Multi-attribute Graph Inference for Social Relationship Recognition

Eduardo V. Sousa and Douglas G. Macharet
VeRLab – Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Emails: {eduardo.vieira, doug}@dcc.ufmg.br

***Abstract*—In many real-world scenarios, people in the environment are interacting with each other, for example, engaged in a conversation, and consequently forming relationship clusters. The understanding of such social relationships is a very useful information for different types of intelligent systems, like an autonomous robot or video surveillance. In this work, we propose a novel Multi-attribute Graph Inference (MAGI) framework built upon a set of pre-trained neural networks in charge of extracting semantic attributes, which are aggregated using a graph model. The set of aggregated features is used for inferring the social relation and a human interpretable attribute graph is generated, indicating how much of each feature was used in the process. The methodology was evaluated in well-known datasets, surpassing the state-of-the-art results for social relation recognition.**

I. Introduction

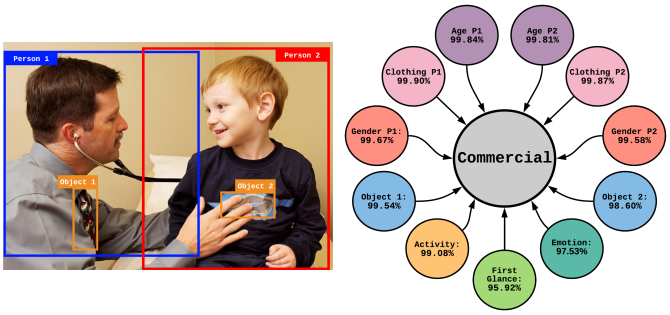
The increasing use of autonomous systems in different parts of society makes it important to propose techniques that will allow them to behave properly in such scenarios. In this sense, a topic that has gained much attention recently is related to human analysis and behavior understanding.

A key factor to better comprehend human behavior is being capable of recognizing the characteristics that embed most common social relationships, which can be defined as the connections between people with recurring meaningful interactions [1]. Although this is an important topic for automated human analysis it remains little explored, and recent works focus on identifying social relation [2]–[4] and relationship traits [5], [6] from visual data.

Social psychology research has shown that humans recognize social relation with the help of appearance cues such as age, gender, and clothing [7], however, this is not an easy task. Besides being affected by usual computer vision issues such as scale, appearance and pose variations [3], the area has its own challenges, like the ambiguity present in some social situations where similar arrangements may represent different types of relationships accordingly to contextual cues, for example, the presence of office or household items [2].

These challenges are depicted in Figure 1, that contains an example of a relationship that could be misinterpreted. Therefore, to infer social relation is necessary not only physical appearance attributes (e.g., age, gender, clothing) but also higher-level information (e.g., action, emotion, context) [3], [6]. Some approaches proposed to solve this problem are based on appearance features extracted from the persons involved [4] and others try to incorporate contextual information by analyzing nearby objects [3]. However, they lack a way to integrate personal attributes with contextual information.

This work was supportted CAPES - Finance Code 001, FAPEMIG, and CNPq. We also thank NVIDIA for the donation of a GeForce Titan XP GPU.



(a) Social relationship. (b) Attribute graph.

Fig. 1: Example of a correctly classified commercial relation between a doctor and a child, and the attribute graph.

In this work, we propose a Multi-attribute Graph Inference (MAGI) framework to evaluate and classify social relationships from visual data. The framework employs pre-trained neural networks to extract semantic attributes, chosen based on [4], which are aggregated using a graph propagation model and refined with an attention mechanism. The network learns representations capable of capturing the interactions between these attributes, which are used for inferring the social relation. The set of aggregated features are used to generate a human interpretable attribute graph, indicating how much of each feature was used in the process (Fig. 1b).

In summary, the main contributions of this paper are: (i) A novel neural network architecture capable of handling individual semantic attributes and context information; (ii) the generation of human-readable attribute graphs that help understanding the decision process; and (iii) new state-of-the-art results for both PISC [3] and PIPA-relation [4] datasets.

The remainder of this paper is organized as follows: Section II presents a literature review regarding social relationship understanding. The proposed methodology is detailed in Section III, and evaluated against state-of-the-art approaches, results of which are shown and discussed in Section IV. Finally, Section V concludes with future research directions.

II. Related Work

A. Social Relation Recognition

Achieving human-level performance on social relation recognition is a difficult task, especially because of the subjective aspect intrinsic to every human related problem. For this reason, recent works [3], [4] employs social psychology theories, which provides knowledge about aspects of social relationships in a objective and general manner.

One of the first methods proposed to classify social relationships [4] made use of face and body images to extract 12 semantic attributes and measured their impact on the recognition rates. The model was backed by Bugental's domain-based theory [7], dividing social life in 5 domains, which are used to derive 16 social relationships. Later, [3] developed a method based on the relational theory [8], defining hierarchical social relationships from coarse-to-fine levels. The model addresses environment information in the form of a context picture involving both persons and also locating objects from the surroundings. Finally, the most recent researches employed graph models [2], [9] to take advantage of prior knowledge, and also to represent complex relationships between objects and individuals of interest.

Considering social relation traits, [6] presented an architecture capable of extracting facial features such as gender, expression, pose and age-related cues. The network performs pairwise reasoning, identifying high-level traits such as friendliness, warm, and dominance. In [10] it is applied a Siamese network to extract features from pairs of face images, and later combined them with a semantic augmentation structure.

Recently, [11] employed a hybrid deep neural network pre-trained for face recognition, which also incorporates scene high-level features. A similar approach was proposed by [10], applying a Siamese network to extract features from pairs of face images, and later combining them with a semantic augmentation structure.

B. Graph Neural Networks

The first works to combine deep learning and graph structures were [12], [13], and later other concepts like convolutions were extended to graph networks [14] in the form of multi-layer CNNs for node classification. Also, other graph network variants has emerged like the Graph LSTM [15] and the GGNN [16] which was adapted with GRU cells to propagate massages through the graph.

The first works to combine deep learning and graph structures were [12], [13], and later concepts like convolutions were extended to graph networks [14] in the form of multi-layer CNNs for node classification. In [17] visual and textual information graphs are build to perform image-text retrieval, scene graph generation is also another common application [18] where graphs can be used to represent entities and their relationships. For group action recognition, [19] generates a graph encoding appearance and positioning data. Finally, graphs can also be applied to feature learning, as done by [20] for relative attribute learning and [21] for group emotion and event recognition.

III. Methodology

We propose a Multi-attribute Graph Inference (MAGI) framework inspired by [2], [4], [21], composed of three modules (Fig. 2). The first module takes as input an image and both person bounding boxes, then it detects objects and extracts multiple traits from the individuals, context, and objects. The second one takes these features and resizes them to the same dimensions, so they can be fed to a Graph Neural Network that employs a gated recurrent unit to aggregate the information contained in them using a graph structure. Finally, the last module applies an attention mechanism to reinforce the most distinctive features forwarding them to the classifier, which generates individual scores and applies different fusion techniques to get the final scores and output a prediction.

A. Multiple Attribute Extraction Module (MAE)

Given an input image $\mathbf{I}$ containing a relation $\mathbf{R}$ between two individuals, and the coordinates of their bounding boxes, p_1 and p_2. First, we crop the boxes regions obtaining $\mathbf{I}_{p1}$ and $\mathbf{I}_{p2}$, along with a context sub-image $\mathbf{I}_c$, defined by the smallest region from $\mathbf{I}$ containing both boxes.

The next step consists in detecting from $\mathbf{I}$ a set of object proposal regions $\mathbf{O} = \{o_1, o_2, ..., o_{\mathbf{K}}\}$, where $\mathbf{K}$ is the number of different objects classes within the image. This is done using a Faster-RCNN [22] with a high confidence score (e.g., 0.7) to avoid false positive object detections. The regions patches $\mathbf{O}$ are finally extracted resulting in a set $\mathbf{I}_o$ of object images and therefore for each relation we get $\mathbf{R} = \{\mathbf{I}_{p1}, \mathbf{I}_{p2}, \mathbf{I}_c, \mathbf{I}_o\}$.

Then, we resize $\{\mathbf{I}_{p1}, \mathbf{I}_{p2}\}$ to 227×227 pixels and $\{\mathbf{I}_c, \mathbf{I}_o\}$ to 224×224 pixels, and extract a number $\mathbf{N}$ of different traits from the relation set $\mathbf{R}$. Let $f_{(i,j)}$ be the feature vector where $i = 1, 2, ..., \mathbf{N}$ represents the type of the attribute and j is the total number of vectors with type i that can be extracted from $\mathbf{R}$. The features are combined into a set $\mathbf{R}_f = \{f_{(1,1)}, f_{(1,2)}, ..., f_{(i,j)}\}$ for the relation.

In this work, we use a maximum value of $\mathbf{N} = 6$, but any other amount of attributes would fit in the model. More precisely, for each person image patch $\mathbf{I}_{bi}$ we build a set $\mathbf{F}_{bi} = \{f_{age}, f_{clothing}, f_{gender}\} \in \mathbb{R}^{3\times4096}$ of individual features using the pre-trained models provided by [4].

From $\{\mathbf{I}_{p1}, \mathbf{I}_{p2}\}$ and $\mathbf{I}_c$, we extract features using the First Glance model, which we implemented as described in [3], generating $f_{firstglance} \in \mathbb{R}^{4096}$. Also from $\mathbf{I}_c$, we extract group emotion features $f_{emotion} \in \mathbb{R}^{1024}$, obtained with the Inception-V2 model provided by [21]. The network was pre-trained on the ImageNet1K and later fine-tuned for their GroupEmoW database, focused on group emotion recognition.

Still from $\mathbf{I}_c$, we employ the CNN-CRF model [23], trained to recognize 504 activity classes on a dateset composed with 126,102 images, to extract $f_{activity} \in \mathbb{R}^{1024}$. With these three feature vectors we form $\mathbf{F}_c = \{f_{firstglance}, f_{emotion}, f_{activity}\}$. Finally, from $\mathbf{I}_o$ we use the SENet-154 model, pre-trained on the ImageNet-1K database, to build $\mathbf{F}_o \in \mathbb{R}^{\mathbf{K}\times2048}$, obtaining the relation feature set $\mathbf{R}_f = \{\mathbf{F}_{b1}, \mathbf{F}_{b2}, \mathbf{F}_c, \mathbf{F}_o\}$.

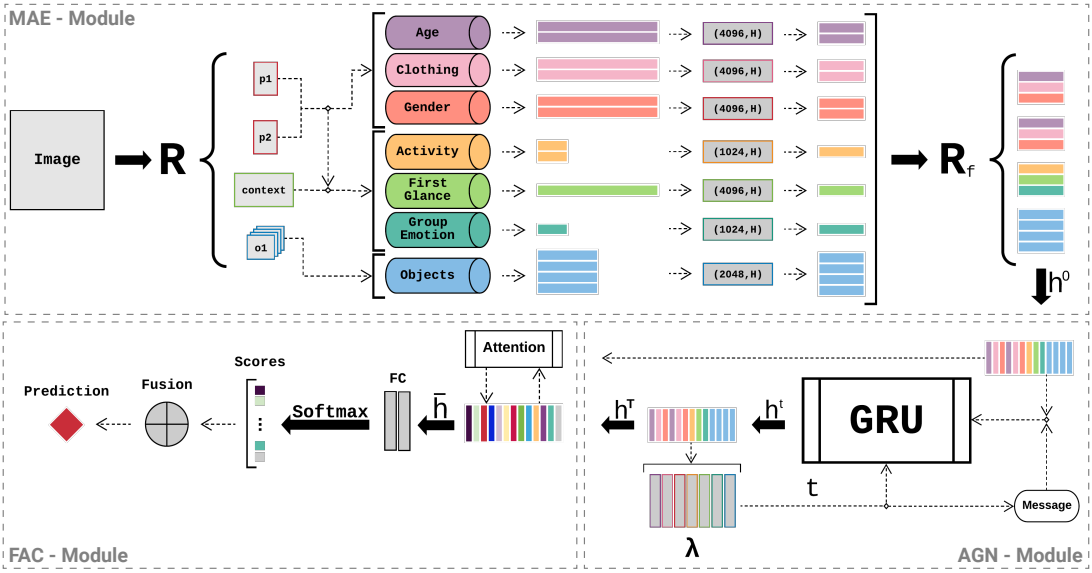


Fig. 2: Overview of the Multi-attribute Graph Inference framework. The MAE module receives the relation image and extracts multiple attributes with pre-trained models. The AGN is composed of a gated graph neural network that aggregates features and forward them to the FAC module, which applies an attention mechanism and fuses nodes predictions obtaining a classification.

Since each attribute type has different sizes, we use a FC layer to resize them to the hidden state dimension $\mathbf{H}$ via

$$\bar{f}_{(i,j)} = \mathbf{ReLU}(W_i f_{(i,j)} + b_i), \tag{1}$$

where W_i and b_i are the learned matrix of weights and bias term shared among inputs of type i. The output vectors $\bar{f}_{(i,j)} \in \mathbb{R}^{\mathbf{H}}$ form a set $\mathbf{R}_{\bar{f}}$ of features with the same size, which are used to represent the initial hidden state h_0 of the nodes on our AGN module (Sec. III-B). We evaluate different hidden state dimensions in Section IV.

B. Attribute Graph Network Module (AGN)

Here we aggregate the features received from the MAE module (Sec. III-A) with a process called message passing. At the end of $\mathbf{T}$ time steps through this process, the final hidden states obtained are passed to the next module that classifies the input relation $\mathbf{R}$. We apply a similar mechanism to the one presented by [16] for the Gated Graph Neural Network (GGNN) to propagate node massages at each iteration through the graph, learning node-level representations.

The GGNN consists of an adaptation from traditional graph networks with fully differential recurrent networks, generating a model that is suitable for non-sequential outputs. The chosen recurrent models are the Gated Recurrent Units (GRUs), and the main differences are a fixed number $\mathbf{T}$ of unrolling steps on the recurrence and the use of backpropagation through time to compute gradients. The propagation model is defined as

$$z^t_{(i,j)} = \sigma(W_z m^t_{(i,j)} + U_z h^{t-1}_{(i,j)}), \tag{2}$$
$$r^t_{(i,j)} = \sigma(W_r m^t_{(i,j)} + U_r h^{t-1}_{(i,j)}), \tag{3}$$
$$\bar{h}^t_{(i,j)} = \tanh(W_h m^t_{(i,j)} + U_h(r^t_{(i,j)} \odot h^{t-1}_{(i,j)})), \tag{4}$$
$$h^t_{(i,j)} = (1 - z^t_{(i,j)}) \odot h^{t-1}_{(i,j)} + z^t_{(i,j)} \odot \bar{h}^t_{(i,j)}, \tag{5}$$

where $z^t_{(i,j)}$ and $r^t_{(i,j)}$ are the update and reset gates, respectively. They control how much of the previous state content will be used for the computation of $h^t_{(i,j)}$, which is the hidden state at the time step t and $\bar{h}^t_{(i,j)}$ is the updated candidate. The trainable parameters of this model are W_z, U_z, W_r, U_r, W_h and U_h, while σ denotes the logistic sigmoid function and $\odot$ represents element-wise multiplication.

In this sense, the AGN plays the role of the propagation model, where the first step consists in generating an undirected graph where each resized feature $\bar{f}_{(i,j)} \in \mathbf{R}_f$ represents the node's initial hidden state $h^0_{i,j}$, and all distinct nodes are connected, forming a complete graph. Since the number of object features varies depending on the image, we might have different graph morphologies for each picture.

Next, we generate $m^t_{(i,j)} \in \mathbb{R}^{\mathbf{H}}$ for each node, formed by the aggregation of the hidden states from the neighbor nodes at the previous time step $t-1$. We employ a two-layer MLP along the graph's edges, denoted by $\boldsymbol{\lambda}(\cdot)$, within the aggregation function, so each node can get information from its neighbors, propagating the initial features deeper into the graph at each time step. The message and $\boldsymbol{\lambda}$ are denoted by

$$m^t_{(i,j)} = \frac{1}{\mathbf{G}-1} \sum_{\substack{(q,p) \\ (q,p)\neq(i,j)}} \boldsymbol{\lambda}(h^{t-1}_{(q,p)}), \tag{6}$$

$$\boldsymbol{\lambda}(h^t_{(q,p)}) = \tanh(W_q \mathbf{ReLU}(U_q h^t_{(q,p)})), \tag{7}$$

where $\mathbf{G} > 1$ is the total number of nodes, W_q and U_q are learned matrices of weights associated with cue type q and the bias terms of each layer are omitted for notation simplicity. These parameters control the information flow between distinct attributes, differently from other works that generate a fixed adjacency matrix for each dataset based on class co-occurrences.

This way, our method can learn the best use for the attributes from each neighbor node, also making it possible to insert a variable number of features for the same attribute type q, since the model is not constrained by a fixed adjacency matrix like other approaches. After $\mathbf{T}$ iterations, the model have learned complex representations for the relationships between input attributes, and it forwards them to the final module (Sec. III-C).

C. Feature Attention and Classification Module (FAC)

The last module acts as the output model, while also applying a feature attention mechanism inspired by [2], [3], whose function is to control how much of each feature generated by the AGN is going the be used for the final classification.

Given the initial hidden state $h^0_{(i,j)}$ and the final state $h^{\mathbf{T}}_{(i,j)}$ for each feature $\bar{f}_{(i,j)}$, we combine them by applying a low-rank bilinear pooling [24], generating $d_{(i,j)}$ as follows

$$d_{(i,j)} = \tanh(U_i h^0_{(i,j)}) \odot \tanh(V_i h^{\mathbf{T}}_{(i,j)}), \tag{8}$$

where U_i, V_i are learned weight matrices shared by the features of the same type i, and bias terms are omitted. Next we apply a function to compute the attention coefficient $a_{(i,j)}$ for each hidden feature $h^{\mathbf{T}}_{(i,j)}$. This can be implemented by an FC layer, with a sigmoid activation function to generate values in the range $(0, 1)$ by

$$a_{(i,j)} = \sigma(W_s d_{(i,j)} + b_s), \tag{9}$$

where W_s and b_s are the matrix of learned weights and bias term. Finally, we compute the weighted feature vector by multiplying the attention coefficients $a_{(i,j)}$ for each hidden feature vector $h^{\mathbf{T}}_{(i,j)}$, generating the weighted hidden features

$$\bar{h}_{(i,j)} = a_{(i,j)} h^{\mathbf{T}}_{(i,j)}. \tag{10}$$

The output layers receives the weighted feature vector $\bar{h}_{(i,j)}$, and uses it to generate class scores $s_{(i,j)}$ for the input relation $\mathbf{R}$. This is done with a two-layer MLP followed by the $\mathbf{softmax}$ function

$$s_{(i,j)} = \mathbf{softmax}(W \mathbf{ReLU}(U \bar{h}_{(i,j)})), \tag{11}$$

where W and U are weights matrices shared across all nodes and bias terms are omitted for notation simplicity.

Since our model generates a set $\mathbf{Z} = \{z_i \mid i = 1, 2, ...\mathbf{G}\}$ of score vectors for each relation $\mathbf{R}$, we consider 3 functions to fuse them into a final prediction [3]: $\mathbf{max}(\mathbf{Z})$, $\mathbf{mean}(\mathbf{Z})$, and the log-sum-exp denoted by $\mathbf{lse}(\mathbf{Z})$ and computed as

$$\mathbf{lse}(\mathbf{Z}) = \log\left(1 + \sum_{i=1}^{\mathbf{G}} \exp(z_i)\right). \tag{12}$$

The main difference here is that we fuse the scores after the $\mathbf{softmax}$ function, where each feature score s_i contributes to the final prediction. Each method generates different predictions, and we discuss the results in Section IV.

IV. Experiments

A. Datasets

We evaluated our approach on two publicly available datasets: (i) **People in Social Context (PISC)** [3]: This dataset consists of 22,670 images and 76,568 manually annotated labels from 9 types of social relationship. Considers two recognition tasks: (a) 3-relationship (coarse-level) recognition; (b) 6-relationship (fine-level) recognition. , and (ii) **People in Photo Album (PIPA-Relation)** [4], [25]: The original PIPA [25] dataset is composed by 37,107 photos with 63,188 instances of 2,356 identities. This extension [4] inserts 26,915 person pair annotations. Divides social life into 5 domains and considers 16 social relations based on these domains.

B. State-of-the-art baselines

The proposed methodology was compared against the following competing methods on the PISC dataset: **Dual-Glance** [3], **Graph Reasoning Model** [2], and **Multi-Granularity Reasoning** [9]. Following [3], we also report the results for **Union-CNN**, **Pair-CNN**, **Pair-CNN+BBox+Union**, and **Pair-CNN+BBox+Global**.

Considering the PIPA-Relation dataset, we selected to evaluate the approach against the **Dual-Glance** [3], **Graph Reasoning Model** [2], and **Multi-Granularity Reasoning** [9] since they are the best performing ones on the PISC dataset. We also show the results of the **Double-stream CNN** [4], the proposer of the extended dataset.

C. Implementation details

During the process of obtaining attribute features, we freeze the parameters of all employed models and extract the output from the following layers. For f_{age}, $f_{clothing}$, f_{gender} and $f_{activity}$ we extract from the *'fc7'* layer, for $f_{emotion}$ we use the *'global_pool'* layer and object features are obtained from the *'pool5/7x7_s1'* layer. Finally, for $f_{firstglance}$ we extract the output from the penultimate fully connected layer.

The loss function for our model is the weighted cross entropy and we use AdamW [26] as optimizer, with a 10^{-4} learning rate and 10^{-5} weight decay. We also employ dropout with 0.3 probability on the final FC layers.

D. Quantitative results

We compute per-class recall and mean average precision (mAP) metrics on the PISC dataset and accuracy for the PIPA dataset. Tables I and II show that our model was capable of extracting the high-level information necessary to achieve a new state-of-the-art result on both datasets.

On the **PISC** dataset, our method obtained a mAP of 71.8% for the fine-grained data and 84.1% for the coarse data. When compared to the GRM, our model achieves a better performance improvement on the 6-relationship split. This may happen not only because of the higher room for improvement on the fine set, but also because our method can scale with the complexity of the dataset by attaching new attributes and aggregating them into higher-level features.

TABLE I: Comparison of the recall-per-class and mean average precision (mAP) of the proposed methodology against the state-of-the-art on the PISC dataset [2], [3], [9].

Methods	3-relationship (coarse)				6-relationship (fine)						
	Intimate	Non-Intimate	No Rel.	mAP	Friends	Family	Couple	Professional	Commercial	No Rel.	mAP
Union-CNN [3]	72.1	81.8	19.2	58.4	29.9	58.5	70.7	55.4	43.0	19.6	43.5
Pair-CNN [3]	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 [3]	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 [3]	70.5	80.0	53.7	70.5	32.2	61.7	72.6	60.8	44.3	51.0	54.6
Dual-Glance [3]	73.1	**84.2**	59.6	79.7	35.4	**68.1**	**76.3**	70.3	**57.6**	60.9	63.2
Graph Reasoning Model [2]	**81.7**	73.4	65.5	82.8	59.6	64.4	58.6	76.6	39.5	67.7	68.7
Multi-Granularity Reasoning [9]	-	-	-	-	64.6	67.8	60.5	**76.8**	34.7	70.4	70.0
Ours	81.4	65.9	**79.7**	**84.1**	**71.5**	61.0	49.6	74.5	37.0	**75.0**	**71.8**

TABLE II: Comparison of the accuracy of the proposed methodology against the state-of-the-art on the PIPA-Relation dataset [2], [3], [9].

Methods	Accuracy (%)
Double-stream CNN [4]	57.2
Dual-Glance [3]	59.6
Graph Reasoning Model [2]	62.3
Multi-Granularity Reasoning [9]	64.4
Ours	**66.9**

The methodology also provides an improvement over the previous models on the **PIPA** dataset, which is formed by 16 classes. This is a very relevant result once [4] combines a set of 12 attributes on their model, which 4 of them are the same as the ones we use, but our model performs significantly better. These numbers suggest the effectiveness of our main concept of learning representations for the interactions between these attributes, generating better features.

E. Model variations

Here we present the results of multiple variations on our model on both splits of the PISC dataset. Initially, we evaluate different hidden state dimensions $\mathbf{H}$ for the AGN module. Next, we measure the effects of the number $\mathbf{T}$ of time steps and different fusion methods on the final prediction scores. Finally, we analyze the contribution of each feature type.

TABLE III: Comparison of mAP (%) for different hidden state dimensions ($\mathbf{H}$) and time steps ($\mathbf{T}$) on the **PISC** dataset.

Level	H				T				
	128	256	512	1024	0	1	2	3	4
Coarse	**84.07**	83.72	83.68	83.62	82.40	**84.07**	83.64	83.58	83.52
Fine	69.86	**71.81**	70.86	70.75	66.59	70.31	71.14	**71.81**	70.85

1) Hidden state: The hidden state dimension relates to the capacity of our model. A small $\mathbf{H}$ value may not be able to represent the output features in a distinctive way and a size too big can slow down the training and increases the chance of overfitting. Table III shows that for 6-relationship, the best $\mathbf{H}$ was 256, and increasing or decreasing from that point degrades the model's performance. For the 3-relationship set, the model dealt better with a smaller size, and this difference between both values is probably related to the task complexity.

2) Time step: The value of $\mathbf{T}$ defines the number of iterations in the message passing process for each forward pass. This process controls how much individual node features are being combined before they can be sent to the final module. A time step of 0 means that the input features are not aggregated at all, they are simply forwarded to the FAC module and directly classified. For the 6-relationship, we found that the best number for $\mathbf{T}$ is 3 (Table III). Lower values lead to a performance decrease by not aggregating features enough and higher values start to generate noise. For the coarse split, a smaller value had a better outcome, with the mAP decreasing for values of $\mathbf{T}$ beyond 1, possibly because coarse data does not need such high-level features.

3) Attention mechanism: Each attribute type correlates differently with each relation class, for example, age is probably more important to identify parents and their children relationships than clothing or gender. In this sense, the function of the attention module is to learn how much and when each generate feature influences the relation, and apply coefficients to improve the performance, as shown on Table IV.

4) Score fusion: After the propagation process made by the AGN module and the computation of the attention scores, each output feature may generate different scores, leading to divergent predictions.

TABLE IV: Evaluation of the attention mechanism and mAP(%) comparison for each score fusion method on the **PISC** dataset.

Level	No Attention			Attention		
	max	mean	lse	max	mean	lse
Coarse	**83.98**	83.48	83.64	**84.07**	83.58	83.84
Fine	71.76	71.62	**71.78**	**71.81**	71.63	71.73

In some cases, taking the $\mathbf{mean}$ from all scores can lead to the best final prediction, by mitigating small variations from outputs that diverge from the general tendency. Similarly, the $\mathbf{lse}$ function also consider all the other outputs, but is more sensible to variations. However, an output with a much higher score value for specific classes, can provide some confidence level over the prediction, thus employing the $\mathbf{max}$ function can improve the overall precision of the network.

For our model, this was the scenario on both splits of the **PISC** dataset, as shown on Table IV, the **max** function performed best, producing the highest mAP. This results means that one of the node features, specifically, can provide better predictions than the others, which suggests that features may still preserve some of the characteristics that distinguish them from each other.

5) Attributes: Here we measure the performance of each attribute, considering only the output scores of features from the same attribute for each sample, obtaining sets of predictions based on cue type for the entire dataset. Next, we compute the mean average precision for these predictions, making it possible to compare them individually after the aggregation process.

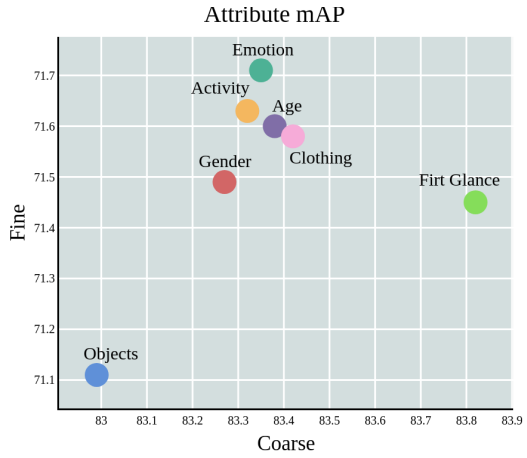


Fig. 3: Overall mAP(%) for each attribute type on the **PISC** dataset. First glance features stands out for coarse relationships, while fine-grained data benefits more from emotion.

The results are in line with the initial dataset proposal, as the coarse data identifies only 3 classes regarding to intimacy, that can be mostly inferred with positioning information, which is captured only by First Glance features, resulting in a higher precision. However, to classify fine relationships we need more specific information such as age, clothing, gender, activity and even high-level features like the emotion, reducing the gap between the precision of all attributes. Finally, the object information performs poorly because they are a secondary type of information, describing the environment and not relations, which can be only be useful when combined with other information.

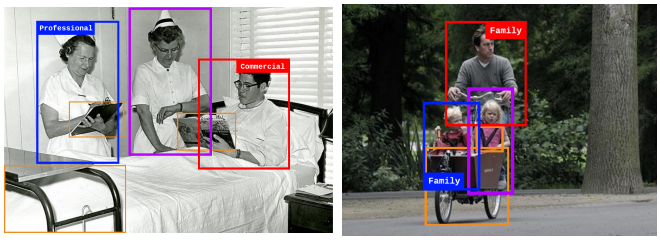
Also, in this experiment none of the attributes alone performed poorly than the $\mathbf{T} = 0$ model, and neither was capable of reaching the best mAP of the full framework. This happens because even attributes with lower performance can sometimes make correct predictions while the best ones are incorrect. This is the main benefit for using multiple attributes, however some of them may carry redundant or irrelevant information, which end up adding noise to the model.

F. Qualitative results

Finally, we present some examples of classifications from the fine-grained relationships on the **PISC** dataset. The initial example presented in Figure 1 shows a correctly classified commercial relation, between a doctor and a child. In this case, only two objects were detected on the image, and the stethoscope was not one of them, which could have been an important cue. Also, without this object, the image could be easily mistaken for a family relation between father and son. The proposed MAGI framework was capable of identifying the correct relation based especially on the doctor's formal clothes, as the clothing attribute for person one has the highest score, and also the object one, which is the tie, has a higher score than the other object. The attribute graph (Fig. 1b) suggests that age, gender, and activity were also important in the decision making, while emotion and first glance features have not contributed much to this case.

Next, we present some results for multi-relationship examples, comparing predictions between the person marked in purple with two distinct individuals in the same picture, marked in red and blue.

Figure 4 shows pictures that were correctly classified on both relations. On Fig. 4a, the relation between patient and nurse are classified due to action features, and the commercial relation is detected mostly because of similar clothing. The detected objects also were important for the classification. Finally, on Fig. 4b, sister and father were correctly classified mostly by age and positioning attributes.



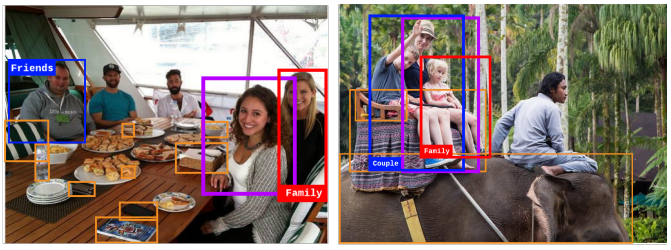
(a) (b)

Fig. 4: Correctly classified multi-relationships. (a) Professional and commercial relations. (b) Family relationships.

Figure 5a shows a correct friends classification and an incorrect family prediction due to the proximity between both girls. Fig. 5b shows a correct classification for the daughter and an incorrect classification for the son, where the network took into account mostly gender and action information.

V. Conclusion and Future Work

We propose the Multi-attribute Graph Inference framework for social relation recognition, which can explore the synergy between different types of information, capturing how they interact and learning meaningful representations. Our method consists of a graph propagation model to aggregate attribute data, combined with an attention mechanism and late fusion methods to refine the generated features.



(a) (b)

Fig. 5: Incorrectly classified multi-relationships. (a) Correct friends relation and incorrect family relationship. (b) Incorrect family relation classified as couple.

During the experiments we evaluated a diverse number of attributes and identified how they contribute to our model predictions. The results outperform the state-of-the-art and highlight the importance of different traits for the social relation recognition problem.

Future work can explore methods to select the most relevant features from the inputs, reducing the noise generated by the propagation model. Combinations between fully connected and fixed prior knowledge graphs can be an interesting approach to tackle this problem, applying constraints to the model. Another direction is to adapt the framework to process temporal data from videos, which is just starting to be explored in the social relation recognition.

REFERENCES

[1] K. J. August and K. S. Rook, *Social Relationships*. New York, NY: Springer New York, 2013, pp. 1838–1842.

[2] Z. Wang, T. Chen, J. Ren, W. Yu, H. Cheng, and L. Lin, "Deep Reasoning with Knowledge Graph for Social Relationship Understanding," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 1021–1028.

[3] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Dual-Glance Model for Deciphering Social Relationships," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2669–2678.

[4] Q. Sun, B. Schiele, and M. Fritz, "A Domain Based Approach to Social Relation Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 435–444.

[5] X. Guo, L. F. Polanía, J. Garcia-Frias, and K. E. Barner, "Social relationship recognition based on a hybrid deep neural network," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1–5.

[6] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning social relation traits from face images," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3631–3639, 2015.

[7] D. Bugental, "Acquisition of the algorithms of social life: A domain-based approach," *Psychological bulletin*, vol. 126, pp. 187–219, 04 2000.

[8] A. Fiske, "The four elementary forms of sociality: Framework for a unified theory of social relations," *Psychological review*, vol. 99, pp. 689–723, 11 1992.

[9] M. Zhang, X. Liu, W. Liu, A. Zhou, H. Ma, and T. Mei, "Multi-Granularity Reasoning for Social Relation Recognition From Images," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 1618–1623.

[10] H. Yan and C. Song, "Semantic three-stream network for social relation recognition," *Pattern Recognition Letters*, vol. 128, pp. 78 – 84, 2019.

[11] X. Guo, L. F. Polanía, J. Garcia-Frias, and K. E. Barner, "Social relationship recognition based on a hybrid deep neural network," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1–5.

[12] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *IEEE International Joint Conference on Neural Networks.*, vol. 2, 2005, pp. 729–734 vol. 2.

[13] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[14] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2017.

[15] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing, "Interpretable structure-evolving lstm," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel, "Gated graph sequence neural networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[17] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in *IEEE Winter Conference on Applications of Computer Vision*, March 2020.

[18] M. Raboh, R. Herzig, J. Berant, G. Chechik, and A. Globerson, "Differentiable scene graphs," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[19] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Z. Meng, N. Adluru, H. J. Kim, G. Fung, and V. Singh, "Efficient relative attribute learning using graph neural networks," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[21] X. Guo, L. Polania, B. Zhu, C. Boncelet, and K. Barner, "Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases," in *IEEE Winter Conference on Applications of Computer Vision*, March 2020.

[22] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *TPAMI*, vol. 39, pp. 1137–1149, 2015.

[23] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5534–5542, 2016.

[24] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016.

[25] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, "Beyond frontal faces: Improving Person Recognition using multiple cues," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4804–4813.

[26] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7