Pattern Recognition Letters 116 (2018) 205–211

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Fusion-Attention Network for person search with free-form natural language[☆]

Zhong Ji[*], Shengjia Li, Yanwei Pang

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Available online 22 October 2018

MSC:
41A05
41A10
65D05
65D17

Keywords:
Person search
Natural language description
Attention network

ABSTRACT

In the task of searching persons from surveillance videos or large scale image dataset, it is more challenging to utilize free-form natural language to retrieve persons than using images and attributes. Thus, to deal with the challenges brought from the complexity of free-from natural language and visual-description mapping, we propose to strengthen the role of textual descriptions by means of fusion and attention mechanisms to make the discriminative words visually sensitive. Specifically, we develop an end-to-end fusion-attention structure, called Description-Strengthened Fusion-Attention Network (DSFA-Net) to tackle the challenging task. Specifically, DSFA-Net has a fusion sub-network and an attention sub-network, where three attention mechanisms are applied. Extensive experiments are performed on the large-scale CUHK-PEDES, which demonstrate the superiority of DSFA-Net.

1. Introduction

Person is one of the most important targets in intelligent perception and surveillance scenarios, especially with the increasing amount of cameras and the rapid development of autonomous driving. Many hot research topics about person are developed, such as pedestrian detection [2], person re-identification (re-id) [17–19,30], person attribute recognition [10,26], image caption [3,27], action recognition [29], and tracking of humans [9]. Very recently, an emerging challenging task, person description search [15], is proposed to address a more practical situation. It aims at retrieving relevant person instances with natural language description as a query in a large-scale person dataset, which is crucial to facilitate human accessible solutions for suspect identification.

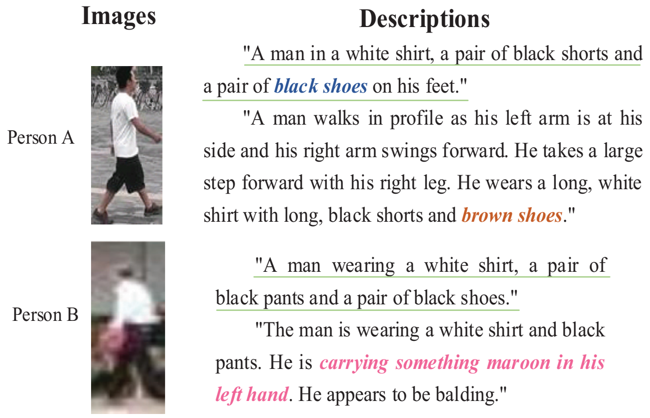
Instead of employing person image and attribute [12,14,30], searching person with natural language description has many merits [15]. For example, it requires no query image that is necessary in person re-id task. This is a practical relaxation, since an eye-witness description is much easier to obtain than a relevant image. Besides, comparing with attributes, it is capable of comprehensively describing the details of person appearance instead of being restricted to a few predefined attributes.

Although person description search has significantly practical value, there are many challenges. First, it is actually a cross-modal retrieval problem, more exactly, a text-based image retrieval task. The distribution and representation of text and image are inconsistent, which leads to a heterogeneity gap [24], making it hard to identify correspondences and measure similarities. Second, the relations between person appearance and language description are highly complex. Actually, person description search can be viewed as a fine-grained cross-modal retrieval task since it confines itself in a specific category of person. Therefore, its descriptions are usually fine-grained words, requiring the embedding of visual and description to capture subtle relationships among words. Take a popular person attribute recognition dataset RAP [13], for example, there are 72 attributes for describing a person, including body parts, gender, age, accessories, clothing style, actions, viewpoints, spatial-temporal information, etc. It is quite challenging to map the diversely visual appearance with these attributes, not to mention verbal descriptions. Third, the verbal descriptions are usually free-form and subjective. On one hand, different descriptions may represent the same person appearance; on the other hand, similar descriptions may describe quite different person appearances just because of the difference of one word. Some illustrations are shown in Fig. 1. Therefore, the discriminative words in verbal descriptions should be selected and paid more attentions. Fourth, person images are usually in low resolution. As person description search is usually applied in real-world surveillance scenarios, some per-

[☆] **Conflicts of interest:** There is no conflict of interest. We have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, 'Fusion-Attention Network for person search with free-form natural language'.

[*] Corresponding author.
E-mail address: jizhong@tju.edu.cn (Z. Ji).

https://doi.org/10.1016/j.patrec.2018.10.020
0167-8655/© 2018 Elsevier B.V. All rights reserved.



"A man in a white shirt, a pair of black shorts and a pair of *black shoes* on his feet."

"A man walks in profile as his left arm is at his side and his right arm swings forward. He takes a large step forward with his right leg. He wears a long, white shirt with long, black shorts and *brown shoes*."

"A man wearing a white shirt, a pair of black pants and a pair of black shoes."

"The man is wearing a white shirt and black pants. He is *carrying something maroon in his left hand*. He appears to be balding."

Fig. 1. Illustrations of person description search in CUHK-PEDES dataset. Each image has two free-from natural language descriptions, which contain diverse information and even distinct descriptions for the same person, such as brown shoes and black shoes for Person A, and carrying something maroon in his left hand in the second description for Person B. On the other hand, the first descriptions for Person A and B are similar. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

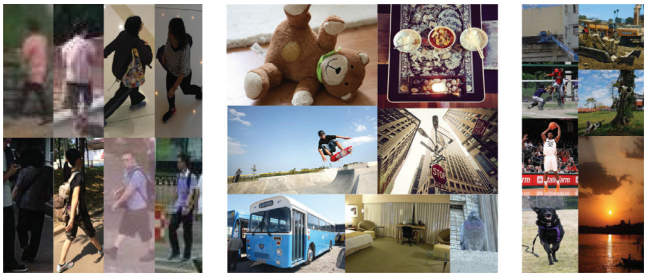

(a) CUHK-PEDES (b) MS-COCO (c) Filcker30k

Fig. 2. Image example comparison between person description search dataset (CUHK-PEDES) and related datasets (MS-COCO and Filcker30k). Relatively, images in CUHK-PEDES are low-resolution and fine-grained.

son images are captured in a fair far distance, which results to quite low-resolution images. These ambiguous images lead to poor quality of visual representations, which makes the mapping between description and image an exceptionally difficult task. Fig. 2 shows some image examples in person description search dataset (CUHK-PEDES) and related datasets (MS-COCO and Filcker30k). MS-COCO and Filcker30k are popular datasets in the field of image-text matching, and their images have high resolutions and belong to various categories, such as toys, rooms, architectures, persons, vehicles, animals, and views. Different from them, images in CUHK-PEDES are low-resolution and fine-grained. Thus, for the image-text matching tasks on MS-COCO and Filcker30k, the keywords in descriptions may play an important role in finding the correct images directly due to the significant differences between different categories. However, persons have great similarities in clothing, actions, and so on, which makes it hard to choose the discriminative keywords for deciding an accurate matching from the textual descriptions to the correct person images. That is to say, descriptions are required to be specially strengthened and analyzed.

As an emerging direction, there exists few research work on person description search. As a pioneering work, in 2017, Li et al. [15] release a dataset named CUHK-PEDES, and address this task by presenting an approach named Recurrent Neural Network with Gated Neural Attention (GNA-RNN). Its purpose is to learn a mapping between a description and a person image by capturing word-image relations with a Long Short-Term Memory (LSTM) network. Particularly, it utilizes a unit-level attention module to decide which visual units should be given more attention according to the input word, as well as a word-level gate to weight the importance of different words. Although good performance has been achieved, however, it still does not well highlight some challenges mentioned above, especially the third challenge. That is to say, it underestimates the significance of description information in the whole network, which may lead to weak relations between descriptions and images. We argue that more attentions should be paid on descriptions to capture a precise and reliable mapping between descriptions and person images since the verbal descriptions are highly free-form and complex. To this end, we propose a Description-Strengthened and Fusion-Attention Network (DSFA-Net) to cope with the challenges, especially to strengthen the description information in the mapping mechanism. Our DSFA-Net consists of two parts: a fusion sub-network and an attention sub-network, as illustrated in Fig. 3. The fusion sub-network fuses the information of descriptive words and image features to provide strong initial matching relations. Then, the attention sub-network performs a further matching operation by employing a fusion-guided attention mechanism to reliably map the relations and a description-guided attention mechanism to strengthen the role of some discriminative description words.

The contributions of this work are three-fold. First, we formulate person description search task in a fusion-attention recurrent neural network. It consists of a fusion sub-network and an attention sub-network, which are integrated into an end-to-end network to ensure a global optimal solution. Second, in the fusion sub-network, we apply an image-guided attention mechanism, which exploits visual content to guide the weight of description words. Then we utilize an LSTM network to model contextual relations. It guarantees a strong initial matching between description words and image features. Third, in the attention sub-network, we apply the fusion information obtained in the fusion sub-network as an input to further direct word attention to the corresponding visual pattern. This fusion-guided attention mechanism ensures reliable mapping relations. Besides, we propose a description-guided attention mechanism to provide weights for different description words to strengthen the role of the discriminative words.

The rest of the paper is organized as follows. Related studies are introduced in the next section. We describe the proposed DSFA-Net in Section 3. Experimental results and analyses are given in Section 4, and Section 5 concludes this paper.

2. Related work

2.1. Cross-modal retrieval

Person search with free-form natural language is actually employing textual description to search person images, which belongs to the field of cross-modal retrieval [7,16,21]. We can also view it as a fine-grained cross-modal retrieval task. In recent years, some promising methods have been put forward for cross-modal retrieval, especially the visual-textual retrieval and mapping. Approaches rise from linear embedding approaches, such as CCA [7] to deep frameworks, such as [21,23,26,27]. Generally speaking, most deep models utilize Convolutional Neural Networks (CNN) for image feature extraction and Recurrent Neural Networks (RNN) for visual-textual mapping. For example, Mao *et al.* [21] learn a common embedding space between the two modalities via the combination of CNN and RNN, trained with Softmax loss regularized by the L2 norm. Reed et al. [26] learn a cross-modal mapping function at the character level by training convolutional and recurrent components (CNN-RNN) end-to-end. Niu et al. [23] present to replace the chain-structured RNNs with a tree-structured LSTM as text encoder to learn the hierarchical relations between images and sentences. Wang et al. [28] propose a two-branch neural net-

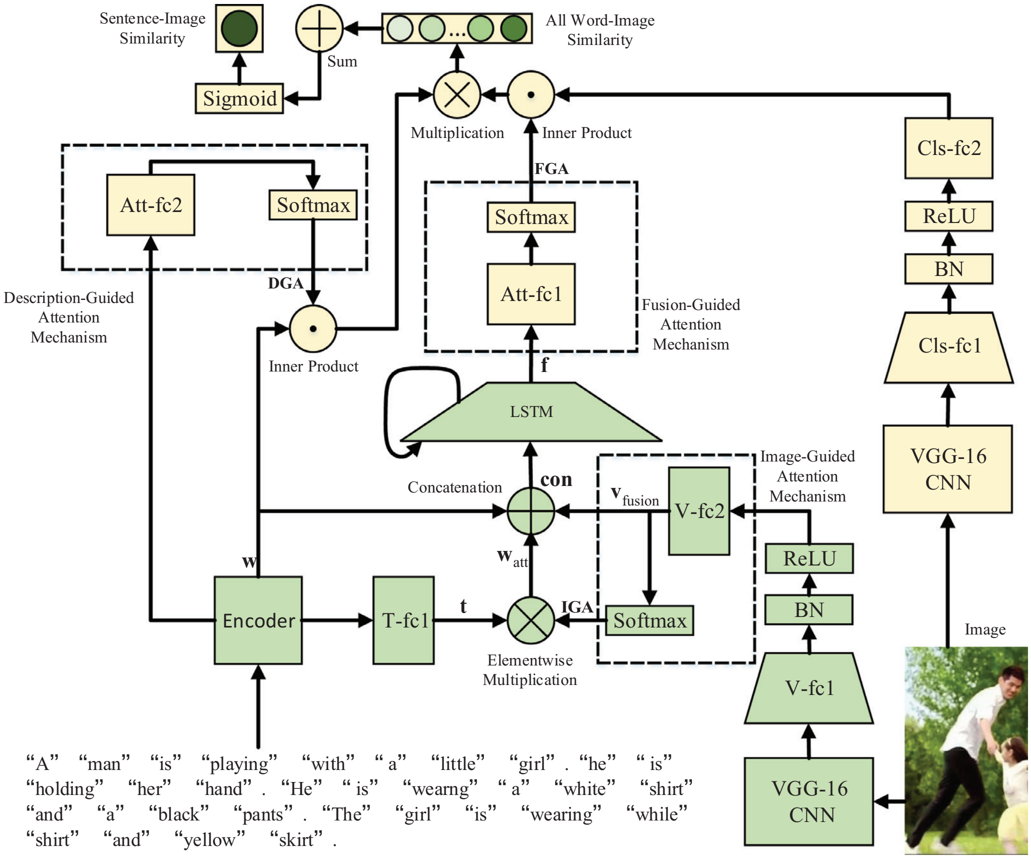

Sentence-Image Similarity
Sum
All Word-Image Similarity
Sigmoid
Multiplication
Inner Product
FGA
Cls-fc2
ReLU
BN
Cls-fc1
Att-fc2
Softmax
DGA
Description-Guided Attention Mechanism
Inner Product
Softmax
Att-fc1
Fusion-Guided Attention Mechanism
f
LSTM
VGG-16 CNN
Concatenation
con
v fusion
V-fc2
Image-Guided Attention Mechanism
ReLU
BN
w
w att
Encoder
T-fc1
t
IGA
Softmax
Elementwise Multiplication
V-fc1
Image
VGG-16 CNN
"A" "man" "is" "playing" "with" " a" "little" "girl" . "he" " is" "holding" "her" "hand" . "He" " is" "wearng" " a" "white" "shirt" "and" "a" "black" "pants" . "The" "girl" "is" "wearing" "while" "shirt" "and" "yellow" "skirt" .

Fig. 3. Flowchart of DSFA-Net. It consists of a fusion sub-network (in green) and an attention sub-network (in yellow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

work with multiple layers of linear projections followed by nonlinearities, which can learn joint embedding of images and texts. Recently, researchers show great interest in deep attention model in this field. For example, Nam et al. [22] propose Dual Attention Networks, which utilizes visual and textual attention mechanisms to capture fine-grained interaction effects between vision and text. Qi et al. [25] propose a visual-language relation attention model to explore not only the local fine-grained matchings between image regions and keywords, but the relations in visual and textual contexts as well.

As a fine-grained cross-modal retrieval task, person search with free-form description is proposed by Li et al. in 2017. They also employ the framework of CNN+RNN. In addition, they apply unit-level and word-level gated neural attention mechanisms on it to determine which visual units should be given more attention according to the input word. In this paper, we follow this idea, however, we present some novel attention mechanisms to further fuse the multi-modal information and strengthen the role of descriptions.

2.2. Person attribute recognition

Person attribute recognition aims at automatically recognizing attributes such as gender, age, clothing style and accessory from person images. Its inverse process is to retrieve the person images with person attributes, which is highly close to person search with descriptions. However, comparing with free-form language description, person attributes are fixed and scarce in descripting the pedestrian. In addition, attributes are all useful words and do not contain the words for sentence structure and grammar. It is another important task in intelligent surveillance, and attracts increasing attentions in recent years. Early work focuses on applying independent attribute classifiers to recognize attributes. For example, Zhu et al. [32] and Deng et al. [5] employ algorithms of AdaBoost and SVM to train independent attribute classifiers, respectively. Li et al. [12] propose a DeepMAR framework based on CNN to recognize multiple attribute simultaneously with a loss function jointly considering all attributes. Recently, it is a popular idea to leverage the framework of CNN+RNN and attention model. For example, Neural PAR [10] is one of the first work to make use of CNN+RNN framework for person attribute recognition. It formulates the task as an end-to-end image to attribute description problem, where the attribute words are concatenated into different attribute sentences to well contextualize the potential relationships among them. And a neural network model is trained based on CNN and LSTM to learn the complex relations between visual features and their corresponding attributes. Based on the idea that person analysis requires a comprehensive feature representation from multi-levels and scales, HPnet [20] develops an attention-based deep neural network by multi-directionally feeding the multi-level attention maps to different feature layers. In this way, it captures multiple attentions from low-level to semantic-level and exploits multi-scale selectiveness of attentive features.

3. Proposed approach

The inherent challenges make person description search a hard task. Since the relations between description and person image are highly complex, and the verbal descriptions are free-form and subjective, the key solution to this task is to build a reliable cross-modal mapping as well as strengthen the role of description. Whats more, as demonstrated in [15], the importance of different word types is quite different. The authors show that nouns provide most information followed by the adjectives, and the verbs. Of course, prepositions and articles have little information. Therefore, we should treat the words differently, and lay more attention to those discriminative keywords, especially to strength those unique and personalized words. To this end, we develop an end-to-end neural network, called Description-Strengthened and Fusion-Attention Network (DSFA-Net) to address these challenges.

As illustrated in Fig. 3, the proposed DSFA-Net consists of two sub-networks. Specifically, the fusion sub-network first applies an Image-Guided Attention (IGA) mechanism to initially match the word-image relations. Its output together with feature of description word and images form a concatenation vector, which is served as input to an LSTM network. The LSTM network makes a deep fusion for these information and builds long-term contextual relations for description sentences and images. Its output is used as one of the inputs to the attention sub-network, which generates the attention vectors by employing a Fusion-Guided Attention (FGA) mechanism. Its purpose is to weight the matching degrees between words and image local patterns. In this way, the affinity between the description sentence and the person image at certain word can be obtained. In addition, to strengthen the role of description sentences, we present a Description-Guided Attention (DGA) mechanism to further offer attention weights to the word-level affinities. Finally, the sentence-image affinity is obtained by summing all word-level affinities in a sentence. The details are introduced in the following.

3.1. Fusion sub-network

Person images and their corresponding description sentences are used as inputs for the fusion sub-network. First, a 512D visual vector $\mathbf{v}_{fusion}$ of a person image is extracted by a VGG-16 CNN network, followed by two extra 512D fully-connected layers (V-fc1 and V-fc2). It is worth noticing that the VGG-16 CNN network is pre-trained, and only V-fc1 and V-fc2 will be trained during the end-to-end training for the whole network. Meanwhile, the words in the corresponding free-form sentence are fed into the fusion network one by one. The words are encoded into 512D word vectors, which is represented by $\mathbf{w}$.

Then, an IGA mechanism is applied on visual vector $\mathbf{v}_{fusion}$ by employing a Softmax function to obtain an attention vector named IGA vector. With an elementwise multiplication operation, IGA vector acts on $\mathbf{t}$ that is obtained by feeding $\mathbf{w}$ into a fully-connected layer, the output is called attention word vector $\mathbf{w}_{att}$. In this way, IGA offers an initial matching for word and image by weighting the importance of words guided by image content. The motivation of IGA step is to strengthen role of description sentence in the fusion stage.

Finally, visual vector $\mathbf{v}_{fusion}$, word vector $\mathbf{w}$ and attention word vector $\mathbf{w}_{att}$ are concatenated together as a concatenation vector $\mathbf{con} = \left[\mathbf{v}_{fusion}, \mathbf{w}, \mathbf{w}_{att}\right]$, which is input into an LSTM network. LSTM is a special type of RNN and is skilled in learning temporal relations of sequential data. In the method of concatenation, we make sure that the vector contains original features and initial matching feature before fusion. The LSTM fuses the input information during the training with input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$, output gate $\mathbf{o}_t$, memory cell $\mathbf{c}_t$ and hidden state $\mathbf{h}_t$ in the following way:

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}_i^c \cdot \mathbf{con}_t + \mathbf{W}_i^h \cdot \mathbf{h}_{t-1} + \mathbf{b}_i\right),\\
\mathbf{f}_t &= \sigma\left(\mathbf{W}_f^c \cdot \mathbf{con}_t + \mathbf{W}_f^h \cdot \mathbf{h}_{t-1} + \mathbf{b}_f\right),\\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_o^c \cdot \mathbf{con}_t + \mathbf{W}_o^h \cdot \mathbf{h}_{t-1} + \mathbf{b}_o\right),\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot T\left(\mathbf{W}_c^c \cdot \mathbf{con}_t + \mathbf{W}_c^h \cdot \mathbf{h}_{t-1} + \mathbf{b}_c\right),\\
\mathbf{h}_t &= \mathbf{o}_t \odot T(\mathbf{c}_t),
\end{aligned} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{b}$ are trained parameters, σ represents Sigmoid function, $T()$ represents Tanh function, $\odot$ represents the element-wise multiplication. The LSTM network makes a deep fusion for images and their corresponding description information and builds long-term contextual relations for description sentences and images. Its output is fusion vector $\mathbf{f}$.

Besides, peephole LSTM [6], CIFG LSTM [8] and GRU [4] are variations for LSTM, which can also replace the standard LSTM in the fusion sub-network. Specifically, Peephole LSTM aims at strengthening the effects of memory cell on the input gate, forget gate and output gate and gains more information from memory cell. CIFG LSTM couples the input and forget gate as one to gate the input and the memory cell. GRU does not have separate memory cells and makes each recurrent unit to capture long-term dependencies adaptively. We will report their respective performance in the experiments.

3.2. Attention sub-network

There are three parts in attention sub-network. The first is the visual representation part, which has the same network structure as that in the fusion sub-network. However, both visual representation parts are trained separately. The second part is a fully-connected layer (Att-fc1) followed by a Softmax layer. This part acts as a Fusion-Guided Attention (FGA) mechanism. That is to say, it takes fusion vector $\mathbf{f}$ as input, and outputs a FGA vector, which is a probability vector used as attention to the visual content with an inner product. The FGA vector decides which visual pattern responses should be summed up to calculate the affinity value. The affinity between description word $\mathbf{w}$ and the person image is obtained by

$$r = \sum_{j=1}^{m} d_j \cdot \alpha_j \quad s.t. \sum_{j=1}^{m} \alpha_j = 1 \tag{2}$$

where m is the dimensionality of image feature vector, d_j represents the image feature, and α_j represents the weight of each image feature.

The third one in attention sub-network is a Description-Guided Attention (DGA) part, which is constructed by a fully-connected layer (Att-fc2) followed by a Softmax layer. Although FGA part can build tight relations for description and person image, however, since the verbal descriptions are usually free-form and subjective, it is necessary to further strengthen the role of description. According to the user study [15], different words carry different roles in building the description and image affinity. For example, the words such as ‘a’, ‘the’, ‘to’, ‘with’ and ‘is’ are common words for sentences, and are not always useful in building the relations. Therefore, we propose to feed the word vector into the Description-Guided Attention (DGA) mechanism to generate the weight for each word. It is realized by performing an inner product between DGA vector the word vector $\mathbf{w}$, which not only just retains word features but also obtains 1D weights. Then, its output q is used a gate to get the final word-image affinity s:

$$s = q \cdot r \tag{3}$$

Finally, the sentence-person image affinity is obtained by aggregating all the word-image affinities:

$$S = \sigma\left(\sum_{k=1}^{N} s^k\right) \tag{4}$$

where N is the number of the words in a sentence, and σ represents the Sigmoid function.

3.3. Training settings

Except for the pre-trained VGG-16 part, the proposed DSFA-Net is trained with batched Adam algorithm end-to-end and without fine-tuning. Each training batch has 16 sentence-image pairs, and all the fully-connected layers are 512D. The training instances are selected randomly with corresponding sentence-image pairs as positive instances and non-corresponding pairs as negative instances. The ratio between the positive and negative instances is 1:3.

Meanwhile, we choose the cross-entropy loss function to train our DSFA-Net,

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y^i logS^i + \left(1 - y^i\right)log\left(1 - S^i\right)\right] \tag{5}$$

where N denotes the number of instances, S^i denotes the predicted affinity of the ith instance, and $y^i = 1$ denotes there is a corresponding sentence-image pair.

4. Experimental results and analysis

4.1. Datasets and settings

Extensive experiments are carried out on the recently released CUHK-PEDES dataset [15] to validate the effectiveness of the proposed DSFA-Net. CUHK-PEDES is the only dataset in this field. It contains 40,206 images from 13,003 different pedestrians, which are selected from some popular person re-identification datasets, such as CUHK03 and Market-1501. Each person has an average of 3.1 images, and each image is described by 2 sentences. The average sentence length is 23.5 words, which is quite longer than the 5.18 words of MS-COCO dataset.

According to the setting in ([15]), we choose 34,054 images with 68,126 captions as training set, 3078 images with 6156 captions as validation set, and 3074 images with 6148 as test set. Specifically, both the validation and test sets have 1000 persons. The *top-k* accuracy is utilized to evaluate the performance of person search with free-form language sentence, which is the possibility that the true match appears in the top k of the rank list. That is to say, if any images of the corresponding person are among the top k images, it is a successful search:

$$top\text{-}k = \frac{c}{a} \tag{6}$$

where c denotes the times of the successful searches, and a denotes the total times of searches.

Except for the pre-trained VGG-16 part, the proposed DSFA-Net is trained with batched Adam without fine-tuning over one Nvidia GTX1080Ti GPU in a Torch platform. Each training batch has 16 sentence-image pairs. All the fully-connected layers are 512D. The layer number of LSTM is one, and its learning rate is set to 4×10^{-4}, which starts to decay at 50,000 iterations. The learning rate for VGG-16 is set to 1×10^{-5} and do not decay. We save the model checkpoint every 500 iterations and use 3000 images to evaluate validation loss. Training time is around 16 hours and testing time is around 35 minutes.

4.2. Comparison results

We choose 6 state-of-the-art approaches for comparison, which are iBOWIMG [31], CNN-RNN [26], GMM+HGLMM [11], NeuralTalk [27], QAWord [15] and, and GNA-RNN [15]. Specifically, iBOWIMG [31] is a model for visual question answering. CNN-RNN [26] learns a matching function between images and descriptions at the character level by training convolutional and recurrent components end-to-end. GMM+HGLMM [11] first applies the fusion of GMM and Hybrid Gaussian–Laplacian Mixture Model (HGLMM) to learn Fisher vector representations of sentences, where HGLMM is a weighted geometric mean of the Laplacian distribution and the Gaussian distribution. Then, the generated Fisher vector and VGG image features are associated to achieve the text-to-image mapping. The NeuralTalk [27] looks for the latent alignment between segments of sentences and image regions in a joint embedding space for sentence generation. QAWord [15] is a modification of LSTM Q + norm I method [1], where the element-wise multiplication between the question and image features and the multi-class classifier are replaced with concatenation operation and binary classifier, respectively. In addition, the network structure is also changed into the form of one LSTM layer and two additional fully-connected layers. GNA-RNN [15] is a recurrent neural network with gated neural attention, which achieves good performance in this task. For a fair comparison, the image features for all comparative methods are from VGG-16 network pre-trained model.

Table 1
Comparison results of the proposed DSFA-Net and state-of-the-art methods on the CUHK-PEDES dataset.

	top-1	top-10
iBOWIMG [31]	8.00	30.56
CNN-RNN [26]	10.48	36.66
GMM+HGLMM [11]	15.03	42.27
NeuralTalk [27]	13.66	41.72
QAWord [15]	11.62	42.42
GNA-RNN [15]	19.05	53.64
DSFA-Net(ours)	**20.33**	**54.90**

Table 1 shows the top-1 and top-10 accuracies of our DSFA-Net and the comparative methods. It can be observed that the proposed DSFA-Net achieves the best performance. It outperforms the second-best method, GAN-RNN, in 1.28% and 1.26% absolute percentage points for top-1 and top-10, respectively.

4.3. Ablation studies

This section investigates the effects of attention sub-network and fusion sub-network in the DSFA-Net by ablation studies on GNA-RNN [15]. That is to say, we apply the two sub-networks to GNA-RNN respectively to demonstrate that GNA-RNN is beneficial

Table 2
Ablation studies of attention sub-network and fusion sub-network on GNA-RNN.

	top-1	top-5	top-10
GNA-RNN [15]	19.05	41.44	53.64
GNA-RNN+ Attention sub-network	19.80	41.18	53.07
GNA-RNN+ Fusion sub-network	19.42	42.05	54.23
DSFA-Net(ours)	**20.33**	**42.40**	**54.90**

As shown in Table 2, GNA-RNN with our proposed attention sub-network in top-1 accuracy is 19.80%, which outperforms GNA-RNN by 0.75%. However, top-5 and top-10 accuracies are 41.18% and 53.07%, respectively, which slightly drop by 0.36% and 0.57%. Our attention sub-network makes each word more discriminative and enlarges the priorities among useful and useless words. However, the weak fusion vectors without text strengthening offered by GNA-RNN cannot capture the words contributing to the image efficiently so that some useful words may be considered as useless words. Thus, the matchings decided by a few useful words establish better top-1 performances and the matchings decided many useful words gain worse top-5 and top-10 results.

On the other hand, GNA-RNN with our proposed fusion sub-network has an overall performance improvement. Compared with GNA-RNN, it improves by 0.37%, 0.61%, and 0.59% in top-1, top-5,

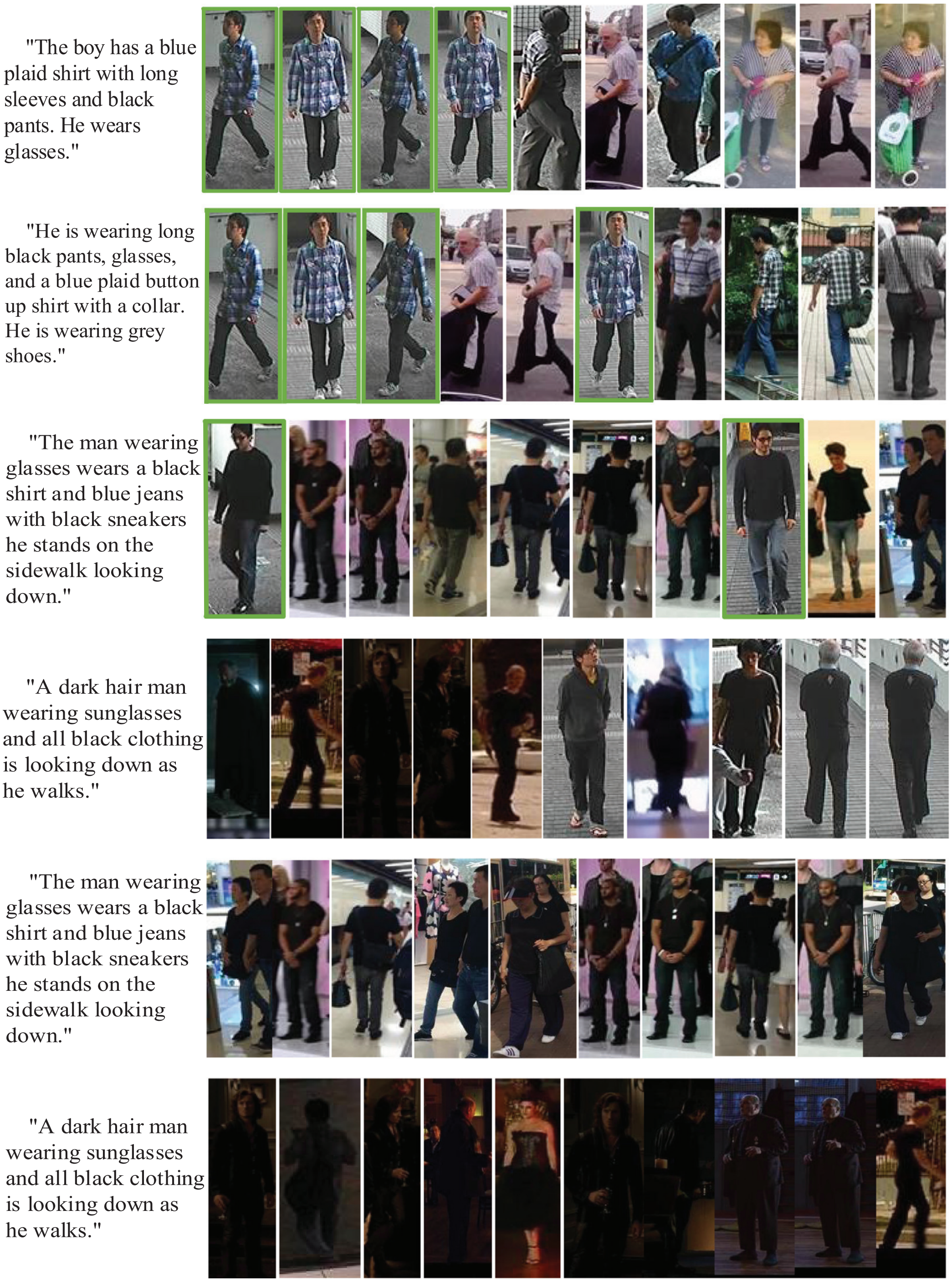

"The boy has a blue plaid shirt with long sleeves and black pants. He wears glasses."
"He is wearing long black pants, glasses, and a blue plaid button up shirt with a collar. He is wearing grey shoes."
"The man wearing glasses wears a black shirt and blue jeans with black sneakers he stands on the sidewalk looking down."
"A dark hair man wearing sunglasses and all black clothing is looking down as he walks."
"The man wearing glasses wears a black shirt and blue jeans with black sneakers he stands on the sidewalk looking down."
"A dark hair man wearing sunglasses and all black clothing is looking down as he walks."

Fig. 4. Examples of top-10 person search results with free-form natural language by DSFA-Net with CIFG LSTM (Rows 1–4) and GNA-RNN (Rows 5–6). Correct images are marked by green rectangles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and top-10, respectively. It can be observed that, relatively speaking, attention sub-network is helpful to enhance the performance of the top-1, while Fusion sub-network is more helpful to enhance the performances of the top-5 and top-10. When combine them together, i.e., the proposed DSFA-Net, achieves the overall performance improvements on three cases. Moreover, DSFA-Net beats GNA-RNN + Attention sub-network and GNA-RNN + Fusion sub-network, which proves that both sub-networks benefit from each other.

4.4. Further analysis

LSTM has many variants, each of which makes different performances on different tasks. It is not certain wether a LSTM variant is better than others. Thus, it is necessary to test our proposed DSFA-Net with some LSTM variants. We choose Peephole LSTM [6], GRU [4] and CIFG LSTM [8] to replace the standard LSTM in DSFA-Net. The results are shown in Table 3. We can observe that DSFA-Net achieves the best performance in top-1, while DSFA+CIFG LSTM

Table 3
Variants of DSFA-Net. DSFA refers DSFA-Net without traditional LSTM.

	top-1	top-5	top-10
DSFA+LSTM (DSFA-Net)	**20.33**	42.40	54.90
DSFA+ Peephole LSTM	20.04	42.26	54.39
DSFA+ GRU	19.36	40.92	52.77
DSFA+ CIFG LSTM	20.14	**43.22**	**55.11**

achieves the best performances in top-5 and top-10. Therefore, we can conclude that CIFG LSTM is the best LSTM variant for DSFA-Net, and followed by the standard LSTM, while Peephole LSTM and GRU have inferior role against LSTM. It is worth mentioning that GRU is a popular simplified LSTM, but it may be weak in capturing fine-grained features.

We use the model of DSFA with CIFG LSTM to provide some search examples, as illustrated in Fig. 4. Queries of rows 1 and 2 are two descriptions for the same person. We can observe that the correctly corresponding images are all ranked in the top positions. Failure examples are shown in rows 3 and 4, which are tow search results for the same person. Our model finds two correct images out of four in row 3, and misses all the correct images in top-10 in row 4. However, although these images are not the desired ground-truth persons, they still match the language descriptions well. It is worth pointing that the descriptions are quite subjective. Take descriptions for row 3 and 4 for examples. Although they are for the same person, their appearance descriptions are rather different: one is a black shirt and blue jeans and the other is all black clothing. The cause of this phenomenon is the blue jeans in dark environment confuses peoples judgement and description. Thus, in row 4, the subjective and inaccurate description leads to the wrong results according to the ground-truth though some searched images are visually matched. From another point of view, we can find the main difference for query 3 and 4 is blue and black. And the corresponding results reflect this difference, which proves that our proposed model is effective and sensitive.

Further, we compare ours results in row 3–4 with those of GNA-RNN, which is shown in rows 5 and 6. First, as shown in row 3 and 5, DSFA-Net finds two correct person images, while GNA-RNN find no one and ignores wearing glasses information. Meanwhile, in row 4 and 6, GNA-RNN ignores wearing sunglasses information and are visually darker as whole, which proves that they build closer relations to black than black clothing. In contrast, DSFA-Net finds the images with a more brighter background and captures glasses information, which perfectly close to the word of sunglasses.

5. Conclusion

In this paper, we propose an end-to-end fusion-attention structure (DSFA-Net) on person search with free-form natural language. Due to the complexity of free-form natural language, the main idea of DSFA-Net is to strengthen the textual descriptions to make keywords more discriminative in constructing the visual-description matching by employing a fusion and an attention sub-networks. Extensive experiments have proved the superiority of DSFA-Net.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grants of 61771329, 61472273 and 61632018.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Zitnick, D. Parikh, J. Dean, Vqa: visual question answering, Int. J. Comput. Vis. 123 (1) (2017) 1–28.
[2] J. Cao, Y. Pang, X. Li, Learning multilayer channel features for pedestrian detection, IEEE Trans. Image Process. 26 (2017) 3210–3220.
[3] H. Chen, G. Ding, S. Zhao, J. Han, Temporal-difference learning with sampling baseline for image captioning, Association for the Advancement of Artificial Intelligence, 2018.
[4] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv:1412.3555v1 (2014).
[5] Y. Deng, P. Luo, C. Chen, X. Tang, Pedestrian attribute recognition at far distance, in: ACM International Conference on Multimedia, 2014, pp. 789–792.
[6] F. Gers, J. Schmidhuber, Recurrent nets that time and count, in: IEEE-INNS-ENNS International Joint Conference on Neural Networks, 3, 2000, pp. 189–194.
[7] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, Int. J. Comput. Vis. 106 (2) (2014) 210–233.
[8] K. Greff, R. Srivastava, J. Koutnik, B. Steunebrink, J. Schmidhuber, Lstm: a search space odyssey, IEEE Trans. Neural Netw. Learn.Syst. 28 (10) (2017) 2222–2232.
[9] J. Han, E. Pauwels, D. Zeeuw, Employing a rgb-d sensor for real-time tracking of humans across multiple re-entries in a smart environment, IEEE Trans. Consum. Electron. 58 (2) (2012) 255–263.
[10] Z. Ji, W. Zheng, Y. Pang, Deep pedestrian attribute recognition based on lstm, in: IEEE International Conference on Image Processing, 2017, pp. 151–155.
[11] B. Klein, G. Lev, G. Sadeh, L. Wolf, Associating neural word embeddings with deep image representations using fisher vectors, in: Internaltional Conference on Computer Vision and Pattern Recognition, 2015, pp. 4437–4446.
[12] D. Li, X. Chen, K. Huang, Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios, in: IAPR Asian Conference on Pattern Recognition (ACPR), 2016, pp. 111–115.
[13] D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A richly annotated dataset for pedestrian attribute recognition, arXiv:1603.07054v1(2016b).
[14] J. Li, C. Xu, W. Yang, C. Sun, Spa: spatially pooled attributes for image retrieval, Neurocomputing 257 (2017) 47–58.
[15] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: Internaltional Conference on Computer Vision and Pattern Recognition, 2017, pp. 5187–5196.
[16] H. Liu, F. Wang, X. Zhang, F. Sun, Weakly-paired deep dictionary learning for cross-modal retrieval, Pattern Recognit. Lett. (2018).
[17] J. Liu, Z. Zha, X. Chen, Z. Wang, Y. Zhang, Dense 3d-convolutional neural network for person re-identification in videos, ACM Trans. Multimed. Comput. Commun.Appl. 14 (2018).
[18] J. Liu, Z. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, T. Mei, Multi-scale triplet cnn for person re-identification, in: ACM on Multimedia Conference, 2016, pp. 192–196.
[19] J. Liu, Z. Zha, H. Xie, Z. Xiong, Y. Zhang, Ca3net: contextual-attentional attribute-appearance network for person re-identification, in: ACM on Multimedia Conference, 2018.
[20] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-net: attentive deep features for pedestrian analysisg, in: IEEE International Conference on Computer Vision, 2017, pp. 350–359.
[21] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks(m-rnn), arXiv:1412.6632v1 (2014).
[22] H. Nam, J. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: International Conference on Computer Vision and Pattern Recognition, 2017, pp. 2156–2164.
[23] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Hierarchical multimodal lstm for dense visual-semantic embedding, in: IEEE International Conference on Computer Vision, 2017, pp. 1899–1907.
[24] Y. Peng, J. Qi, X. Huang, Y. Yuan, Ccl: cross-modal correlation learning with multigrained fusion by hierarchical network, IEEE Trans. Multimed. 20 (2) (2018) 405–420.
[25] J. Qi, Y. Peng, Y. Yuan, Cross-media multi-level alignment with relation attention network, arXiv:1804.0953v1 (2018).
[26] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: International Conference on Computer Vision and Pattern Recogintion, 2016, pp. 49–58.
[27] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: International Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[28] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5005–5013.
[29] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, L. Shao, Action recognition using 3d histograms of texture and a multi-class boosting classifier, IEEE Trans. Image Process. 26 (10) (2017) 4648–4660.
[30] C. Zhao, K. Chen, Z. Wei, Y. Chen, D. Miao, W. Wang, Multilevel triplet deep learning model for person re-identification, Pattern Recognit. Lett. (2018), doi:10.1016/j.patrec.2018.04.029.
[31] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, R. Fergus, Simple baseline for visual question answering, arXiv:1512.02167v1 (2015).
[32] J. Zhu, S. Liao, Z. Lei, D. Yi, S. Li, Pedestrian attribute classification in surveillance: database and evaluation, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 331–338.