THE UNIVERSITY OF TENNESSEE KNOXVILLE

PERSON RE-IDENTIFICATION USING VISUAL ATTENTION

Alireza Rahimpour, Liu Liu, Ali Taalimi, Yang Song, Hairong Qi

Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA

ICIP 2017

Introduction

❑ Person re-identification is the problem of matching the same individuals across multiple cameras, or across time within a single camera.

Challenges: camera view point changes, clothing similarity, background clutter & occlusions, cross view lighting variation.

Motivation: Attention Mechanism for Person Re-id

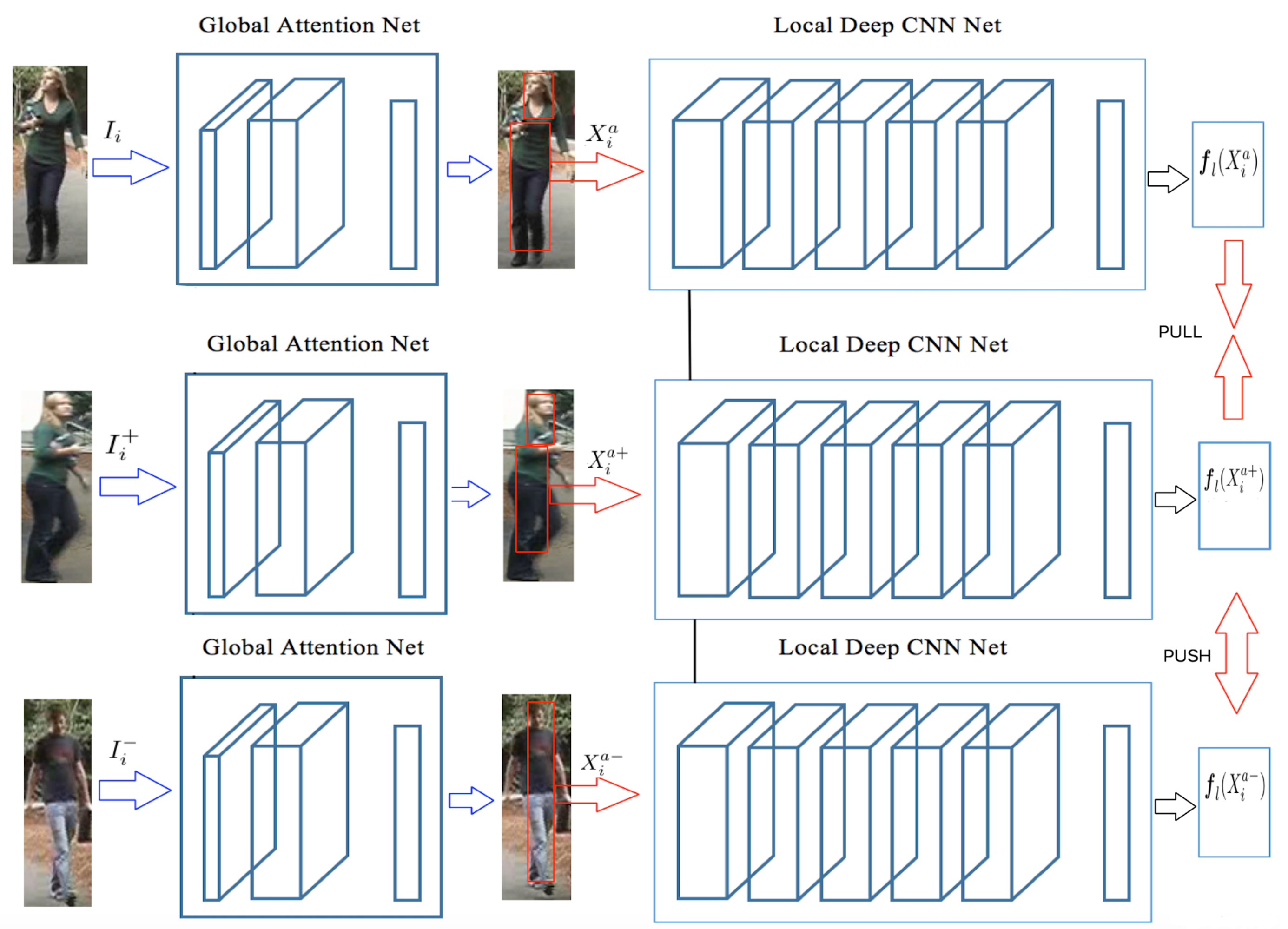
❑ Humans do not focus their attention on an entire scene at once when they want to identify another person.

❑ Instead, they pay attention to different parts of the scene (e.g., the person's face) to extract the most discriminative information.

Our Objective

Being able to focus on a certain important regions of the person's image with **high resolution** while perceiving the surrounding image in **low resolution**.

Model Architecture



Global Attention Net

We use the entropy of the output vector $\boldsymbol{h}_g(I)$ as a measure of saliency:

$$H = \sum_{l=1}^{C} \boldsymbol{h}_g^l \log(\boldsymbol{h}_g^l)$$

In order to find the attention map we then compute the norm of the gradient of the entropy H with respect to the feature vector $\boldsymbol{g}_{i,j}$ associated with the input region (i, j) in the input image:

$$A_{i,j} = \left\| \nabla_{\boldsymbol{g}_{i,j}} H \right\|_2 \qquad \text{where,} \qquad \boldsymbol{g}_{i,j} = f_g(I_{i,j}) \in \mathbb{R}^D$$

Hence, the whole attention map would be $\boldsymbol{A} \in \mathbb{R}^{s_1 \times s_2}$ for the whole image.

Using the attention map **A**, we select a set of k input region positions (i, j) corresponding to the $A_{i,j}$s with the k largest values.

We denote the selected set of positions and the selected patches by p^a and X^a, respectively:

$$p^a \in [1, s_1] \times [1, s_2] \text{ such that } \#p^a = k.$$

$$X^a = \{x_{i,j} | (i,j) \in p^a\}$$

❑ After selecting the salient patches (X^a) within the input image, the local deep network (L) will be applied only on those patches.

In the test time, the local feature representation $f_l(X^a)$ and the global feature representation $f_g(I)$ can be fused to create a refined representation of the whole image:

$$f_r(x) = \begin{cases} f_l(x_{i,j}), & \text{if } x_{i,j} \in X^a \\ f_g(x_{i,j}), & \text{otherwise,} \end{cases} \qquad \text{where } (i,j) \in [1, s_1] \times [1, s_2].$$

❑ Cost function for N triplet images:

$$J = \frac{1}{N} \sum_{i=1}^{N} [\left\| f_l(X_i^a) - f_l(X_i^{a+}) \right\|_2^2 - \left\| f_l(X_i^a) - f_l(X_i^{a-}) \right\|_2^2 + \alpha]_+$$

❑ Exploiting the gradient of the entropy as the saliency measure for our attention network encourages selecting the **input regions which have the maximum effect on the uncertainty of the model predictions**.

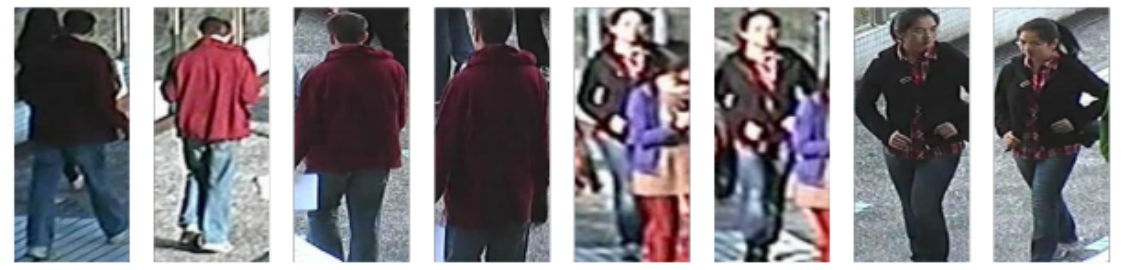
Experiments and Results

❑ The CUHK01 dataset contains 971 persons captured from two camera views in a campus environment. Each person has four images with two from each camera.

Comparison of performance of the proposed GAN to the state-of-the-art on CUHK01 dataset.

Method	Rank1	Rank5	Rank10	Rank20
FPNN (li2014,CVPR)	22.87	58.20	73.46	86.31
SDALF (farenzena,2010,CVPR)	9.90	41.21	56.00	66.37
eSDC (zhao,2013,CVPR)	22.84	43.89	57.67	69.84
KISSME (kostinger,2012,CVPR)	29.40	57.67	72.42	86.07
Partb-reid (cheng,2016,CVPR)	53.7	84.3	**91.0**	96.3
GAN-L	54.6	83.6	89.4	90.2
GAN	**64.2**	**86.4**	90.6	**96.9**

❑ The CUHK03 dataset contains 13164 images of 1360 identities. All pedestrians are captured by six cameras, and each person's image is only taken from two camera views.



Comparison of performance of the proposed GAN to the state-of-the-art on CUHK03 dataset.

Method	Rank1	Rank5	Rank10	Rank20
Imp-reid (ahmed,2015,CVPR)	54.74	86.50	**93.88**	**98.10**
FPNN (li,2014,CVPR)	20.65	51.50	66.50	80.00
SDALF (farenzena,2010,CVPR)	5.60	23.45	36.09	51.96
eSDC (zhao, 2013,CVPR)	8.76	24.07	38.28	53.44
KISSME (kostinger, 2012,CVPR)	14.17	48.54	52.57	70.53
GAN-L	60.5	82.2	88.8	91.5
GAN	**61.2**	**89.1**	91.3	93.9

Conclusion

✓The proposed model learns to focus selectively on parts of the input image for which the networks' output is most sensitive to.

✓Unlike the previous works, the proposed attention model can be trained with back propagation and it does not require a policy network (such as reinforcement learning) for training. Also, there is no need of using the Long Short Term Memory (LSTM) for the attention model which makes the training process even easier.

✓ Thanks to the computational efficiency resulting from the attention architecture, we would be able to train deeper neural networks and use large high resolution input images in order to obtain higher accuracy in the re-identification task.

✓Our attention-based Re-id eliminates the adverse effect of irrelevant parts in the image (e.g., background clutter).