Visual-Semantic Graph Reasoning for Pedestrian Attribute Recognition

Qiaozhe Li,[1,3] **Xin Zhao,**[1,3] **Ran He,**[2,3,4] **Kaiqi Huang**[1,3,4]
[1]CRISE, CASIA, [2]CRIPAC & NLPR, CASIA
[3]University of Chinese Academy of Sciences
[4]CAS Center for Excellence in Brain Science and Intelligence Technology
liqiaozhe2015@ia.ac.cn, {xzhao,rhe,kqhuang}@nlpr.ia.ac.cn

Abstract

Pedestrian attribute recognition in surveillance is a challenging task due to poor image quality, significant appearance variations and diverse spatial distribution of different attributes. This paper treats pedestrian attribute recognition as a sequential attribute prediction problem and proposes a novel visual-semantic graph reasoning framework to address this problem. Our framework contains a spatial graph and a directed semantic graph. By performing reasoning using the Graph Convolutional Network (GCN), one graph captures spatial relations between regions and the other learns potential semantic relations between attributes. An end-to-end architecture is presented to perform mutual embedding between these two graphs to guide the relational learning for each other. We verify the proposed framework on three large scale pedestrian attribute datasets including PETA, RAP, and PA-100k. Experiments show superiority of the proposed method over state-of-the-art methods and effectiveness of our joint GCN structures for sequential attribute prediction.

Introduction

Pedestrian attribute recognition aims to make prediction of a set of attributes as the semantic descriptions of a pedestrian image. It has recently drawn a remarkable amount of attentions due to its promising applications in face verification (Kumar et al. 2009), person retrieval (Siddiquie, Feris, and Davis 2011), and person re-identification (Layne et al. 2012; Wang et al. 2018). Although it's easy to state, however, recognizing pedestrian attributes in real-world surveillance scenarios can be extremely challenging due to three factors: (1) Some attributes only relate to a small part of regions, which could be affected by ambiguous details caused by limited image resolution; (2) The appearance of pedestrian images is diversified caused by pose variation, viewpoint change, occlusion, background distraction, etc., which make it difficult to learn reliable image representations; (3) A pedestrian image usually contains multiple correlated attributes and each attribute may locate at different part regions according to its semantic characteristics. All these factors make it difficult to learn an effective attribute recognition model.

To deal with the above mentioned problems, it's desirable to jointly explore the spatial and semantic relations of attributes. However, such relations have not been fully exploited in traditional methods. Earlier methods solve the attribute recognition problem by learning a separate binary classifier for each of the attributes (Deng et al. 2014; Sudowe, Spitzer, and Leibe 2015). In such way, the relationships between attributes are simply ignored. To associate attributes with their corresponding regions, some methods employ auxiliary pose or part supervision information to learn part-based models (Zhang et al. 2014; Li et al. 2016c) and others formulate recognition as a weakly supervised localization problem using attention mechanism (Liu et al. 2017b; Zhu et al. 2017; Sarafianos and Kakadiaris 2018). While better recognition accuracy can be achieved with spatial context learning, these methods fail to capture semantic relations of attributes.

To model attribute relations at semantic level, some methods employ probabilistic graphical models (Chen, Gallagher, and Girod 2012) or structured inference models (Hu et al. 2016). Based on statistical hand-crafted features or a holistic deep representation model, these methods fail to consider spatial distribution of different attributes. The recognition task can also be formulated as a sequential prediction process by adopting sequential encoder-decoder architecture (Wang et al. 2017) to model the high-order dependencies among attributes. Although benefiting from the sequential prediction framework, the pairwise relations between attributes may not be described. Besides, the encoder-decoder model is deep considering RNN unrolling, which is difficult to optimize as the length of prediction sequence increases.

In this work, pedestrian attribute recognition is also formulated as a sequential prediction problem. A graph-based reasoning framework is proposed to jointly model spatial and semantic relations of region-region, attribute-attribute, and region-attribute. For each pedestrian image, the image regions and corresponding semantic attributes are respectively represented as nodes in a spatial graph and a directed semantic graph. Unlike existing methods which employ RNNs (Wang et al. 2016; Liu et al. 2017a; Wang et al. 2017) to characterize latent high-order dependencies, pairwise relations can be captured by performing message passing inside each graph using the Graph Convolution Network (GCN) (Kipf and Welling 2017). To better explore relations between regions and attributes, the output representations of

the two graphs are mutually embedded as additional inputs to guide the relational learning for each other. It's achieved by two sub-networks, which decompose the mutual embedding into two separate feed-forward streams to avoid the existence of closed loops. In the first sub-network, the spatial graph is first introduced to capture similarity and topological relations between image regions. The spatial context representation, which is obtained by performing average pooling over the outputs of all region nodes, is then embedded into semantic space to guide relational learning between attributes on the directed semantic graph. In the second sub-network, a directed semantic graph is first adopted to model semantic dependencies of attributes along the prediction path. The output of each attribute node is embedded into spatial graph to perform semantic-aware feature learning for next attribute. The two sub-networks are aggregated together to perform joint spatial and semantic relational learning for sequential attribute recognition.

The contributions of this paper are: (1) A visual-semantic graph reasoning framework is proposed to jointly model spatial and semantic relations for sequential pedestrian attribute prediction. (2) A novel end-to-end architecture is presented based on spatial and semantic graphs, which not only capture spatial relations between regions and potential semantic relations between attributes by performing graph convolutions inside each graph, but also model the relations between regions and attributes by performing mutual embedding between the two graphs to guide the relational learning for each other. (3) The proposed method is evaluated on 3 large-scale pedestrian attribute benchmarks including PETA (Deng et al. 2014), RAP (Li et al. 2016a) and PA-100k (Liu et al. 2017b). Experiments show superiority of the proposed method over state-of-the-art methods and effectiveness of our joint GCN structures for sequential attribute prediction.

Related Work

Pedestrian Attribute Recognition

Pedestrian attribute has been applied in a variety of vision tasks (Layne et al. 2012; Jaha and Nixon 2014; Wang et al. 2018). Earlier pedestrian attribute models (Zhu et al. 2013; Deng et al. 2014) treated multiple attributes independently and trained a sperate classifier for each of the attributes. Later, CNNs have been introduced for image feature learning and joint multi-attribute classification (Sudowe, Spitzer, and Leibe 2015; Li, Chen, and Huang 2015). These deep methods are based on holistic image representations, and may have limited capability to recognize attributes covering fine-grained details. Motivated by part-based models (Bourdev, Maji, and Malik 2011), (Zhang et al. 2014) and (Li et al. 2016c) employed body-part detectors for feature representations to recognize human attributes. However, the part detectors trained with auxiliary images make strong assumptions on image qualities, which may introduce additional noise when applied in surveillance data. Some methods formulate attribute recognition as a weakly supervised localization problem using attention mechanism. (Liu et al. 2017b) proposed multi-directional attention modules to learn attention-strengthened features at multiple levels and scales. (Sarafianos and Kakadiaris 2018) extended the work of (Zhu et al. 2017) by adding penalties on attention masks which have high prediction variance and introducing a weighted loss function for an attention aggregation model. Although recognition accuracy has been significantly improved, the generated attention masks are regularized by typical convolution operations, which make the semantic relations between attributes less interpretable.

On the other hand, semantic relations between attributes have also been studied. (Chen, Gallagher, and Girod 2012) explored the mutual dependencies between cloth attributes using the Conditional Random Fields (CRFs). However, such graphical inference model fails to take spatial context into consideration, thus may not be able to describe the underlying spatial relations between attributes. Moreover, the attribute classifiers and CRF are optimized separately instead of being unified into an end-to-end inference model. Motivated by recent success of sequential multi-label prediction models (Wang et al. 2016; Liu et al. 2017a), (Wang et al. 2017) proposed a RNN encoder-decoder based framework to jointly learn image level context and attribute level sequential correlation. As the RNN encoder-decoder mainly characterizes high-order dependencies of attributes, the pairwise relations may not be described.

Graphical Models

Reasoning on pairwise relations has been proved to be beneficial to a variety of vision tasks including object recognition (Gkioxari et al. 2018; Chen et al. 2018), video understanding (Ma et al. 2018), and action recognition (Gkioxari, Girshick, and Malik 2015). In multi-label image classification problem, CRF is often applied to model the dependencies between multiple labels (Xue et al. 2011; Li, Zhao, and Guo 2014; Li et al. 2016b). However, these methods also fail to explore label relations in spatial layout. Instead of performing iterative mean-field approximation, simpler graph-based neural networks (Scarselli et al. 2009; Li et al. 2015; Marino, Salakhutdinov, and Gupta 2017) have been proposed to unify the inference procedure into an end-to-end model. In this paper, the Graph Convolutional Network (GCN) (Kipf and Welling 2017), which was originally proposed to perform semi-supervised classification in language processing, is utilized for relational reasoning. The GCN serves as basic layers for the proposed framework to jointly model the spatial and semantic relations between pedestrian attributes.

Approach

In this paper, pedestrian attribute recognition is treated as a sequential attribute prediction problem. Different from existing methods which employ RNNs to describe latent high-order dependencies of attributes, pairwise relations can be modeled via graph convolutions in the proposed graph-based reasoning framework. In this section, the Graph Convolutional Network will be first introduced. The proposed framework and its components will be introduced in the following.

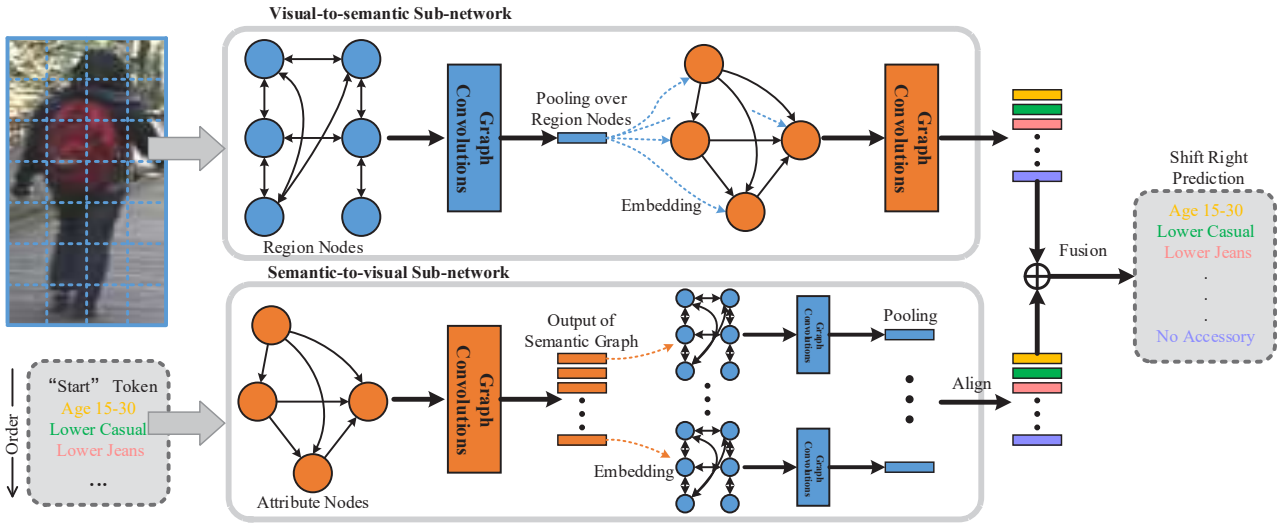

Figure 1: Overview of the proposed visual-semantic graph reasoning framework. It consists of two sub-networks, which perform mutual embedding between the spatial and semantic graphs in complementary ways to achieve joint visual-semantic reasoning.

Graph Convolutional Network

The Graph Convolutional Network (GCN) is proposed to perform reasoning on graphs. Unlike standard convolutions which operate on a local regular grid, the graph convolutions compute the response of a node based on its neighbors defined by the graph relations. Thus, message passing is performed inside the graphs with the graph convolutions. Let $\mathbf{Z} \in \mathbb{R}^{N \times d}$ denotes the input features of the graph, where N denotes the number of entities, d is the number of the feature channels. Mathematically, graph convolutions for one layer can be represented as:

$$\mathbf{G} = \mathbf{AZW} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ represents the adjacency matrix of the graph, which can either be pre-defined or learned from data. $\mathbf{W}$ is the weight matrix of the layer. To achieve effective training, two non-linear operations including Layer Normalization (Ba, Kiros, and Hinton 2016) and ReLU activation are usually applied after each convolutional layer before the features are forwarded to the next layer.

Visual-semantic Graph Reasoning

Given a pedestrian image I, our goal is to make prediction of its attributes $\mathbf{y}_1, ..., \mathbf{y}_K$. It's intrinsically a multi-label classification problem since a pedestrian image can be annotated by multiple attributes. Since attribute recognition is formulated as a sequential prediction problem, the joint probability over $\mathbf{y}_1, ..., \mathbf{y}_K$ can be modeled in the chain rule as,

$$P(\mathbf{y}_1, ..., \mathbf{y}_K | I) = \prod_{k=1}^{K} P(\mathbf{y}_k | I, \mathbf{y}_0, ..., \mathbf{y}_{k-1}) \tag{2}$$

In our framework, the relations between image regions are modeled on a spatial graph and the dependencies between attributes are modeled on a semantic graph with directed edges. The joint reasoning is performed in a mutually guided way with two sub-networks, which will be introduced in details in the following.

Visual-to-semantic Sub-network. In this sub-network, the spatial graph is first used to learn image feature representations to capture spatial relations between different body parts of the whole pedestrian image. The learned spatial context is then embedded into semantic space to guide relational learning between attributes on the directed semantic graph. In spatial graph, each node corresponds to one specific image part region and the relations between regions are modeled by edges. Two types of spatial relations including similarity relations and topological relations are described. They are respectively modeled by two sub-graphs, which share the nodes but employ different edges.

The first sub-graph measures visual similarity in image feature space. Assuming that the input of the spatial graph $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M)^{\mathrm{T}}$ consists of the visual features extracted from a convolutional neural network, where M denotes the number of locations of the convolutional feature maps and $\mathbf{x}_i$ corresponds to i-th image region. The pairwise similarity between every two part regions can be computed by the function,

$$\mathbf{F}_s(\mathbf{x}_i, \mathbf{x}_j) = \varphi_s(\mathbf{x}_i)^{\mathrm{T}} \varphi'_s(\mathbf{x}_j) \tag{3}$$

where $\varphi_s(\mathbf{x}) = \mathbf{w}_s \mathbf{x}$ and $\varphi'_s(\mathbf{x}) = \mathbf{w}'_s \mathbf{x}$ denote two different linear transformations of the visual features. The weight matrices $\mathbf{w}_s \in \mathbb{R}^{d \times d}$ and $\mathbf{w}'_s \in \mathbb{R}^{d \times d}$ can be learned through back propagation. After being computed using Eq.(3), the similarity adjacency matrix is normalized using a softmax function along each row,

$$\mathbf{A}_{s_a}(i, j) = \frac{\exp(\mathbf{F}_s(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{j=1}^{M} \exp(\mathbf{F}_s(\mathbf{x}_i, \mathbf{x}_j))} \tag{4}$$

Besides the similarity relations between parts, the topological structures are also accounted for by connecting one part

with its neighbor regions. The edge values capturing the topological relations between body parts can be computed as,

$$\mathbf{A}_{s_l}(i,j) = \frac{\exp(-d_{ij}/\Delta)}{\sum_{j=1}^{M}\exp(-d_{ij}/\Delta)} \tag{5}$$

where d_{ij} denotes pixel-level distance between the two parts, and Δ is the scaling factor. Softmax is also performed as normalization on each row so that the sum of all edge weights connected to one part is 1.

With the edge weights of both sub-graphs, the outputs of the spatial graph are computed by combining the two sub-graphs together,

$$\mathbf{G}_s = \mathbf{A}_{s_a}\mathbf{X}\mathbf{W}_{s_a} + \mathbf{A}_{s_l}\mathbf{X}\mathbf{W}_{s_l} \tag{6}$$

where $\mathbf{W}_{s_a} \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_{s_l} \in \mathbb{R}^{d\times d}$ are weight matrices for two sub-graphs. After convolution operations, average pooling is performed over all nodes of spatial graph to obtain the spatial context representation $\mathbf{g}_s \in \mathbb{R}^d$.

Conditioned on spatial context, the directed semantic graph is employed to perform relational learning in semantic space. In this graph, each node corresponds to one specific semantic attribute. Given a pedestrian image, we use $\mathbf{R} = (\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_K) \in \mathbb{R}^{f\times(K+1)}$ to denote the embedding matrix of its ground-truth semantic attributes aligned according to a prediction order, where each column $\mathbf{r}_i \in \mathbb{R}^f$ is an embedding vector and $\mathbf{r}_0$ denotes the "start" token. To make use of order information, positional encoding (Gehring et al. 2017) is performed by embedding the absolute position of attributes $\mathbf{P} = (\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_K) \in \mathbb{R}^{f\times(K+1)}$, where $\mathbf{p}_i \in \mathbb{R}^f$. They are combined together to obtain semantic attribute representations on an ordered prediction path $\mathbf{E} = (\mathbf{e}_0, \mathbf{e}_1, ..., \mathbf{e}_K) \in \mathbb{R}^{f\times(K+1)}$, where $\mathbf{e}_k = \mathbf{r}_k + \mathbf{p}_k$. Further, spatial context is embedded into each node by the function,

$$\mathbf{C} = \mathbf{E} \oplus (\mathbf{U}_s\mathbf{g}_s) \tag{7}$$

where $\mathbf{U}_s \in \mathbb{R}^{f\times d}$ denotes the learnable projection matrix. $\oplus$ operation is computed by adding the embedding vector to each column of matrix $\mathbf{E}$. The i-th column $\mathbf{c}_i \in \mathbb{R}^f$ of $\mathbf{C} \in \mathbb{R}^{f\times(K+1)}$ is the input representation of i-th node. To ensure the prediction of current attribute only has relations with previously known outputs, the i-th node is only connected with nodes whose subscript $\leq i$. For those connected edges, the edge weights can be computed as,

$$\mathbf{F}_{\hat{e}}(\mathbf{c}_i, \mathbf{c}_j) = \varphi_{\hat{e}}(\mathbf{c}_i)^{\mathrm{T}}\varphi'_{\hat{e}}(\mathbf{c}_j) \tag{8}$$

where $\varphi_{\hat{e}}(\cdot)$ and $\varphi'_{\hat{e}}(\cdot)$ are linear transformation functions with weight matrices $\mathbf{w}_{\hat{e}} \in \mathbb{R}^{f\times f}$ and $\mathbf{w}'_{\hat{e}} \in \mathbb{R}^{f\times f}$. Similarly, $\mathbf{A}_{\hat{e}} \in \mathbb{R}^{(K+1)\times(K+1)}$ is also computed by normalizing the connected edge weights along each row, and its upper triangular elements will be 0. The convolutions on semantic graph can be represented as,

$$\mathbf{G}_{\hat{e}} = \mathbf{A}_{\hat{e}}\mathbf{C}^{\mathrm{T}}\mathbf{W}_{\hat{e}} \tag{9}$$

where $\mathbf{W}_{\hat{e}} \in \mathbb{R}^{f\times f}$ denotes the weight matrix.

After performing convolutions on semantic graph, the output representations $\mathbf{G}_{\hat{e}} \in \mathbb{R}^{(K+1)\times f}$ are used for sequential attribute prediction. The output supervision is obtained by right shifting the input sequence by one position, which means the output of k-th node $\mathbf{g}^k_{\hat{e}} \in \mathbb{R}^f$ is used to predict (k+1)-th attribute according to the prediction order.

Semantic-to-visual Sub-network. In this sub-network, the directed semantic graph is first adopted to capture semantic relations between attributes. At each prediction step, the output of current attribute node is embedded into spatial graph to perform semantic-aware feature learning to predict the next attribute. In this sub-network, attribute nodes are represented by $\mathbf{E} = (\mathbf{e}_0, \mathbf{e}_1, ..., \mathbf{e}_K) \in \mathbb{R}^{f\times(K+1)}$ and its edge values can be computed as,

$$\mathbf{F}_e(\mathbf{e}_i, \mathbf{e}_j) = \varphi_e(\mathbf{e}_i)^{\mathrm{T}}\varphi'_e(\mathbf{e}_j) \tag{10}$$

where the weights of functions $\varphi_e(\cdot)$ and $\varphi'_e(\cdot)$ are different from Eq.(8). Similarly, current attribute is only connected with previous existed attributes in the adjacency matrix. After obtaining the normalized adjacency matrix $\mathbf{A}_e \in \mathbb{R}^{(K+1)\times(K+1)}$, the convolution operations can be represented as,

$$\mathbf{G}_e = \mathbf{A}_e\mathbf{E}^{\mathrm{T}}\mathbf{W}_e \tag{11}$$

where $\mathbf{W}_e \in \mathbb{R}^{f\times f}$ denotes the weight matrix. $\mathbf{G}_e \in \mathbb{R}^{(K+1)\times f}$ are outputs of the attribute nodes, in which each row $\mathbf{g}^k_e \in \mathbb{R}^f$ denotes the semantic representation at k-th step.

The learned semantic representation is embedded into each region node of the spatial graph. Thus, the input representations of region nodes at k-th prediction step can be computed as,

$$\mathbf{D}^k = \mathbf{X} \oplus (\mathbf{U}_e\mathbf{g}^k_e)^{\mathrm{T}} \tag{12}$$

where $\mathbf{U}_e \in \mathbb{R}^{d\times f}$ is used to project $\mathbf{g}^k_e$ into image feature space. The edge values capturing the similarity between region nodes can be represented as,

$$\mathbf{F}^k_{\hat{s}}(\mathbf{d}^k_i, \mathbf{d}^k_j) = \varphi_{\hat{s}}(\mathbf{d}^k_i)^{\mathrm{T}}\varphi'_{\hat{s}}(\mathbf{d}^k_j) \tag{13}$$

where $\mathbf{d}^k_i \in \mathbb{R}^d$ is i-th row of $\mathbf{D}^k$ representing the i-th image region embedded with current semantic representation. In such way, visual features related to different attributes can be extracted by learning different relations between regions guided by semantic context. The weight matrices of functions $\varphi_{\hat{s}}(\cdot)$ and $\varphi'_{\hat{s}}(\cdot)$ are shared across entire sequential prediction process in consideration of computation efficiency. The convolution operations on spatial graph can be represented as,

$$\mathbf{G}^k_{\hat{s}} = \mathbf{A}^k_{\hat{s}_a}\mathbf{D}^k\mathbf{W}_{\hat{s}_a} + \mathbf{A}_{s_l}\mathbf{D}^k\mathbf{W}_{\hat{s}_l} \tag{14}$$

where the first term characterizes similarity relations between regions guided by semantic context. In second term, the adjacency matrix $\mathbf{A}_{s_l}$ remains unchanged to preserve the topological structure. After convolution operations, average pooling is performed over all region nodes of $\mathbf{G}^k_{\hat{s}}$ to obtain the representation $\mathbf{g}^k_{\hat{s}} \in \mathbb{R}^d$ at k-th step. Totally K+1 representations $\mathbf{g}^1_{\hat{s}}, ..., \mathbf{g}^{K+1}_{\hat{s}}$ can be obtained by embedding each $\mathbf{g}^k_e$ into the spatial graph.

Training and Inference

The two sub-networks are unified into an end-to-end model to jointly perform visual-semantic graph-based reasoning. At each prediction step, the current output representations of the two sub-networks are concatenated together to predict the next attribute. The output supervision is obtained by right shifting the ground-truth attributes aligned according to the pre-defined prediction order and adding an additional "end" token at the last position. The conditional probability at k-th predication step can be represented by,

$$P(\mathbf{y}_k|I, \mathbf{y}_0, ..., \mathbf{y}_{k-1}) \propto \exp(\mathbf{W}_y \begin{bmatrix} \mathbf{g}_{\hat{e}}^{k-1} \\ \mathbf{g}_{\hat{s}}^{k-1} \end{bmatrix}) \quad (15)$$

where $\mathbf{W}_y \in \mathbb{R}^{(C+1)\times(d+f)}$ is the weight matrix and C is the total number of attributes.

The entire network is trained with BCE loss after normalizing the probability scores with softmax function. During the training process, the computation of output representations and back-propagation can be parallelized across all nodes in the graph, which is in contrast with traditional RNN encoder-decoder based methods which perform information propagation through the entire long-range sequences. To boost the performance of sequential prediction methods, some policies have been proposed to explore prediction paths, such as beam search (Wang et al. 2016) and order ensemble (Wang et al. 2017). In this paper, the later scheme is adopted since it shows more reliable results compared with the former. Besides, the prediction orders are also defined similar as (Wang et al. 2017).

In the testing stage, the proposed model performs sequential multi-attribute prediction given each pedestrian image. At each prediction step, current attribute is predicted conditioned on visual information of image I and previously predicted attributes $\mathbf{y}_k^* = \arg\max P(\mathbf{y}_k|I, \mathbf{y}_0, ..., \mathbf{y}_{k-1})$. Once predicted, current attribute is added as nodes in the semantic graph and used for next prediction. This procedure is repeated until the "end" token is met or the model reaches its maximum prediction length.

Experiments

Datasets. The proposed method is evaluated on three publicly available pedestrian attribute datasets: (1) The PEdesTrian Attribute (PETA) dataset (Deng et al. 2014) consists of 19, 000 person images collected from 10 small-scale person datasets. Each image is labelled with 61 binary attributes and 4 multi-class attributes. This paper follows the same experimental protocal as (Deng et al. 2014; 2015). The whole dataset is randomly divided into three non-overlapping partitions: 9500 for training, 1900 for verification, and 7600 for evaluation. 35 attributes whose positive ratios are higher than 5% are used for evaluation. (2) The Richly Annotated Pedestrian (RAP) attribute dataset (Li et al. 2016a) contains 41,585 images drawn from 26 indoor surveillance cameras. Each image is labelled with 69 binary attributes and 3 multi-class attributes. Following the official protocol (Li et al. 2016a), the whole dataset is split into 33,268 training images and 8,317 test images.

Dataset	Method	Metric				
		mA	Acc	Pre	Recall	F1
PETA	MRFr2	75.60	-	-	-	-
	ELF-mm	75.21	43.68	49.45	74.24	59.36
	FC7-mm	76.65	45.41	51.33	75.14	61.00
	FC6-mm	77.69	48.31	54.06	76.49	63.35
	ACN	81.15	73.66	84.06	81.26	82.64
	Deep-Mar	82.89	75.07	83.68	83.14	83.41
	HP-net	81.77	76.13	84.92	83.24	84.07
	JRL	**85.67**	-	86.03	85.34	85.42
	VeSPA	83.45	77.73	86.18	84.81	85.49
	MsVAA	84.59	<u>78.56</u>	<u>86.79</u>	<u>86.12</u>	<u>86.46</u>
	Ours	<u>85.21</u>	**81.82**	**88.43**	**88.42**	**88.42**
RAP	MRFr2	-	-	-	-	-
	ELF-mm	69.94	29.29	32.84	71.18	44.95
	FC7-mm	72.28	31.72	35.75	71.78	47.73
	FC6-mm	73.32	33.37	37.57	73.23	49.66
	ACN	69.66	62.61	<u>80.12</u>	72.26	75.98
	Deep-Mar	73.79	62.02	74.92	76.21	75.56
	HP-net	76.12	65.39	77.33	78.79	78.05
	JRL	<u>77.81</u>	-	78.11	78.98	78.58
	VeSPA	77.70	<u>67.35</u>	79.51	<u>79.67</u>	<u>79.59</u>
	MsVAA	-	-	-	-	-
	Ours	**77.91**	**70.04**	**82.05**	**80.64**	**81.34**
PA100k	Deep-Mar	72.70	70.39	82.24	80.42	81.32
	HP-net	<u>74.21</u>	<u>72.19</u>	<u>82.97</u>	<u>82.09</u>	<u>82.53</u>
	Ours	**79.52**	**80.58**	**89.40**	**87.15**	**88.26**

Table 1: Comparisons against 10 state-of-the-art methods on three datasets. 1st and 2nd best results in **bold** fonts and <u>underlined</u>, respectively.

51 binary attributes are used to evaluate the recognition performance. (3) The PA-100k Dataset (Liu et al. 2017b) consists of 100,000 pedestrian images from 598 outdoor scenes. Each image is described with 26 commonly used attributes. The whole dataset is split into training, validation and test sets with a ratio of 8:1:1 (Liu et al. 2017b). For both PETA and RAP datasets, the multi-class attributes are converted into binary attributes as in (Deng et al. 2014; 2015; Li et al. 2016a).

Performance Metrics. Five metrics are adopted to evaluate attribute recognition performance. (1) Class-based: For each attribute class, the classification accuracy of positive and negative samples are computed respectively and then averaged as the recognition score for this attribute. mA is then computed by averaging the recognition scores over all attributes (Deng et al. 2014). (2) Instance-based: Accuracy, precision, recall and F1-score are used to measure instance-based (Li et al. 2016a) attribute recognition results. For accuracy, precision and recall, we first compute the scores of predicted attributes against the groundtruth for each test image and then average the scores over all test images. The F1-score is computed based on precision and recall. Different from mA that assume independence between attributes, instance-based evaluation metrics consider the inter-attribute correlation.

Compared Methods. The proposed method is compared against 10 state-of-the-art models. (1) MRFr2 (Deng et al. 2015) exploits the context of neighbouring images by Markov Random Field for mining the visual appearance

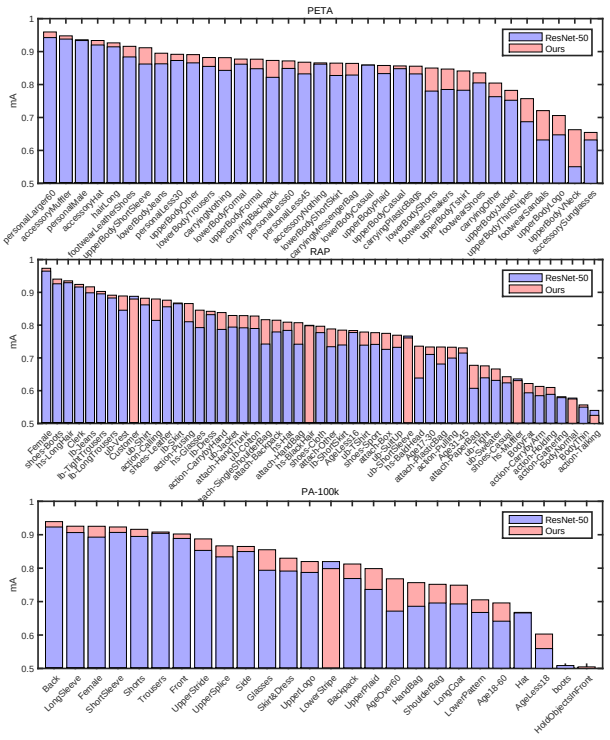

Figure 2: Mean Accuracy (mA) for all attributes on three datasets. Baseline and our method are marked with blue and red bars respectively.

proximity relations between different images; (2) ELF-mm (Gray and Tao 2008) employs SVM classifier with Ensemble of Localized Features (ELF) for attribute recognition; (3)-(4) FC7-mm and FC6-mm replace the hand-crafted ELF features with CNN features (FC7 and FC6 output of the AlexNet); (5) Attributes Convolutional Network (ACN) (Sudowe, Spitzer, and Leibe 2015) jointly trains a monolithic CNN model for all attributes, which allows to share weights and transfer knowledge among different attributes; (6) DeepMAR (Li, Chen, and Huang 2015) is a joint attribute learning model which considers inter-attribute correlation by weighted cross entropy loss function; (7) HydraPlus Network (HP-net) (Liu et al. 2017b) is an attention based method that employs multi-directional attention modules to train multi-level and multi-scale attention-strengthened features for pedestrian analysis; (8) Joint Recurrent Learning (JRL) model (Wang et al. 2017) considers attribute recognition as a sequence-to-sequence mapping problem and employs RNN encoder-decoder to jointly learn image level context and attribute level sequential correlation. (9) View-sensitive Pedestrian Attribute (VeSPA) model (Sarfraz et al. 2017) jointly learns a coarse view predictor and view-dependent image features for attribute inference. (10) MsVAA (Sarafianos and Kakadiaris 2018) is based on visual attention aggregation on multi-scales, combined with additional penalties on attention masks and a weighted loss function.

Implementation Details. In this paper, ResNet-50 is used to extract convolutional features for pedestrian images. The output of last convolutional layer (output of "Res_5c" block) is used as the visual input to spatial graph. The original pedestrian images are resized to 128×256 pixels. Stochastic gradient descend algorithm (Sutskever et al. 2013) is employed for training, with momentum of 0.9, and weight decay of 0.0005. The batch size is set to 32. The initial learning rate is set to 10^{-3} for the first 20 epoches, and decreased to 10^{-4} for the second 20 epoches. The model is implemented with pytorch.

Dataset	Method	Metric				
		mA	Acc	Pre	Recall	F1
PETA	ResNet-50	81.27	76.69	87.33	82.76	84.99
	$\mathbf{G}_s$	81.65	76.80	<u>87.94</u>	83.44	85.63
	$\mathbf{G}_e$	83.85	79.65	87.05	86.09	86.56
	$\mathbf{G}_{s\to e}$	<u>84.96</u>	<u>81.10</u>	87.91	<u>87.84</u>	<u>87.87</u>
	$\mathbf{G}_{e\to s}$	84.32	80.92	87.63	86.76	87.19
	Ours	**85.21**	**81.82**	**88.43**	**88.42**	**88.42**
RAP	ResNet-50	75.12	66.67	<u>81.66</u>	76.52	79.00
	$\mathbf{G}_s$	75.54	67.35	81.34	76.37	78.78
	$\mathbf{G}_e$	75.95	68.74	80.94	78.63	79.76
	$\mathbf{G}_{s\to e}$	<u>76.86</u>	69.57	80.93	<u>79.54</u>	<u>80.23</u>
	$\mathbf{G}_{e\to s}$	76.54	<u>69.76</u>	80.86	79.04	79.94
	Ours	**77.91**	**70.04**	**82.05**	**80.64**	**81.34**
PA100k	ResNet-50	76.31	76.76	88.62	83.22	85.84
	$\mathbf{G}_s$	77.05	76.82	88.75	83.45	86.02
	$\mathbf{G}_e$	78.54	78.93	88.45	86.41	87.42
	$\mathbf{G}_{s\to e}$	<u>79.03</u>	<u>79.54</u>	<u>88.87</u>	<u>86.75</u>	<u>87.80</u>
	$\mathbf{G}_{e\to s}$	78.54	78.95	88.53	86.18	87.34
	Ours	**79.52**	**80.58**	**89.40**	**87.15**	**88.26**

Table 2: Effect of each component of the proposed network. 1^{st} and 2^{nd} best results in **bold** fonts and <u>underlined</u>.

Comparison to the State-of-the-Arts

Table 1 shows evaluations on three datasets. The proposed method shows the best performance on all three datasets measured by five evaluation metrics, except it achieves second best score in mA on PETA dataset. Though JRL model reports a minor gain in mA on PETA dataset, however, our method outperforms JRL in precision, recall and F1-score with a significant margin (2.40%, 3.08%, and 3.00%, respectively). MsVAA method has achieved second best scores on 4 instance-based metrics, despite adopting a stronger ResNet-101 baseline. On RAP dataset, ACN model presents the second best score in precision but with a much lower recall. It indicates that ACN tends to miss some attributes in recognition by simply adopting a holistic representation model. In contrast, the proposed method achieves significant improvement on all instance-based metrics due to the visual-semantic graph-based reasoning. JRL and VeSPA show competitive results on this dataset. On PA-100k dataset, the proposed method has achieved greater performance improvement. It further demonstrates the effectiveness of the proposed method in modelling the spatial and semantic relations of attributes on graphs.

Ablation Study

Quantitative Evaluation. Besides the comparison with state-of-the-art methods, we also conduct experiments to evaluate the effectiveness of each component of the pro-

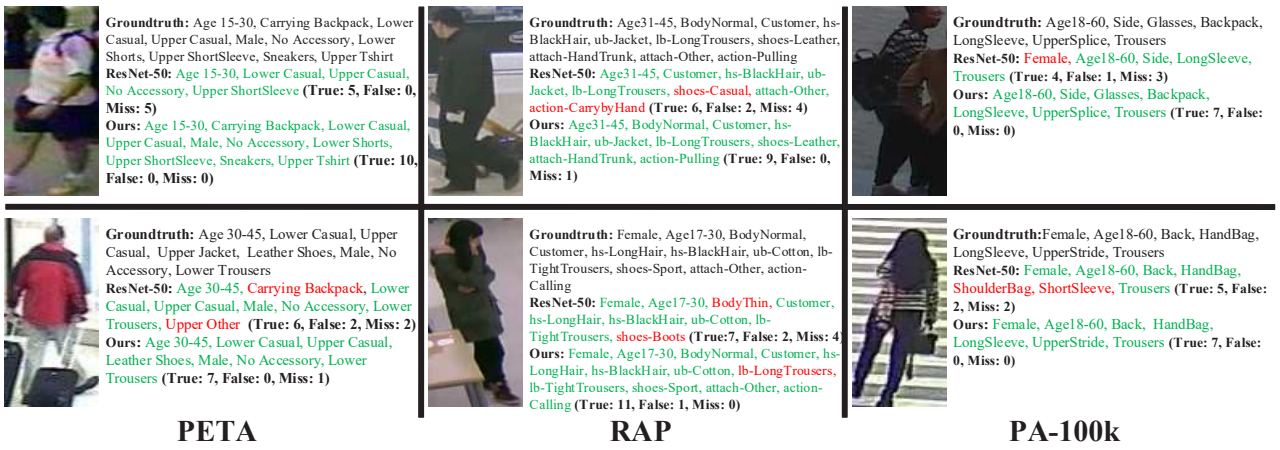

Figure 3: Qualitative evaluation of some pedestrian images from three datasets. The correct and wrong predictions are marked in green and red, respectively.

posed network, which is shown in Table 2. (1) The original ResNet-50 is used as the baseline which is fine-tuned on each of the datasets. (2) A spatial graph is added on top of the final convolutional layer of ResNet-50 to explore the spatial relations between different image regions. The output representation of spatial graph is then average-pooled and followed by a classification layer. (3) A directed semantic graph is adopted to model semantic dependencies between attributes during training, which directly use ResNet-50 FC features as visual input. (4)-(5) Each of the two sub-networks is used to perform attribute recognition independently. These models are compared with the proposed method in the same experimental settings.

Fig.2 reports the overlapped histograms of mean Accuracy (mA) on all three datasets for all attributes by baseline ResNet-50 model and the proposed method. The bars are sorted in descending order according to the larger mA between the two methods at one attribute. It is evident that the proposed method has achieved significant performance gain on most attributes on all three datasets. For some attributes which either only cover small parts of the images ("Sandals" and "V-neck" in PETA, "BaldHead" and "Attach-PaperBag" in RAP) or require deduction from contextual information ("action-Calling" in RAP, "AgeOver60" in PA-100k), the improvement is particularly prominent. This can be contributed to the effectiveness of visual-semantic graph reasoning in attribute recognition.

Qualitative Evaluation. Fig.3 shows instance-based recognition results of some pedestrian images from three datasets. For each examplar image, the correct and wrong predicted attributes are respectively marked in green and red color. The true, false and missed numbers of predicted attributes are also given. Results show that the baseline model is prone to miss some attributes in recognition, especially those describing fine-grained pedestrian details. This is consistent with the reported relatively low recall rate of baseline model in qualitative evaluation. On the other hands, by performing visual-semantic graph-based reasoning, the proposed method has recognized more detailed attributes while making less mistakes. In both RAP exemplars, the baseline model has failed to recognize the person's actions. This may caused by its limited capability to infer with contextual clues. In first exemplar of PA-100k, the person is partially occluded. This brings disturbance to baseline model, which predicts sex incorrectly and fails to recognize the attributes of "Glasses", "Backpack", and "UpperSplice". In second image of PA-100k, the person's appearance is distracted by zebra crossing from background, which leads to incorrect recognition of upper-body attributes. In contrast, all attributes are corrected predicted with the proposed method.

Conclusions

A visual-semantic graph reasoning framework is proposed to jointly model spatial and semantic relations for sequential attribute prediction. It is achieved by performing mutual embedding between the spatial and semantic graphs to guide the relational learning for each other using two sub-networks. Experimental results on PETA, RAP and PA-100k pedestrian attribute datasets demonstrate that the proposed graph-based reasoning framework significantly outperforms a wide range of state-of-the-art methods. Moreover, the proposed method is shown to be more effective in recognizing some hard attributes against a variety of challenge factors.

Acknowledgement

This project is partial supported by the National Key Research and Development Program of China (Grant No. 2016YFB1001005), the National Natural Science Foundation of China (Grant No. 61673375 and Grant No.61602485), and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006 and Grant No. 173211KYSB20160008).

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. In *arXiv preprint arXiv:1607.06450*.

Bourdev, L.; Maji, S.; and Malik, J. 2011. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 1543–1550.

Chen, X.; Li, L.-J.; Fei-Fei, L.; and Gupta, A. 2018. Iterative visual reasoning beyond convolutions. In *CVPR*.

Chen, H.; Gallagher, A.; and Girod, B. 2012. Describing clothing by semantic attributes. In *ECCV*, 609–623.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian attribute recognition at far distance. In *ACM Multimedia*, 789–792.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2015. Learning to recognize pedestrian attribute. In *arXiv preprint arXiv:1501.00901*.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. In *ICML*, 1243–1252.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *CVPR*.

Gkioxari, G.; Girshick, R.; and Malik, J. 2015. Contextual action recognition with r* cnn. In *ICCV*, 1080–1088.

Gray, D., and Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 262–275.

Hu, H.; Zhou, G.-T.; Deng, Z.; Liao, Z.; and Mori, G. 2016. Learning structured inference neural networks with label relations. In *CVPR*, 2960–2968.

Jaha, E. S., and Nixon, M. S. 2014. Soft biometrics for subject identification using clothing attributes. In *IJCB*, 1–6.

Kipf, T. N., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *ICCV*, 365–372.

Layne, R.; Hospedales, T. M.; Gong, S.; and Mary, Q. 2012. Person re-identification by attributes. In *BMVC*.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Li, D.; Zhang, Z.; Chen, X.; Ling, H.; and Huang, K. 2016a. A richly annotated dataset for pedestrian attribute recognition. In *arXiv preprint arXiv:1603.07054*.

Li, Q.; Qiao, M.; Bian, W.; and Tao, D. 2016b. Conditional graphical lasso for multi-label image classification. In *CVPR*, 2977–2986.

Li, Y.; Huang, C.; Loy, C. C.; and Tang, X. 2016c. Human attribute recognition by deep hierarchical contexts. In *ECCV*, 684–700.

Li, D.; Chen, X.; and Huang, K. 2015. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, 111–115.

Li, X.; Zhao, F.; and Guo, Y. 2014. Multi-label image classification with a probabilistic label enhancement model. In *UAI*.

Liu, F.; Xiang, T.; Hospedales, T. M.; Yang, W.; and Sun, C. 2017a. Semantic regularisation for recurrent image annotation. In *CVPR*, 4160–4168.

Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; and Wang, X. 2017b. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 350–359.

Ma, C.-Y.; Kadav, A.; Melvin, I.; Kira, Z.; AlRegib, G.; and Graf, H. P. 2018. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*.

Marino, K.; Salakhutdinov, R.; and Gupta, A. 2017. The more you know: Using knowledge graphs for image classification. In *CVPR*, 20–28.

Sarafianos, N., and Kakadiaris, I. A. 2018. Deep imbalanced attribute classification using visual attention aggregation. *arXiv preprint arXiv:1807.03903*.

Sarfraz, M. S.; Schumann, A.; Wang, Y.; and Stiefelhagen, R. 2017. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In *BMVC*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The graph neural network model. *IEEE Transactions on Neural Networks* 20(1):61–80.

Siddiquie, B.; Feris, R. S.; and Davis, L. S. 2011. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 801–808.

Sudowe, P.; Spitzer, H.; and Leibe, B. 2015. Person attribute recognition with a jointly-trained holistic cnn model. In *ICCV Workshops*, 87–95.

Sutskever, I.; Martens, J.; Dahl, G.; and Hinton, G. 2013. On the importance of initialization and momentum in deep learning. In *ICML*, 1139–1147.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2285–2294.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2017. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, 531–540.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*.

Xue, X.; Zhang, W.; Zhang, J.; Wu, B.; Fan, J.; and Lu, Y. 2011. Correlative multi-label multi-instance image annotation. In *ICCV*, 651–658.

Zhang, N.; Paluri, M.; Ranzato, M.; Darrell, T.; and Bourdev, L. 2014. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 1637–1644.

Zhu, J.; Liao, S.; Lei, Z.; Yi, D.; and Li, S. 2013. Pedestrian attribute classification in surveillance: Database and evaluation. In *ICCV Workshops*, 331–338.

Zhu, F.; Li, H.; Ouyang, W.; Yu, N.; and Wang, X. 2017. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*, 2027–2036.