Norm-Aware Embedding for Efficient Person Search

Di Chen[1,3] Shanshan Zhang[1] Jian Yang[1] Bernt Schiele[2]

[1]Nanjing University of Science and Technology, Nanjing, China [2]Max Planck Institute for Informatics, Saarbrücken, Germany

Task & Challenge

Person Search:

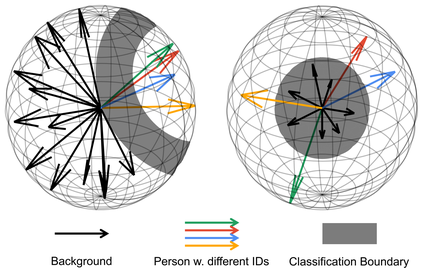
- Find a query person among a set of scene images
- A hybrid task of *pedestrian detection* and *re-identification*

The Challenge:

- Contradictory objectives of detection (find person commonness) and re-ID (find person uniqueness)
- Relatively low-quality alignment for RCNN is harmful to re-ID.

Motivation

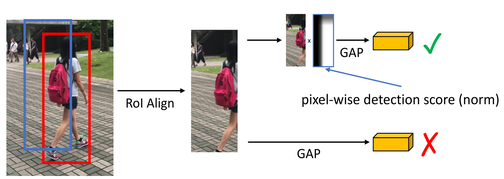
1. Use feature norm for detection; angle for re-ID



Left: For L_2 normalized embeddings, the inter-class angle distances for different persons are squeezed by backgrounds.

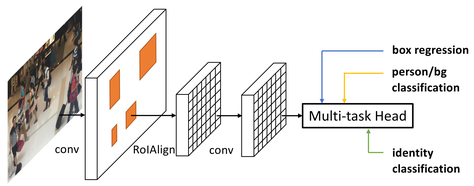
Right: Norm-aware embeddings separate persons and background by norms and discriminate person identities by angles, thus the constrain on inter-class distances is relaxed.

2. Pixel-wise det score as attention

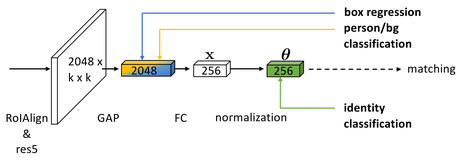


Network Architecture

Overall:

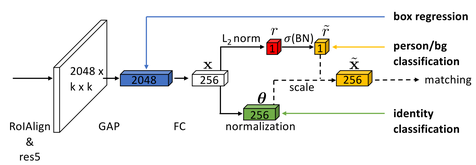


Head design for baseline[1]:



- **box regression** and **classification** remain the same as Faster R-CNN
- an additional L_2 normalized fully-connected layer is added upon the top layer for **identity embedding generation**

Head design for NAE:



Decompose feature into norm and angle

$$\mathbf{x} = r \cdot \boldsymbol{\theta}$$

Re-scale norm with BN and sigmoid

$$\tilde{r} = \sigma\left(\frac{r - \mathbb{E}[r]}{\mathrm{Var}[r] + \epsilon} \cdot \gamma + \beta\right)$$

Assemble as the norm-aware embedding

$$\tilde{\mathbf{x}} = \tilde{r} \cdot \boldsymbol{\theta}$$

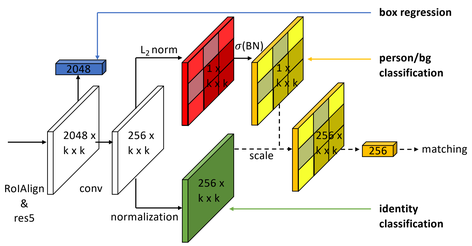
When matching

$$\mathrm{sim}(\tilde{\mathbf{x}}_q, \tilde{\mathbf{x}}_g) = \tilde{\mathbf{x}}_q^T \tilde{\mathbf{x}}_g = \tilde{r}_g \cdot \tilde{\boldsymbol{\theta}}_q^T \tilde{\boldsymbol{\theta}}_g$$

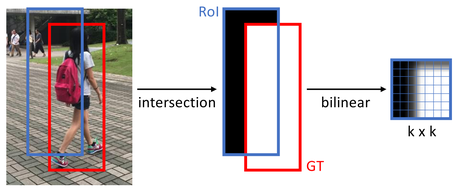
Sharing the same formation as Class Weighted Similarity (CWS)[2]

NAE+: Pixel-wise Extension

Segmentation-like classification before global average pooling:



Pseudo-label generation



Training Losses

- RCNN detection loss: Binary cross-entropy loss over the embedding norm

$$\mathcal{L}_{\mathrm{det}} = -y \log(\tilde{r}) - (1 - y) \log(1 - \tilde{r})$$

$$\mathcal{L}_{\mathrm{det+}} = -\frac{1}{k^2} \sum_{i=1}^{k^2} y_i \log(\tilde{r}_i) + (1 - y_i) \log(1 - \tilde{r}_i)$$

- RCNN re-ID loss: OIM loss[1] over the embedding angle

$$\mathcal{L}_{\mathrm{reID}} = -\mathbf{1}_{\mathrm{ID}=i} \cdot \log\left(\frac{e^{\mathbf{w}_i \boldsymbol{\theta}}}{\sum_{j=1}^{N+M} e^{\mathbf{w}_j \boldsymbol{\theta}}}\right)$$

- Other losses remain the same as Faster RCNN's
- Final loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rpn\text{-}det}} + \mathcal{L}_{\mathrm{rpn\text{-}reg}} + \mathcal{L}_{\mathrm{det}} + \mathcal{L}_{\mathrm{reg}} + \mathcal{L}_{\mathrm{reID}}$$

For NAE+, replace $\mathcal{L}_{\mathrm{det}}$ with $\mathcal{L}_{\mathrm{det+}}$

Dataset

- **CUHK-SYSU**[1]
  - 11206 images, 5532 identities for training
  - 2900 query persons, 6978 gallery images
  - Several gallery subsets with different sizes
- **PRW**[2]
  - 5704 images, 482 identities for training
  - 2057 query persons, 6112 gallery images
  - Search among the whole gallery

Ablation Study

Disentangled investigation on detection and re-ID

Detector	Recall	AP	Re-idetntifier	mAP	top-1
OIM-base	89.3	79.7	OIM-base	84.4	86.1
			NAE	90.0	91.8
NAE	92.6	86.8	OIM-base	85.9	87.6
			NAE	91.5	92.4
GT	100	100	OIM-base	90.7	91.2
			NAE	93.5	94.0

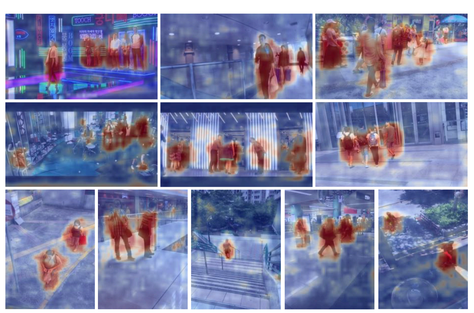
Conclusions

1. Detection quality of NAE is better
2. With the same detection, NAE is more discriminative for re-identification.

Efficacy of CWS ✓

Method	mAP	top-1	ΔmAP	Δtop-1
OIM-base	84.4	86.1		
OIM-base w/ CWS	87.1	88.5	+2.7	+2.4
NAE	91.5	92.4		
NAE w/o CWS	89.9	91.3	-1.6	-1.1

NAE+ visualization



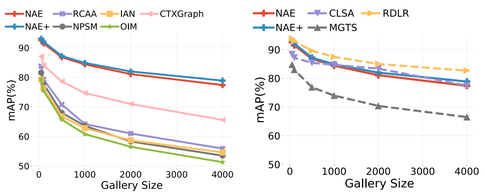
Comparison to the state-of-the-arts

Quantitative results:

- Best among one-step methods
- Comparable to two-step methods

	Method	CUHK-SYSU mAP	CUHK-SYSU top-1	PRW mAP	PRW top-1
one-step	OIM [44]	75.5	78.7	21.3	49.9
	IAN [42]	76.3	80.1	23.0	61.9
	NPSM [26]	77.9	81.2	24.2	53.1
	RCAA [2]	79.3	81.3	-	-
	CTXGraph [47]	84.1	86.5	33.4	73.6
	QEEPS [30]	88.9	89.1	37.1	76.7
	OIM-base (ours)	84.4	86.1	34.0	75.9
	NAE (ours)	91.5	92.4	43.3	80.9
	NAE+ (ours)	**92.1**	**92.9**	**44.0**	**81.1**
two-step	DPM+IDE [61]	-	-	20.5	48.3
	CNN+MGTS [3]	83.0	83.7	32.6	**72.1**
	CNN+CLSA [21]	87.2	88.5	38.7	65.0
	FPN+RDLR [16]	**93.0**	**94.2**	**42.9**	70.2

mAP vs. gallery size:



Speed comparison

GPU (TFLOPs)	MGTS	QEEPS	NAE	NAE+
K80 (4.1)	1269	-	663	606
P6000 (12.6)	-	300	-	-
P40 (11.8)	-	-	158	161
V100 (14.1)	-	-	83	98

References

[1] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In CVPR, 2017.

[2] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In CVPR, 2017.

Code is available at: https://github.com/DeanChan/NAE4PS