Decoupled Self-attention Module for Person Re-identification

Chao Zhao, Zhenyu Zhang, Jian Yan, Yan Yan
Nanjing University of Science and Technology
zhch@njust.edu.cn

Abstract

Person re-identification aims to identify the same person from different cameras, which needs to integrate whole-body information and capture global correlation. However, convolutional neural network is able to only capture short-distance information because of the size of filters. Self-attention is introduced to capture long-distance correlation, but inner-product similarity calculation in self-attention mingles semantic response and semantic difference together. Semantic difference is more important for person re-identification, because it is robust to illumination without the effect of semantic response. However, we find the scale of norms measuring semantic response is much larger than angle measuring semantic difference by decoupling inner-product similarity into norms and angle. To balance the importance of semantic response and semantic difference in self-attention, we propose the decoupled self-attention module for person re-identification to make the most of self-attention. Extensive experiments show that the decoupled self-attention module obtains significant performance with easier convergence and stronger robustness.

Background

- traditional self-attention mechanism does not always work and its performance is **not stable** for person Re-ID.
- **Semantic response** and **semantic difference** have totally different meanings, but they are **mingled** together in similarity calculation in self-attention, making the loss of independence.

Motivation

- Activation value in a location on feature maps can reflect whether this location contains some kind of semantic information, and semantic similarity can tell us whether there is the same semantic information in two different locations.
- Semantic difference is more important for person Re-ID because of its robustness to illumination and noise from background, but the scale of the norms and the angle is obviously different.

$$Sim(F_i, G_j) = \|F_i\| \|G_j\| \cos\langle F_i, G_j \rangle$$

Method

The working mechanism of self-attention

The traditional form of self-attetion (SA) is

$$SA(F, G, H) = softmax\left(FG^{\top}\right) H$$

The similarity between F_i and G_j is

$$Sim(F_i, G_j) = F_i G_j^{\top} = \langle F_i, G_j \rangle$$

Thus, SA has a form of

$$Y = \begin{bmatrix} \langle F_1, G_1 \rangle & \cdots & \langle F_1, G_{hw} \rangle \\ & \ddots & \vdots \\ \langle F_{hw}, G_1 \rangle & & \langle F_{hw}, G_{hw} \rangle \end{bmatrix} \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_{hw} \end{bmatrix}$$

For location i($1 \le i \le hw$), the feature after SA will be

$$Y_i = \sum_{j=1}^{hw} (F_i G_j^{\top}) H_j = \sum_{j=1}^{hw} \langle F_i, G_j \rangle H_j$$
$$= \|F_i\| \sum_{j=1}^{hw} \|G_j\| \cos\langle F_i, G_j \rangle H_j$$

The "attention" location is the location which has high activation value, and has a high correlation with those locations having high activation values. Concretely, the norms of F_i and G_j measure response (activation) degree in locations I and j on all channels, yet $\cos\langle F_i, G_j \rangle$ measures semantic difference (correlation).

The decoupled self-attention module

In order to balance the scale of norms and angle in inner-product similarity, and better model semantic response and semantic correlation for person re-identification, we introduce a generic form of the decoupled self-attention module as follows:

$$Y_i = \sum_{j=1}^{hw} softmax(f(\|F_i\|, \|G_j\|) \cdot g(\cos\langle F_i, G_j \rangle)) H_j$$

The functions f and g is introduced to generalize the form of similarity, and balance the scale of norms and cosine of the angle.

Experiments

TABLE I
COMPARISON OF DIFFERENT FORMS OF THE DECOUPLED SELF-ATTENTION MODULE WITH THE BACKBONE NETWORK OF RESNET50 ON MARKET-1501 AND DUKEMTMC-REID. WE COMBINE DIFFERENT ACTIVATION FUNCTIONS OF NORMS AND ANGLE.

Models	Market-1501				DukeMTMC-ReID			
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
ResNet Baseline	87.6	95.4	98.0	71.9	79.4	88.9	94.3	60.6
LogNorm+Cosine	**92.3**	97.0	98.7	75.2	**82.8**	90.8	94.8	62.6
ScaledNorm+Cosine	92.0	96.4	98.8	75.4	82.4	**91.3**	**95.0**	**64.1**
NonNorm+Cosine	91.8	**97.3**	**98.9**	**75.8**	82.7	91.1	94.8	63.5
LogNorm+Sqcosine	91.8	96.6	98.5	75.4	81.9	90.4	94.0	62.8
ScaledNorm+SqCosine	91.3	96.9	98.7	75.4	82.4	**91.4**	**94.9**	**63.9**
NonNorm+SqCosine	**92.3**	**97.1**	**98.8**	**75.8**	**82.7**	90.8	**94.9**	63.6

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART MODELS ON MARKET-1501 AND DUKEMTMC-REID, WHERE ECN IS A DOMAIN ADAPTATION MODEL.

Models	Market		DukeMTMC	
	rank-1	mAP	rank-1	mAP
ECN [35]	75.1	43.0	63.3	40.4
ECN+Decoupled module	**75.5**	**43.8**	**63.9**	**40.5**
OSNet [36]	94.8	84.9	88.6	73.5
OSNet+Decoupled module	**95.1**	**85.2**	**88.8**	**75.3**
MHN [12]	94.8	85.2	89.5	77.5
MHN+Decoupled module	**95.1**	**85.9**	**89.8**	**78.1**

TABLE III
COMPARISON OF DIFFERENT FORMS OF THE DECOUPLED SELF-ATTENTION MODULE WITH THE BACKBONE NETWORK OF DENSENET121 ON MARKET-1501 AND DUKEMTMC-REID. WE COMBINE DIFFERENT ACTIVATION FUNCTIONS OF NORMS AND ANGLE.

Models	Market-1501				DukeMTMC-ReID			
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
DenseNet Baseline	90.5	96.1	98.4	73.8	81.2	90.1	94.3	61.4
LogNorm+Cosine	90.9	96.5	98.5	74.7	82.6	**90.8**	94.4	62.2
ScaledNorm+Cosine	**91.1**	**96.8**	**98.7**	**74.9**	82.2	**90.8**	94.3	**63.6**
NonNorm+Cosine	**91.1**	96.6	98.6	**74.9**	**82.8**	90.3	**94.8**	63.1
LogNorm+Sqcosine	90.7	96.5	98.5	74.5	81.8	90.9	94.5	63.0
ScaledNorm+SqCosine	**91.7**	**96.7**	**98.6**	**75.9**	**82.1**	**91.1**	94.5	63.2
NonNorm+SqCosine	91.0	96.3	98.4	75.1	**82.1**	90.6	**94.6**	**63.4**

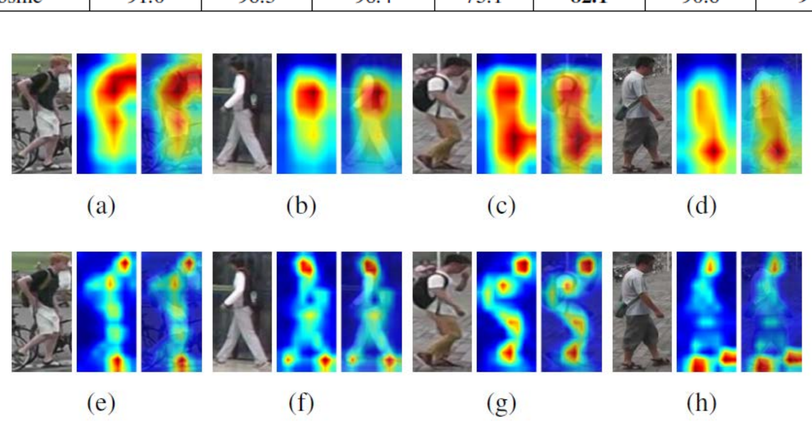

Fig 1: The first row is activation maps without the decoupled self-attention module, and the second row is activation maps with the decoupled self-attention module.

Acknowledgements

This work was supported by the National Science Fund of China under Grant No. 61806094, No. U1713208.