the corresponding ground-truth.

person; *person-age-gender*) to measure the accuracy of generating a full SRG with correct age and gender nodes and relationship edges. The directed edges of the graph depends on the direction of the relationship, say, from father to child. For bi-directional relationships like colleagues, we show only one edge for better visualization.

We report results for different variations of our model and compare with the baselines. *Our MN-CNN module only*, is a variation of our model without the GRUs by using concatenated *PPairAtt* and *RshipAtt* as input to the relationship and domain prediction task specific layers and $f_{c_{age}}$ and $f_{c_{gender}}$ to the age and gender prediction task layers respectively. *Our SRG-GN without scene*, is our final model without the scene context features $f_{c_{scene}}$, in the *RshipAtt*. *Our SRG-GN*, is the final model as shown in Figure 2.

Results on PIPA-relation dataset: In Table 1, we provide the accuracy for our first setup, *SRRec*. Our *MN-CNN module* improves on the Fine-tuned model by 3.5% for the task of social relationship recognition. This clearly indicates the importance of using the semantic attributes, scene and activity features over the visual features pre-trained on Imagenet. Our final model, *SRG-GN*, outperforms *only MN-CNN* by 3.81%, which explains the capability of our message passing scheme for generating social relationship graphs. This technique helps to retain significant information from the nearby nodes and edges in a social relationship

graph and thus gives better results. SRG-GN performs better than the primal-dual graph baseline as the latter localizes objects using visual cues with an exchange of information between multiple classes of objects unlike our problem.

Table 3 shows the performance of our model on the second setup of Social Relationship Graph Generation, *SRGen*. We achieve an accuracy of 27.64% using our final model. The accuracy for the *Our SRG-GN without scene* is 7.4% lower than *Our SRG-GN*, which empirically proves that context information plays a major role in generating a coherent social relationship graph.

Results on PISC dataset: Table 4 compares the mean-average precision evaluated on the PISC dataset for Social Relationship Recognition (SRRec). Our final model with mean pooling and 2 time steps notably outperforms the state-of-the-art model on PISC dataset by $\sim 8.5\%$. Our final model improves only slightly in precision over our SRG-GN model without scene. One possible reason is that the scene context in PISC dataset has similar contextual information for relationships unlike in the PIPA-relation dataset.

We report the precision of each of the 6 relationship labels in Table 2. Our SRG-GN model improves in precision over the MN-CNN-only model for the classes *couple* and *commercial*. The class *friends* has lower precision, indicating that other classes are sometimes wrongly classified as “friends”. Due to imbalance in the training dataset, we in-

| MODEL | mAP | Family | Couple | Commercial | No-Relation | Professional | Friends |
|--------------------------|-------------|-------------|--------------|-------------|-------------|--------------|-------------|
| Our MN-CNN module only | 60.2 | 75.0 | 57.1 | 62.5 | 59.9 | 80.6 | 26.0 |
| Our SRG-GN without Scene | 69.2 | 80.0 | 77.7 | 88.8 | 61.7 | 81.8 | 24.5 |
| Our SRG-GN (final model) | 71.6 | 80.0 | 100.0 | 83.3 | 62.5 | 78.4 | 25.2 |

Table 2: Detection results for 6-relationship labels on PISC dataset.

| MODEL | Accuracy |
|--------------------------|---------------|
| Our SRG-GN without Scene | 20.24% |
| Our SRG-GN (final model) | 27.64% |

Table 3: Accuracy for the task of Social Relationship Graph Generation (*SRGGen*) on PIPA-relation graph dataset. Chance-level accuracy is 0.52% = (1/16 * 1/6 * 1/2)



Figure 4: Wrong relationship predictions from the SRG-GN model on the PISC dataset. The relationships in yellow are the ground-truth, the relationships in red are the incorrect predictions. Only the relationships marked as red in an image are incorrectly predicted by our model.

Introduce a weighted cross entropy loss to penalize the classes with few samples; this improves performance significantly.

4.4.2 Qualitative Results

The Social Relationship graph (SRG) is a rich semantic graph with attribute and relationship information for the people in a given scene. Our SRG contains ground-truth information about the class and bounding-box labels of the objects in the image. Through our SRG-GN, we predict the social relationships, age and gender attributes of the people in a given scene.

Figure 3 shows qualitative results on PIPA-relation graph dataset to compare the SRG generated from our model and the ground truth. In the first example, the SRG-GN correctly predicts the relationships between the given people. As shown in the graph, all nodes (persons) have “friends” relationship between them which are correctly predicted

| MODEL | mAP |
|--------------------------|--------------|
| Pair-CNN+BBox | 54.3% |
| Pair-CNN+BBox+Union | 56.9% |
| Pair-CNN+BBox+Global | 54.6% |
| Pair-CNN+BBox+Scene | 51.7% |
| Dual-Glance | 63.2% |
| Our MN-CNN module only | 60.2% |
| Our SRG-GN without Scene | 69.2% |
| Our SRG-GN (final model) | 71.6% |

Table 4: Mean-Average Precision (mAP) for the task of Social Relationship Recognition (*SRRec*) on PISC dataset.

by our model. Gender attributes also correspond to the ground-truth, but the age attributes are incorrectly predicted as “middle-age” instead of “young-adult”. The model correctly predicts more complex relationships like “sports-team members” which has a lot more contextual information than other relationships like “grandma-grandchild” which it falsely predicts as “mother-child” due to ambiguity in such relationships.

Figure 5 gives examples of the correct predictions on PISC dataset. Our model predicts multiple relationship instances in an image, such as a group of players are correctly labeled as “professional”. Figure 4 shows examples for misclassified relationships. For instance, the model falsely detects the relationship in bottom-left image as “family”, when they are more likely to be friends due to information from adjacent nodes and edges. There is ambiguity between “professional” and “commercial” in some cases due to similar global and scene context for these classes.

5. Ablative Analysis

In this section, we examine the performance of our *SRG-GN* model variations on the PIPA-Relation graph dataset.

5.1. Model Variations

We evaluate the importance of scene context in predicting relationships in our final graph inference framework. As shown in Section 4.4, adding scene context significantly improves the performance on both tasks of *SRRec* and *SRGGen*. Intuitively, we can infer that scene information can be important in many different situations. For instance, given a party scene, the group of people are more likely to be friends than colleagues, and a group of athletes running on a track are much more likely to be sports team members than band members. In Figure 6(a), we present



Figure 5: Correct predictions from our final model on the PISC dataset.

(a). SRG results from the SRG-GN model and SRG model without scene



(b). SRG results from the SRG-GN model and only MN-CNN model

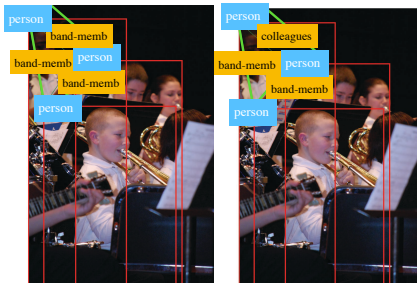


Figure 6: Qualitative analysis of our model variations on PIPA-Relation. The left results are from our final model, *SRG-GN*. The top-right result is from *SRG-GN without Scene*, while the bottom-right result is from the *only MN-CNN model*.

an example to highlight the importance of using whole image scene context for accurate predictions. Our *SRG-GN without scene* incorrectly predicts the two people as sports team members, but if we look at the whole scene together it increases the chances of them being colleagues and not related to sports. Without scene context, identifying the relationships between two people can be sometimes ambiguous. This clearly explains the motivation behind using scene context as an important feature in the *SRG-IN* module.

We also examine how predicting relationships in isolation from the *only MN-CNN* module has lower accuracy than the combined model with the *SRG-IN* module. For example, a group of people performing on the stage should all very likely be band members, and our model exploits this information for overall inference, whereas the *only MN-CNN* module predicts the triplets in the social relationship graph independently. In Figure 6(b), our final model correctly predicts the relationships as band-members due to the message information from the adjacent group of relationships in an image. Without this message passing network, the *MN-CNN* module only considers information from the pair of people between whom relationship has to be pre-

| Pooling | # time steps | Accuracy |
|-------------|--------------|---------------|
| max | 1 | 50.41% |
| max | 2 | 52.16% |
| max | 3 | 51.27% |
| mean | 1 | 50.89% |
| mean | 2 | 53.56% |
| mean | 3 | 52.08% |

Table 5: Ablation study for different time-steps and pooling techniques on the PIPA-relation graph dataset.

dicted. Thus, the *SRG-IN* module uses contextual information from the nearby nodes and edges in a graph to improve individual predictions.

5.2. Pooling and Time-Step variations

We evaluate our *SRG-GN* model on the PIPA-relation with different number of time steps and pooling techniques. From Table 5, it can be observed that mean-pooling is more effective in passing useful information between hidden states than max-pooling. Also, there is a $\sim 1.5\%$ decrease in accuracy on increasing the time steps as it starts passing noisy information between states with more false detections in the social relationship graph.

6. Conclusion

We introduced a novel end-to-end-trainable network for generating social relationship graphs from images using GRUs. Previous work on generating graphs dealt with relationships between objects, whereas our work tackles the more challenging problem of inferring social relationships. Experimental results show the importance of using attribute and contextual features with message passing in a graph. Our model outperforms the state-of-the-art for recognizing social relationships, and performs well for generating social relationship graphs. This work can be extended for more complex tasks, such as predicting social intentions.

Acknowledgements

This work was supported by NRF grant no. NRF2015-NRF-ISF001-2541 (KTM and CT) and A*STAR SERC SSF grant no. A1718g0048 (AG and KTM).

References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016.
- [3] Daphne Blunt Bugental. Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, 126(2):187–219, 2000.
- [4] Ishani Chakraborty, Hui Cheng, and Omar Javed. 3d visual proxemics: Recognizing human interactions in 3d from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3406–3413, 2013.
- [5] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
- [6] Ruogu Fang, Kevin D Tang, Noah Snavely, and Tsuhan Chen. Towards computational models of kinship verification. In *17th IEEE International Conference on Image Processing (ICIP), 2010*, pages 1577–1580. IEEE, 2010.
- [7] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. Object detection meets knowledge graphs. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1661–1667. AAAI Press, 2017.
- [8] Chris Frith. Role of facial expressions in social interactions. *Philosophical transactions of the royal society of London B: Biological sciences*, 364(1535):3453–3458, 2009.
- [9] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7211–7221, 2018.
- [10] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, June 2015.
- [11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [12] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2650–2659, 2017.
- [13] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1270–1279, 2017.
- [14] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. Gated graph sequence neural networks. *International Conference on Learning Representations (ICLR)*, 2016.
- [15] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.
- [16] Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. Multi-stream fusion model for social relation recognition from videos. In Klaus Schoeffmann, Thanarat H. Chalidabhongse, Chong Wah Ngo, Supavadee Aramvith, Noel E. O’Connor, Yo-Sung Ho, Moncef Gabbouj, and Ahmed El-gammal, editors, *MultiMedia Modeling*, pages 355–368, Cham, 2018. Springer International Publishing.
- [17] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20–28. IEEE, 2017.
- [18] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Person recognition in personal photo collections. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3862–3870, 2015.
- [19] Vignesh Ramanathan, Bangpeng Yao, and Li Fei-Fei. Social role discovery in human events. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2475–2482, 2013.
- [20] Joseph Peter Robinson, Ming Shao, Yue Wu, Hongfu Liu, Timothy Gillis, and Yun Fu. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2624–2637, 2018.
- [21] 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, 2009.
- [22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.
- [23] Eliot R Smith and Jamie DeCoster. Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and social psychology review*, 4(2):108–131, 2000.
- [24] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 435–444, 2017.
- [25] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3233–3241. IEEE, 2017.
- [26] Gang Wang, Andrew Gallagher, Jiebo Luo, and David Forsyth. Seeing people in social context: Recognizing people and social relationships. In *European conference on computer vision*, pages 169–182. Springer, 2010.
- [27] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge

- graph for social relationship understanding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1021–1028. AAAI Press, 2018.
- [28] Danfei Xu, Yuke Zhu, Christopher Bongo Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. pages 3097–3106, 2017.
- [29] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.
- [30] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639, 2015.
- [31] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.