Learning a Key-Value Memory Co-Attention Matching Network for Person Re-Identification

Yaqing Zhang, Xi Li,* Zhongfei Zhang
Zhejiang University, Hangzhou, China
{yaqing, xilizju, zhongfei}@zju.edu.cn

Abstract

Person re-identification (Re-ID) is typically cast as the problem of semantic representation and alignment, which requires precisely discovering and modeling the inherent spatial structure information on person images. Motivated by this observation, we propose a Key-Value Memory Matching Network (KVM-MN) model that consists of key-value memory representation and key-value co-attention matching. The proposed KVM-MN model is capable of building an effective local-position-aware person representation that encodes the spatial feature information in the form of multi-head key-value memory. Furthermore, the proposed KVM-MN model makes use of multi-head co-attention to automatically learn a number of cross-person-matching patterns, resulting in more robust and interpretable matching results. Finally, we build a setwise learning mechanism that implements a more generalized query-to-gallery-image-set learning procedure. Experimental results demonstrate the effectiveness of the proposed model against the state-of-the-art.

Introduction

As an important and challenging problem in computer vision, person re-identification (Re-ID) focuses on effectively matching persons across non-overlapping cameras, and has a wide range of applications such as cross-camera person tracking and target person search. Typically, a key challenge for Re-ID is how to precisely align person images in semantics. To address this problem, many efforts have been devoted to taking into account auxiliary prior information for misalignment reduction such as pose-driven approaches(Su et al. 2017; Sarfraz et al. 2017) and part-segmented models(Qi et al. 2018; Kalayeh et al. 2018). Usually, such prior information is expensive to obtain in practice, and relies on scenario-specific settings with a certain inflexibility. Motivated by this observation, we concentrate on designing an end-to-end learning scheme with the capability of adaptively discovering the inherent matching structures in a totally data-driven fashion.

As shown in Fig. 1, precise semantic matching for the paired images is confronted with many difficulties like local-position-sensitive correspondence (e.g., bag). Therefore, it is necessary to set up an effective local-position-aware person representation as well as its associated matching mechanism. In this paper, we propose a Key-Value Memory Matching Network (KVM-MN) that comprises the modules of key-value memory representation and key-value co-attention matching. Specifically, the key-value memory representation module is in charge of encoding the spatial feature information in the form of multi-head key-value memory, where the key indicates the presence index of a feature component while the value stands for its detailed feature attribute information. In this way, a person image is naturally represented as a set of local-position-aware key-value pairs. Based on these key-value pairs, the matching problem for a given person image pair is converted to that of dense cross-matching between the person-specific key-value pairs. In order to make cross-matching more robust and interpretable, the key-value co-attention matching module further makes use of multi-head co-attention to automatically learn a collection of cross-person-matching patterns. Consequently, the above two modules are jointly learned in a unified end-to-end framework. For the sake of effective matching network learning, we build a setwise learning mechanism that supports the joint learning task with a query image and a set of gallery images. The mechanism is capable of adaptively selecting the matched gallery image from the set with a certain matching confidence, resulting in a more flexible learning pipeline against pairwise or triplet training (Cheng et al. 2016a; Ahmed, Jones, and Marks 2015).

In summary, the main contributions of this work are threefold. Firstly, we propose a dense local-position-aware key-value memory representation, which effectively encodes the spatial structure information on person images. Secondly, we present a Key-Value Memory Matching Network that fully utilizes multi-head co-attention to adaptively discover a set of inherent cross-person-matching patterns over key-value pairs. Thirdly, we introduce a setwise learning mechanism that performs the query-to-gallery-image-set learning procedure in a more flexible manner.

Related Works

Matching in Re-ID: Many recent works on person re-identification focus on finding the spatial relationships of the paired images. People have adopted a bunch of techniques to align the paired images for comparison, fixed or learning-based. The approaches with fixed alignment models often

*Corresponding author

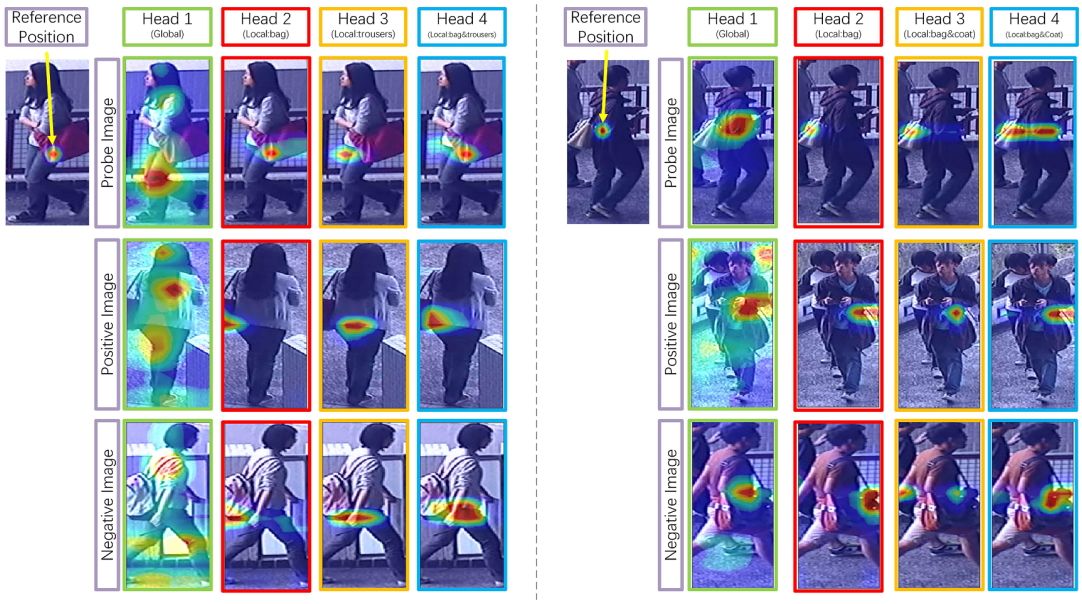

Figure 1: Illustration of multi-head co-attention of a specific location ($H = 4$). For a given reference location, we obtain different matching patterns of different heads using co-attention. Specifically, Head 1 tends to attend global information of the image while Heads 2,3,4 are likely to find local correspondences in this trained model. Also, when a reference position stays near multiple feature components, the model may attend them using different heads to achieve more robust matching structures and reduce confusions. Best viewed in color.

involve a pre-trained network to extract key information for alignment, such as poses (Su et al. 2017; Sarfraz et al. 2017; Liu, Zhao, and Wu 2018; Zhao et al. 2017a), and part segmentations (Suh et al. 2018; Kalayeh et al. 2018). These approaches may suffer from domain biases due to different settings between scenes and cameras and lead to incorrect matchings for person re-identification. Learning-based approaches for alignment try to enable the network to find the matching structures adaptively from training data. Most of these methods involve image matching with spatial constraints that ensure matchings occuring at the neighbor of the reference region, and they are roughly divided into neighborhood-based matching (Ahmed, Jones, and Marks 2015; Lin et al. 2017; Zhang et al. 2016), row-based matching (Li et al. 2014; Zhang et al. 2017), and sub-region-based matching (Zhou et al. 2017; Chang, Hospedales, and Xiang 2018). In this paper, we focus on a more generic matching structure that does not rely on any constraints of spatial relationships and can capture the long-range precise dependencies between paired images.

Metric learning in Re-ID: Metric learning is another core process in the person re-identification problem. In person re-identification, metric learning mainly focuses on minimizing the intra-personal variance while maximizing the inter-personal margin, including classification-based approaches and margin-based approaches. In classification-based approaches (Zhang et al. 2016; Ahmed, Jones, and Marks 2015; Qian et al. 2017), people employ the learned metric function to classify whether a given image pair belongs to the same person or not. The margin-based methods learn discriminative feature representations with generic matching metrics like L2 distance. They design contrastive loss (Chen et al. 2017; Guo and Cheung 2018) or triplet loss (Cheng et al. 2016b; Paisitkriangkrai, Shen, and van den Hengel 2015; Zhao et al. 2017b) to learn the feature representations based on the distance between positive and negative image pairs. Recent efforts mainly focus on improving the ranking loss by introducing harder samples or by optimizing the margins between positive and negative pairs (Hermans, Beyer, and Leibe 2017; Cheng et al. 2016a). However, the existing metric learning methods based on pairwise loss and triplet loss often suffer from slow convergence and overfitting (Sohn 2016). In this paper, we focus on building and learning from a flexible non-parametric matching structure between the probe image and all the images in gallery set and obtain a robust similarity metric.

Key-Value Memory Matching Network

Person re-identification (Re-ID) aims at finding the person of the same identity given a probe image. In this paper, we seek to adaptively find the inherent and fine-grained matching structure between person images to help similarity measurement. To achieve this goal, we propose a Key-Value Memory Matching Network (KVM-MN), as shown in Fig. 2. The proposed network mainly consists of two modules: key-value memory representation and key-value co-attention matching. In this section, we first describe how to represent the image as a set of multi-head key-value vector pairs that encode multiple aspects of spatial structure information. Based on these representations, a robust attention-based matching mechanism is built upon the key-value memory representations. We make use of multi-head co-attentions to adaptively learn a

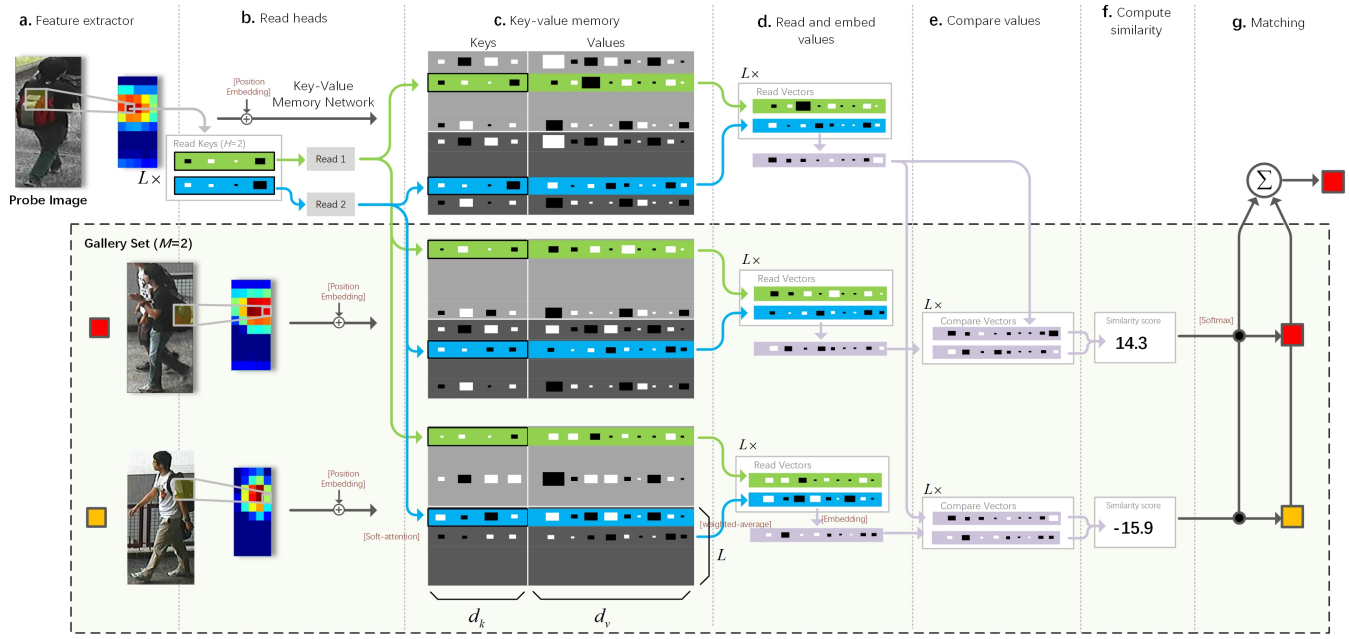

Figure 2: Key-Value Memory Matching Network (KVM-MN) for person re-identification. Each image can be represented using a key-value matrix pair corresponding to the extracted feature maps. Given a common head from the probe view, we address and read required values related to the key for further comparison and similarity measurement. The probe image is matched with all the images in the gallery set by using Softmax over similarity scores and finding the correct match.

set of matching patterns between persons and output the final similarity. The proposed KVM-MN could be further benefited using a set-based learning schema to fully capture the matching structure between the probe and a set of gallery images.

Position-aware Key-Value Memory Representation

Given a person image $\mathbf{I}$, we are interested in how to represent its local feature effectively in order to conduct dense cross-matching. In KVM-MN, we investigate the effectiveness of key-value memory network for structure representation. For the extracted convolutional feature maps $\mathbf{F} \in \mathbb{R}^{H\times W\times d}$ of $\mathbf{I}$, we build a set of key-value vectors pairs $\{(k_l, v_l)\}$ that encode the local structure information. Specifically, the key is denoted as the index of a specific feature component and the value represents its detailed attributes and appearances.

Position embedding: To capture local-position-sensitive correspondences, we first need to inject some information on the absolute and relative positions of the local feature f_l. Inspired by the idea of position embedding (Gehring et al. 2017; Vaswani et al. 2017) used for word embeddings, we add sinusoidal position embedding p with length d to each location:

$$\begin{aligned} p_l\,(2i) &= \sin(x/10000^{4i/d}) \\ p_l\,(2i+1) &= \cos(x/10000^{4i/d}) \\ p_l\,(2i+{}^{d}\!/_{2}) &= \sin(y/10000^{4i/d}) \\ p_l\,(2i+1+{}^{d}\!/_{2}) &= \cos(y/10000^{4i/d}) \end{aligned} \tag{1}$$

where (x, y) corresponds to the absolute position in the feature maps of the l-th location and $i = 1, 2, ..., d/4$ is to represent the dimension. We use a vector of length $d/2$ to encode the x and y positions respectively and add the feature vector to the feature representation: $f_l \leftarrow f_l + p_l$.

Multi-head key-value representation: Using the position embedding technique, it is much convenient for key-value memory networks to encode the spatial feature information. To comprehensively capture multiple feature components, we employ a multi-head key-value memory network that projects the original feature vectors into different key-value subspaces to capture various underlying local patterns. This is achieved by simply using feed-forward networks parameterized by a set of weight matrices $\{W_{k,h} \in \mathbb{R}^{d\times d_k}\}_{h=1}^{H}$ and $\{W_{v,h} \in \mathbb{R}^{d\times d_v}\}_{h=1}^{H}$ with totally H heads:

$$\begin{aligned} K_h &= \varphi(FW_{k,h}) \\ V_h &= \varphi(FW_{v,h}) \end{aligned} \tag{2}$$

where φ denotes the ReLU activation after Batch Normalization (Ioffe and Szegedy 2015) and $F \in \mathbb{R}^{L\times d}$ is the unrolled representation of $\mathbf{F}$.

Multi-Head Co-Attention for Dense Matching

Given the paired images $(\mathbf{I}^A, \mathbf{I}^B)$ from different views, we extract their position-aware key-value memory representation in Equ. 2, written as $\{(K_h^A, V_h^A)\}_{h=1}^{H}$ and $\{(K_h^B, V_h^B)\}_{h=1}^{H}$ respectively. Based on the memory representations, the matching problem of $(\mathbf{I}^A, \mathbf{I}^B)$ is converted to that of dense matching between these paired key-value memory matrices.

To make the matching process more effective and efficient, in this paper, we explore the capacity of attention

mechanisms to precisely associate the paired images. Different from conventional attention models (Xu et al. 2015; Sharma, Kiros, and Salakhutdinov 2015), the proposed method instead searches attentional regions in both images to find the matching using a common head. Various heads formulate a set of matching patterns that connect the paired images in different ways. So we call the attention mechanism **"co-attention"** as it seeks to find "where to compare" to associate the paired images. Some typical co-attentions are illustrated in Fig. 1.

Co-attention: As shown in Fig. 2, co-attention aims at mapping a head $k_h^r \in \mathbb{R}^{d_k}$ and paired memory matrices to an output pair. We employ content addressing to obtain attentional distributions over the L memory locations and read the required values by weighted averaging over values:

$$\begin{aligned} r_h^A &= \text{softmax}\left(K_h^A k_h^r/\mu\right)^T V_h^A \\ r_h^B &= \text{softmax}\left(K_h^B k_h^r/\mu\right)^T V_h^B \end{aligned} \tag{3}$$

where $\mu = \sqrt{d_k}$ is a scalar to prevent the dot-product value to be large.

Fusing multi-head co-attention: By Equ. 3, we can obtain co-attentions of H heads written as $r_1^A, ..., r_H^A$ and $r_1^B, ..., r_H^B$. They are concatenated and once again embedded to produce the output using the weight matrix $W_o \in \mathbb{R}^{Hd_v \times d_o}$:

$$\begin{aligned} o^A &= \varphi\left({W_o}^T \left[r_1^A, ..., r_H^A\right]\right) \\ o^B &= \varphi\left({W_o}^T \left[r_1^B, ..., r_H^B\right]\right) \end{aligned} \tag{4}$$

where $[\cdot]$ indicates the concatenation operation, and the pair (o^A, o^B) encodes the multi-head co-attention of a set of reference keys $\{k_h^r\}$.

Comparing co-attentions: The co-attention representations o^A and o^B of Equ. 4 encode the visual representations with common heads. To compare the values, we employ a metric function to capture the interactions between them with $W_c \in \mathbb{R}^{2d_o \times d_c}$:

$$c = \varphi\left(W_c^T \left[o^A, o^B\right]\right) \tag{5}$$

where $c \in \mathbb{R}^{d_c}$ encodes the comparison representations given the heads $\{k_h^A\}$.

Scoring with dense matching: To perform dense cross-image matching, we use all the keys in $\{K_h^A\}$ to produce paired multi-head co-attention vectors and their comparison vector using Equ. 3 to Equ. 5, resulting in a comparison matrix denoted as $C \in \mathbb{R}^{L \times d_c}$, where the l-th row of C represents the comparison representation based on the heads $\{k_{h,l}^A\}_h$. We then perform a simple feed-forward network producing the final similarity score using the concatenation of comparison vectors:

$$f(\mathbf{I}^A, \mathbf{I}^B) = \text{FFN}\left([c_1, ..., c_L]\right) \tag{6}$$

where $\text{FFN}()$ is a simple neural network using two fully connected layers to output the final similarity score.

Matching Networks for Similarity Learning

Finally, we build a setwise learning mechanism that greatly boosts the training process of the similarity function. Inspired by the idea of Matching Networks (Vinyals et al. 2016) for one-shot learning, we further introduce a much flexible learning framework to learn the similarity metric as shown in Fig. 2. In detail, we construct a specific "training sample" as a combination of training images. Each combination consists of a probe image $\mathbf{I}^A$ and a gallery set $\{\mathbf{I}^{B_i}\}_{i=1}^M$, where the gallery set consists of totally M images of different identities labeled as 1 to M respectively including probe identity labeled as $p \in \{1, ..., M\}$. Given the probe image $\mathbf{I}^A$, the model adaptively interacts with gallery images $\{\mathbf{I}^{B_i}\}$ according to the similarity scores produced by $f(\mathbf{I}^A, \mathbf{I}^{B_i})$ so as to correctly classify the probe image $\mathbf{I}^A$ into one of the M classes in the gallery set. To achieve this goal, we first define the label distribution $\hat{y}^A \in \mathbb{R}^M$ of the probe sample in the gallery set:

$$\hat{y}^A = \sum_{i=1}^{M} \text{softmax}\left(f(\mathbf{I}^A, \mathbf{I}^{B_i})\right) y^{B_i} \tag{7}$$

where y^{B_i} is the one-hot label distribution of image $\mathbf{I}^{B_i}$ labeled as i, and $\text{softmax}(f(\mathbf{I}^A, \mathbf{I}^{B_i})) = \exp(f(\mathbf{I}^A, \mathbf{I}^{B_i}))/\sum_{j=1}^{M} \exp(f(\mathbf{I}^A, \mathbf{I}^{B_j}))$ is the matching probability of $\mathbf{I}^A$ in the gallery image set. Finally, we train the whole network by minimizing the cross-entropy error over the ground-truth label distribution y^A and the predicted distribution $\hat{y}^A$. We formulate the loss of each training sample as follows:

$$\ell = -\sum_{i=1}^{M} y^A(i) \log\left(\hat{y}^A(i)\right) \tag{8}$$

where y^A is the ground truth one-hot label distribution in the gallery set.

Experiments

To demonstrate the effectiveness of the proposed matching network, we evaluate our method on three popular datasets. Using the Tensorflow(Abadi et al. 2015) framework, it takes around 24 hours to train the network thoroughly on two NVIDIA GTX 1080Ti GPUs and 2.1ms for computing the similarity between the paired images. In this section, we first compare our method with state-of-the-art approaches on the three datasets. Also, we conduct ablation study to examine the effectiveness of each component.

Datasets

The method is evaluated on three public datasets, namely CUHK03 (Li et al. 2014), CUHK01 (Li, Zhao, and Wang 2012), and Market-1501 (Zheng et al. 2015). For clarity, we illustrate the settings and evaluation protocols of all the datasets to be fairly compared with other approaches in Table 2.

As for CUHK03, we train and evaluate the network using the "labeled" set. Following the standard process, we use two settings on the CUHK01 dataset with different train/test splits, as shown in Table 2. To avoid accidental results, experiments conducted on the CUHK03 and CUHK01 datasets

Table 1: Top recognition rate (%) of the various methods over CUHK03 labeled dataset with 100 test IDs, CUHK01 dataset with 100 test IDs, CUHK01 dataset with 486 test IDs with rank = 1, 5, 10, and over Market-1501 with rank = 1 and mAP. The method with "*" means that it involves extra dataset for training.

Method	CUHK03 (labeled)			CUHK01 (100 test IDs)			CUHK01 (486 test IDs)			Market-1501	
	r = 1	r = 5	r = 10	r = 1	r = 5	r = 10	r = 1	r = 5	r = 10	r=1	mAP
KVM-MN (ours)	**94.0**	**99.6**	**99.8**	**96.9**	**99.97**	**99.99**	**84.4**	**95.9**	**98.6**	**91.5**	**78.0**
KVM-MN-noAug (ours)	91.7	98.7	99.4	94.2	99.2	99.5	82.0	94.4	97.3	89.1	74.8
Spindle Net*(Zhao et al. 2017a)	88.5	97.8	98.6	-	-	-	79.9	94.4	97.1	76.9	-
PDC*(Su et al. 2017)	88.7	98.6	99.2	-	-	-	-	-	-	84.1	63.4
MSCAN(Li et al. 2017)	74.2	94.3	97.5	-	-	-	-	-	-	80.3	57.5
PartAligned(Zhao et al. 2017b)	85.4	97.6	99.4	88.5	98.4	99.6	75.0	93.5	95.7	81.0	63.4
DCSL(Zhang et al. 2016)	80.2	97.7	99.2	89.6	97.8	98.9	76.5	94.2	97.5	-	-
JSTL*(Xiao et al. 2016)	75.3	-	-	-	-	-	66.6	-	-	-	-
ImprovedDL(Ahmed, Jones, and Marks 2015)	54.7	86.5	93.9	65.0	89.0	94.0	47.5	71.6	80.3	-	-
KISSME(Koestinger et al. 2012)	14.2	37.5	52.2	29.4	60.2	74.4	-	-	-	-	-
FPNN(Li et al. 2014)	20.7	50.9	67.0	27.9	59.6	73.5	-	-	-	-	-

Table 2: Datasets and settings in our experiments.

Dataset	**CUHK03**	**CUHK01**	**Market-1501**
# identities	1360	971	1501
# images	13,164	3,884	32668
# cam./ ID	2	2	6
# train IDs	1,160	871;485	750
# test IDs	100	100;486	751
evaluation protocol	CMC	CMC	top-1, mAP

are repeated with 10 random splits, and the results are reported by taking the average accuracies on these test splits. Market-1501 is a much larger person re-identification dataset that involves more misalignments, occlusions, and other variations.

Training the Network

We employ Inception-V1 (Szegedy et al. 2015) as our deep architecture for extracting the feature maps, and the pretrained model on the ILSVRC-2012-CLS image classification dataset (Russakovsky et al. 2014) is downloaded from TF-slim library [1] . Using the model, we extract the feature maps named as "Mixed_4f" of the input image. As a result, given input image shape of $224 \times 112 \times 3$, the shape of feature maps which we obtain is $14 \times 7 \times d$ with the output stride of 16 after 1×1 convolutions upon the output feature maps.

In our method, we use stochastic gradient descent for updating the weights of the network with the momentum of 0.9. We set the base learning rate as 0.01 and weight decay as 0.0002 to train around 250K steps with a batch size of 20 until the model converges. A stepping function is applied to decay the learning rate to 0.001 and 0.0001 after 60% and 80% of total training steps. As for KVM-MN, we use $d = 512$, $H = 4$ and $d_k = d_v = 64$ for the dimension of the feature vector, number of heads, and dimensions of key-value vectors respectively. The lengths of co-attention output vector d_o and comparison vector d_c is set equal to d. Unlike some other efforts to improve accuracy using extra dataset or re-ranking techniques, we **do not** apply any extra data or re-ranking techniques to examine the effectiveness of the proposed KVM-MN.

Collecting training samples: To train the network, we use $M = 15$ images in the gallery set for a selected probe image, where the negative samples are randomly collected from 100 nearest neighbors of the probe image measured by L2 distance and a positive sample is chosen from different views of the probe. As a result, we collect around 500K combinations as training samples.

Data augmentation: To generate more training images and increase the robustness of the trained model, we perform data augmentation including random rotation of the angle randomly sampled in $[-1/16\pi, 1/16\pi]$ and random erasing (Zhong et al. 2017) during training.

Overall Performances

Our KVM-MN is compared with several person re-identification methods in recent years in Table 1. As a whole, the proposed method outperforms state-of-the-art methods using the powerful key-value memory-based matching structures as well as the robust learning method. Specifically, the competitive Spindle Net (Zhao et al. 2017a) and PDC (Su et al. 2017) both employ pose estimation models to find better body part alignments. Instead, the proposed method does not use any auxiliary models and can still beat them by over 4% in all of the reported evaluations. Some approaches learn part detectors adaptively from training data, including MSCAN(Li et al. 2017) and PartAligned (Zhao et al. 2017b). Compared with their models, the proposed model is capable of handling fine-grained matching structure via multi-head key-value representations, and thus achieves better performance. In addition, we see that even without data augmentation ("KVM-MN-noAug"), the results on all the datasets can beat the state-of-the-art.

Besides, we show some visualization results in Fig. 1 and Fig. 3 for better understanding the matchings based on key-value memory representations. It is observed that the proposed method can find semantic-level correspondences between images, such as head, bags, or shoes, even when the image is not perfectly detected (last row of the second group). Also, by employing multi-head representations, we find some interesting matching structures as shown in Fig. 1.

[1] https://github.com/tensorflow/models/tree/master/research/slim

Figure 3: Visualization of co-attentions. Each column shows the co-attention regions of a specific head of probe at the position (x, y) in the probe feature map, as shown in the top row. Best viewed in color.

Ablation Experiments

To further demonstrate the effectiveness of each component and examine the sensitivity to hyper-parameters, we design a set of ablation experiments in Table 3. Due to the limited computing resources, we generate a subset that totally contains 100K training samples for training and testing all the ablation experiments on one of the training/test split on the CUHK03 dataset. We establish a baseline denoted as "base" of KVM-MN in Table 3 and all the ablation experiments follow the same setups except some specific settings to examine the effectiveness of the component.

Effectiveness of multi-head co-attentions. We first evaluate on the number of heads to check the effectiveness of multi-head memory representation for co-attentions. We use $H = 1, 2, 4, 8$ and let $d_k = d_v = d/H$ to keep the total computing cost approximately the same. From the row (A) in Table 3, we observe that key-value memory representation with multiple heads performs better than using a single head, but the performance drops with too many heads. To evaluate the effectiveness of key-value structure representation, we further compare with the memory representation with $K = V$ and find the accuracy drops significantly. This suggests that the key-value structure can help to find much better matching structures.

Effectiveness of matching networks for training. The number of images in the gallery set during training has a significant influence on the final accuracy, as observed in row (D). When the gallery set is small, it is more likely to misclassify hard negative images, and the performance is much worse. To further investigate the influence of hard negative samples in training, we select the harder negative samples for $M = 2$ as "hard" other than randomly sampled negatives "rand", the performance increases by a large margin (68.4% $\rightarrow$ 78%). In the meanwhile, when the scale of the gallery set gets large ($H = 30$), the model converges slowly in our experiments, and it is expensive and hard to obtain a better performance. We also perform pairwise metric learning to classify positive and negative image pairs as denoted by "Pair" and we find that training with the matching networks performs better than the traditional pairwise metric learning methods (88.0% vs. 85.4%). The model converges fast and we obtain a competitive result by only training 10K steps.

Besides, we find that increasing the dimension of feature vectors improves the accuracy from row (B). In row (C), we observe that position embedding has a small impact on the overall appearance, mainly due to the capacity of convolutional neural networks in encoding information on the relative positions. We further observe in the row (E) that the proposed network could benefit a lot using data augmentation techniques as they can help the model find more stable long-range dependencies by disturbing the input images.

Conclusion

In this paper, we propose and evaluate a novel framework called Key-Value Memory Matching Network (KVM-MN) for person re-identification. The proposed KVM-MN builds an effective local-position-aware image representation using multi-head key-value memory representation as well as captures inherent matching structures using a novel co-attention mechanism. The network also benefits from a flexible setwise learning mechanism to learn an effective similarity metric by matching with a set of images. The overall performances on popular datasets, the visualization on the learned matching structures, and the ablation experiments all demonstrate the effectiveness of the components in the KVM-MN.

Table 3: Ablation experiments on CUHK03 (labeled) dataset of one training/test split. The symbol "+" or "-" denotes whether a specific component is enabled or disabled. Full names of the components in the table: "PosEmb": position embedding; "#Steps": number of training steps; "#Samples": number of training samples; "AUG": data augmentation; "RE": random erasing; "RR": random rotation. The blanks are identical to those of the "Based" model.

	Multi-head co-attention					**Training**			**AUG**		**Accuracy**		
	d	H	d_k	d_v	PosEmb	M	#Step	#Sample	RE	RR	r=1	r=5	r=10
Base	256	4	64	64	+	10	100K	100K	-	-	87.1	98.9	99.6
(A)		1	256	256							85.3	98.1	98.7
		2	128	128							86.9	99.0	99.2
		8	32	32							86.2	99.1	99.4
		4	$K = V$								83.5	96.9	99.2
(B)	512	4	128	128							87.3	98.4	99.5
	1024	4	256	256							**88.3**	**99.1**	**99.8**
(C)					-						86.8	98.7	99.5
(D)						Pair					85.4	98.0	99.1
						2-rand					68.4	93.8	97.6
						2-hard					78.0	96.6	99.3
						5					83.5	98.6	99.6
						15					**88.0**	**99.2**	99.6
						30					87.6	**99.2**	**99.8**
							10K				83.2	98.8	99.4
(E)									+		88.2	98.3	99.3
									+	+	**90.4**	**99.5**	**99.9**
(F)	512	4	128	128		15	250K	500K	+	+	**93.6**	**99.7**	**99.9**

Acknowledgments This work is supported in part by NSFC (61672456, U1509206, 61472353, and 61751209), Zhejiang Provincial Natural Science Foundation of China (LR19F020004), ZhiJiang Lab (2018EC0ZX01-2), the fundamental research funds for central universities in China (2017FZA5007), the National Basic Research Program of China (2015CB352302), Zhejiang University K.P.Chao's High Technology Development Foundation, Zhejiang provincial engineering research center on network media data cloud processing and analysis technologies, Tencent AI Lab Rhino-Bird Joint Research Program (No. JR201806), and the funding from HIKVision and ZJU Converging Media Computing Lab.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; et al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*.

Chang, X.; Hospedales, T. M.; and Xiang, T. 2018. Multi-level factorisation net for person re-identification. *arXiv:1803.09132*.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. A multi-task deep network for person re-identification. In *AAAI*.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016a. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016b. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. *arXiv:1705.03122*.

Guo, Y., and Cheung, N.-M. 2018. Efficient and deep person re-identification using multi-level similarity. *arXiv:1803.11353*.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Kalayeh, M. M.; Basaran, E.; Gokmen, M.; Kamasak, M. E.; and Shah, M. 2018. Human semantic parsing for person re-identification. *arXiv:1804.00216*.

Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*.

Li, D.; Chen, X.; Zhang, Z.; and Huang, K. 2017. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*.

Li, W.; Zhao, R.; and Wang, X. 2012. Human reidentification with transferred metric learning. In *ACCV*.

Lin, W.; Shen, Y.; Yan, J.; Xu, M.; Wu, J.; Wang, J.; and Lu, K. 2017. Learning correspondence structures for person re-identification. *TIP*.

Liu, Y.; Zhao, Q.; and Wu, Z. 2018. Pooling body parts on feature maps for misalignment robust person re-identification. In *ISBA*.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2015. Learning to rank in person re-identification with metric ensembles. In *CVPR*.

Qi, L.; Huo, J.; Wang, L.; Shi, Y.; and Gao, Y. 2018. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv:1804.03864*.

Qian, X.; Fu, Y.; Jiang, Y.-G.; Xiang, T.; and Xue, X. 2017. Multi-scale deep learning architectures for person re-identification. *arXiv:1709.05165*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2014. Imagenet large scale visual recognition challenge. *IJCV*.

Sarfraz, M. S.; Schumann, A.; Eberle, A.; and Stiefelhagen, R. 2017. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. *arXiv:1711.10378*.

Sharma, S.; Kiros, R.; and Salakhutdinov, R. 2015. Action recognition using visual attention. *arXiv:1511.04119*.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*.

Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2017. Pose-driven deep convolutional model for person re-identification. In *ICCV*.

Suh, Y.; Wang, J.; Tang, S.; Mei, T.; and Lee, K. M. 2018. Part-aligned bilinear representations for person re-identification. *arXiv:1804.07094*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *NIPS*.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Zhang, Y.; Li, X.; Zhao, L.; and Zhang, Z. 2016. Semantics-aware deep correspondence structure learning for robust person re-identification. In *IJCAI*.

Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; and Sun, J. 2017. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv:1711.08184*.

Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; and Tang, X. 2017a. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*.

Zhao, L.; Li, X.; Zhuang, Y.; and Wang, J. 2017b. Deeply-learned part-aligned representations for person re-identification. In *ICCV*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2017. Random erasing data augmentation. *arXiv:1708.04896*.

Zhou, S.; Wang, J.; Wang, J.; Gong, Y.; and Zheng, N. 2017. Point to set similarity based deep feature learning for person re-identification. In *CVPR*.