This ICCV Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Fusing Two Directions in Cross-domain Adaption for Real Life Person Search by Language

Kai Niu[1,3] Yan Huang[1,4] Liang Wang[1,2,3,4]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
[4]Artificial Intelligence Research, Chinese Academy of Sciences (CAS-AIR)
kai.niu@cripac.ia.ac.cn {yhuang, wangliang}@nlpr.ia.ac.cn

Abstract

Person search by language is an important application in video surveillance. The existing huge visual-semantic discrepancy and the cross-domain difficulty of emerging pedestrian images with new identities while no language description for training in real life application make this problem non-trivial to be addressed. In this paper, we first propose a concise and effective framework for image-sentence alignment to deal with the visual-semantic discrepancy. Second, we innovatively fuse the two opposite directions, i.e., source → target and target → source, for cross-domain adaption. Extensive experiments have validated the significant superiority of the proposed method on both source domain and target domain, and we have obtained the state-of-the-art performance and won the 1st place in competition.

1. Introduction

Person search by language [10, 12] is an important task in intelligent surveillance [4], which requires discriminative cross-modal representations to distinguish different people. It is difficult to directly measure the similarities between images and natural language descriptions due to the existing huge visual-semantic discrepancy. And in many practical situations, the problem that the training data and testing data are in different domains makes it even harder to accurately search for the matched person.

Although much progress [3] has been achieved for matching images and sentences accurately, it is still non-trivial to address the problem of person search by language, due to the less discriminative situation among images of different pedestrians. Specifically, different from the conventional image-sentence matching problem with images having various topics, scenes and styles, all images in the problem of person search by language belong to the pedestrian category, only having fine-grained differences and being much harder to distinguish. Therefore, we have to consider the characteristic of person search rather than only depending on the cross-modal matching approaches. Many solutions [17, 15] to the problem of image-based person search employ the pedestrian identities for classification, which can obtain more discriminative features for distinguishing pedestrians. And appropriately combining the objectives for person search and image-sentence matching may contribute to better visual-semantic embeddings further.

Beyond the visual-semantic discrepancy, there is another problem that the newly emerging pedestrian images have new identities but no language description for training, *i.e.*, cross-domain difficulty in real life application. To address this problem, domain adaption is necessary for narrowing the cross-domain gap. Many effective solutions [16, 14] have focused on transferring from the source domain to the target domain, but neglect the opposite direction. In fact, these two directions can contribute complementarily to the final fusion and obtain a model that better addresses the cross-domain problem.

In summary, this paper first introduces a general framework for dealing with the problem of person search by language, which considers identity classification as well as cross-modal matching for better visual-semantic embeddings. Second, a Cross-domain Bi-directional Adaption (CBA) method is proposed to alleviate the cross-domain difficulty by innovatively fusing the two opposite adaption directions, which facilitates the practical application of person search by language. Specifically, in the source → target (S → T) direction, the model is first trained to have pedestrian identity classification and cross-modal matching abilities in

the source domain. Then the visual encoder is transferred to handle the visual feature extraction in target domain, and the textual encoder is adjusted accordingly afterwards. In the opposite target → source (T → S) direction, we first train the visual encoder using the pedestrian identities in the target domain[1]. Then the target-domain-trained visual encoder is fine-tuned for matching images with sentences semantically in the source domain. Our solution has obtained the state-of-the-art performance in both domains and top rank in competition. The main contributions are as follows:

- We introduce a concise and effective image-sentence matching framework to deal with the problem of person search by language.
- To address the cross-domain difficulty in real life application, our solution innovatively fuses the two opposite directions for better cross-domain adaption.
- We have obtained the state-of-the-art performance in both domains and won the 1st place in competition.

2. Related Work

2.1. Person Search by Language

Li *et al*. [10] propose the first large-scale dataset, CUHK PErson DEScription dataset (CUHK-PEDES), for the problem of person search by language. They also provide the Recurrent Neural Network with Gated Neural Attention (GNA-RNN) mechanism model with unit-level attentions and word-level gates to determine the cross-modal affinity. Niu *et al*. [12] achieve better cross-modal similarity evaluation by a Multi-granularity Image-text Alignments (MIA) model which combines three different granularities hierarchically. Different from them, the major difficulty for real life person search by language is the cross-domain problem.

2.2. Cross-domain Person Re-identification

Wang *et al*. [15] simultaneously learn global identity and local attribute information through an identity inferred attribute space and introduce an attribute consistency scheme for performing unsupervised adaptation on the unlabelled target data. Generative models are also proved to be effective for domain adaption, for instance, Wei *et al*. [16] propose the Person Transfer Generative Adversarial Network to bridge the domain gap by making the transferred person images show similar styles with the target dataset while keeping the appearance and identity cues. More than only employing the source → target direction, we additionally consider the opposite target → source direction and further fuse them for better cross-domain adaption.

[1]The final testing protocol lies on using a sentence to retrieve the matched pedestrian images, but there is no sentence available for training in the target domain. Pedestrian identities in the target domain can be used as external cues in a weakly supervised configuration to initialize the visual encoder for preliminary pedestrian identification ability.

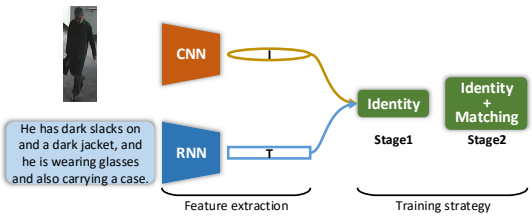

Figure 1. The proposed general framework for the problem of person search by language (best viewed in colors).

3. Proposed Approach

We first present a general framework for the problem of person search by language. Then we consider the cross-domain difficulty in real life application, and innovatively fuse the two opposite directions for cross-domain adaption.

3.1. Image-sentence Alignment Framework

To address the cross-modal problem of person search by language, we first need to extract visual and textual features, respectively. After that, cross-modal image-sentence alignment is important for alleviating the existing huge visual-semantic discrepancy. Specifically, the proposed general framework is shown in Fig. 1, and there are mainly two parts, *i.e.*, feature extraction and training strategy. We employ the convolutional neural networks (CNN) [7, 5] as visual encoder for image feature extraction, and obtain the visual feature $\mathbf{I} \in \mathbb{R}^D$. As for textual encoder, the recurrent neural networks (RNN) [6, 2] are used to obtain the feature vector $\mathbf{T} \in \mathbb{R}^D$ of a sentence. And D indicates the dimension of the shared visual-semantic space.

Based on the features $\mathbf{I}$ and $\mathbf{T}$, we follow [12] and employ the Identity and Matching objectives for training, which are suitable for initialization and fine-tuning, respectively. The Identity objective regards different IDs in training set as the number of categories, and classifies images and sentences to the corresponding ID category:

$$L_{ID} = L_I + L_T,$$
$$\mathbf{P}_I = softmax(FC_s(\mathbf{I})), \quad L_I = -log(\mathbf{P}_I(ID)),$$
$$\mathbf{P}_T = softmax(FC_s(\mathbf{T})), \quad L_T = -log(\mathbf{P}_T(ID)).$$

The shared Identity fully-connected (FC) layer, FC_s, is for mapping the features of images and sentences into the same space, whose dimension is the number of different identities in training set. $\mathbf{P}.(ID)$ is the predicted probability of the correct person ID, and $L.$ is the loss value after negative logarithm. As for the image-sentence alignment, the hinge-based triplet Matching objective is employed:

$$L_M = \sum_{\hat{\mathbf{T}}} max[0, \alpha - S(\mathbf{I}, \mathbf{T}) + S(\mathbf{I}, \hat{\mathbf{T}})] + \sum_{\hat{\mathbf{I}}} max[0, \alpha - S(\mathbf{T}, \mathbf{I}) + S(\mathbf{T}, \hat{\mathbf{I}})],$$

where $(\mathbf{I}, \mathbf{T})$ and $(\mathbf{T}, \mathbf{I})$ mean the matched image-sentence pairs, and $(\mathbf{I}, \hat{\mathbf{T}})$, $(\mathbf{T}, \hat{\mathbf{I}})$ indicate the mismatched pairs.

$S(\cdot,\cdot)$ means cosine similarity function and α is for the margin between the matched and mismatched pairs.

We employ a 2-stage training strategy for training the visual and textual encoders. In stage-1, we only use the Identity objective and focus on training the textual encoder and the shared Identity FC layer from scratch while fixing the pre-trained visual encoder. Then in stage-2, we train the whole network including the visual encoder with both Identity and Matching objectives, *i.e.*, $L_2 = L_{ID} + L_M$.

3.2. Cross-domain Bi-directional Adaption

Due to that images in training/validation sets and testing set are from different datasets, *i.e.*, cross-domain problem, we have to carry out cross-domain adaption for narrowing the gap between the two datasets. Many existing methods [16, 14] only consider transferring from the source domain to the target domain (S $\rightarrow$ T), *i.e.*, single direction, while neglecting the opposite direction (T $\rightarrow$ S). In fact, these two opposite directions can contribute complementarily to the final fusion and obtain a fused model that better addresses the cross-domain problem. Therefore, we propose a Cross-domain Bi-directional Adaption (CBA) method which fuses three sub-models coming from the two opposite directions for better cross-domain adaption.

3.2.1 Source $\rightarrow$ Target

Based on the proposed framework in Sec. 3.1, we employ the ResNet [5] which is pre-trained on the ImageNet [13] dataset as visual encoder for fine-tuning. As for the textual encoder, a bi-directional gated recurrent unit network (Bi-GRU) is used to extract the textual features.

We then carry out the cross-domain adaption from the source domain (CUHK-PEDES [10] dataset) to the target domain (MSMT17 [16] dataset) step-by-step. First, we train the visual and textual encoders on the CUHK-PEDES dataset through our 2-stage training strategy. Second, we employ the Domain Adaptive Re-Identification [14] method to transfer the CUHK-PEDES-trained visual encoder to the MSMT17 dataset for cross-domain adaption. Third, we use the adaptive visual encoder to re-train the textual encoder on the CUHK-PEDES dataset through the 2-stage training strategy once again. It is worth noting that the visual encoder is fixed in the re-training process. After these three steps, we finally obtain the **Model-A** of S $\rightarrow$ T adaption.

3.2.2 Target $\rightarrow$ Source

We employ the Osnet [17] model which is pre-trained on the MSMT17 [16] dataset as initialization for visual encoder. For better describing pedestrians, we additionally use the pre-trained DeepMAR [8] method to obtain the attribute features as external cues. It is worth noting that the DeepMAR method is directly employed for obtaining the attribute features without fine-tuning. Then we combine the visual context feature $\mathbf{I}_c$ (output from Osnet), and attribute feature $\mathbf{I}_a$ (output from DeepMAR) by feature concatenation. And a concatenation FC layer is employed for adjusting the feature space of the final visual feature $\mathbf{I}$. For sentences, a bi-directional long short term memory network (Bi-LSTM) is used to extract the textual feature $\mathbf{T}$. At last, we train the whole model (excluding DeepMAR for attribute feature extraction) on the CUHK-PEDES [10] dataset by our 2-stage training strategy.

There are two sub-models, **Model-B-1** and **Model-B-2**, addressing the cross-domain problem in the T $\rightarrow$ S direction. Sharing the foregoing overall structure, these two sub-models have some differences in the attribute feature extraction for diversity and complementarity. Specifically, the DeepMAR [8] method for Model-B-1 is trained on the RAP [9] dataset, while Model-B-2 on the Pa100K [11] dataset.

3.2.3 Similarity Fusion

The final image-sentence similarity results are obtained by fusing the similarity matrixes from the three foregoing sub-models, which alleviate the cross-domain difficulty by considering the two opposite adaption directions comprehensively. We denote the similarity matrixes from the Model-A, Model-B-1 and Model-B-2 as $\mathbf{S}_A$, $\mathbf{S}_{B1}$ and $\mathbf{S}_{B2}$, respectively. The final image-sentence similarity matrix $\mathbf{S}_F$ is

$$\mathbf{S}_F = \alpha \times \mathbf{S}_A + \beta \times \mathbf{S}_{B1} + \gamma \times \mathbf{S}_{B2}.$$

In the following parts, we set $\alpha = 1.7$, $\beta = 1.0$ and $\gamma = 1.4$ for obtaining the 'Final Model'.

4. Experiments and Analysis

4.1. Datasets and Protocols

We evaluate our solution based on the CUHK-PEDES dataset [10] (source domain, Validation set in competition) and the MSMT17 [16] dataset (target domain, Validation-2 and Final Test sets in competition)[2]. We measure performance by R@1, R@5 and R@10 criteria, *i.e.*, recall rates at the top-1, 5 and 10 results. Following the protocols in [10], a successful search is achieved if any image of the corresponding pedestrian is among the top-k images.

4.2. Ablation Study and Performance Comparison

Training Strategy Analysis. We carry out experiments on the base model (trained on CUHK-PEDES and no cross-domain adaption), and the results are shown in Tab. 1.

Performance Comparison. Extensive experiments for evaluation of ablation models on the CUHK-PEDES and MSMT17 datasets, and comparisons with other state-of-the-art methods are shown in Tab. 2. We have **won the 1st place** in WIDER Face and Person Challenge[3], and the detailed performance comparisons are shown in Tab. 3.

[2]Detailed descriptions of the datasets in competition are in this Link.
[3]Challenge Homepage.

Table 1. Objective analysis in the proposed framework. 'Id' and 'Mat' are for the Identity and Matching objectives, respectively. 'Id + Mat' means we simply sum the two objectives together.

Objectives	CUHK-PEDES			MSMT17		
	R@1	R@5	R@10	R@1	R@5	R@10
Id	29.03	51.63	62.98	11.86	25.73	36.54
Mat	50.15	73.28	81.72	22.37	42.84	54.90
Id + Mat	52.70	75.16	82.91	25.23	46.95	**59.41**
2-stage	**54.98**	**76.46**	**83.98**	**26.33**	**47.75**	58.91

Table 2. Evaluation of ablation models and comparisons with other state-of-the-art methods. A, B-1 and B-2 are for Model-A, Model-B-1 and Model-B-2, respectively.

Methods	CUHK-PEDES			MSMT17		
	R@1	R@5	R@10	R@1	R@5	R@10
S $\rightarrow$ T (A)	53.95	76.01	83.31	27.88	50.50	62.21
T $\rightarrow$ S (B-1)	49.67	73.44	81.83	28.13	50.30	60.86
T $\rightarrow$ S (B-2)	50.13	73.29	81.46	27.88	51.45	62.66
GNA-RNN [10]	19.05	-	53.64	-	-	-
GLIA [1]	43.58	66.93	76.26	-	-	-
MIA [12]	53.10	75.00	82.90	-	-	-
Final Model	**57.84**	**78.33**	**85.43**	**33.68**	**58.01**	**67.52**

Table 3. Results on the final test set in competition.

Rank	Team	Score (R@1)
1	**Ours (Final Model)**	**23.07%**
2	ZJU-Challenger	13.14%
3	SummerWalrus	12.75%

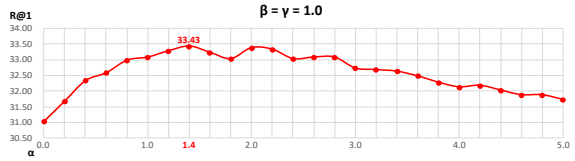

Figure 2. Analysis of hyper-parameters (best viewed in colors).

Hyper-parameters. In Fig. 2, we set $\beta = \gamma = 1.0$ for Model-B-1 and Model-B-2 in the T $\rightarrow$ S direction, and analyze α for Model-A in the S $\rightarrow$ T direction, to show the effect of the two adaption directions in similarity fusion.

5. Conclusion

In this paper, we first present a concise and effective framework to deal with the cross-modal problem of person search by language. Second, a Cross-domain Bi-directional Adaption (CBA) method is proposed to alleviate the cross-domain difficulty for facilitating the real life application of this task, which innovatively fuses the two opposite directions for cross-domain adaption. Experimental results show the significant superiority of the proposed solution on both source and target domains, and we have obtained the state-of-the-art perfromance and top rank in competition.

Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015), Capital Science and Technology Leading Talent Training Project (Z181100006318030), Beijing Science and Technology Project (Z181100008918010), HW2019SOW01, and CAS-AIR. This work is also supported by grants from NVIDIA and the NVIDIA DGX-1 and Saturn V Supercomputers.

References

[1] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018. 4

[2] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST*, 2014. 2

[3] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 1

[4] I. Haritaoglu, D. Harwood, and L. S. Davis. W/sup 4: real-time surveillance of people and their activities. *TPAMI*, 22(8):809–830, 2000. 1

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 2

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 2

[8] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, 2015. 3

[9] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv:1603.07054*, 2016. 3

[10] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *CVPR*, 2017. 1, 2, 3, 4

[11] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 3

[12] K. Niu, Y. Huang, W. Ouyang, and L. Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *arXiv:1906.09610*, 2019. 1, 2, 4

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 3

[14] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang. Unsupervised domain adaptive re-identification: Theory and practice. *arXiv:1807.11334*, 2018. 1, 3

[15] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, 2018. 1, 2

[16] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 2, 3

[17] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. *arXiv:1905.00953*, 2019. 1, 3