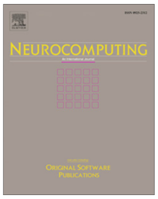
Neurocomputing 463 (2021) 388–399

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Divide-and-Merge the embedding space for cross-modality person search

Chengji Wang, Zhiming Luo *, Zhun Zhong, Shaozi Li *

Department of Artificial Intelligence, Xiamen University, China

ARTICLE INFO

Article history:
Received 9 August 2020
Revised 8 March 2021
Accepted 15 August 2021
Available online 19 August 2021
Communicated by Zidong Wang

Keywords:
Text-based person search
Convolutional Neural Network (CNN)
Divide-and-Merge
Minimum Redundancy
Maximum Relevance

ABSTRACT

This study considers the problem of text-based person search, which aims to find the corresponding person of a given text description in an image gallery. Existing methods usually learn a similarity mapping of local parts between image and text, or embed the whole image and text into a unified embedding space. However, the relevance between local and the whole is largely underexplored. In this paper, we design a Divide-and-Merge Embedding (DME) learning framework for text-based person search. DME explicitly 1) models the relations between local parts and global embedding, 2) incorporates local details into global embedding. Specifically, we design a Feature Dividing Network (FDN) to embed the input into K locally guided semantic representations by self-attentive embedding, each representation depicts a local part of the person. Then, we propose a Relevance based Subspace Projection (RSP) method for merging diverse local representations to a compact global embedding. RSP helps the model to obtain discriminative embedding by jointly minimizing the redundancy of local parts and maximizing the relevance between local parts and global embedding. Extensive experimental results on three challenging benchmarks, i.e., *CUHK-PEDES*, *CUB* and *Flowers* datasets, have demonstrated the effectiveness of the proposed method.

1. Introduction

With the urgent need for public security, person re-identification [1–3] plays an important role in intelligent surveillance systems. It can be used for person search and suspect criminal tracking. As a crucial complement to visual-based (image and video) person re-identification, text-based person search in the large-scale database has attracted remarkable attention. Given a text description of an interested person, the text-based person search aims to find the matched person from the image gallery. This task has unique properties that are different from traditional person search. On the one hand, it does not require any images of the query person that are commonly used in the image-based person search. Image instances are more difficult to obtain than text descriptions in many cases. On the other hand, the natural language descriptions can depict a person more accurately and comprehensively than individual attributes that are used in the attribute-based person search. The text descriptions are easier to obtain than manually annotated attributes.

Text-based person search is a challenging task due to the complexity and diversity of natural language. In this task, a pedestrian image is described by several text descriptions. The image and text expresses objects differently that causes large modality discrepancy. So, it is hard to obtain modality-invariant embeddings and measure the cross-modal similarity between image and text. Li et al. [4] first introduces the task of text-based person search. They compute the affinity scores between the image and each word in the text, and aggregate these scores to get the final matching score. After that, many methods have been proposed for addressing this challenging task. These methods either try to align the local parts between image and text by attention mechanism [5–8], or coarsely learn a global representation by jointly embedding the image and text into a common feature space [9,10,8,11]. Despite their great achievements, these methods commonly regard the local parts and global embedding as two separated components, and largely overlook the relevance between local parts and global embedding. It results in global features losing important local details.

To address the shortcomings of these methods, we design a Divide-and-Merge Embedding (DME) learning framework. As illustrated in Fig. 1, the proposed DME first extracts diverse local representations of a person, and then merges these representations to obtain a global embedding. In our method, we mainly focus on two aspects: 1) how to extract robust representations of local parts and avoid producing meaningless information; 2) how to effectively merge these diverse local representations and obtain a discriminative global embedding.

* Corresponding authors.
E-mail addresses: zhiming.luo@xmu.edu.cn (Z. Luo), szlig@xmu.edu.cn (S. Li).

https://doi.org/10.1016/j.neucom.2021.08.058
0925-2312/© 2021 Elsevier B.V. All rights reserved.

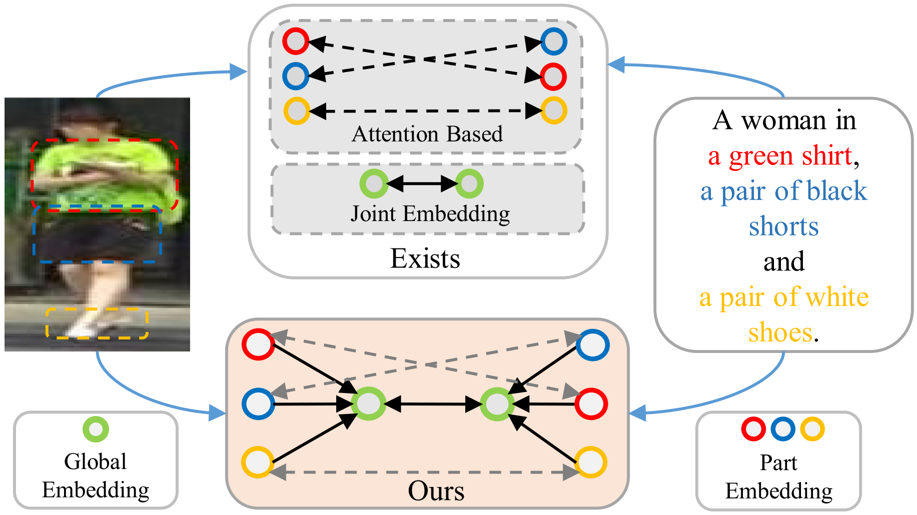


Fig. 1. Text-based person search aims to find the person by a text description. Existing methods either attempt to build a learning framework to compute the affinity scores of parts or learn a joint embedding in a shared feature space. However, they ignore the relationships of local parts and global embedding, which makes the embedding less discriminative. In this study, we design a Divide-and-Merge Embedding learning framework, which first embeds the input into several locally guided semantic representations, and then effectively merges local details into global embedding by explicitly considering their relationships.

For the first aspect, we design a Feature Dividing Network (FDN) which introduces self-attentive embedding to embed the inputs into K locally guided representations. Each local representation latently attends on different semantic parts of the input. In order to bridge the modality-gap, we impose a cross-modality part matching loss that can jointly reduce the distances of matched image-text pairs while enlarging that of unmatched pairs. *For the second aspect*, we aim to map locally guided representations into a compact subspace. To achieve this goal, we propose a Relevance based Subspace Projection (RSP) method, which jointly minimizes the redundancy of locally guided representations and maximizes the relevance between locally guided representations and global embedding. In training, the model is optimized with the cross-modality matching loss and identification loss. In testing, the global embedding is directly used to measure the similarity between image and text instances.

The main contributions of this study are summarized as follows:

1. We design a Divide-and-Merge Embedding Learning (DME) framework, which enables the model to explore the relevance between the global embedding and local parts and thus generates more discriminative representations for matching text and image examples.
2. We demonstrate that self-attentive embedding can successfully be applied to cross-modality person search. By leveraging the Feature Dividing Network (FDN), the DME can mine ample local details of a person.
3. We propose a Relevance based Subspace Projection (RSP) method, it can effectively merge local representations into a global embedding. RSP can encourage the global embedding to contain more useful local details while avoiding redundancy information.
4. Experiment results on three public datasets demonstrate the effectiveness of the proposed method.

2. Related work

In this section, we will briefly review the related work about our studies, including text-based cross-modality person search, attention-based embedding, and subspace learning.

2.1. Text-based person search

As a complement to image-based or attribute-based person re-identification, Li et al. [4] introduce the task of text-based person search and propose a recurrent neural network with gated neural attention (GNA-RNN) for tackling the task. Lin et al. [5] further improve the performance by sampling hard negative samples for learning discriminative feature representations. They propose a novel Cross-Modal Cross-Entropy (CMCE) loss to efficiently screen easy incorrect matching. Chen et al. [7] propose a patch-word matching framework. They compute the affinity between the words and local spatial patches, and then select the best patch-word matching affinity to represent the image-word affinity. In [6], the authors not only identify the global image representation with the supervision of the overall description but also implicitly associate image regions with mined local linguistic features by the corresponding noun phrases. Zhang et al. [10] attempt to learn the joint discriminative embeddings for both image and text in a shared latent space. Recently, Liu et al. [8] propose a novel A-GANet for text-based person search, which exploits the object relationships from the text queries and gallery images by graph attention network. PWA [12] extracts the visual parts with the assistant of human poses and tries to align visual parts with noun phrases. Aggarwal et al. [11] perform human attributes recognition and cross-modality matching simultaneously.

Although these methods have a great achievement, they commonly regard the local parts and global representation as two separate components or merely learning a unified global embedding, and largely ignore the relationships between local parts and the global embedding. In this paper, we design a Divide-and-Merge (DME) learning framework that obtains discriminative person representations by modeling the relationships between local parts and the global embedding.

2.2. Attention-based embedding

Many works [13–18] have used attention mechanisms to learn better feature representations. Most of the attention methods [14–16] can be seen as semantic alignment, the parts with similar semantics will have big weights. They need to obtain two independent representations and then align them based on their semantics. Different from semantic alignment based methods, some works [13,17,18] design an alignment free paradigm. They directly generate the attention weights by a sub-network. Current text-based person search works [4,5,7,6,12] are focused on align the semantics between image and text that needs to fuse cross-modality features. [4,5] focus on computing image-word affinity scores. [7,6] directly partition the image into local patches by their spatial position. They compute the patch-word affinity scores. Some works [4,5,7,6] utilize the cross-modality attention mechanism to align cross-modality semantics.

Our feature dividing network follows the alignment free paradigm. It directly applies two fully connected layers to generate the weights of each part. Different from previous text-based person search methods, we only take single-modality features as input and do not perform cross-modality feature fusion.

2.3. Subspace learning

Subspace learning aims at projecting the original high dimensional feature into a meaningful low-dimensional subspace. The most popular methods PCA [19], LDA [19], LPP [20] and neighborhood preserving embedding (NPE) [21] devote to preserve the statistical properties of features. Wang et al. [22] unifies subspace learning and feature selection into the same learning framework. Works [23–26] joint embed image and text into a shared subspace and perform image-text matching in this space. Peng et al. [27] firstly proposes the Minimum Redundancy Maximum Relevance rule (mRMR) to do feature selection. The mRMR minimizes redundancy of selected features and maximizes relevance between the selected features and class label. In this study, we further extend the mRMR to text-based person search task and propose a new Relevance based Subspace Projection method.

3. Approach

Our proposed Divide-and-Merge Embedding Learning framework is shown in Fig. 2. Firstly, a visual CNN and a Bi-LSTM are used to extract modality-specific features from images and texts, respectively. Then, we utilize a Feature Dividing Network to transform the modality-specific features into K locally guided representations by self-attentive embedding [13]. To learn diverse and meaningful local representations and reduce the modality-gap, we design a cross-modality part matching loss. It constructs one-to-one matching relationships between local representations from matched image-text pairs. After that, a Relevance based Subspace Projection module compresses these representations into compact global embeddings. Finally, a multi-task cross-modality loss function is used to optimize the whole network. In the following subsections, we will describe each part in detail.

3.1. Modality-specific feature encoder

Image Encoder: In this study, we adopt a visual CNN pre-trained on ImageNet [28] as the image-modality feature encoder. We take all the layers before the final pooling layer as the backbone network, and add a 1 × 1 convolutional layer of 512 filters with *ReLU* activation function. For an image x_i of size 224 × 224, we can obtain the local feature $f(x_i) \in R^{7\times7\times512}$, and then reshape it to $f(x_i) \in R^{49\times512}$.

Sentence Encoder: For a sentence z_j with T words, we firstly use the Word2Vec matrix to project each word in a 300d vector. Then we feed them into a Bi-LSTM with 512 hidden units, and fetch the representation of each word by adding the hidden state from two directions. As the same with image encoder, we also apply a convolutional layer with 512 kernel to get the text local features $f(z_j) \in R^{T\times512}$.

3.2. Feature dividing network

After getting the modality-specific features $f(*)$, we use a feature dividing network to group different local features into diverse and meaningful representations for describing different local parts of a person. Due to it is hard to define which local features need to be grouped explicitly, we implicitly implement this intuition by employing the self-attentive embedding [13] to compute K attention coefficients, then obtain the corresponding representations through a linear combination. In addition, we propose a cross-modality part matching loss to enforce the diversity among the local representations and increase the local similarity between matched image-text pairs.

For the local feature $f(*) \in R^{H\times512}$, we use a 2-layer MLP without bias to compute the attention map $\alpha \in R^{K\times H}$:

$$\alpha = \mathrm{softmax}(w_2 \tanh(w_1 f^T(*))), \tag{1}$$

where $w_1 \in R^{A\times512}, w_2 \in R^{K\times A}$, and A is the number of hidden units. The softmax function is performed along the row dimension to make the summation of each of the K attention coefficients equal up to one.

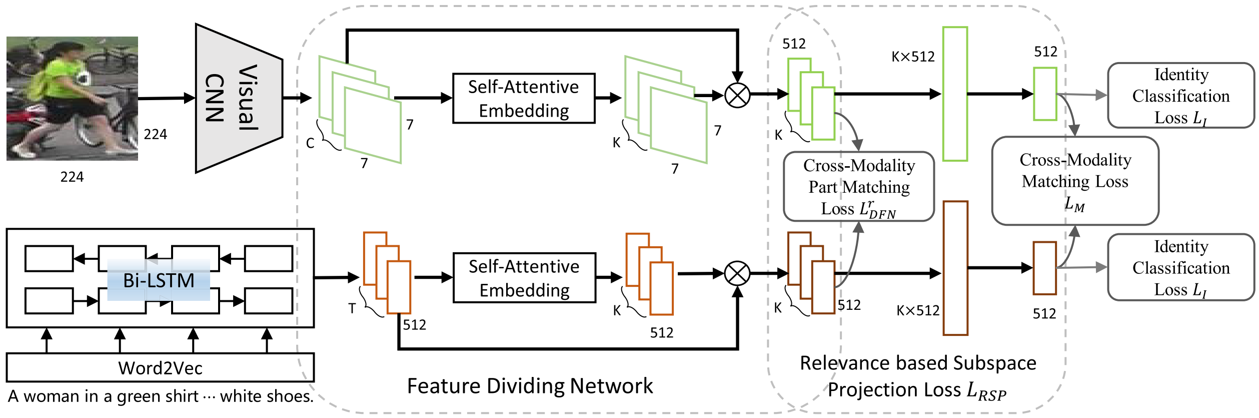

224
224
Visual CNN
C
7
7
Self-Attentive Embedding
K
7
7
512
K
K×512
512
Identity Classification Loss L_I
Cross-Modality Part Matching Loss L^r_{DFN}
Cross-Modality Matching Loss L_M
Bi-LSTM
Word2Vec
A woman in a green shirt ··· white shoes.
T
512
Self-Attentive Embedding
K
512
K
512
K×512
512
Identity Classification Loss L_I
Feature Dividing Network
Relevance based Subspace Projection Loss L_{RSP}

Fig. 2. The overall architecture of our Divide-and-Merge Embedding Learning framework. It mainly contains two parts: 1) The Feature Dividing Network is to extract diverse local representations, each one depicts a local part of a person; To learn meaningful and modality-invariant representations, we design a cross-modality part matching loss to align the local representations from two modalities. 2) We merge the local representations to a compact global embedding. The Relevance based Subspace Projection loss encourages the global embedding to contain more useful local details. 3) Both cross-modality matching loss and identity classification loss are employed to train the model.

After getting the attention map α, we multiply it with the local feature $f(*)$ followed by an L2-norm to obtain K locally-guided representations $\phi(*) \in R^{K\times 512}$:

$$\phi(*) = \frac{\alpha f(*)}{\|\alpha f(*)\|}. \tag{2}$$

In this study, we denote the locally guided representations of image x_i as $\{\phi^1(x_i), \dots, \phi^K(x_i)\}$ and the representations of text z_j as $\{\phi^1(z_j), \dots, \phi^K(z_j)\}$.

Cross-modality Part Matching. To learn diverse and meaningful local representations, we construct one-to-one relationships between image representations $\{\phi^k(x_i)\}$ and text representations $\{\phi^h(z_j)\}$. Besides, the relationships between image and text are mutually symmetric. To model the relationships, as is shown in Fig. 3, we firstly establish the similarity matrix S by using cosine distance, where

$$s_{kh} = \max(0, \frac{\phi^k(x_i)\phi^h(z_j)}{|\phi^k(x_i)| \cdot |\phi^h(z_j)|}). \tag{3}$$

For the matched pairs, we assume the image representation $\phi^k(x_i)$ only has one corresponded text representation $\phi^h(z_j)$ and vice versa, that means the matrix S is an orthogonal matrix. For an orthogonal matrix, the rows of matrix S are unit vectors and orthogonal to each other, and the columns of matrix S are unit vectors and orthogonal to each other. So, we have $SS^T = I$, where $I \in R^{K\times K}$ is the identity matrix. In here, we construct two matrices: S_1, the row-wise normalized matrix S; S_2, the column-wise normalized matrix S; as is shown in Fig. 3. We require $S_1S_1^T = I$ and $S_2^TS_2 = I$,

$$L^r_m = \frac{1}{K^2}\sum_{k,h}^{K}(S_1S_1^T - I)^2 + (S_2^TS_2 - I)^2. \tag{4}$$

While for unmatched image-text pairs, previous relationships will not be satisfied. In the ideal case, we hope that the similarity vectors for each $\phi^k(x_i)$ and $\phi^k(z_j)$ are uniformly distributed, which means each element in S_1 and S_2 is close to $1/K$.

$$L^r_u = \frac{1}{K^2}\sum_{k,h}^{K}\sum_{i\in\{1,2\}}(S_i - \frac{1}{K})^2 \tag{5}$$

The cross-modality part matching loss can be denoted as

$$L^r_{DFN} = \frac{1}{N^2}\sum_{i,j}^{N} y_{ij} \cdot L^r_m + (1 - y_{ij}) \cdot L^r_u \tag{6}$$

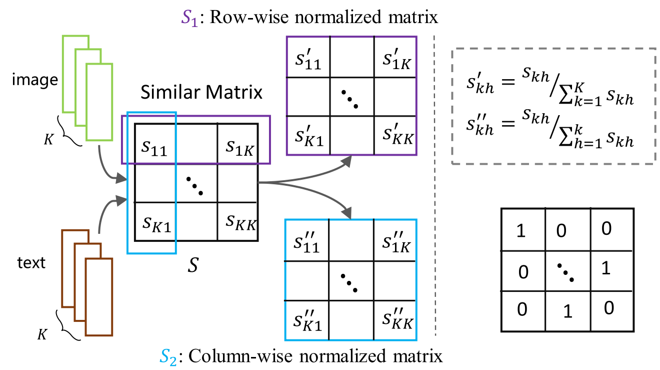

S_1: Row-wise normalized matrix
image
K
Similar Matrix
s_{11}
s_{1K}
s_{K1}
s_{KK}
S
text
K
s'_{11}
s'_{1K}
s'_{K1}
s'_{KK}
s''_{11}
s''_{1K}
s''_{K1}
s''_{KK}
$s'_{kh} = {}^{s_{kh}}/_{\sum_{k=1}^{K} s_{kh}}$
$s''_{kh} = {}^{s_{kh}}/_{\sum_{h=1}^{k} s_{kh}}$
1 0 0
0 1
0 1 0
S_2: Column-wise normalized matrix

Fig. 3. We design a local part matching loss to reduce the modality-gap. It not only reduces the redundancy of the learned local parts from the same modality and also increases the correlation between cross-modality features by making S to be an orthogonal matrix.

where $y_{ij} = 1$ for matched image-text pairs, and otherwise $y_{ij} = 0$. N is the batch size.

3.3. Relevance based subspace projection

The previous FDN can produce diverse local representations of a person. However, it will result in an extremely high-dimension feature vector if directly concatenate them together.

To alleviate this issue, we propose a Relevance-based Subspace Projection (RSP) to compress the K representations into a low-dimensional global embedding,

$$\begin{aligned} g(x_i) &= w_i * \left[\phi^1(x_i), \dots, \phi^K(x_i)\right], \\ g(z_j) &= w_t * \left[\phi^1(z_j), \dots, \phi^K(z_j)\right], \end{aligned} \tag{7}$$

where w_i and w_t are the parameters of the subspace projection, and $[*]$ is the concatenation operation. In this study, the feature dimension of the global embedding and local representation is the same.

To encourage the global embedding $g(x_i)$ and $g(z_j)$ maintain the essential information of each local representation and minimize the redundancy of local representations, we optimize the parameters w_i and w_t in RSP following the Minimum Redundancy Maximum Relevance (mRMR) rule [27]. The original mRMR is designed for selecting a feature subset in the classification problem. It mainly consists of the following two properties: (1) maximizing relevance between selected features m_k and target label c, and (2) minimizing redundancy between the next selected feature m_h and m_k. The optimization equation of original mRMR is,

$$min\sum_k\left[-R(m_k, c) + \frac{1}{|M|}\sum_{h,h\neq k}RD(m_h, m_k)\right], \tag{8}$$

where $M = \{m_k\}$ is the selected feature set, and R and RD are the criteria for measuring relevance and redundancy respectively.

Instead of doing the feature selection as in the Eq. (8), we reformulate it for optimizing the feature projection parameter w_i and w_t. Taking the image modality as example, the corresponding objective is as following,

$$min\sum_k^K\left[-R(\phi^k(x_i), g(x_i)) + \frac{1}{K}\sum_{h,h\neq k}^{K}RD(\phi^h(x_i), \phi^k(x_i))\right], \tag{9}$$

where R measures the relevance between the global embedding and the local representation, RD calculates the redundancy between two local representations. We also can get the objective function for the text modality in the same manner.

However, the Eq. (9) only considers the relevance within the same modality, and the $g(x_i)$ is a learned global embedding not a fixed label. These will cause the ambiguous embedding problem. To solve this problem, we extend the Eq. (9) to the cross-modality form from the following two aspects, (1) increasing the inter-modality relevance between $g(x_i)$ and $g(z_j)$ of matched image-text pairs; (2) considering the intra-modality feature redundancy and inter-modality feature correlation simultaneously. We can get the cross-modality form of Eq. (9) as

$$\begin{aligned} min - \sum_k^K\left[R(\phi^k(x_i), g(x_i)) + R(\phi^k(z_j), g(z_j))\right] - R(g(x_i), g(z_j)) \\ + RD\Big(\{\phi^k(x_i)\}, \{\phi^k(z_j)\}\Big). \end{aligned} \tag{10}$$

Notice that the Eq. (10) consists of three terms. The first term is *the intra-modality relevance* as the same as in Eq. (9). The second term is related to *the cross-modality relevance*. The last term is *the cross-modality redundancy*. In the next, we will give the computation details of these three items.

In order to maximize the *intra-modality relevance* between local representations $\phi^k(*)$ and its corresponding global embedding, we convert it into minimizing the cross-entropy between them. Firstly, we first transform each local representation and the global embedding to a probability distribution. Then, we maximize their relevance by minimizing

$$L_{imr} = \frac{1}{K}\sum_{k=1}^{K}\left[H\left(\phi^k(x_i), g(x_i)\right) + H\left(\phi^k(z_j), g(z_j)\right)\right], \tag{11}$$

where $H(p,q) = \Theta(p)\cdot log\frac{1}{\Theta(q)}$, and Θ is the softmax function.

To maximize the *cross-modality relevance* between matched image-text pairs $\{g(x_i), g(z_j)\}$ against unmatched pairs, we implement it by making the similarity of matched pairs bigger than unmatched ones. Therefore, we can minimize

$$\begin{aligned} L_{cmr} = & max\left(cos(g(x_i), g(z_h)) - cos(g(x_i), g(z_j)), 0\right) \\ & + max\left(cos(g(z_j), g(x_h)) - cos(g(z_j), g(x_i)), 0\right), \end{aligned} \tag{12}$$

where x_h and z_h are the hardest negative image and text for z_j and x_i within a mini-batch, respectively. $cos(\cdot,\cdot)$ is the cosine similarity.

For the third *cross-modality redundancy* term, we implement it by increasing the correlation between cross-modality local representations. As discussed in the cross-modality part matching loss L^r_{DFN}, we construct a one-to-one relationship between image representation $\phi^k(x_i)$ and $\phi^h(z_j)$, which means the row-normalize similarity matrix S_1 and column-wise normalized S_2 should have similar distributions for increasing correlation. Besides, the S_1 and S_2 tend to be more sparse-like matrices for reducing the redundancy of local parts. Therefore, we minimize following loss to reduce the cross-modality redundancy,

$$L_{cmd} = -\frac{1}{K^2}\sum_{k,h}^{K} ln(1 - S^d) \tag{13}$$

where $S^d = \sqrt{(S_1 - S_2)^2}$.

By combining these three terms together, the final objective of our proposed relevance based subspace projection can be represented by

$$L_{RSP} = \frac{1}{N}\sum_{i,j}^{N} L_{imr} + L_{cmr} + L_{cmd}. \tag{14}$$

3.4. Training and testing

The main goal of the training procedure is to optimize the parameters, which can increase the similarity of matched pairs and enlarge the distance of unmatched pairs. With the effectiveness of cross-modality projection in learning discriminative image-text embedding [10], we also utilize it for optimization. The loss function mainly consists of two parts: **the identity classification** and **the cross-modality matching**.

Give a mini-batch with N image-text pairs. The i-th image-text pair (x_i, z_i) have the same person ID. We first compute the cross-modality feature projection [10] between image and text by:

$$\begin{aligned} \hat{g}(x_i) & = g^T(x_i)\bar{g}(z_i)\cdot\bar{g}(z_i) \\ \hat{g}(z_i) & = g^T(z_i)\bar{g}(x_i)\cdot\bar{g}(x_i), \end{aligned} \tag{15}$$

where $\bar{g}(x_i) = \frac{g(x_i)}{\|g(x_i)\|}$ and $\bar{g}(z_i) = \frac{g(z_i)}{\|g(z_i)\|}$ are the normalized feature, respectively.

The first **identity classification** part encourages that the projected feature vectors $\{\hat{g}(x_i)), \hat{g}(z_j)\}$ of the same identify are similar and discriminate from features of other identities. It is achieved by a shared cross-modality norm-softmax loss function:

$$\begin{aligned} L_I = & \frac{1}{N}\sum_{i,j}^{N} \\ & -\left[\log(\frac{\exp(W^T_{y_i}\hat{g}(x_i))}{\sum_h^{N_{id}}\exp(W^T_h\hat{g}(x_i))}) + \log(\frac{\exp(W^T_{y_j}\hat{g}(z_j))}{\sum_h^{N_{id}}\exp(W^T_h\hat{g}(z_j))})\right], \end{aligned} \tag{16}$$

where W^T_h is the cross-modality shared classifier for each different ID, we restrict $\|W_h\| = 1$. There are N_{id} IDs in the training set.

The second **cross-modality matching** part enforces the cross-modality matching probability distribution to be close to the ground truth. Given a mini-batch with N images and texts, we can obtain N different image-text pairs $\{(x_i, z_j), y_{ij}\}$ for each image x_i. $y_{ij} = 1$ means (x_i, z_j) is a matched pair, otherwise unmatched. The normalized probability of matching x_i to z_j is computed by

$$p_{ij} = \frac{exp(g^T(x_i)\bar{g}(z_j))}{\sum_{h=1}^{N} exp(g^T(x_i)\bar{g}(z_h))}, \tag{17}$$

and the ground truth probability is

$$q_{ij} = \frac{y_{ij}}{\sum_h y_{ih}}. \tag{18}$$

We can compute the normalized probability of matching z_j to all x_i and its corresponding ground truth in the same manner.

By minimizing the KL divergence between these two probability distributions, we then have the cross-modality matching loss as

$$L_M = \frac{1}{N}\left[\sum_{i}^{N} KL(p_i\|q_i) + \sum_{j}^{N} KL(p_j\|q_j)\right]. \tag{19}$$

Finally, the model is jointly optimized by the identity classification loss (L_I), the cross-modality matching loss (L_M), the objective of relevance based subspace projection loss (L_{RSP}) and the cross-modality restriction (L^r_{DFN}):

$$L = L_I + L_M + \lambda L_{RSP} + \gamma L^r_{DFN}, \tag{20}$$

where λ and γ are two hyper parameters.

In the **testing phrase**, we firstly extract the text embedding $g(z_j)$ for a given probe textual description z_j. Then, we get the matching score of z_j with each image x_i in the gallery set by computing their cosine similarity

$$Score(x_i, z_j) = \frac{g(x_i)\cdot g(z_j)}{\|g(x_i)\|\cdot\|g(z_j)\|}. \tag{21}$$

4. Experiments

In this section, we evaluate the proposed Divide-and-Merge embedding learning framework (DME) on a challenging text-based person search dataset (*CUHK-PEDES* [4]) and two popular cross-modality retrieval datasets. We first provide the experimental details in Section 4.1. Then, we compare our DME with the state-of-the-art methods in Section 4.2. The analysis of our method is provided in Section 4.3. We discuss the details of the model in Section 4.4. Finally, we show some retrieval results to further demonstrate the effectiveness of our method in Section 4.6.

4.1. Experimental setup

4.1.1. Datasets

CUHK-PEDES dataset [4] is a large-scale text-based person search dataset which contains 40,206 images from 13,003 identities, and each image is associated with two different text descriptions. Following [4,10], we randomly split the dataset into the training, validation, and testing sets with non-overlapping between person identities. The training set has 34054 images, 11003 persons, and 68126 textual descriptions. The validation set has 3078 images, 1000 persons and 6158 textual descriptions. The testing set has 3074 images, 1000 persons and 6156 textual descriptions. We also test our methods on two public text-based image retrieval datasets *Caltech-UCSD Birds (CUB)* [29] and *Flowers* [29]. The CUB dataset contains 200 different categories and 11788 bird images, each image has 10 text descriptions. The Flowers dataset consists of 8189 flower images of 102 categories with 10 descriptions for each image.

4.1.2. Evaluation metrics

To evaluate the performance, we report results with Rank-K (K = 1, 5, 10) for CUHK-PEDES. Given a textual description, the Rank-K indicates that a correct item of the query sample is retrieved in the top-k results. We rank the gallery images with a query text based on their matching score. In addition, we also adopt mean Average Precision (mAP) to evaluate the performance. mAP is the mean of the average precision scores for each query. For comparison purposes, we report Rank-1 accuracy on Image-to-Text task and AP@50 on Text-to-Image task on CUB and Flowers datasets as same as other works [10,32]. The AP@50 report the average matching percentage of top-50 retrieved images of all test classes.

4.1.3. Implementation details

All experiments are implemented based on TensorFlow [36] with a GTX 2080Ti GPU. The vocabulary includes 12,000 words, the dimension of word embedding is 300. We randomly initialize the Word2Vec matrix. The dimension of the global embedding is 512. We initialize the weights of visual CNN [37,38] with the model pre-trained on ImageNet [28] and the newly added layers by Xavier initialization [39]. To avoid over-fitting, we also add dropout [40] layers after the input and output of LSTM [41] with a rate of 0.3. During the training phase, the Adam optimizer [42] is utilized for optimizing the model with a learning rate of $2e-4$. All the models are trained with 50 epochs in total. The size of the images are resized to $224 \times 224 \times 3$. We set the mini-batch size to 32 for CUHK-PEDES. In default, we set the values of λ to 1.0, γ to 0.5 and $K = 8$. For CUHK-PEDES, we use MobileNet [37] and ResNet-50 [38] as the backbone of visual CNN. For CUB and Flowers datasets, we use ResNet-152 [38] as the backbone of visual CNN and then training the whole model for 35 epochs. We set the values of λ to 1.0, γ to 0.5 and $K = 6$.

4.2. Comparing with state-of-the-art

4.2.1. Results on CUHK-PEDES dataset

We compare the proposed DME with existing state-of-the-of methods on the CUHK-PEDES dataset. The text-to-image matching results are reported in Table 1. We can see that methods trained with MobileNet or ResNet-50 always produce higher results than the methods trained with VGG-16. With the same backbone, our DME consistently outperforms state-of-the-of methods. Specifically, our DME obtains 56.32% in Rank-1 accuracy and 55.45% in Rank-1 accuracy when using ResNet-50 and MobileNet as the backbone, respectively. When the backbone is ResNet-50, our method surpasses the second-best method (PMA [12]) by 2% in Rank-1 accuracy. When removing the proposed FDN and RSP, our DME reduces to CMPM-CMPC [10]. In this paper, we implement CMPM-CMPC in two ways: 1) using default parameters proposed in CMPM-CMPC; and 2) using the same parameters as our method. The difference between the parameters of CMPM-CMPC and this paper is that 1) Changing the Word2Vec dimension from 512-d to 300-d; 2) Adding the hidden states of forward and backward directions of Bi-LSTM, the original CMPM-CMPC concatenates the hidden states of forward and backward directions; 3) Applying a 1×1 convolutional layer with 512 kernels to reduce the feature dimension, the image and text use separate convolutional layer. As we can see, these modifications make us achieve better results. More details can be seen in Fig. 4. When using the same setting of parameters, our method improves the Rank-1 accuracy of CMPM-CMPC from 52.31% to 55.45%. This verifies the effectiveness of the proposed modules. The cross-attention based methods (GNA-RNN [4], IATV [5], PWM-ATH [7], GLA [6] and PMA [12]) obtain the image-text embedding by fusing features from two modality. Instead, our DME directly computes embeddings for image and text, respectively. Compared to A-GANet [8] which exploits 200 objects, our DME only requires to extract 8 local representations. The CMAAM [11] uses manually annotated person attributes. Different from it, our DME does not use any additional labels except the provided identity annotations.

4.2.2. Results on the CUB and flowers dataset

Table 2 is the comparison of image-to-text and text-to-image retrieval results on the CUB [29] and Flowers [29] datasets. Considering that our proposed method is based on CMPM-CMPC [10], we report the results with the same metrics for a fair comparison.

As can be seen, our modified CMPM-CMPC* achieves a Rank-1 accuracy of 66.4% and an AP@50 of 69.1% on the CUB dataset,

Table 1
Comparison of DME with state-of-the-art methods on CUHK-PEDES. * indicates method reproduced with the parameters of this paper.

Method	Backbone	Text-to-Image		
		Rank-1	Rank-5	Rank-10
CNN-RNN [30]	VGG-16	8.07	–	32.47
Neural Talk [31]		13.66	–	41.72
GNA-RNN [4]		19.05	–	53.64
IATV [5]		25.94	–	60.48
PWM-ATH [7]		27.14	49.45	61.02
PMA [12]		47.02	68.54	78.06
CMPM-CMPC [10]	MobileNet	49.37	–	79.27
CMPM-CMPC * [10]		52.31	75.41	83.10
CMAAM [11]		55.13	76.14	83.77
DME (Ours)		**55.45**	**76.22**	**83.85**
GLA [6]	ResNet-50	43.58	66.93	76.26
A-GANet [8]		53.14	74.03	81.95
PMA [12]		53.81	73.54	81.23
DME (Ours)		**56.32**	**77.23**	**84.71**
TIMAM [32]	Resnet101	54.51	77.56	84.78

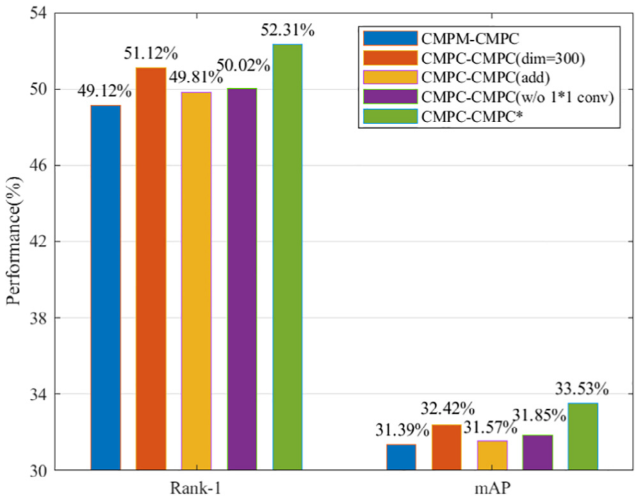


Fig. 4. Performance (Rank-1 accuracy and mAP) with different settings of CMPM-CMPC [10]. Comparing to original CMPM-CMPC [10], we have made three improvements: 1) we set the dimension of Word2Vec to 300; 2) we add the outputs of Bi-LSTM; 3) we apply a 1 × 1 convolutional layer with 512 kernels to reduce the feature dimension.

and 68.9% and 69.7% on the Flowers dataset. Compared to CMPM-CMPC*, our proposed DME increases the Rank-1 accuracy of 2.9% and 1.9%, and the AP@50 of 2.7% and 2.3% on the two datasets, respectively. Besides, we can observe that the proposed DME outperforms the other comparing methods. Specifically, our method surpasses the previous the-state-of-art TIMAM [32] by 1.7% and 1.8% for the image-to-text task, and 1.5% and 1.3% on the text-to-image task, respectively.

4.3. Ablation experiments

Our method consists of two parts, Feature Dividing Network (FDN) and Relevance based Subspace Projection (RSP). To optimize our model, the identity classification loss (L_I) and the cross-modality matching loss (L_m) are used. To demonstrate the effectiveness of each component, in Table 3, we evaluate the effectiveness of each component of DME by training the model with different component combinations.

From Table 3, we can obtain the following observations. First, the identification loss (L_I) achieves higher results than the cross-modal matching loss (L_M). When using them together, the performance can be significantly improved. We regard the combination of them as the baseline model of our method. We gradually add FDN and RSP into the baseline to validate their effectiveness. Second, FDN consistently improves the baseline in all metrics. This is benefited from local details introduced by FDN. Third, when additionally using RSP based on FDN, a large improvement can be obtained. Specifically, using RSP increases the Rank-1 accuracy from 53.35% to 55.45%. Fourth, the model trained with both FDN and RSP clearly outperforms the baseline model (55.45% vs 52.31% in Rank-1 accuracy), demonstrating the effectiveness of the proposed modules.

4.4. Evaluations

4.4.1. Comparing the cross-modality part matching loss with MMD and l_2 distance

In the proposed FDN, we perform a cross-modality part matching loss (CPM) between image and text (L^r_{DFN}). It is introduced to reduce the gap between the modalities of text and image. We also want it to make the FDN learn corresponding semantic information between image and text. To demonstrate its effectiveness, we first evaluate the DME with different γ, then we compare the CPM with two commonly used methods: maximum mean discrepancy (MMD) [43,44] and l_2 distance.

Gamma γ. In Fig. 5, we evaluate the impact of the parameter (γ) of cross-modality part matching. When $\gamma = 0$, the cross-modality part matching is ignored during FDN. We can see that the Rank-1 accuracy drops by 1.4% and mAP drops by 0.8% when $\gamma = 0$. When injecting the cross-modality part matching into the system ($\gamma > 0$), the Rank-1 accuracy and mAP are consistently improved. The best performance is achieved when $\gamma = 0.5$.

Comparison to MMD and l_2 distance. We compare the CPM with MMD and l_2 distance, the results are shown in Fig. 6. Our CPM and MMD do not establish any clear correspondences of the learned semantic parts between image and text, they constrain the distance of the whole set. The l_2 distance restricts the first semantic part of the image corresponding to the first semantic part of text, and the same way of others. From Fig. 6, we have three observations: 1) the CPM is slightly better than MMD; 2) both CPM and MMD are better than l_2 distance; 3) all of them are better than the model which does not use CPM. Based on these observations, we have: 1) adding restrictions to local semantic parts can improve the performance of the model; 2) forced one-to-one relationships (l_2 distance) are not optimal.

4.4.2. Analysis of relevance based subspace projection

In this section, we will analyze our proposed Relevance based Subspace Projection (RSP) in detail. We also compare the RSP with a popular method on CUHK-PEDES.

Table 2
Comparison of image-to-text (Rank-1(%)) and text-to-image (AP@50(%)) retrieval results on the CUB and Flowers dataset. * indicates method reproduced with the parameters of this paper.

Method	CUB		Flowers	
	Image-to-Text	Text-to-Image	Image-to-Text	Text-to-Image
	Rank-1	AP@50	Rank-1	AP@50
Bow [33]	44.1	39.6	57.7	57.3
Word2Vec [34]	38.6	33.5	54.2	52.1
Word CNN [30]	51.0	43.3	60.7	56.3
Word CNN-RNN [30]	56.8	48.7	65.6	59.6
GMM + HGLMM [35]	36.5	35.6	54.8	52.8
Triplet [5]	52.5	52.4	64.3	64.9
Latent Co-attention [5]	61.5	57.6	68.4	70.1
CMPM-CMPC [10]	64.3	67.9	68.9	69.7
CMPM-CMPC* [10]	66.5	69.1	70.5	72.3
TIMAM [32]	67.7	70.3	70.6	73.3
DME (Ours)	**69.4**	**71.8**	**72.4**	**74.6**

Table 3
Investigation of different components of the proposed DME on the CUHK-PEDES dataset.

Method				Text-to-Image			
L_I	L_M	FDN	RSP	Rank-1	Rank-5	Rank-10	mAP
	✓			44.77	69.30	78.75	28.30
✓				46.80	70.73	79.00	30.23
✓	✓			52.31	75.41	83.30	33.53
✓	✓	✓		53.35	75.42	83.56	34.49
✓	✓	✓	✓	55.45	75.89	83.70	35.62

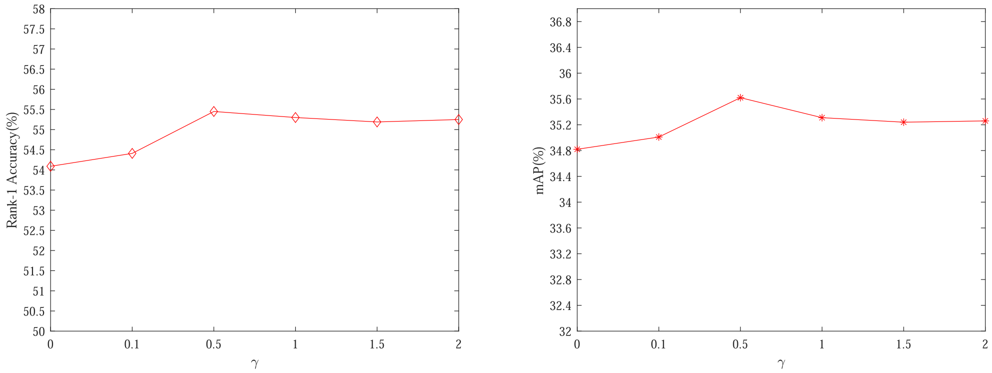


Fig. 5. Evaluation of different γ on CUHK-PEDES (fixed $\lambda = 1.0$, $k = 8$).

Lambda λ. In Fig. 7, we investigate the impact of the weight of RSP. When $\lambda = 0$, the model reduces to the baseline trained with FDN. We can find that introducing RSP into the model can consistently improve the results. Especially, the highest accuracy is obtained when $\lambda = 1$.

Component analysis. To further investigate RSP, we conduct ablation studies by removing one of the components in RSP. In Fig. 8, RSP w/o L_{imr} refers to RSP without the term that maximizes intra-modality relevance, and RSP w/o L_{cmr} indicates RSP without the term that maximizes cross-modality relevance, RSP w/o L_{cmd} refers to RSP without the term that minimizes cross-modality redundancy. From Fig. 8, we can see that removing any one of the components will reduce the performance, especially the L_{imr}. These results indicate the importance of each component in RSP for learning discriminative global embedding.

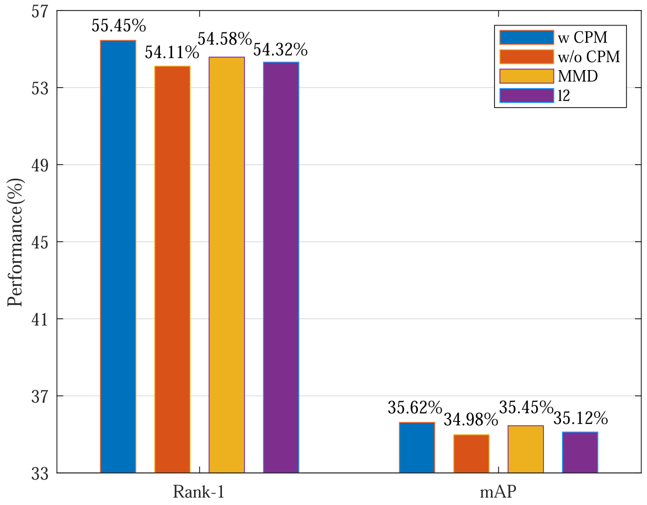


Fig. 6. Performance (Rank-1 accuracy and mAP) comparison of different cross-modality part matching loss.

Comparison to Re-Weighting. The proposed RSP compresses diverse local representations into a low-dimensional global embedding based on the subspace projection. To further demonstrate its effectiveness, we compared it with a Re-Weighting scheme, which performs a global weighted sum pooling over the local representations. In the Re-weighting, we apply a fully connected layer on each local representation $\{\phi^k(*), k = 1, \ldots K\}$, and get the weight $c_k \in R$. Then, we utilize a softmax function to normalize their sum to 1. Finally, we use the weighted global sum pooling to obtain the final global embedding. Here, we reported the comparison between RSP and Re-Weighting in Fig. 9. We can observe a rapid performance degradation when replacing the RSP with Re-Weighting. It demonstrates the superiority of our proposed RSP.

4.4.3. Intra-modality vs. cross-modality redundancy

To verify the effectiveness of the proposed cross-modality redundancy (**cmd**) in the RSP, we replace the **cmd** by an intra-modality redundancy (**imd**). The **imd** only computes the redundancy within each modality by restricting the Gram matrix of locally guided representations to be an identity matrix. From Fig. 10, we can see that **cmd** clearly outperforms **imd** in the two metrics, demonstrating the benefit of constraining the redundancy across cross-modality features.

4.4.4. Influence of K.

To evaluate the influence of the number of locally guided representations, we conduct experiments by varying K in the range of $[2, 12]$ on CUHK-PEDES. Results are shown in Fig. 12. The Rank-1 accuracy and mAP are first increased with K. However, assigning a large value to K may bring unrelated information, such as noisy and background, and thus reduces the performance. The best results are obtained when K is around 8.

4.5. Loss curves

The model is trained with 106400 iterations, we record the values of the losses every 10 iterations and provide the curves of each

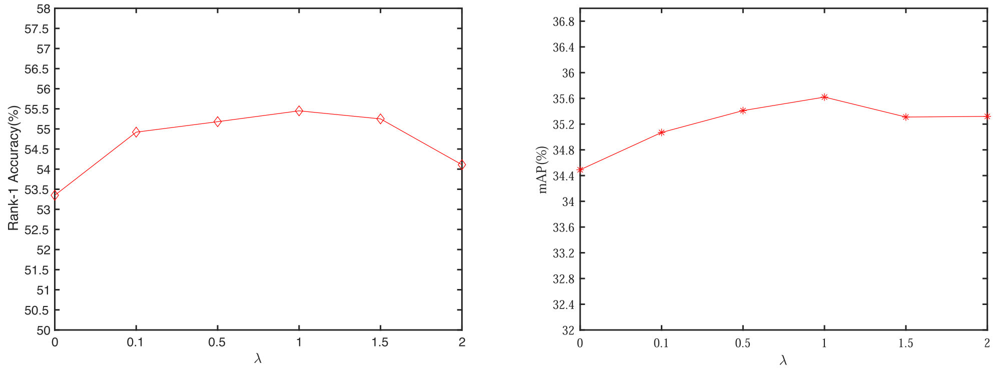


Fig. 7. Experiment results with different λ on CUHK-PEDES (fixed $\gamma = 0.5$, $k = 8$).

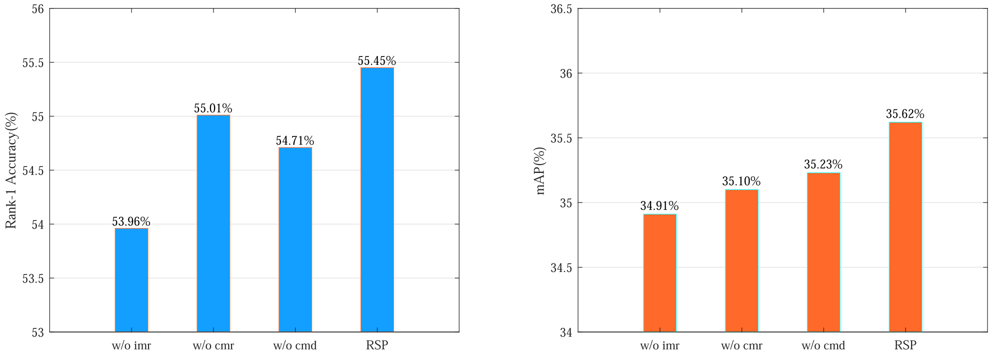


Fig. 8. Evaluation of the effectiveness of each component in RSP on the CUHK-PEDES dataset.

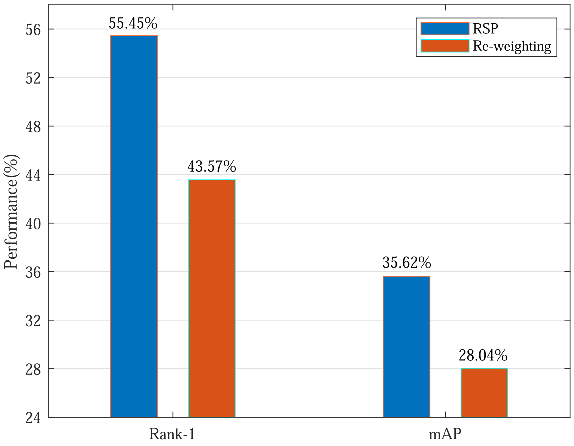


Fig. 9. Performance comparison of using RSP against Re-Weighting to obtain the final global embedding.

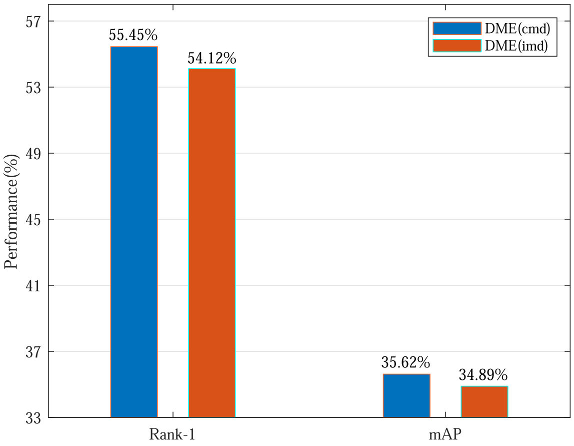


Fig. 10. Comparison of Cross-Modality redundancy (**cmd**) and Intra-Modality Redundancy (**imd**).

loss in Fig. 13. We can see that 1) the training losses continues to decrease; 2) the model converges quickly and stably.

4.6. Visualization

In Fig. 11, we illustrate the retrieval results of text-based person search. The true samples are outlined with orange. We compare DME with two methods: 1) the baseline which refers to DME without using Feature Dividing Network and Relevance based Subspace Projection; 2) DME without Relevance based Subspace Projection. For comparison, we show the results of three queries. The first query and the other two queries belong to different persons. The last two queries share the same ID but have different text descriptions. We observe that DME always ranks more true persons in the top of the ranking list than other two methods. Besides, we find an interesting observation from the retrieval results of person-2. The second image in the second row is a mismatched sample which has a very similar appearance with the candidate person. All three methods rank this sample in the top of the ranking list when using the description of person-2(a). However, when providing a more detailed text description, *i.e.*, 'carrying a white bag in his hand' of person-2(b), the proposed DME gives lower similarity for this mismatched sample while the other two methods still give high similarities. This demonstrates that the proposed Relevance based Subspace Projection can help the model to learn discriminative features from the relation between person and object.

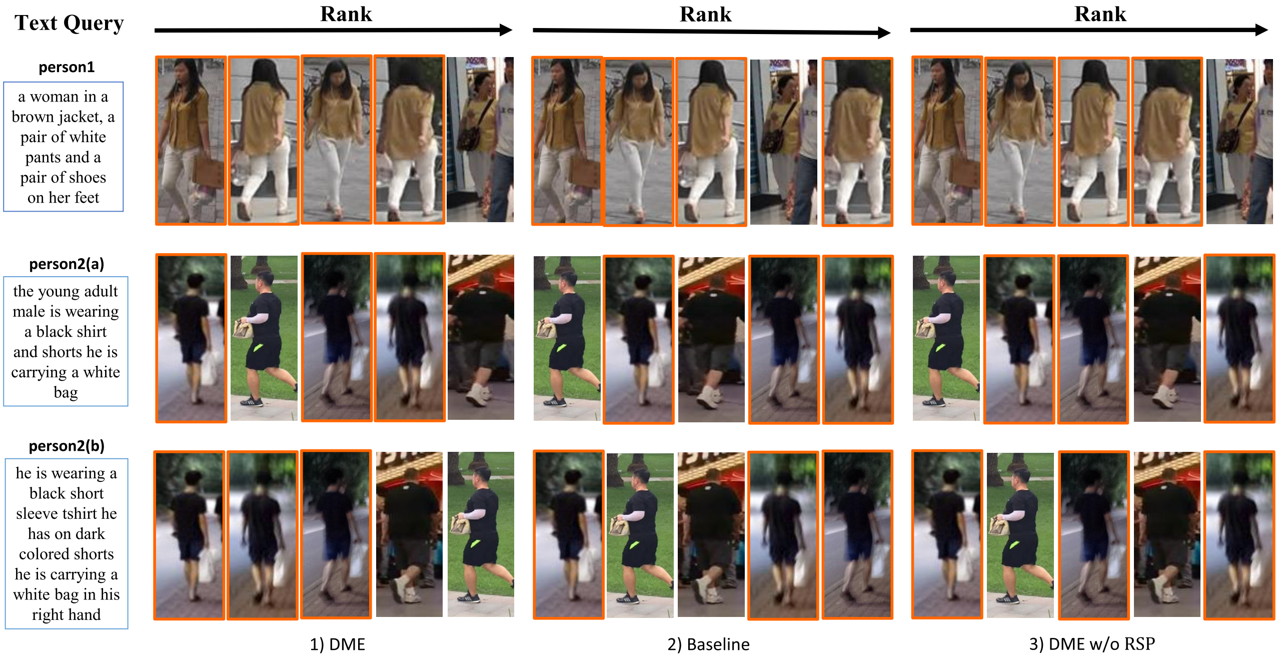

Text Query
Rank
Rank
Rank
person1
a woman in a brown jacket, a pair of white pants and a pair of shoes on her feet
person2(a)
the young adult male is wearing a black shirt and shorts he is carrying a white bag
person2(b)
he is wearing a black short sleeve tshirt he has on dark colored shorts he is carrying a white bag in his right hand
1) DME
2) Baseline
3) DME w/o RSP

Fig. 11. Examples of text-image retrieval results on CUHK-PEDES. The results are sorted by their cosine distances. Sample drawn in the range box represents the true person.

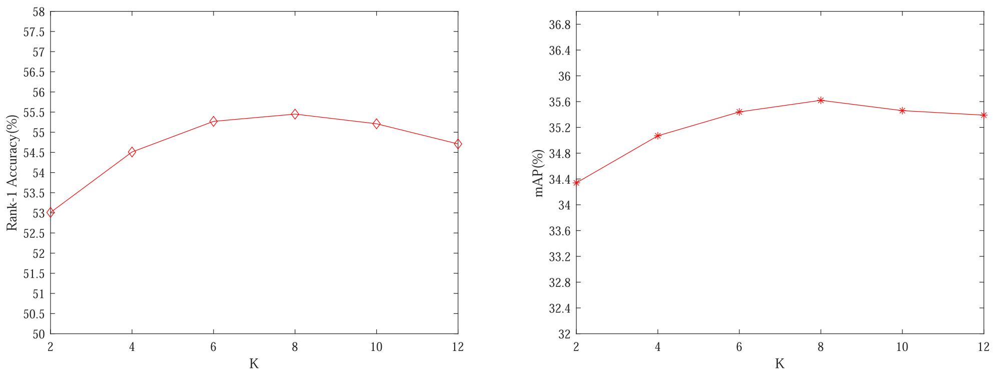
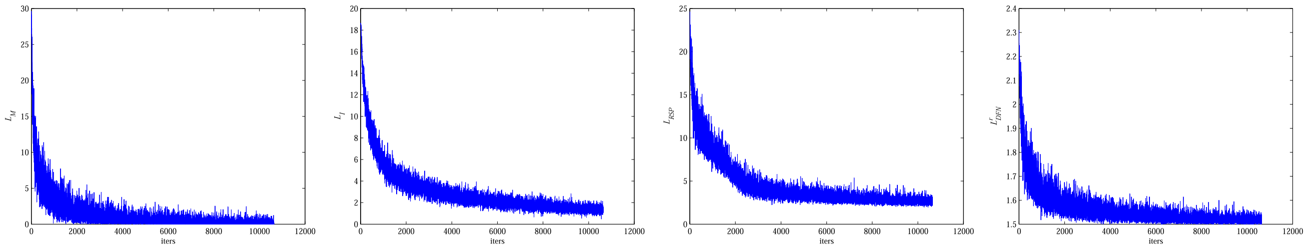


Fig. 12. Performance (Rank-1 and mAP) with different values of K (fixed $\gamma = 0.5$, $\lambda = 1.0$). The model trained with $K = 8$ achieves best performance. Experiments are evaluated on CUHK-PEDES.



(a) The curve of cross-modality matching loss. (b) The curve of identity classification loss. (c) The curve of relevance based subspace projection loss. (d) The curve of cross-modality part matching loss.

Fig. 13. The curves of each loss in the training procedure.

5. Conclusion and future works

Heterogeneous feature embedding is a crucial step in cross-modality person search. In this study, we propose a Divide-and-Merge Embedding (DME) learning framework for text-based person search. DME enables the model to generate embeddings of local parts (Feature Dividing Network) and effectively merge them into a global representation (Relevance based Subspace Projection). Consequently, the global representation contains discriminative details of local parts and can be used to compare samples of text and images directly. Experiments on three cross-modal retrieval datasets demonstrate the benefit of our method and show that our method can produce state-of-the-art performance.

For the future work, we would like to improve our model from the following three aspects: 1) The attributes are a mid-level feature. We will introduce attribute learning to tag the learned local representations. 2) The image quality has a significant impact on performance. We will investigate the impact of image quality on the text-based person search and try to design a solution. 3) The interaction between humans and objects plays an important role in the person search. We will exploit these interactions to improve performance.

CRediT authorship contribution statement

Chengji Wang: Conceptualization, Methodology, Software, Validation, Writing - original draft. **Zhiming Luo:** Supervision, Formal analysis, Writing - review & editing. **Zhun Zhong:** Investigation, Writing - review & editing. **Shaozi Li:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (No. 61876159, 61806172, 62076116, & U1705286); the Fundamental Research Funds for the Central Universities, Xiamen University (No. 20720200030).

References

[1] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, D.-S. Huang, Deep learning-based methods for person re-identification: A comprehensive review, Neurocomputing 337 (2019) 354–371.

[2] K. Chen, Y. Chen, C. Han, N. Sang, C. Gao, Hard sample mining makes person re-identification more efficient and accurate, Neurocomputing 382 (2020) 259–267.

[3] J. Jiang, K. Jin, M. Qi, Q. Wang, J. Wu, C. Chen, A cross-modal multi-granularity attention network for rgb-ir person re-identification, Neurocomputing..

[4] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5187–5196.

[5] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-aware textual-visual matching with latent co-attention, in: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 1908–1917.

[6] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, X. Wang, Improving deep visual representation for person re-identification by global and local image-language association, Proceedings of European Conference on Computer Vision (2018) 54–70.

[7] T. Chen, C. Xu, J. Luo, Improving text-based person search by spatial matching and adaptive threshold, in: Proceedings of Winter Conference on Computer Vision, 2018, pp. 1879–1887.

[8] J. Liu, Z.-J. Zha, R. Hong, M. Wang, Y. Zhang, Deep adversarial graph attention convolution network for text-based person search, in: Proceedings of ACM International Conference on Multimedia, 2019, pp. 665–673.

[9] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, Y. Shen, Dual-path convolutional image-text embedding, arXiv:1711.05535..

[10] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, Proceedings of European Conference on Computer Vision (2018).

[11] S. Aggarwal, V.B. RADHAKRISHNAN, A. Chakraborty, Text-based person search via attribute-aided matching, in: The IEEE Winter Conference on Applications of Computer Vision, 2020..

[12] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided joint global and attentive local matching network for text-based person search..

[13] Z. Lin, M. Feng, C.N.D. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, in: Proceedings of International Conference on Learning Representations, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of Conference and Workshop on Neural Information Processing Systems, 2017, pp. 5998–6008.

[15] C. Gao, R. Yao, J. Zhao, Y. Zhou, F. Hu, L. Li, Structure-aware person search with self-attention and online instance aggregation matching, Neurocomputing 369 (2019) 29–38.

[16] H. Zhang, I. Goodfellow, D.N. Metaxas, A. Odena, Self-attention generative adversarial networks (2019) 7354–7363.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 1.

[18] Y. Song, M. Soleymani, Polysemous visual-semantic embedding for cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[19] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 7 (1997) 711–720.

[20] X. He, P. Niyogi, Locality preserving projections, in: Proceedings of Conference and Workshop on Neural Information Processing Systems, 2004, pp. 153–160.

[21] X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: Proceedings of IEEE International Conference on Computer Vision, vol. 2, IEEE, 2005, pp. 1208–1213..

[22] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (10) (2015) 2010–2023.

[23] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of ACM International Conference on Multimedia, 2017, pp. 154–162.

[24] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Generalized semi-supervised and structured subspace learning for cross-modal retrieval, IEEE Transactions on Multimedia 20 (1) (2017) 128–141.

[25] N.C. Mithun, R. Panda, E.E. Papalexakis, A.K. Roychowdhury, Webly supervised joint embedding for cross-modal image-text retrieval, in: Proceedings of ACM International Conference on Multimedia, 2018, pp. 1856–1864.

[26] D. Semedo, J. Magalhães, Cross-modal subspace learning with scheduled adaptive margin constraints, in: Proceedings of ACM International Conference on Multimedia, 2019, pp. 75–83.

[27] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.

[28] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Feifei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[29] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[30] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 49–58.

[31] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, Proceedings of CVPR (2015) 3156–3164.

[32] N. Sarafianos, X. Xu, I.A. Kakadiaris, Adversarial representation learning for text-to-image matching, The IEEE International Conference on Computer Vision (2019).

[33] Z.S. Harris, Distributional structure, Word 10 (2–3) (1954) 146–162.

[34] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119..

[35] B. Klein, G. Lev, G. Sadeh, L. Wolf, Associating neural word embeddings with deep image representations using fisher vectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4437–4446.

[36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th ({USENIX}) symposium on operating systems design and implementation ({OSDI}) 16), 2016, pp. 265–283..

[37] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications..

[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[39] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (1) (2014) 1929–1958.

[41] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the Association for Computational Linguistics, vol. 2, 2016, pp. 207–212..

[42] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of International Conference on Learning Representations, 2014.

[43] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A.J. Smola, A kernel method for the two-sample-problem, in: Proceedings of ACM International Conference on Multimedia, 2007, pp. 513–520.

[44] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Scholkopf, A.J. Smola, A kernel two-sample test, Journal of Machine Learning Research 13 (1) (2012) 723–773.

Chengji Wang received the M.S. Degree from the Cognitive Science Department, Xiamen University, Xiamen, China, in 2018. He is currently working towards his Ph. D. at Xiamen University. His research interests include Person Re-identification and Multimedia Information Retrieval.

Zhiming Luo received the B.S. degree from the Cognitive Science Department, Xiamen University, Xiamen, China, in 2011; the Ph.D. degree in computer science with Xiamen University and University of Sherbrooke, Sherbrooke, QC, Canada, in 2017. His research interests include traffic surveillance video analytics, computer vision, and machine learning.

Zhun Zhong received the M.S. Degree in Computer Science and Technology from China University of Petroleum, Qingdao, China, in 2015. He is currently working towards his Ph.D. at Xiamen University. He is also a joint Ph.D. student at University of Technology Sydney. His research interests include person re-identification and domain adaptation.

Shaozi Li received the B.S. degree from Hunan University, and the M.S. degree from Xi'an Jiaotong University, and the Ph.D. degree from National University of Defense Technology. He currently serves as the Chair and Professor of Cognitive Science Department of Xiamen University, the Vice Director of Technical Committee on Collaborative Computing of CCF, the Vice Director of the Fujian Association of Artificial Intelligence. He is also the senior Member of IEEE, ACM and China Computer Federation (CCF). His research interests cover Artificial Intelligence and Its Applications, Moving Objects Detection and Recognition, Machine Learning, Computer Vision, Multimedia Information Retrieval, etc. He has directed and completed more than twenty research projects, including several National 863 Programs, National Nature Science Foundation of China, Ph.D. Programs Foundation of Ministry of Education of China.