Deep Neural Networks with Inexact Matching for Person Re-Identification

Arulkumar Subramaniam, Moitreya Chatterjee, Anurag Mittal
Indian Institute of Technology Madras

Objectives

Problem definition : Person Re-Identification is the task of identifying a particular person across multiple images captured from the same/different cameras, from different viewpoints, at the same/different points in time.

Contributions : Two Deep Convolutional Neural Network (DCNN) architectures:

1. A Siamese network with novel Normalized correlation (Inexact matching) layer and Wider search space
2. A Fused model using Normalized correlation layer and a state of the art exact matching technique (Ahmed et al. CVPR-2015)

Challenges in Person Re-Identification

Viewpoint Change | Illumination Variation | Partial Occlusion

Prior approaches

Deep learning based approaches:

Yi et al.(ICPR-2014) : Two-input network → 3 stages of convolution (shared weights) → Cosine similarity score

Li et al.(CVPR-2014) : Two-input network → 1 stage convolution (non-shared weights) → take the product of the responses obtained from first set of convolutions corresponding to the two inputs → Maxout grouping → 1 stage convolution (shared weights) → Softmax classifier (same / different)

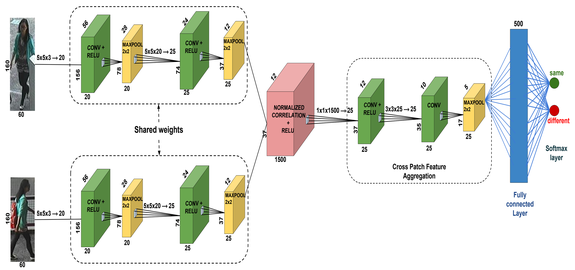
Ahmed et al.(CVPR-2015) : Siamese network → 2 stages of convolution (shared weights) → Cross-input neighborhood matching → 2 stages of convolution (non-shared weights) → Softmax classifier (same / different)

Solutions proposed in our work

Challenge	Solution
Illumination variation	Normalized correlation between pixel's patch neighborhood
Pose / viewpoint variations	Wider search space
Partial occlusion	Normalized correlation (Inexact matching) + Wider search space

Code: `https://github.com/InnovArul/personreid_normxcorr`

NormXcorr model (∼1.12M parameters)



Normalized correlation (Inexact Matching) layer

Given two corresponding input feature maps $X(12 \times 37)$ and $Y(12 \times 37)$ after first two convolution layers, we compute the normalized correlation as follows.

- We start with every pixel of X located at (x, y), where x is along the width and y along the height (denoted as $X(x, y)$).
- We then create two matrices. The first is a 5×5 matrix representing the 5×5 neighborhood of $X(x, y)$, while the second is the corresponding 5×5 neighborhood of Y centered at (a, b), where $1 \leq a \leq 12$ and $y - 2 \leq b \leq y + 2$.
- We perform **inexact matching** over a **wider search space**, by computing a Normalized Correlation between the two patch matrices.

Given two matrices, E and F, whose elements are arranged as two N-dimensional vectors, Normalized Correlation is given by:

$$normxcorr(E, F) = \frac{\sum_{i=1}^{N}(E_i - \mu_E) * (F_i - \mu_F)}{(N-1) * \sigma_E * \sigma_F},$$

where μ_E, μ_F denotes the means of the elements of the 2 matrices E and F respectively, while σ_E, σ_F denotes their respective unbiased standard deviation (a small $\epsilon = 0.01$ is added to the unbiased standard deviation to avoid division by 0).

- The mean of a N-dimensional vector E, $\mu_E = \frac{\sum_{i=1}^{N} E_i}{N}$
- The unbiased standard deviation of a N-dimensional vector E, $\sigma_E = \sqrt{\frac{\sum_{i=1}^{N}(E_i - \mu_E)^2}{N-1}}$

For every pair of feature maps X and Y, this gives us 60 feature maps of dimension 12×37 each. For all 25 pairs of maps that are input to the Normalized Correlation layer, we obtain an output of 1500, 12×37 maps.

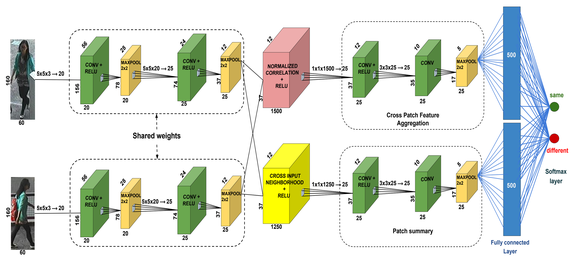
Gradient of Normalized correlation

The gradient for the Normalized Correlation layer is calculated as:

$$\frac{\partial normxcorr(E,F)}{\partial E_i} = \frac{1}{(N-1) * \sigma_E} * \left(\frac{F_i - \mu_F}{\sigma_F} - \frac{normxcorr(E,F) * (E_i - \mu_E)}{\sigma_E} \right) \quad (1)$$

$$\frac{\partial normxcorr(E,F)}{\partial F_i} = \frac{1}{(N-1) * \sigma_F} * \left(\frac{E_i - \mu_E}{\sigma_E} - \frac{normxcorr(E,F) * (F_i - \mu_F)}{\sigma_F} \right) \quad (2)$$

Fused model (∼2.22M parameters)



Fused Model

We fuse our novel Normalized correlation layer with CrossInput Neighborhood(CIN) layer (Ahmed et al.[1]) to overcome occasional false matches due to wider search space.

CrossInput Neighborhood(CIN) layer: Given two corresponding input feature maps $X(12 \times 37)$ and $Y(12 \times 37)$ after first two convolution layers, for every pixel of X located at (x, y), where $1 \leq x \leq 12$ and $1 \leq y \leq 37$, the CIN is calculated as ,

$$CIN(x, y) = X(x, y)\mathbb{1}(5, 5) - \mathcal{N}(Y(x, y))$$

where $\mathcal{N}(Y(x, y))$ is the 5×5 neighborhood of pixel $Y(x, y)$. Unlike Normalized correlation, CIN is an asymmetric operation and needs to be computed in both directions $(X \rightarrow Y)$ and $(Y \rightarrow X)$.

Results

Table: CUHK03 (labeled & detected) datasets

Method	r = 1	r = 10	r = 20
Fused Model (ours)	**72.43**	**95.51**	**98.40**
Norm X-Corr (ours)	64.73	92.77	96.78
Ensembles	62.1	92.30	97.20
LOMO+MLAPG	57.96	94.74	98.00
Ahmed et al.	54.74	93.88	98.10
LOMO+XQDA	52.20	92.14	96.25
Li et al.	20.65	68.74	83.06

Method	r = 1	r = 10	r = 20
Fused Model (ours)	**72.04**	**96.00**	**98.26**
Norm X-Corr (ours)	67.13	94.49	97.66
LOMO+MLAPG	51.15	92.05	96.90
Ahmed et al.	44.96	83.47	93.15
LOMO+XQDA	46.25	88.55	94.25
Li et al.	19.89	64.79	81.14

Table: CUHK01 test100 & test486 datasets

Method	r = 1	r = 10	r = 20
Fused Model (ours)	**81.23**	**97.39**	**98.60**
Norm X-Corr (ours)	77.43	96.67	98.40
Ahmed et al.	65.00	93.12	97.20
Li et al.	27.87	73.46	86.31
KISSME	29.40	72.43	86.07
LDML	26.45	72.04	84.69

Method	r = 1	r = 10	r = 20
Fused Model (ours)	**65.04**	**89.76**	**94.49**
Norm X-Corr (ours)	60.17	86.26	91.47
CPDL	59.5	89.70	93.10
Ensembles	51.9	83.00	89.40
Ahmed et al.	47.50	80.00	87.44

Table: QMUL GRID Dataset

Method	r = 1	r = 5	r = 10	r = 20	Deep Learning Model
Fused Model (ours)	**19.20**	38.40	**53.6**	**66.4**	Yes
Norm X-Corr (ours)	16.00	32.00	40.00	55.2	Yes
KEPLER	18.40	**39.12**	50.24	61.44	No
LOMO+XQDA	16.56	33.84	41.84	52.40	No
PolyMap	16.30	35.80	46.00	57.60	No
MtMCML	14.08	34.64	45.84	59.84	No

Ablation study

Dataset	NormXcorr(ours), 5x5 search			NormXcorr(ours), 5x12 search			Ahmed et al.[1], 5x5 search			Ahmed et al.[1], 5x12 search		
	r=1	r=10	r=50	r=1	r=10	r=50	r=1	r=10	r=50	r=1	r=10	r=50
CUHK03 labeled	62.43	92.22	99.60	**64.73**	92.77	99.60	54.74	**93.30**	**99.70**	57.60	90.63	99.17
CUHK03 detected	63.12	92.76	99.20	**67.13**	**94.49**	**99.73**	44.96	83.47	99.40	54.31	90.24	99.18
CUHK01 test100	72.3	95.80	99.60	**77.43**	**96.67**	99.29	65.00	94.00	**99.90**	69.70	95.03	99.13
CUHK01 test486	56.79	84.43	95.95	**60.17**	**86.26**	**96.44**	47.50	80.25	96.30	49.31	81.48	95.95

Illumination invariance

Viewpoint change

Partial occlusion

Occasional false matches

References

[1] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3908–3916, 2015.

Acknowledgements

This work is partly supported by travel grant from Google to Arulkumar Subramaniam.