Pattern Recognition 122 (2022) 108217

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-task framework based on feature separation and reconstruction for cross-modal retrieval

Li Zhang, Xiangqian Wu*

School of computer science and technology, Harbin Institute of Technology, China

ARTICLE INFO

Article history:
Received 1 July 2020
Revised 19 July 2021
Accepted 31 July 2021
Available online 2 August 2021

MSC:
00-01
99-00

Keywords:
Cross-modal retrieval
Feature separation
Image reconstruction
Text reconstruction

ABSTRACT

Cross-modal retrieval has become a hot research topic in both computer vision and natural language processing areas. Learning intermediate common space for features of different modalities has become one of mainstream methods. In this paper, we propose a novel multi-task framework based on feature separation and reconstruction (mFSR) for cross-modal retrieval based on common space learning methods, which introduces feature separation module to deal with information asymmetry between different modalities, and introduces image and text reconstruction module to improve the quality of feature separation module. Extensive experiments on MS-COCO and Flickr30K datasets demonstrate that feature separation and specific information reconstruction can significantly improve the baseline performance of cross-modal image-caption retrieval.

1. Introduction

With rapid development of multi-media, there is a tremendous amount of information on the Internet, such as image, text, video, audio, etc. Obtaining useful information among different modalities in massive data by hand becomes more and more difficult. Naturally, we need a powerful method to help us obtain texts, images or videos that we need. Cross-modal retrieval takes one modality of data as the query to retrieve relevant data of another modality. For example, we can use texts to retrieve interested images (just like what we do on Google Image Search), or use images to retrieve the corresponding texts. Of course, the modality is not restricted to image and text, the other modalities such as speech, physical signals and video can also be used as a component of cross-modal retrieval.

According to our statistics, research on cross-modal retrieval originate from 2006 [1]. From then on, the research about cross-modal retrieval is in the ascendant. The main problem of cross-modal retrieval is "modal gap", which means that representations of different modalities are inconsistent and lie in different feature spaces, so it is extremely challenging to measure similarities among them. There have been various methods proposed for addressing this problem by analyzing the rich correlations contained in cross-modal data, such as: common space learning method, graph-based method, neighbor analysis, multi-modal topic model, etc. [2] Due to the good performance of deep learning in feature extraction, common space learning method, which learning an intermediate common space for features, and measuring the similarities among them in one common space, becomes the current mainstream method.

In this paper, we also focus on common space learning method. The existing methods, such as VSE++ [3] and DSVE [4], mainly pay attention to learn an intermediate common space for features of image and text, and measure the similarities among them in common space. Intuitively, the image features obtained by image pipeline and text features obtained by text pipeline should not be completely consistent. For example, as shown in Fig. 1, the features from image space contain information about the image modality itself, the high-level semantic information, low-level information (such as texture, color) or some image-specific information, etc. In the same way, the features from text space contain information about the text modality itself, the high-level semantic information, named entities and some text-specific information. Thus it can be seen that, images and texts, which representing the same semantic information, are usually asymmetrical. To solve asymmetrical relationship between different modal data, much work has been done. Yang et al. [5] propose learning shared semantic space with correlation alignment (S^3CA) to solve this problem. But S^3CA mainly aims at cross-modal retrieval on datasets with classification labels. Moreover, Gu et al. [6] are also aware of this asymmetry rela-

* Corresponding author.
E-mail addresses: zhangli92@hit.edu.cn (L. Zhang), xqwu@hit.edu.cn (X. Wu).

https://doi.org/10.1016/j.patcog.2021.108217
0031-3203/

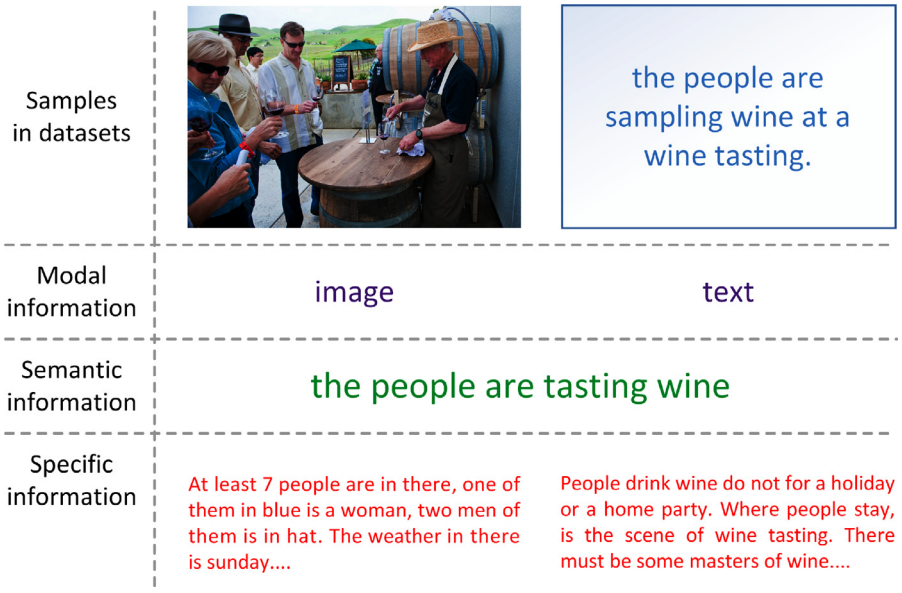

Fig. 1. A sketch map of the asymmetry between different modalities (*e.g.* image and text). Naturally, we decompose the feature vectors from different modal space into three parts: modal information, semantic information and specific information. The sample pair of image and caption come from MS-COCO dataset.

tionship between different modal data, they proposed generative cross-modal feature learning framework (GXN). GXN adopts the high-level features of image modality and text modality for cross-modal retrieval, uses the low-level features of image modality to generate text, and uses the low-level features of text modality to generate images, through multi-tasks joint learning, to improve the performance of cross-modal retrieval branch. However, in GXN, cross-modal retrieval branch and two generative branches (image-to-text and text-to-image) are lack of interaction and weakly correlated, which will cause the features used in cross-modal retrieval not only contain high-level features, but also inevitably mix with low-level features, which affect the performance of cross-modal retrieval.

Inspired by the above work, we propose a novel multi-task framework based on feature separation and reconstruction (mFSR) for cross-modal retrieval. mFSR decomposes the feature vectors from different modal spaces into three parts (as shown in Fig. 1):

(1) Modal information, which characterizes the source of feature vector. The modal information from feature vector of same modality should be as close as possible to each other. On the contrary, they should just try to stay away.

(2) Semantic information, which characterizes the high-level semantics represented by feature vector. The semantic information of feature vectors, which from different modalities with the same semantics or related semantics (i.e. pair of samples in datasets), should be as close as possible to each other. On the contrary, they should just try to stay away. In this paper, we use image semantic information vector and text semantic information vector for cross-modal retrieval.

(3) Specific information, which characterizes the specifics of sample from particular modality represented by feature vector. For example, the feature vector of image modality will have detailed information specific to the image (the specifics of different images are obviously different), and this information does not exist in the feature vector of corresponding text modality.

As for semantic information, mFSR utilities the loss functions commonly used in cross-modal retrieval to train image and text pipeline. With regard to specific information, mFSR constructs a specific loss function to achieve the purpose of feature separation by forcing the similarity between the image (text) semantic information vector and the image (text) specific information vector to be as low as possible. About modal information, mFSR constructs a modal loss function to require that the modal information of all images (text) should be as consistent as possible, and the modal information of any image and that of any text should be as inconsistent as possible. At last, mFSR constructs image and text reconstruction tasks to combine three different information of image and text respectively, and improves the performance of cross-modal retrieval task through multi-task joint learning. The contributions of this paper are as follow:

(1) We introduce feature separation into traditional cross-modal retrieval task to deal with information asymmetry between different modalities, and use different loss functions to supervise different parts of the feature vectors.

(2) We introduce image and text reconstruction tasks to combine three different information of image and text respectively, and improves the performance of cross-modal retrieval task through multi-task joint learning.

(3) We conduct extensive experimentation on MS-COCO and Flickr30K datasets. Our empirical results demonstrate that feature separation and specific information reconstruction can significantly improve the baseline performance of cross-modal image-text retrieval.

2. Related works

2.1. Cross-modal retrieval

In recent years, cross-modal retrieval has made rapid progress in common space learning methods. Ma et al. [7] proposed convolutional neural networks (m-CNNs), which is an end-to-end framework with convolutional architectures to exploit image representation, word composition, and the matching relations between the two modalities. Wang et al. [8] proposed a method for learning joint embeddings of images and text using a two-branch neural network with multiple layers of linear projections followed by non-linearities. Faghri et al. [3] introduced hard negative mining to common loss functions used for cross-modal retrieval and yields significant gains. Engilberge et al. [4] proposed a two-path neural network, named DVSE, with a visual path which leverages space-

aware pooling mechanisms and a textual path which are jointly trained from scratch. Song and Soleymani [9] proposed Polysemous Instance Embedding Networks (PIE-Nets) that compute multiple and diverse representations of an instance by combining global context with locally-guided features, and tied-up two PIE-Nets and optimized them jointly in the multiple instance learning framework. Liu et al. [10] designed a modality classification network with an adversarial loss, which classifies an embedding into either the image or text modality. In addition, they designed multi-stage training procedure to enforce the image and text embedding distributions to be similar by adversarial learning. Niu et al. [11] proposed a heuristic re-ranking method called Adaptive Metric Fusion (AMF) for image-text matching. AMF can obtain a better metric by adaptively fusing metrics based on cross-modal reciprocal encoding and query replacement gap, which can be implemented in an unsupervised way without requiring any human interaction or annotated data, and can be easily applied to any initial ranking result. Wang et al. [12] created a fusion layer to extract intermediate modes and proposed a concise way to update the loss function that makes it easier for neural networks to handle difficult problems. Zhang et al. CMRN designed a decipherable cross-modal multi-relationship aware reasoning network (CMRN) to extract multi-relationship and to learn the correlations between image regions. CMRN introduced spatial relation encoder to perform reasoning on the image graphs and adopted contextual text encoder to learn distinctive textual representations. Wu et al. [13] proposed Dual-View Semantic Inference (DVSI) network to leverage both local and global semantic matching in a holistic deep framework. For the local view, a region enhancement module is proposed to mine the priorities for different regions in the image. For the global view, the overall semantics of image is summarized for global semantic matching to avoid global semantic drift. The two views are unified together for final image-text matching. Owing to good performance of DVSE and VSE++ in cross-modal retrieval tasks with a compact structure, we design our multi-modal feature extraction with reference to them.

2.2. Feature separation

According to Linear Feature Space conjecture [14], deep representations, when well-trained, tend to do a better job at disentangling the underlying factors of variation. The conjecture has been applied in many fields, such as: facial attribute editing, GAN-based image generation, image encryption, etc. Zhou et al. [15] proposed GeneGAN, which can learn object transfiguration from two unpaired sets of images: one set containing images that have that kind of object, and the other set being the opposite, with the mild constraint that the objects be located approximately at the same place. He et al. [16] proposed AttGAN, which applys an attribute classification constraint to the generated image to just guarantee the correct change of desired attributes, i.e. to ãchange what you want. Duan et al. [17] proposed an image encryption method to generate a visually same image as the original one by sending a meaning-normal and independent image to a corresponding well-trained generative model to achieve the effect of disguising the original image. As far as we know, few people applied Linear Feature Space conjecture in cross-modal retrieval. Inspired by the conjecture, in this paper, we introduce feature separation into traditional cross-modal retrieval task to deal with information asymmetry between different modalities.

2.3. Reconstruction of image and text

Image generation has a wide range of applications in many fields, such as GAN-based image generation tasks, which need to restore feature maps or feature vectors to three-channel RGB images. Reed et al. [18] developed text-conditional convolutional GAN to translate visual concepts from characters to pixels. Zhu et al. [19] proposed cycleGAN to translate images from a source domain to a target domain in the absence of paired examples. cycleGAN adapts the architecture for generative networks from Johnson et al. [20] which have shown impressive results for style transfer and super-resolution. As for text generation, which has been widely used in NLP, such as neural machine translation, automatic poetry generation, etc. Text generation is also used in cross-modal feature learning/embedding for cross image-text applications, such as image caption, visual question answer, etc., which need to restore feature vectors to sentences through RNN-like structures. Artetxe et al. [21] propose a method to train a neural machine translation system in a completely unsupervised manner. Zhang et al. [22] propose a memory-augmented neural model for Chinese poem generation. Our work is also part of reconstruction of image and text, but it is different from the existing work. In order to overcome the asymmetry of different modal samples, we use three different information generated by image pipeline to reconstruct image, and use three different information generated by text pipeline to reconstruct text.

3. Approach

Figure 2 show the overall architecture of the proposed mFSR. Common cross-modal retrieval usually only involve multi-modal feature extraction module, and then map the feature vector to a common space. Our method introduces feature separation and reconstruction, and distinguishes the modal information, semantic information and specific information from feature vector, so that the semantic information can be better used for cross-modal retrieval tasks.

The overall implementation process of mFSR is as follow. Arbitrary pair of image and caption are input into multi-modal feature extraction module to obtain visual and literal representations (Section 3.1). Then, two representations will pass feature separation module to obtain the modal information, semantic information and specific information of image and caption respectively (Section 3.2). Finally, three different information of image and caption input GAN and RNN to reconstruct image and caption respectively (Section 3.3). After that, we train mFSR via our loss functions (Section 3.4), and show some implementation details (Section 3.5).

3.1. Multi-modal feature extraction

We follow common space learning methods to extract the feature of image and text. Formally, given arbitrary image-caption pair (x, y), where x is the image and $y = (w_0, \cdots, w_{T-1})$ is the corresponding tokens (one-hot encoding) of caption. We encode a caption by embedding each token w_t into a distributed representation using $\mathbf{W}_e w_t \in \mathbb{R}^K$, where $K = 300$ and $\mathbf{W}_e$ is a word2vec embedding matrix. Then, we use GRU as our base text pipeline [23], to turn such variable length sequence of words into meaningful, fixed-sized literal representation $l \in \mathbb{R}^d$ (just take the last step output of GRU as the literal representation of whole sentence), where $d = 3072$ in our experiment. As for image encoding, in order to accommodate variable size images and to benefit from the performance of very deep architecture, we rely on fully convolutional residual ResNet-152 [24] (pre-trained on ImageNet) as our base image pipeline. As in previous work we extract image features directly from FC7, the penultimate fully connected layer, to obtain visual representation $v \in \mathbb{R}^d$. The dimensionality of the image embedding is 2048 for ResNet152. In short, the formulation of multi-modal feature extraction is as follows.

$$v = FC_1(Res(x; \theta_{Res}); \theta_{FC_1}) \tag{1}$$

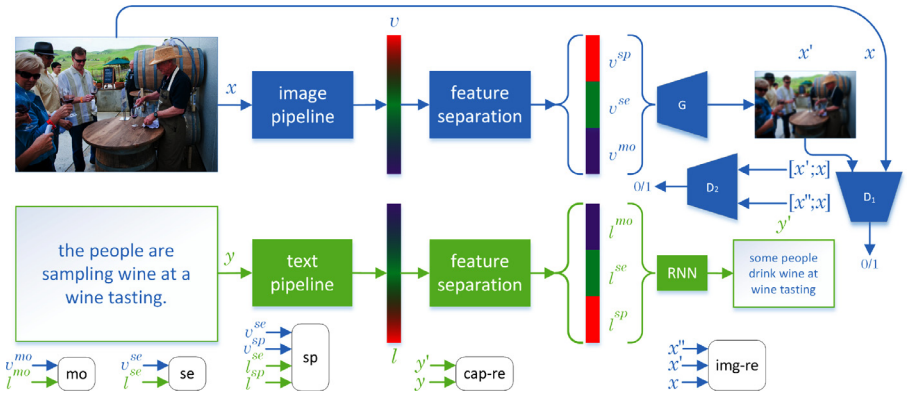

Fig. 2. Arbitrary image x inputs into image pipeline to obtain a visual representation $v \in \mathbb{R}^d$; likewise, corresponding caption y inputs into text pipeline to obtain a literal representation $l \in \mathbb{R}^d$. Then v (resp. l) will pass feature separation module to obtain the modal information, semantic information and specific information of image (resp. caption). Finally, the modal loss (*mo*) is calculated with the modal information v^{mo} and l^{mo}, the semantic loss (*se*) is calculated with the semantic information v^{se} and l^{se}, and the specific loss (*sp*) is calculated with semantic information v^{se}, l^{se} and specific information v^{sp}, l^{sp}. The image reconstruction loss (*img-re*) is calculated with x, x and x. The caption reconstruction loss (*cap-re*) is calculated with y and y.

$$\begin{aligned} l' &= GRU(\mathbf{W}_e y; \theta_{GRU}) \\ l &= l'(-1, :) \end{aligned} \tag{2}$$

Where *Res*, FC_1, and *GRU* refer to pre-trained ResNet-152, fully connected layer and GRU encoder, as well as θ_{Res}, θ_{FC_1} and θ_{GRU} represent corresponding parameters respectively.

3.2. Feature separation

From multi-modal feature extraction module, we get visual representation v and literal representation l. Owing to the asymmetry between different modality, we introduce feature separation module to decompose the feature vectors from different modal spaces into three parts: modal information (v^{mo}, l^{mo}), semantic information (v^{se}, l^{se}) and specific information (v^{sp}, l^{sp}). In order to cut v (resp. l) into three vectors, we send visual representation v into MLP_v, and send literal representation l into MLP_l.Through linear transformation of learned parameters of multilayer perceptron layers, the modal part, semantic part and specific part can be separated from visual representation v (resp. literal representation l). In order to adapt the similarity measure of feature vectors in loss functions in subsequent optimization and inference section, above three parts need to be normalized. After that, we can get modal information (v^{mo}, l^{mo}), semantic information (v^{se}, l^{se}) and specific information (v^{sp}, l^{sp}) of visual representation v and literal representation l respectively. In short, the formulation of feature separation module is as follows.

$$\begin{aligned} [m^{mo}; m^{se}; m^{sp}] &= \frac{MLP_m(m; \theta_{MLP_m})}{\|MLP_m(m; \theta_{MLP_m})\|_2} \\ \forall m &\in \{v, l\} \end{aligned} \tag{3}$$

Where $\|.\|_2$ refers to L2 norm. θ_{MLP_m} refers to the corresponding parameter of fully connected layer.

3.3. Image and text reconstruction

At last, mFSR constructs image and text reconstruction tasks to combine three different information of image and text respectively, and improves the performance of cross-modal retrieval task through multi-task joint learning. In order to combine three different information of image and text, we introduce image and text reconstruction respectively. The better quality of reconstruction, the more significant effect of the feature separation. Through this way, the cross-modal retrieval of semantic information will be promoted to achieve better results.

As for image reconstruction, our goal is to encourage three different information of visual representation to generate an image that is similar to the ground-truth one. Intuitively, generative adversarial network (GAN) [25] can be used to generate images, which consists of a discriminator and generator. The generator and discriminator of DCGAN [26] are treated as our generator G and discriminator D_1 respectively. Since D_1 can only distinguish between real and fake images, it cannot guarantee the consistency of the generated image with ground-truth in content. Inspired by R^2GAN [27], we introduced D_2 to determine whether the generated image is consistent with ground-truth in content. More formally, the process of data flow is as follow.

$$\begin{aligned} x &= G([v^{mo}; v^{se}; v^{sp}]; \theta_G) \\ p &= D_1(n; \theta_{D_1}) \quad \forall n \in \{x, x\} \\ q &= D_2([n; x]; \theta_{D_2}) \quad \forall n \in \{x, x''\}, x'' \sim G \setminus x' \end{aligned} \tag{4}$$

Where p and q refer to the probability of n belongs to real image and the specific image. θ_G, θ_{D_1} and θ_{D_2} refer to the parameters of generator G, discriminator D_1 and D_2 respectively. The discriminator D_1, D_2 and the generator G losses will be mentioned in Section 3.4.

As for text reconstruction, our goal is to encourage three different information of literal representation to generate a sentence that is similar to the ground-truth one. In particular, we decode $[l^{mo}; l^{se}; l^{sp}]$ into a sentence with RNN. More formally, the process of data flow is as follow.

$$y = FC_2(RNN(\mathbf{W}_e c, [l^{mo}; l^{se}; l^{sp}]; \theta_{RNN}); \theta_{FC_2}) \tag{5}$$

Where y refers to the output probability distribution of reconstructed caption. FC_2 and *RNN* refer to fully connected layer and RNN decoder, as well as θ_{FC_2} and θ_{RNN} represent corresponding parameters respectively.

Like traditional RNN-based text generation models, we train our model on cross-entropy loss, which will be mentioned in Section 3.4.

3.4. Optimization

In this section, given image-caption dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with N sample pairs, we will discuss in detail how to optimize our mFSR by minimize a learning objective as follow.

$$\mathcal{L} = \mathcal{L}_{mo} + \mathcal{L}_{se} + \mathcal{L}_{sp} + \mathcal{L}_{img-re} + \mathcal{L}_{cap-re} \tag{6}$$

We will describe each loss term below.

Modal loss As for modal information of images and captions, we design modal loss $\mathcal{L}_{mo}$, which let arbitrary v^{mo} (resp. l^{mo}) be

as close as possible to their mean value $\bar{v}^{mo} = \frac{1}{N}\sum_{i=1}^{N} v_i^{mo}$ (resp. $\bar{l}^{mo} = \frac{1}{N}\sum_{i=1}^{N} l_i^{mo}$), as well as let $\bar{v}^{mo}$ and $\bar{l}^{mo}$ as far away as possible. Therefore, we define $\mathcal{L}_{mo}$ as follow.

$$\mathcal{L}_{mo} = \frac{1}{N}\sum_{i=1}^{N}\left(\|v_i^{mo} - \bar{v}^{mo}\|_2 + \|l_i^{mo} - \bar{l}^{mo}\|_2\right) + e^{-\|\bar{v}^{mo} - \bar{l}^{mo}\|_2} \tag{7}$$

Semantic loss As for semantic information of images and captions, we follow hard negative mining strategy proposed by Faghri et al.[3]. For each positive pair in the mini-batch, a hard negative sample is selected in this batch as the one that has the highest similarity to the query image (resp. caption) while has no association with query. Given arbitrary batch of image-caption $\mathcal{B} = \left\{(x_j, y_j)\right\}_{j\in B}$, our semantic loss $\mathcal{L}_{se}$ is as follow.

$$\mathcal{L}_{se} = \frac{1}{|B|}\sum_{j\in B}\left(\max_{k\in B\setminus j}\left[v_j^{se}, l_j^{se}, l_k^{se}\right]_+ + \max_{k\in B\setminus j}\left[l_j^{se}, v_j^{se}, v_k^{se}\right]_+\right) \tag{8}$$

$$\left[v_j^{se}, l_j^{se}, l_k^{se}\right]_+ = \max\left[(\alpha - \langle v_j^{se}, l_j^{se}\rangle + \langle v_j^{se}, l_k^{se}\rangle), 0\right] \tag{9}$$

Where $\left[l_j^{se}, v_j^{se}, v_k^{se}\right]_+$ is similar. α serves as a margin parameter, and $\langle .,. \rangle$ refers to similarity function of two vectors, which is usually inner product.

Specific loss As for specific information of images and captions, we design specific loss $\mathcal{L}_{sp}$, which let the correlation between arbitray v^{sp} (resp. l^{sp}) and v^{se} (resp. l^{se}) as small as possible. Therefore, we define $\mathcal{L}_{sp}$ as follow.

$$\mathcal{L}_{sp} = \frac{1}{N}\sum_{i=1}^{N}(e^{v_i^{spT} v_i^{se}} + e^{l_i^{spT} l_i^{se}}) \tag{10}$$

Image reconstruction loss As for image reconstruction loss $\mathcal{L}_{img-re}$, we utilize GAN loss. The discriminator loss $\mathcal{L}_{D_1}$, $\mathcal{L}_{D_2}$ and generator loss $\mathcal{L}_G$ are defined as follows.

$$\mathcal{L}_{D_1} = \mathbb{E}_{x\sim p_{image}}[log(1 - D_1(x))] + \mathbb{E}_{x'\sim G}[logD_1(x')] \tag{11}$$

$$\mathcal{L}_{D_2} = \mathbb{E}_{x'\sim G}[log(1 - D_2([x'; x]))] + \mathbb{E}_{x''\sim G\setminus x'}[logD_2([x''; x])] \tag{12}$$

$$\begin{aligned}\mathcal{L}_G \quad &= \tfrac{1}{2}(\mathbb{E}_{x'\sim G}[log(1 - D_1(x'))] \\ &+\mathbb{E}_{x'\sim G}[logD_2([x'; x])] + \mathbb{E}_{x''\sim G\setminus x'}[log(1 - D_2([x''; x]))])\end{aligned} \tag{13}$$

Caption reconstruction loss As for text reconstruction loss $\mathcal{L}_{cap-re}$, we utilize cross-entropy loss defined as:

$$\mathcal{L}_{cap-re} = -\sum_{w_t\in y,\ p(w_t)\in y'} \log p(w_t|w_{0:t-1}) \tag{14}$$

3.5. Implementation details

The dimensionality of the word embedding that are input to the GRU is set to 300. We do not restrict the sentence length for it has little effort on the GPU memory. Our *GRU* (resp. *RNN*) is trained from scratch and has four stacked hidden layers of dimension 3072. We set the dimensionality of the joint embedding space to 3072. Owing to the limitation of hardware, we set batch size at 24 in single RTX 2080Ti (11G), while VSE++ got R@1=64.6 (caption retrieval) in the condition of batch size of 128. If setting the same batch size as the baseline, we need more GPUs. We use ADAM [28] optimizer. mFSR is trained for at most 30 epochs. Except for fine-tuned models, we start training with learning rate 0.0002 for 15 epochs, and then lower the learning rate to 0.00002 for another 15 epochs. The fine-tuned models are trained by taking a model trained for 30 epochs with a fixed image encoder, and then training it for 15 epochs with a learning rate of 0.00002. Notice that since the size of the training set for different models is different, the actual number of iterations in each epoch can vary. For evaluation on the test set, we tackle over-fitting by choosing the snapshot of the model that performs best on the validation set. The best snapshot is selected based on the sum of the recalls on the validation set.

Table 1
Comparisons of the cross-modal retrieval results on MS-COCO dataset with the state-of-the-art methods.

Method	Image-to-text			Text-to-image			Batch Size
	R@1	R@5	R@10	R@1	R@5	R@10	
Results from papers							
Backbone: Pre-trained VGG, ResNet etc.							
DVSA [29]	38.4	69.9	80.5	27.4	60.2	74.8	100
m-CNN [7]	42.8	73.1	84.1	32.6	68.6	82.8	150
DSPE [8]	50.1	79.7	89.2	39.6	75.2	86.9	1500
RRF-Net [30]	56.4	85.3	91.5	43.9	78.1	88.6	1500
CMPM [31]	56.1	86.3	92.9	44.6	78.8	89.0	128
VSE+ [3]	64.6	90.0	95.7	52.0	84.3	92.0	128
GXN [6]	68.5	-	97.9	56.6	-	94.5	128
DSVE [4]	69.8	91.9	96.6	55.9	86.9	94.0	160
SCO [32]	69.9	92.9	97.5	56.7	87.5	94.8	128
Backbone: Pre-trained Faster R-CNN							
SCAN [33]	72.7	94.8	98.4	58.8	88.4	94.8	128
PVSE [9]	69.2	91.6	96.6	55.2	86.5	93.7	128
CMR-SC [34]	73.8	95.3	98.3	59.9	88.9	94.9	128
SGM [35]	73.4	93.8	97.8	57.5	87.3	94.3	200
GSMN [36]	78.4	96.4	98.6	63.3	90.1	95.7	64
Results in batch size = 24							
VSE+	52.1	82.0	91.3	41.2	76.6	87.8	24
mFSR	**55.5**	**83.9**	**92.3**	**42.9**	**77.8**	**88.1**	24

4. Experiments and analysis

4.1. Dataset

We evaluate our mFSR on MS-COCO [37] and Flickr30K [38] datasets. MS-COCO contains 123,287 images (train+val), each of them annotated with 5 captions. It is originally split into a training set of 82,783 images and a validation set of 40,504 images. For cross-modal retrieval, we use the setting of [29], which contains 113,287 training images with five captions each, 5000 images for validation and 5000 images for testing. Flickr30K contains 31,000 images collected from Flickr website with five captions each. Following the split in [3,29], we use 1000 images for validation and 1000 images for testing and the rest for training.

4.2. Results on MS-COCO

Our mFSR is quantitatively evaluated on a cross-modal retrieval task. Given a query image (resp. caption), the aim is to retrieve the corresponding captions (resp. image). Since MS-COCO contains 5 captions per image, recall at r (ãR@rå, r = 1, 5, 10) for caption retrieval is computed based on whether at least one of the correct captions is among the first r retrieved ones [39]. In our experiments, the task is performed 5 times in 1000-image subsets of the test set and the results are averaged.

Table 1 shows the results on MS-COCO. To facilitate comprehensive comparisons, we provide previously reported results and their batch sizes on this dataset. Due to the limitations of hardware, we cant set the batch size to which most methods set in their experiments, such as 100, 128, 160, etc. We choose VSE++ as our baseline, and re-run their code under our hardware conditions (batch size =24). Our mFSR has a significant improvement in image-to-text (∼3.4 R@1, ∼1.9 R@5 and ∼1.0 R@10) and text-to-image (∼1.7 R@1, ∼1.2 R@5 and ∼0.3 R@10). As for the results

Fig. 3. Visual results of caption retrieval given image queries on MS-COCO test dataset. For each image query, we show the top-5 retrieved captions ranked by the similarity scores predicted by our mFSR. We set the true matches in blue and false matches in red. Our mFSR retrieves the correct results in the top ranked sentences even for image queries of complex and cluttered scenes. The model outputs some reasonable mismatches, *e.g.* (c.4). On the other hand, there are incorrect results such as (b.5), which is possibly due to the style of painting is similar to the texture of the woodwork, and a table appears in the painting. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in Table 1, several comparison methods obtain better performance. Owing to the limitation of hardware, we set batch size at 24 in single RTX 2080Ti (11G), while those methods got better performance in the condition of bigger batch size. If setting the same batch size as them, we need more GPUs. It can be seen that, our mFSR may achieve new state-of-the-art on cross-modal retrieval task under bath size=128, 160, etc.

4.3. Results on Flickr30K

Table 2 shows the results on Flickr30K and a comparison with prior work. We provide previously reported results and their batch sizes on this dataset. We choose VSE++ as our baseline, and re-run their code under our hardware conditions (batch size =24). Our mFSR has a significant improvement in image-to-text (~3.3 R@1, ~2.7 R@5 and ~2.0 R@10) and text-to-image (~3.2 R@1, ~2.5 R@5 and ~1.3 R@10). It can be seen that, our mFSR may achieve new state-of-the-art on cross-modal retrieval task under bath size=128, 160, etc.

Table 2
Comparisons of the cross-modal retrieval results on Flickr30K dataset with the state-of-the-art methods.

Method	Image-to-text			Text-to-image			Batch Size
	R@1	R@5	R@10	R@1	R@5	R@10	
Results from papers							
Backbone: Pre-trained VGG, ResNet etc.							
DVSA [29]	22.2	48.2	61.4	15.2	37.7	50.5	100
m-CNN [7]	33.6	64.1	74.9	26.2	56.3	69.6	150
DSPE [8]	40.3	68.9	79.9	29.7	60.1	72.1	1500
RRF-Net [30]	47.6	77.4	87.1	35.4	68.3	79.9	1500
CMPM [31]	48.3	75.6	84.5	35.7	63.6	74.1	128
VSE+ [3]	52.9	80.5	87.2	39.6	70.1	79.5	128
GXN [6]	56.8	-	89.6	41.5	-	80.1	128
SCO [32]	55.5	82.0	89.3	41.1	70.5	80.1	128
Backbone: Pre-trained Faster R-CNN							
SCAN [33]	67.4	90.3	95.8	48.6	77.7	85.2	128
CMR-SC [34]	69.7	91.7	96.4	54.0	79.7	87.2	128
SGM [35]	71.8	91.7	95.5	53.5	79.6	86.5	200
GSMN [36]	76.4	94.3	97.3	57.4	82.3	89.0	64
Results in batch size = 24							
VSE+	40.1	70.2	79.9	30.1	59.2	69.6	24
mFSR	**43.4**	**72.9**	**81.9**	**33.3**	**61.7**	**70.9**	24

4.4. Ablation study

To investigate the importance of different parts in our method, we conduct the ablation study. We use VSE++ as our baseline, and add feature separation module, modal loss, specific loss, image reconstruction and caption reconstruction respectively to verify the validity of our proposed mFSR. (shown in Table 3).

Effect of feature separation. The first row VSE++ works as the baseline. In all remaining rows, we add feature separation mod-

Fig. 4. Visual results of image retrieval given caption queries on MS-COCO test dataset. For each caption query, we show the top-3 retrieved images ranked by the similarity scores predicted by our mFSR. We set the true matches in blue and false matches in red. In the examples we show, our mFSR outputs some reasonable mismatches, *e.g.* (b.1) and (b.3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Ablation studies on the MS-COCO 1K test set. Results are reported in terms of Recall@K(R@K). *mo* means modal loss, *sp* means specific loss, $img-re$ means image reconstruction loss and $cap-re$ means text reconstruction loss.

Method	Image-to-text			Text-to-image			Batch Size
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE+	52.1	82.0	91.3	41.2	76.6	87.8	24
VSE++ & *mo*	52.5	81.5	91.5	41.6	76.9	87.4	24
VSE++ & *sp*	52.6	82.3	91.4	41.2	76.2	87.9	24
VSE++ & *mo* & *sp*	53.3	83.2	91.8	41.5	77.0	88.0	24
VSE++ & *mo* & *sp* & $img-re$	54.0	83.4	91.7	42.5	77.3	88.3	24
VSE++ & *mo* & *sp* & $cap-re$	55.2	83.7	92.5	42.1	77.1	87.9	24
mFSR	55.5	83.9	92.3	42.9	77.8	88.1	24

ule based on VSE++. In the second and third row, we introduce modal loss *mo* and specific loss *sp* respectively. Compared with VSE++, both VSE++ & *mo* and VSE++ & *sp* improve the performance of baseline, as shown in Table 3. VSE++ & *mo* & *sp* can separate the modal information and specific information from entire feature vectors, and partly promote cross-modal retrieval task. Compared with VSE++, VSE++ & *mo* & *sp* increases the image-to-text R@1 from 52.1 to 53.3, R@5 from 82.0 to 83.2, R@10 from 91.3 to 91.8, and text-to-image R@1 from 41.2 to 41.5, R@5 from 76.6 to 77.0, R@10 from 87.8 to 88.0

Effect of image reconstruction. In fifth row, we add image reconstruction based on VSE++ & *mo* & *sp*. Compared with VSE++, VSE++ & *mo* & *sp* & $img-re$ can reconstruct image by fusing three different information, better promote visual feature separa-

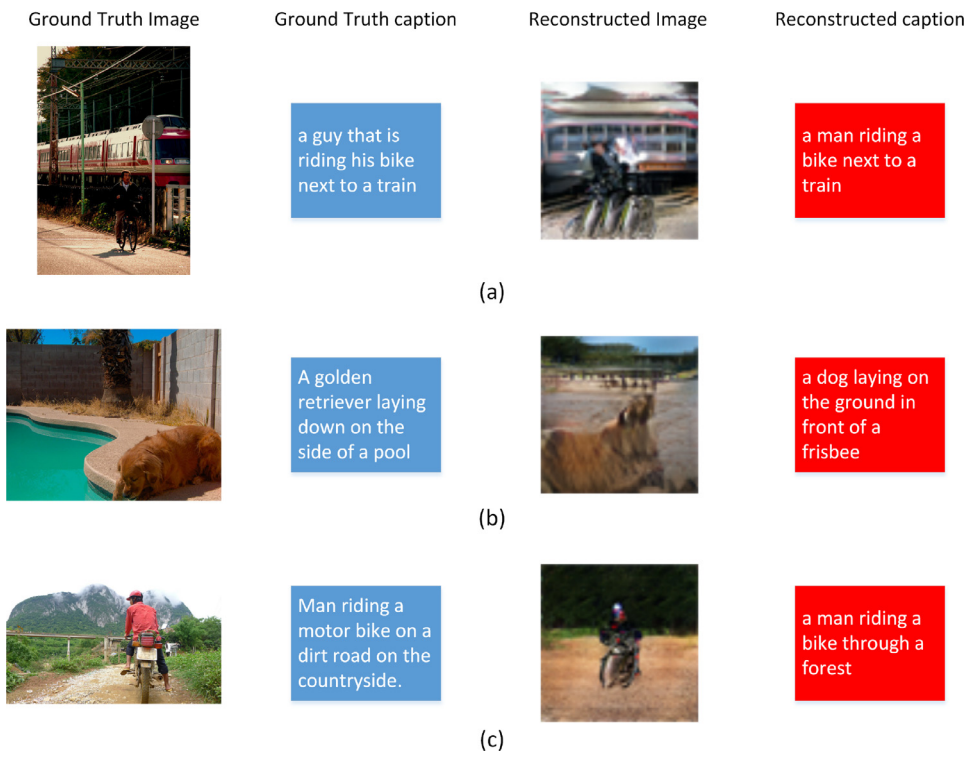

Fig. 5. Visual results of image and text reconstruction on MS-COCO dataset. The first and second columns represent the image text pairs input into the mFSR, respectively. The third and fourth columns represent the reconstruction results of the mFSR, respectively.

tion process, and partly encourage cross-modal retrieval task, especially text-to-image. Compared with VSE++, VSE++ & *mo* & *sp* & $img-re$ increases the image-to-text R@1 from 52.1 to 54.0, R@5 from 82.0 to 83.4, R@10 from 91.3 to 91.7, and text-to-image R@1 from 41.2 to 42.5, R@5 from 76.6 to 77.3, R@10 from 87.8 to 88.3.

Effect of text reconstruction. In sixth row, we add caption reconstruction based on VSE++ & *mo* & *sp*. Compared with VSE++, VSE++ & *mo* & *sp* & $cap-re$ can reconstruct caption by fusing three different information, better promote literal feature separation process, and partly encourage cross-modal retrieval task, especially image-to-text. Compared with VSE++, VSE++ & *mo* & *sp* & $cap-re$ increases the image-to-text R@1 from 52.1 to 55.2, R@5 from 82.0 to 83.7, R@10 from 91.3 to 92.5, and text-to-image R@1 from 41.2 to 42.1, R@5 from 76.6 to 77.1, R@10 from 87.8 to 87.9.

From the above analyses and Table 3, through adding different parts on baseline, we get better results than VSE++.

4.5. Visual results of cross-modal retrieval

Figure 3 shows the visual results of caption retrieval given image queries and Fig. 4 shows the visual results of image retrieval given caption queries on MS-COCO datasets. For each image query, we show the top-5 retrieved captions ranked by the similarity scores predicted by our mFSR. For each caption query, we show the top-3 retrieved images ranked by the similarity scores. We set the true matches in blue and false matches in red.

4.6. Visual results of image and text reconstruction

Figure 5 shows the results of image and text reconstruction. The first and second columns represent the image text pairs input into the mFSR, respectively. The third and fourth columns represent the reconstruction results of the mFSR, respectively.

5. Conclusion and further work

We propose multi-task framework based on feature separation and reconstruction (mFSR) for cross-modal retrieval, and get a significant improvement compared with performance of the baseline method in multiple datasets. Compared with existing methods, we introduces feature separation into cross-modal retrieval task to deal with information asymmetry between different modalities, and introduce image and text reconstruction tasks combined with three different information of image and text respectively, to refine these information, especially semantic information of image and text. At last, the refined semantic information of image and text can improve the performance of cross-modal retrieval task.

The limitation of mFSR is that, it cannot develop good performance over some methods, which backbone are Faster R-CNN. Some work also pointed out that the feature vector generated by Faster R-CNN, can hard be used for image reconstruction [40], which limits the application of our proposed mFSR in the method of extracting image features based on Faster R-CNN. So we do not construct our mFSR based on those methods as a baseline. In the future, we will explore how to reconstruct image using the output of object detection network and apply our mFSR to the latest cross modal retrieval methods, which backbone are faster R-CNN, to get better results.

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.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018YFC0832304 and 2020AAA0106502, by the Natural Science Foundation of China under Grant 62073105, by the Distinguished Youth Science Foundation of Heilongjiang Province of China under Grant JC2018021, by the State Key Laboratory of Robotics and System (HIT) under Grant SKLRS-2019-KF-14 and SKLRS-202003D, and by the Heilongjiang Touyan Innovation Team Program.

References

[1] P. Saragiotis, Cross-modal classification and retrieval of multimodal data using combinations of neural networks, University of Surrey, Guildford, UK, 2006 Ph.D. thesis.
[2] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges, IEEE Trans. Circuits Syst. Video Techn. 28 (9) (2018) 2372–2385, doi:10.1109/TCSVT.2017.2705068.
[3] F. Faghri, D.J. Fleet, J.R. Kiros, S. Fidler, VSE++: improving visual-semantic embeddings with hard negatives, in: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3–6, 2018, BMVA Press, 2018, p. 12.
[4] M. Engilberge, L. Chevallier, P. Pérez, M. Cord, Finding beans in burgers: deep semantic-visual embedding with localization, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, 2018, pp. 3984–3993, doi:10.1109/CVPR.2018.00419.
[5] Z. Yang, Z. Lin, P. Kang, J. Lv, Q. Li, W. Liu, Learning shared semantic space with correlation alignment for cross-modal event retrieval, ACM Trans. Multim. Comput. Commun. Appl. 16 (1) (2020) 9:1–9:22, doi:10.1145/3374754.
[6] J. Gu, J. Cai, S.R. Joty, L. Niu, G. Wang, Look, imagine and match: improving textual-visual cross-modal retrieval with generative models, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, 2018, pp. 7181–7189, doi:10.1109/CVPR.2018.00750.
[7] L. Ma, Z. Lu, L. Shang, H. Li, Multimodal convolutional neural networks for matching image and sentence, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, 2015, pp. 2623–2631, doi:10.1109/ICCV.2015.301.
[8] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 5005–5013, doi:10.1109/CVPR.2016.541.
[9] Y. Song, M. Soleymani, Polysemous visual-semantic embedding for cross–modal retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, 2019, pp. 1979–1988.
[10] R. Liu, Y. Zhao, S. Wei, L. Zheng, Y. Yang, Modality-invariant image-text embedding for image-sentence matching, ACM Trans. Multim. Comput. Commun. Appl. 15 (1) (2019) 27:1–27:19, doi:10.1145/3300939.
[11] K. Niu, Y. Huang, L. Wang, Re-ranking image-text matching by adaptive metric fusion, Pattern Recognit. 104 (2020) 107351, doi:10.1016/j.patcog.2020.107351.
[12] D. Wang, L. Wang, S. Song, G. Huang, Y. Guo, S. Cheng, N. Ao, A. Du, Fusion layer attention for image-text matching, Neurocomputing 442 (2021) 249–259, doi:10.1016/j.neucom.2021.01.124.
[13] C. Wu, J. Wu, H. Cao, Y. Wei, L. Wang, Dual-view semantic inference network for image-text matching, Neurocomputing 426 (2021) 47–57, doi:10.1016/j.neucom.2020.09.079.
[14] Y. Bengio, G. Mesnil, Y.N. Dauphin, S. Rifai, Better mixing via deep representations, in: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16–21 June 2013, 2013, pp. 552–560.
[15] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, W. He, GeneGAN: learning object transfiguration and attribute subspace from unpaired data, CoRR (2017) abs/1705.04932.
[16] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, AttGAN: facial attribute editing by only changing what you want, IEEE Trans. Image Process. 28 (11) (2019) 5464–5478, doi:10.1109/TIP.2019.2916751.
[17] X. Duan, H. Song, E. Zhang, J. Liu, Image camouflage based on generate model, CoRR (2017) abs/1710.07782.
[18] S.E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016, 2016, pp. 1060–1069.
[19] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, 2017, pp. 2242–2251, doi:10.1109/ICCV.2017.244.
[20] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II, 2016, pp. 694–711, doi:10.1007/978-3-319-46475-6_43.
[21] M. Artetxe, G. Labaka, E. Agirre, K. Cho, Unsupervised neural machine translation, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30, - May 3, 2018, Conference Track Proceedings, 2018.
[22] J. Zhang, Y. Feng, D. Wang, Y. Wang, A. Abel, S. Zhang, A. Zhang, Flexible and creative chinese poetry generation using neural memory, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30, - August 4, Volume 1: Long Papers, 2017, pp. 1364–1373, doi:10.18653/v1/P17-1125.
[23] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1724–1734, doi:10.3115/v1/d14-1179.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.
[25] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, 2014, pp. 2672–2680.
[26] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings, 2016.
[27] B. Zhu, C. Ngo, J. Chen, Y. Hao, R2GAN: cross-modal recipe retrieval with generative adversarial network, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 11477–11486, doi:10.1109/CVPR.2019.01174.
[28] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015.
[29] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 664–676, doi:10.1109/TPAMI.2016.2598339.
[30] Y. Liu, Y. Guo, E.M. Bakker, M.S. Lew, Learning a recurrent residual fusion network for multimodal matching, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, 2017, pp. 4127–4136, doi:10.1109/ICCV.2017.442.
[31] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part I, 2018, pp. 707–723, doi:10.1007/978-3-030-01246-5_42.
[32] Y. Huang, Q. Wu, C. Song, L. Wang, Learning semantic concepts and order for image and sentence matching, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, 2018, pp. 6163–6171, doi:10.1109/CVPR.2018.00645.
[33] K. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IV, 2018, pp. 212–228, doi:10.1007/978-3-030-01225-0_13.
[34] H. Chen, G. Ding, Z. Lin, S. Zhao, J. Han, Cross-modal image-text retrieval with semantic consistency, in: L. Amsaleg, B. Huet, M.A. Larson, G. Gravier, H. Hung, C. Ngo, W.T. Ooi (Eds.), Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21–25, 2019, ACM, 2019, pp. 1749–1757, doi:10.1145/3343031.3351055.
[35] S. Wang, R. Wang, Z. Yao, S. Shan, X. Chen, Cross-modal scene graph matching for relationship-aware image-text retrieval, in: IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1–5, 2020, IEEE, 2020, pp. 1497–1506, doi:10.1109/WACV45572.2020.9093614.
[36] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020, IEEE, 2020, pp. 10918–10927, doi:10.1109/CVPR42600.2020.01093.
[37] T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V, 2014, pp. 740–755, doi:10.1007/978-3-319-10602-1_48.
[38] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78.
[39] I. Vendrov, R. Kiros, S. Fidler, R. Urtasun, Order-embeddings of images and language, in: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings, 2016.
[40] J. Cho, J. Lu, D. Schwenk, H. Hajishirzi, A. Kembhavi, X-LXMERT: paint, caption and answer questions with multi-modal transformers, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020, Association for Computational Linguistics, 2020, pp. 8785–8805.

Li Zhang works as a Ph.D. candidate in Harbin Institute of Technology, Harbin, China. His current research interests include deep learning, computer vision and multi-modal fusion.

Xiangqian Wu received the B.Sc., M.Sc. and Ph.D. degree in computer science from Harbin Institute of Technology (HIT), Harbin, China, in 1997, 1999 and 2004, respectively. Dr. Wu works as a lecturer (2004–2006), associate professor (2006–2009) and professor (2009-present) in the School of Computer Science and Technology at HIT. He has published one book and more than 100 papers in several international journals and conferences. Dr. Wu is a principal investigator of dozens of research projects, including the projects of Natural Science Foundation of China (NSFC) and national 863 plan project. He held dozens of patents of China, US and HK. He won the CCF outstanding Doctoral Dissertation Award (2006), the academic achievement award of Hei Longjiang Province (2006), the Nomination of the National Excellent PhD Dissertation Award (2007), the first prize of natural science of Hei Longjiang Province (2011), Program for New Century Excellent Talents in University (2008) and the Elsevier Most Cited Chinese Researcher (2014, 2015 and 2016). His current research interests include image processing, biomedical image analysis and biometrics, etc.