Rich Features Embedding for Cross-Modal Retrieval: A Simple Baseline

Xin Fu, Yao Zhao, *Senior Member, IEEE*, Yunchao Wei, Yufeng Zhao, and Shikui Wei, *Senior Member, IEEE*

***Abstract*—During the past few years, significant progress has been made on cross-modal retrieval, benefiting from the development of deep neural networks. Meanwhile, the overall frameworks are becoming more and more complex, making the training as well as the analysis more difficult. In this paper, we provide a Rich Features Embedding (RFE) approach to tackle the cross-modal retrieval tasks in a simple yet effective way. RFE proposes to construct rich representations for both images and texts, which is further leveraged to learn the rich features embedding in the common space according to a simple hard triplet loss. Without any bells and whistles in constructing complex components, the proposed RFE is concise and easy to implement. More importantly, our RFE obtains the state-of-the-art results on several popular benchmarks such as MS COCO and Flickr 30 K. In particular, the image-to-text and text-to-image retrieval achieve 76.1% and 61.1% (R@1) on MS COCO, which outperform others more than 3.4% and 2.3%, respectively. We hope our RFE will serve as a solid baseline and help ease future research in cross-modal retrieval.**

***Index Terms*—Rich features embedding, image-text matching, deep representation learning, cross-modal retrieval.**

I. Introduction

With the rapid development of information technology, multimedia data such as image, text, video and audio has been widely available on the Internet and contribute the dominant forms of the data. Usually, the data with different modalities are leveraged collectively to describe the same object or event. Therefore, it is of great significance to mine the semantic consistency of multimedia data. In this work, we focus mainly on image-sentence cross-modal retrieval task, which is to find the best matching sentence (image) from the database for a given image (sentence). Owing to semantic gap of multi-modal data, the heterogeneous characteristic has been widely considered as a main challenge for cross-modal retrieval. The popular way for bridging the heterogeneous gap is to learn features embedding. Recently, some feature-enhancement based approaches [1]–[7] propose to learn features embedding for capturing the common characteristics of isomerous data. These methods usually adhibit sophisticated network to strengthen text or image representation when learning features embedding. For instance, Yue *et al.* [8] construct independent semantic spaces by a modality-specific cross-modal similarity measurement approach for different modalities. Zheng *et al.* [3] creatively build a convolutional network amenable for fine-tuning the visual and textual representation. Huang *et al.* [4] propose a novel semantic-enhanced image and sentence matching model, which can improve the image representation by learning semantic concepts and then organizing them in a correct semantic order. Gu *et al.* [5] propose to incorporate generative processes into the cross-modal feature embedding for the first time. Peng *et al.* [6] propose an effective cross-modal correlation learning approach with multi-grained fusion by hierarchical network. Although these methods significantly bridge the heterogeneous gap among different modalities and achieve impressive performance, they make the network more and more complex. For these complex networks, they are not only hard to train but also difficult to identify which modules are really work well. Therefore, our objective is to exploit simple yet effective way to learn features embedding.

This work is motivated by [9], which makes the first attempt to use convolutional neural network (CNN) features for conducting cross-modal retrieval. Specifically, Wei *et al.* [9] have revealed that the performance of cross-modal retrieval will be significantly improved as long as the distinguished representation is learned. They achieve superior results compared with traditional visual features despite barely exploiting off-the-shelf CNN visual features without complex network model. Inspired by [9], we make an assumption: accurate retrieval results may be achieved by simply constructing more powerful feature representations for different modality data. Accordingly, we propose a RFE approach to learn discriminative representation for image and text in common space, as shown in Fig. 1. As for RFE, we target to learn rich features embedding by merging high-level

Manuscript received December 14, 2018; revised April 23, 2019, July 10, 2019, and September 25, 2019; accepted November 22, 2019. Date of publication December 12, 2019; date of current version August 21, 2020. This work was supported in part by the National Key Research and Development of China under Grant 2016YFB0800404, in part by the National Science Foundation of China under Grants U1936212, 61532005, and 61972022, in part by Program of China Scholarships Council under Grant 201807095006, and in part by the Fundamental Research Funds for the Central Universities under Grants 2018JBZ001 and 2018YJS028. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ramazan S. Aygun. *(Corresponding author: Yao Zhao.)*

X. Fu, Y. Zhao, and S. Wei are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: xinfu@bjtu.edu.cn; yzhao@bjtu.edu.cn; shkwei@bjtu.edu.cn).

Y. Wei is with the Beckman Institute, University of Illinois Urbana-Champaign, Champaign, IL 61801 USA (e-mail: yunchao@illinois.edu).

Y. Zhao is with the Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China (e-mail: snowmanzhao@163.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2957948

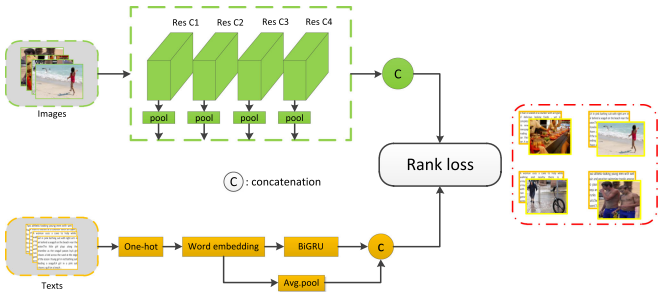

Fig. 1. The motivation of RFE. We aim to learn rich features embedding by merging high-level representation with low-level information to obtain complementary feature representation with employing pooling operation on feature maps. The green and yellow indicate the learned features embedding for images and texts, respectively.

representation with low-level information to obtain complementary feature representation.

It is a common knowledge that different layers carry different information in deep neural network (DNN). For example, we can obtain representation from local feature to semantic feature with the increase of the number of the layers. Only exploiting high-level semantic information will lose details of the low-level local information, which is an important cue for cross-modal retrieval. Therefore, we consider exploiting both high-level semantic information and low-level local details rather than only making use of global information of the last layer. Firstly, local and global features are integrated from sentences and images to obtain isomorphic semantic representation. Then, these pairwise representations will be embedded to a shared space in order to compute similarity of heterogeneous data. Finally, we consider a metric to minimize the gap between semantically similar items from different modalities while maximizing the distance between semantically different items of the same modality.

Our RFE includes three key modules: 1) construct rich features for images; 2) construct rich features for texts; 3) one hard triplet loss for optimization. Concretely, we handle max-pooling on the feature maps of Resnet-C1, Resnet-C2, Resnet-C3, and Resnet-C4 to get the local feature. In order to obtain global semantic feature, we use average-pooling on Resnet-C4 that is the final layer of Resnet. We first concatenate the local feature and global feature as the feature representation of the image. For text encoder pipeline, we then represent the word by word-embedding. Finally, we learn the temporal context information by averaging the forward final hidden state and backward final hidden state, which is called global feature. Since the feature after word-embedding is the low-level word representation, we handle average pooling on word-embedding feature as the local feature of the text. Same as image representation, we combine local and global features to obtain text representation. To bridge the semantic gap between different modal data, we map the two representations into two intermediate spaces that have a natural correspondence. We employ triplet loss to make the paired data as close as possible while the unpaired data as far as possible. Different from other methods [10]–[14], we focus solely on the hardest negative in a mini-batch.

To sum up, our main contributions are three-fold:

- We propose a simple yet effective approach that high-level semantic information and low-level local details are integrated for discriminative image representation.
- We propose a new approach to encoder text by exploiting not only word-level local representation but also sentence-level context global feature.
- Our work achieves 76.1%/61.1% R@1 accuracy for image-to-text retrieval and 72.2%/53.3% R@1 accuracy text-to-image retrieval in the popular MS COCO and Flickr30 K retrieval datasets, which are the new state-of-the-arts.

II. Related Work

A. CCA-Based Methods

As a popular baseline for common space learning, canonical correlation analysis (CCA) [15] is usually employed to find a pair of mapping matrices to maximize the correlation between two kinds of feature representations. Sharma *et al.* [16] propose a generic framework, called generalized multi-view analysis, to map feature representation in different modality spaces into an isomorphic nonlinear space. Gong *et al.* [17] present a three-view CCA method by introducing a semantic view to produce a better separation for multi-modal data belonging to different categories in the learned common space. Moreover, Andrew *et al.* introduce a Deep CCA (DCCA) [18] to learn complex nonlinear transformations for two associated views. Zhang *et al.* [10] develope a general framework to project cross-view data into a unique high-level low-dimensional semantically shared subspace to mine the semantically consistent patterns for cross-view data. Eisenschtat *et al.* [19] employ the Euclidean loss and a tied 2-way architecture to link paired samples from two sources.

B. Ranking-Based Methods

In recent years, training with a ranking loss is one of the most effective methods for cross-modal retrieval. In general, these methods are supervised but do not enforce the assumption that the trained multi-modal data must be paired as needed for CCA-based models (e.g., one image is in pair-correspondence with one text description). Specifically, Yang *et al.* [20] propose a semi-supervised algorithm called ranking with local regression and global alignment to learn a robust Laplacian matrix for multi-modal data ranking. Inspired by the use of hard negatives in structured prediction and ranking loss functions used in retrieval, Faghri *et al.* [21] present a new technique for learning visual-semantic embedding for cross-modal retrieval tasks. Nam *et al.* [2] propose dual attention networks which jointly leverage visual and textual attention mechanisms to capture fine-grained interplay between vision and language. Huang *et al.* [22] propose a multi-modal context-modulated attention scheme to select salient pairwise instances from image and sentence, and a multi-modal long short-term memory (LSTM) network for local similarity measurement and aggregation. Wehrmann *et al.* [23] introduce an efficient character-level inception module which is designed for learning textual semantic embeddings by convolving raw characters in distinct granularity levels. Lee *et al.* [24] present stacked cross attention to discover the full latent alignments using both image regions and words in sentence as

context and infer the image-text similarity, which exploits additional data visual genomes to train bottom-up attention model.

C. Hashing Based Methods

With the explosive growth of high-dimensional cross-modal data, the problem of nearest neighbor search becomes more expensive than ever before. To address this problem, hashing-based approaches for large scale similarity search have attracted considerable interest in the cross-modal retrieval community. Ding *et al.* [25] propose a similarity-preserving based hashing method named collective matrix factorization hashing for cross-view similarity search on multimodal data. Lin *et al.* [26] propose a supervised semantics-preserving hashing method for cross-view retrieval. Jiang *et al.* [27] presented a deep cross-modal hashing method, which is an end-to-end deep learning framework that can perform feature learning and hash-code learning simultaneously. Li *et al.* [28] propose a novel ranking-based hashing framework that maps data from different modalities into a common Hamming space where the cross-modal similarity can be measured using Hamming distance. Liu *et al.* [29] propose a new cross-media hashing scheme to treat the propose categories as the third view and preserve the correlation between heterogeneous instances and categories as well. Xu *et al.* [30] propose a novel discrete cross-modal hashing method to learns discriminative binary codes by retaining the discrete constraints. Li *et al.* [31] propose a self-supervised adversarial hashing approach, which lies among the early attempts to incorporate adversarial learning into cross-modal hashing in a self-supervised fashion. Hu *et al.* [32] propose to process heterogeneous data by making use of using modalitys-pecific models. Zhang *et al.* [33] raise a novel approach called HashGAN for the cross-modal hashing based on the idea of adversarial architecture.

D. Classification-Based Methods

Learning the similarity between images and texts could be also modeled as classification. Ba *et al.* [34] train a two branch network using classification loss to match visual and text data for zero-shot learning. Wei *et al.* [35] propose a modality-dependent cross-media retrieval (MDCR) model, where two couples of projections are learned for different cross-media retrieval tasks instead of one couple of projections. Rohrbach *et al.* [36] use a softmax function to estimate the posterior probability of a phrase over all the available region proposals in an image. Fukui *et al.* [37] systematically investigate multiple feature fusion strategies and find element-wise product to be among the most effective. Wei *et al.* [9] propose a deep semantic matching method to address the cross-modal retrieval problem with respect to samples which are annotated with one or multiple labels. Liu *et al.* [38] propose a deep framework introducing a latent embedding layer to learn joint parameters. Zheng *et al.* [3] propose a dual-path CNN which learns discriminative feature embedding from training image/text pairs. Jabri *et al.* [39] use a softmax function to predict whether the input image and question match with the answer choice for visual question answering (VQA). Wang *et al.* [12] propose adversarial cross-modal retrieval method to learn representation which is both discriminative and modality invariant for cross-modal retrieval. Wang *et al.* [40] propose a joint global and co-attentive representation learning method for image-sentence retrieval. Yang *et al.* [41] a hierarchical multi-clue fusion approach to predict the popularity of point of interest (POI) for cross-modal data. Peng *et al.* [42] construct independent semantic spaces by a modality-specific crossmodal similarity measurement approach for different modalities.

III. Rich Features Embedding

We show the overall architecture of the rich features embedding approach in Fig. 2. It consists of three components, i.e. Rich Image Representation (RIR), Rich Text Representation (RTR) and hard triplet loss for cross-modal retrieval. Deep image net and deep text net are utilized to extract isomorphic semantic representations for images and texts, respectively. Triplet loss is bidirectional max-margin ranking loss adopted for image-text similarity learning. The overall framework is trained by minimizing the following composite loss functions from the two branches using stochastic gradient descent:

$$L_{biTri} = L_{img} + L_{sent} \tag{1}$$

The RFE will be elaborated on details in the following components.

A. Rich Image Representation

To enhance image representation, we propose the RIR approach to fuse high-level global information and low-level local information. The popular image representation methods [3]–[5], [21] only use global average pooling to extract the last layer convolution feature that is semantic feature ignoring the local information of image. As shown in Fig. 2, we try to develop not only semantic feature from the last convolution layer but also local information of each layer. Therefore, it is very important and urgent to know how to develop partial information when given a convolution. In this paper, we employ the simple method that is max-pooling operation universally acknowledged as digging up local information of image. For details, we handle max-pooling on Resnet-C1, Resnet-C2, Resnet-C3 and Resnet-C4 to get the local feature vector. Then, we concatenate all local feature vectors and semantic feature vector as image feature. The image feature I_{fea} of DIRP approach can be formulated as:

$$I_{fea} = Norm_{L2}[Pool_{\max}(C1), Pool_{\max}(C2), Pool_{\max}(C3), Pool_{\max}(C4), Pool_{avg}(C4)] \tag{2}$$

where $Pool_{\max}$ and $Pool_{avg}$ denote max pooling and average pooling, respectively. $C1$, $C2$, $C3$ and $C4$ denote Resnet-C1, Resnet-C2, Resnet-C3 and Resnet-C4, and $Norm_{L2}$ denotes L2 normalization. $[:,:]$ denotes concatenation.

B. Rich Text Representation

To exploit discriminative text representation, we propose RTR approach to make full use of both global temporal context feature and local word-embedding feature, which is shown in Fig. 2. Different from previous methods [4], [5], [21], [43], we explore to mine not only low-level but also high level clues for text representation by exploiting both word-embeddings and bidirectional gated recurrent unit (biGRU) information. Concretely,

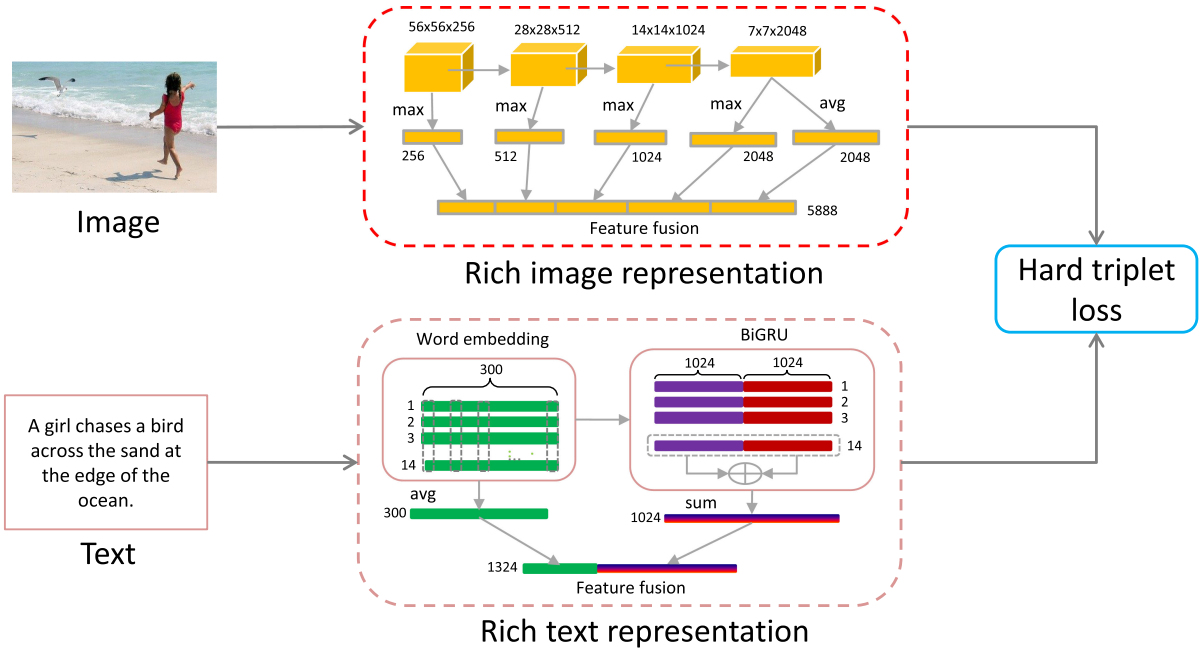

Fig. 2. Overview of the proposed RFE approach for cross-modal retrieval. For deep image net, we concatenate the local feature by max-pooling from Resnet-C1 to Resnet-C4 and global feature by average-pooling the Resnet-C4 as the feature representation of the image. With regard to deep text net, we first represent the word by word-embedding. Then, we learn the global temporal context information by BiGRU. Similar to image representation, we hand average pooling on word embedding feature as the local feature. Finally, we fuse the local feature and global feature to obtain text representation. Based on image and text representation, we exploit hard triplet loss to make the paired data as close as possible and the unpaired data as far as possible.

given sentence s_t where t is the index word in sentence, we use a BiGRU to encode the context for each word. Firstly, we embed each word s_t into a feature vector Emb_t using word embedding. Then, a layer BiGRU is exploited to encode the sentence after embedding all words. Next, we sum the final hidden representation of each direction in BiGRU for global text representation and handle average pooling on word embedding matrix Emb to obtain local text representation. Finally, L2-normalization is used after the concatenation of global feature and local feature to obtain final text representation T_{fea}:

$$Emb_t = Embedding(s_t) \tag{3}$$

$$\overrightarrow{h}_t = \overrightarrow{GRU}(Emb_t, \overrightarrow{h}_{t-1}) \tag{4}$$

$$\overleftarrow{h}_t = \overleftarrow{GRU}(Emb_t, \overleftarrow{h}_{t-1}) \tag{5}$$

$$T_{fea} = [Norm_{L2}(\overrightarrow{h}_t + \overleftarrow{h}_t), Norm_{L2}(Pool_{avg}(Emb))] \tag{6}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ denote forward and backward context feature of tth word in sentence, respectively. $\overrightarrow{GRU}$ denotes forward GRU while $\overleftarrow{GRU}$ denotes backward GRU.

C. Hard Triplet Loss

After encoding image and text, we employ triplet loss function to measure the similarity between sentences and images, which is motivated by VSE++ [21]. The overview of the triplet loss is demonstrated in Fig. 3. Different from vanilla triplet loss, we only choose the hardest negative sample for each anchor within each mini-batch at each iteration, which is potentially more robust to label errors [21]. We take the image and text as anchor to get the following two sets of triplet loss L_{img} and L_{sent}:

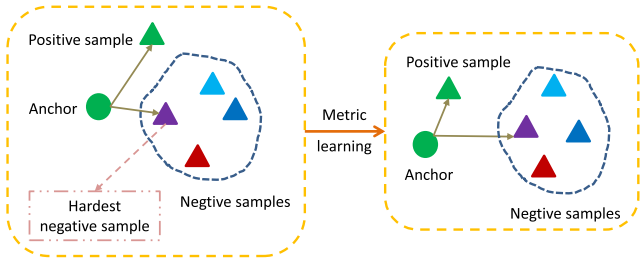

Fig. 3. Overview of the triplet loss. The hardest negative sample is the sample that is the closest to anchor. By minimizing the triplet loss, the distance between anchor and positive sample in the learned embedding space becomes smaller than distance between anchor and negative samples.

$$L_{img} = \sum_{I^x_{fea}, T^z_{fea}} \max\left(mg - dist\left(f^{(I)}_{W_I}(I^x_{fea}), f^{(T)}_{W_T}(T^y_{fea})\right) + dist\left(f^{(I)}_{W_I}(I^x_{fea}), f^{(T)}_{W_T}(T^z_{fea})\right)\right) \tag{7}$$

$$L_{sent} = \sum_{T^{x'}_{fea}, I^{z'}_{fea}} max\left(mg - dist\left(f^{(T)}_{W_T}(T^{x'}_{fea}), f^{(I)}_{W_I}(I^{y'}_{fea})\right) + dist\left(f^{(T)}_{W_T}(T^{x'}_{fea}), f^{(I)}_{W_I}(I^{z'}_{fea})\right)\right) \tag{8}$$

where $dist(X, Y)$ denotes the cosine distance between X and Y in the embedding space. $f^{(T)}_{W_T}$ denotes the embedding text feature by employing a fully-connected on T_{fea} in the common space,

TABLE I
ABLATION STUDY OF MARGIN, BIDIRECTIONAL RETRIEVAL RESULTS ON FLICKR30 K 1000-IMAGE TEST SET

$Flickr30K$	$image-to-sentence$			$sentence-to-image$			SUM
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
mg=0.1	50.7	79.6	86.3	38.4	67.6	77.9	400.5
mg=0.2	**55.7**	**82.6**	**89.2**	**42.5**	**72.1**	**80.7**	**422.8**
mg=0.3	53.9	80.6	87.4	40.9	71.2	80.3	414.3
mg=0.6	50.8	77.6	85.0	38.0	70.0	79.0	400.9

and $f_{W_I}^{(I)}$ denotes the embedding image feature by employing a fully-connected on I_{fea}. The margin of the triplet is mg.

IV. EXPERIMENT RESULTS

A. Datasets and Evaluation Metrics

To evaluate the effectiveness of proposed approach, we conduct extensive experiments on two popular publicly available datasets, Flickr30K [44] and MS COCO [45].

MS COCO: We use the MS COCO caption dataset which is used by many papers [3]–[5], [21] for cross-modal retrieval. We adopt the same splits as reported in [21], which contains 113,287 training images with each five captions, 5,000 images for validation and 5,000 images for testing.

Flickr30 K: This dataset consists of 31,783 images, and each image is accompanied by five descriptive sentences. Following the same protocols as the recent works [3], [4], [21], we randomly split it into a training set with 29,783 images, and use 1000 images for validation and 1000 images for testing.

Evaluation Metrics: In our experiments, we consider both T2I and I2T tasks. We report the performance at Recall@K (K = 1, 5, 10), which is the percentage of queries that at least one correct result is ranked among the top K of the ranked list. We also evaluate another indicator SUM:

$$SUM = \underbrace{R@1 + R@5 + R@10}_{image-to-sentence} + \underbrace{R@1 + R@5 + R@10}_{sentence-to-image} \tag{9}$$

$$R@K = \frac{queries\ returned\ true\ results\ at\ top\ K}{all\ queries} \tag{10}$$

B. Implementation Details

RIR: We adopt the Resnet152 network [46] pre-trained on ImageNet [47] as the backbone of deep image representation. We first resize the image to 256 × 256, and then use random crop of size 224 × 224 for training and center crop of size 224 × 224 for testing. We use kernel 7 × 7 pooling for Resnet C4, kernel 14 × 14 pooling for Resnet C3, kernel 28 × 28 for Resnet C2 and kernel 56 × 56 for Resnet C1 to encode low-level and high-level feature map, respectively. Finally, we concatenate the local feature vectors by max-pooling and global feature vector by average-pooling.

RTR: We use word embeddings initiated by uniform distribution between -0.1 and 0.1 and one-layer biGRU for deep text representation. The dimensionality of the word embeddings is 300, and the hidden size of biGRU is 1024. We only exploit context representation by summing the forward final hidden state and backward final hidden state of BiGRU and low-level representation vector by average-pooling on word-embedding feature matrix. We concatenate context representation and low-level representation to obtain the final text feature.

We take a mini-batch size of 128 pairs and train the model for 36 epochs. The initial learning rate is set to 0.0002 and decreased by a factor of 10 after 15 epochs for Flickr30 K and a factor of 1.2 after 15 epochs for MS COCO. The margin of the triplet loss is set to 0.2, which is motivated by [21]. From Table I, we can obtain the best performance by setting the parameter mg as 0.2. To learn the image-text embeddings in common space, we add the fully connected layer after the text and image representation, respectively. The dimension of the space is set to 1024. Our experiments use the VSE++ [21] code and reranking [48] code, which is implemented based on the publicly available Pytorch [49] deep learning framework. All of our experiments are run on NVIDIA TITAN X PASCAL GPUs.

C. Comparisons With State-of-the-Art Methods

We make extensive comparisons with state-of-the-art cross-modal retrieval approaches on MS COCO in Table II and Table III, and Flickr30 K datasets in Table IV.

Cross-modal Retrieval on MS COCO: Experiment results for 1000-image test set on MS COCO dataset are shown in Table II. From the results, we can observe that our proposed method outperforms all the other state-of-the-art works, including Skip-thought [50], DVSA [51], Fisher Vector [52], m-RNN [53], MNLM [54], m-CNN [55], OEM [56], VQA [1], DSPE [57], sm-LSTM [22], 2WayNet [19], RRF [58], 2Branch [43], VSE++ [21], DPC [3], CHAIN-VSE [23], GXN [5], SCO [4] and CMPM [59]. In particular, SCO achieves the second best performance among the baselines. Even SCO is trained with extra dataset that keeps the nouns, adjectives, verbs and numbers as semantic concepts and eliminate all the semantic-irrelevant words from the sentences. GXN exploits four-path network to learn feature in common space, which achieves the third best performance. As a result, it needs lots of GPU memory when training and testing, and it is hard to train by end-to-end manner. In contrast, the proposed RFE method is very simple and convenient to learn cross-modal feature embeddings. RFE is also effective and efficient for training the cross-modal retrieval network. Compared with those state-of-the-art methods, the proposed RFE improves upon the best performance by over 11.3% for SUM. In addition, we can see that our RFE outperforms prior methods by a relatively large margin at all recall metrics, such as Recall@1, Recall@5 and Recall@10 for image-to-text and text-to-image retrieval.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS, BIDIRECTIONAL RETRIEVAL RESULTS ON MS COCO 1000-IMAGE TEST SET

MS COCO	*image − to − sentence*			*sentence − to − image*			*SUM*
	R@1	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*	
Skip-thought [NIPS2015] [50]	33.8	67.7	82.1	25.9	60.0	74.6	344.1
DVSA [CVPR2015] [51]	38.4	69.9	80.5	27.4	60.2	74.8	351.2
Fisher Vector [CVPR2015] [52]	39.4	67.9	80.9	25.1	59.8	76.6	349.7
m-RNN [ICLR2015] [53]	40.8	71.9	83.2	29.6	64.8	80.5	371.6
MNLM [ICML2014] [54]	43.4	75.7	85.8	31.0	66.7	79.9	382.5
m-CNN [ICCV2015] [55]	42.8	73.1	84.1	32.6	68.6	82.8	384.0
OEM [ICLR2016] [56]	46.7	78.6	88.9	37.9	73.7	85.9	411.7
VQA [ECCV2016] [1]	50.5	80.1	89.7	37.0	70.9	82.9	411.1
DSPE [CVPR2016] [57]	50.1	79.7	89.2	39.6	75.2	86.9	420.7
sm-LSTM [CVPR2017] [22]	53.2	83.1	91.5	40.7	75.8	87.4	431.7
2WayNet [CVPR2017] [19]	55.8	75.2	-	39.7	63.3	-	-
RRF [ICCV2017] [58]	56.4	85.3	91.5	43.9	78.1	88.6	443.8
2Branch [TPAMI2018] [43]	54.9	84.0	92.2	43.3	76.4	87.5	438.3
VSE++ [BMVC2018] [21]	64.6	-	95.7	52.0	-	92.0	-
DPC [3]	65.6	89.8	95.5	47.1	79.9	90.0	467.9
CHAIN-VSE [CVPR2018] [23]	61.2	89.3	95.8	46.6	81.9	90.9	465.7
GXN [CVPR2018] [5]	68.5	-	97.9	56.6	-	94.5	-
SCO [CVPR2018] [4]	69.9	92.9	97.5	56.7	87.5	94.8	499.3
CMPM [ECCV2018] [59]	56.1	86.3	92.9	44.6	78.8	89.0	447.7
RFE-ensemble [ours]	**74.4**	**95.0**	**98.2**	**59.2**	**88.8**	**95.0**	**510.6**
SCAN t-i LSE [ECCV2018] [24]	67.5	92.9	97.6	53.0	85.4	92.9	489.3
SCAN t-i AVG [ECCV2018] [24]	70.9	94.5	97.8	56.4	87.0	93.9	500.5
SCAN i-t LSE [ECCV2018] [24]	68.4	93.9	98.0	54.8	86.1	93.3	494.5
SCAN i-t AVG [ECCV2018] [24]	69.2	93.2	97.5	54.4	86.0	93.6	493.9
SCAN t-i LSE + i-t AVG [ECCV2018] [24]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
SCAN-RFE t-i LSE [ours]	71.9	94.2	97.9	56.7	86.9	93.8	501.4
SCAN-RFE t-i AVG [ours]	74.8	95.3	98.2	59.6	88.0	94.3	510.2
SCAN-RFE i-t LSE [ours]	70.9	94.4	97.9	56.3	86.8	93.6	499.9
SCAN-RFE i-t AVG [ours]	69.4	93.5	97.5	53.0	84.9	92.4	490.7
SCAN-RFE t-i AVG + i-t LSE [ours]	75.6	95.3	**98.6**	60.9	**89.2**	**95.1**	514.7
SCAN-RFE t-i AVG + t-i LSE [ours]	**76.1**	**95.4**	98.4	**61.1**	88.9	**95.1**	**515.3**

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS, BIDIRECTIONAL RETRIEVAL RESULTS ON MS COCO 5000-IMAGE TEST SET

MS COCO	*image − to − sentence*			*sentence − to − image*			*SUM*
	R@1	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*	
VQA [ECCV2016] [1]	23.5	50.7	63.6	16.7	40.5	53.8	248.8
DSPE [CVPR2016] [57]	24.0	50.8	65.1	17.4	42.7	56.7	256.7
OEM [ICLR2016] [56]	23.3	-	65.0	18.0	-	57.6	-
VSE++ [BMVC2018] [21]	41.3	-	81.2	30.3	-	72.4	-
DPC [3]	41.2	70.5	81.1	25.3	53.4	66.4	337.9
GXN [CVPR2018] [5]	42.0	-	84.7	31.7	-	74.6	-
SCO [CVPR2018] [4]	42.8	72.3	83.0	33.1	62.9	75.5	369.5
CMPM [ECCV2018] [59]	31.1	60.7	73.9	22.9	50.2	63.8	302.6
RFE-ensemble [ours]	**47.8**	**77.2**	**86.4**	**34.9**	**65.0**	**76.8**	**388.1**
SCAN i-t LSE [ECCV2018] [24]	46.4	77.4	87.2	34.4	63.7	75.7	384.8
SCAN t-i AVG + i-t LSE [ECCV2018] [24]	50.4	82.2	90.0	38.6	**69.3**	**80.4**	410.9
SCAN-RFE t-i AVG [ours]	52.4	81.1	90.0	38.1	67.4	78.1	407.1
SCAN-RFE t-i LSE [ours]	48.1	78.0	88.1	34.3	64.1	76.0	388.6
SCAN-RFE i-t LSE [ours]	47.2	77.7	87.9	34.0	63.8	75.5	386.1
SCAN-RFE t-i AVG + i-t LSE [ours]	54.6	82.5	90.6	**39.9**	69.1	79.9	416.6
SCAN-RFE t-i AVG + t-i LSE [ours]	**55.1**	**82.7**	**91.0**	39.7	68.9	79.7	**417.1**

In order to verify the robustness of our method on MS COCO dataset, we also test the 5000-image test set and present the comparison results in Tabel III. Compared to the 1000 test data, the 5000 test data is more challenging for improving performance. Therefore, only a few works conduct experiments on the 5000 test data. From the Table III, we can notice that SCO still achieves the second best performance among all state-of-the-art approaches. RFE-ensemble obtains 47.8% in Recall@1, 77.2% in Recall@5 and 86.4% in Recall@10 for image-to-sentence retrieval task, which outperforms the SCO more than 5.0%, 4.9% and 3.4%, respectively. Besides, RFE-ensemble gains a new baseline for sentence-to-image retrieval task with 34.9% in Recall@1, 65.0% in Recall@5 and 76.8% in Recall@10. It is worth mentioning that RFE-ensemble lead to 388.1% for SUM, which acquires significant improvements over the other approaches more than 18.6%.

In order to further verify the effectiveness of our method, we have done experiments on features with rich semantic information [24]. we exploit both global text representation and local text representation with same visual feature from SCAN. Concretely, we concatenate word embedding vector and word feature from BiGRU while the rest of SCAN remains unchanged. From Table II, we can see that the Recall@1 score is improved from 72.7% to 76.1% for image-to-sentence retrieval and 58.8% to 61.1% for text-to-image retrieval. In addition, SCAN-RFE achieves 55.1% and 39.7% at Recall@1 for image-to-sentence

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS, BIDIRECTIONAL RETRIEVAL RESULTS ON FLICKR30 K 1000-IMAGE TEST SET

$Flickr30K$	$image-to-sentence$			$sentence-to-image$			SUM
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
DVSA [CVPR2015] [51]	22.2	48.2	61.4	15.2	37.7	50.5	235.2
Fisher Vector [CVPR2015] [52]	35.0	62.0	73.8	25.0	52.7	66.0	314.5
m-RNN [ICLR2015] [53]	32.7	62.7	72.6	26.2	55.1	69.2	318.5
MNLM [ICML2014] [54]	23.0	50.7	62.9	16.8	42.0	56.5	251.9
m-CNN [ICCV2015] [55]	33.6	64.1	74.9	26.2	56.3	69.6	324.7
VQA [ECCV2016] [1]	33.9	62.5	74.5	24.9	52.6	64.8	313.2
RTP [ICCV2015] [44]	37.4	63.1	74.3	26.0	56.0	69.3	326.1
DSPE [CVPR2016] [57]	40.3	68.9	79.9	29.7	60.1	72.1	351.0
sm-LSTM [CVPR2017] [22]	42.5	71.9	81.5	30.2	60.4	72.3	358.8
2WayNet [CVPR2017] [19]	49.8	67.5	-	36.0	55.6	-	-
RRF [ICCV2017] [58]	47.6	77.4	87.1	35.4	68.3	79.9	395.7
DAN [CVPR2017] [2]	55.0	81.8	89.0	39.4	69.2	79.1	413.5
2Branch [TPAMI2018] [43]	43.2	71.6	79.8	31.7	61.3	72.4	360.0
VSE++ [BMVC2018] [21]	52.9	-	87.2	39.6	-	79.5	-
DPC [3]	55.6	81.9	**89.5**	39.1	69.2	**80.9**	416.2
SCO [CVPR2018] [4]	55.5	82.0	89.3	41.1	70.5	80.1	418.5
CMPM [ECCV2018] [59]	49.6	76.8	86.1	37.3	65.7	75.5	391.0
RFE [ours]	**56.1**	**82.7**	89.4	**42.5**	**72.3**	**80.9**	**423.9**
SCAN t-i LSE [ECCV2018] [24]	61.6	85.4	91.5	43.3	71.9	80.9	434.6
SCAN t-i AVG [ECCV2018] [24]	61.8	87.5	93.7	45.8	74.4	83.0	446.2
SCAN i-t LSE [ECCV2018] [24]	67.7	88.9	94.0	44.0	74.2	82.6	451.4
SCAN i-t AVG [ECCV2018] [24]	67.9	89.0	94.4	43.9	74.2	82.8	452.2
SCAN t-i AVG + i-t LSE [ECCV2018] [24]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
SCAN-RFE t-i LSE [ours]	66.4	90.6	95.6	49.8	77.9	85.9	466.2
SCAN-RFE t-i AVG [ours]	66.7	91.9	96.0	50.9	77.7	85.1	468.3
SCAN-RFE i-t LSE [ours]	68.7	91.5	95.9	48.2	77.1	84.9	466.3
SCAN-RFE i-t AVG [ours]	67.3	89.5	94.4	44.3	73.8	82.8	452.1
SCAN-RFE t-i AVG + i-t LSE [ours]	**72.2**	**93.8**	**97.2**	**53.3**	**80.3**	**87.3**	**484.1**

and sentence-to-image retrieval on MS COCO 5000-image test set in Table III, which outperform SCAN more than 4.7% and 1.1%, respectively.

Cross-modal Retrieval on Flickr30 K: We report results on another small dataset Flickr30 K for cross-modal retrieval to validate the performance of our approach in Table IV. From the Table IV, SCO also achieves the second best performance and DPC achieves the third best performance that exploits dual-path Resnet-50 network to learn image-text Embedding. However, the Resnet-50 network costs heavier GPU memory than the GRU in our approach for deep text representation. We can observe that our method achieves similar performance with the method of DPC in Recall@10 for both image-to-sentence and sentence-to-image retrieval tasks. Nevertheless, we yield results 56.1% in recall@1, 82.7% in recall@5, 42.5% in recall@1 and 72.3% in recall@5 for image-to-sentence and sentence-to-image retrieval tasks, which outperforms large margin when compared with the methods of both DPC and SCO. In general, recall@1 is better than recall@10 in evaluating the effectiveness of a method in real-world scenarios. Furthermore, the proposed simple and effective RFE improves upon the best performance SCO by over 5.4% for summing all Recall metric.

We also report results on Flickr30 K based on rich semantic feature of SCAN for cross-modal retrieval to validate the effectiveness of our approach in Table IV. It can be observed that our SCAN-RFE achieves the state-of-the-art results in the respect of all the evaluation metrics. In particular, SCAN-RFE achieves 72.2% and 53.3% in R@1 for image-to-text and text-to-image tasks, which outperforms SCAN more than 4.8% and 4.7%, respectively.

Fig. 4. Qualitative text-to-image retrieval examples. For each text query, we show the top-3 ranked images. The ground-truth matching images aren't surrounded by red box.

Both experimental results on MS COCO and Flickr30 K datasets well demonstrate the superiority of our approach. Fig. 4 and Fig. 5 show some successful retrieval results on MS COCO datasets by VSE++-RFE, indicating that our methods can learn discriminative feature representation in common space for cross-modal bi-directional retrieval tasks. It can be observed that our approach can obtain the reasonable results for image-text matching.

At Table V, we report the time of training and 1 K testing on val set at Tab 1 based on NVIDIA TITAN X PASCAL GPUs, respectively. Both the training phase and the testing phase are

TABLE V
COMPARISION WITH THE STATE-OF-THE-ART ON TRAINING AND TESTING EFFICIENCY

$TIME$	$Train/AVG(s)$	$Train/SD$	$Test/AVG(s)$	$Test/SD$
VSE++	0.847	0.055	63.69	2.031
VSE++-RFE	0.867	**0.063**	66.9	1.665
SCAN	0.95	0.017	292.23	5.186
SCAN-RFE	**1.028**	0.049	**402.39**	**5.805**

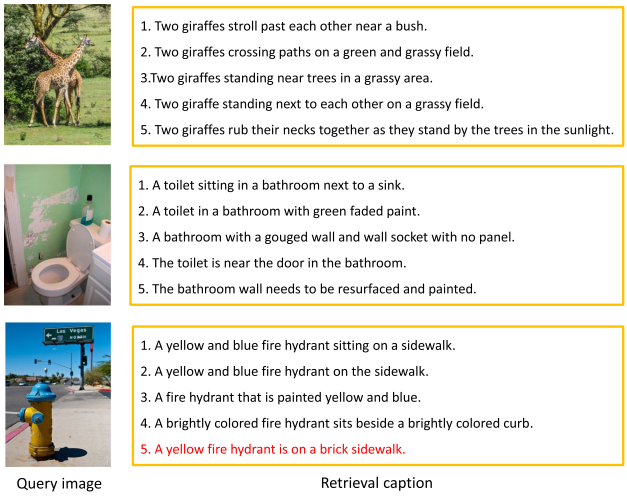

Fig. 5. Qualitative image-to-text retrieval examples. For each image query, we show the top-5 ranked texts. The ground-truth matching texts are in black.

executed for 10 runs, of which each run in the training phase consists of 10 iterations. AVG and SD indicate the average time and the standard deviation of 10 runs, respectively. The testing run time and every training iteration time of SCAN and SCAN-RFE are the average time of four approaches that are *t-i LSE*, *t-i AVG*, *i-t LSE* and *i-t AVG*. As can be seen from the table, our method costs more time than the base model. In addition, the deeper the network structure of base model, the less complexity our approach will increase.

D. Ablation Analysis

We conduct extensive ablation analysis of the proposed RFE to validate the effectiveness of each key module and generate the following baselines including:

VSE++: encoding image by average-pooling Resnet-C4 and text by exploiting context feature of GRU.

avg-4-GRU-emb-avg: encoding image by average-pooling Resnet-C4 and text by concatenating context feature of GRU and average-pooling word-embedding feature.

avg-4-BiGRU: encoding image by average-pooling Resnet-C4 and text by summing forward and backward context feature.

avg-4-BiGRU-emb-max: encoding image by average-pooling Resnet-C4 and text by concatenating context representation from summing forward and backward context and max-pooling word-embedding feature.

avg-4-BiGRU-emb-avg: encoding image by average-pooling Resnet-C4 and text by concatenating context representation from summing forward and backward context and average-pooling word-embedding feature.

avg-4-max-4-BiGRU-emb-avg: encoding image by concatenating max-pooling Resnet-C4 and average-pooling Resnet-C4 and encoding text by concatenating context representation from summing forward and backward context and average-pooling word-embedding feature.

avg-4-max-123-BiGRU-emb-avg: encoding image by concatenating max-pooling Resnet-C1, max-pooling Resnet-C2, max-pooling Resnet-C3 and average-pooling Resnet-C4, and encoding text by concatenating context representation from summing forward and backward context and average-pooling word-embedding feature.

avg-4-max-34-BiGRU-emb-avg: encoding image by concatenating max-pooling Resnet-C3, max-pooling Resnet-C4 and average-pooling Resnet-C4, and encoding text by concatenating context representation from summing forward and backward context and average-pooling word-embedding feature.

avg-1234-BiGRU-emb-avg: encoding image by concatenating max-pooling Resnet-C1, max-pooling Resnet-C2, max-pooling Resnet-C3 and max-pooling Resnet-C4, and encoding text by concatenating context representation from summing forward and backward context and average-pooling word-embedding feature.

avg-1234-max-1234-BiGRU-emb-avg: encoding image by concatenating max-pooling Resnet-C1, max-pooling Resnet-C2, max-pooling Resnet-C3, max-pooling Resnet-C4, average-pooling Resnet-C1, average-pooling Resnet-C2, average-pooling Resnet-C3 and average-pooling Resnet-C4, and encoding text by concatenating context representation from summing forward and backward context and average-pooling word-embedding feature.

avg-4-max-1234-BiGRU-emb-avg: encoding image by concatenating max-pooling Resnet-C1, max-pooling Resnet-C2, max-pooling Resnet-C3, max-pooling Resnet-C4 and average-pooling Resnet-C4, and encoding text by concatenating context representation from summing forward and backward context and average-pooling word-embedding feature.

model-rerank: reranking the retrieval results from the *model* by exploiting test data when testing.

RFE-ensemble: ensemble with the four models from *avg-4-max-4-BiGRU-emb-avg*, *avg-4-max-123-BiGRU-emb-avg*, *avg-4-max-34-BiGRU-emb-avg* and *avg-4-max-1234-BiGRU-emb-avg*.

SCAN-RFE: concatenating word embedding vector and word feature from BiGRU while the rest of SCAN remains unchanged.

SCAN-RFE A + B: ensemble with the two models from SCAN-RFE A and SCAN-RFE B.

1) RTR: With the RTR approach, we aim to learn discriminative text feature for bi-directional image-sentence matching tasks. The main motivation of RTR is to learn both low-level word-embedding feature and high-level context feature of

TABLE VI
ABLATION STUDY OF RICH FEATURE EMBEDDING, BIDIRECTIONAL RETRIEVAL RESULTS ON MS COCO 1000-IMAGE TEST SET

$MS\ COCO$	$image-to-sentence$			$sentence-to-image$			SUM
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
VSE++ [BMVC2018]	64.6	-	95.7	52.0	-	92.0	-
avg-4-GRU-emb-avg	69.9	93.0	96.9	56.5	87.0	93.7	497.0
avg-4-BiGRU	67.4	93.7	97.7	55.6	86.8	93.8	495.0
avg-4-BiGRU-emb-max	68.9	93.7	97.6	56.3	87.0	93.5	497.0
avg-4-BiGRU-emb-avg	70.0	93.3	97.7	55.7	87.7	93.7	498.1
avg-1234-BiGRU-emb-avg	70.8	93.4	97.5	56.7	87.7	93.8	499.9
avg-1234-max-1234-BiGRU-emb-avg	71.3	94.1	97.7	57.0	87.1	93.9	501.1
avg-4-max-4-BiGRU-emb-avg	70.9	94.2	97.6	57.1	87.3	94.1	501.2
RFE-ensemble	**74.4**	**95.0**	**98.2**	**59.2**	**88.8**	**95.0**	**510.6**

TABLE VII
ABLATION STUDY OF RICH FEATURE EMBEDDING, BIDIRECTIONAL RETRIEVAL RESULTS ON MS COCO 5000-IMAGE TEST SET

$MS\ COCO$	$image-to-sentence$			$sentence-to-image$			SUM
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
VSE++ [BMVC2018] [21]	41.3	-	81.2	30.3	-	72.4	-
avg-4-BiGRU-emb-avg-5K	43.6	73.5	83.8	32.3	62.5	74.5	370.2
avg-4-max-4-BiGRU-emb-avg-5K	44.1	74.6	84.5	32.6	62.8	75.0	373.6
avg-4-max-123-BiGRU-emb-avg-5K	42.5	73.2	83.8	32.2	62.3	74.5	368.5
avg-4-max-34-BiGRU-emb-avg-5K	44.2	74.3	84.6	32.9	62.7	74.6	373.3
avg-4-max-1234-BiGRU-emb-avg-5K	44.8	75.0	84.8	32.9	62.9	74.9	375.2
RFE-ensemble	**47.8**	**77.2**	**86.4**	**34.9**	**65.0**	**76.8**	**388.1**

TABLE VIII
ABLATION STUDY OF RICH FEATURE EMBEDDING, BIDIRECTIONAL RETRIEVAL RESULTS ON FLICKR30 K 1000-IMAGE TEST SET

$Flickr30K$	$image-to-sentence$			$sentence-to-image$			SUM
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
VSE++ [BMVC2018] [21]	52.9	-	87.2	39.6	-	79.5	-
avg-4-GRU-emb-avg	**55.7**	82.6	**89.2**	**42.5**	72.1	**80.7**	422.8

biGRU when representing the sentence. In order to convert word-embedding output matrix to a vector, we employ the popular operation average pooling and max pooling, respectively. In addition, we exploit the biGRU to learn backward and forward context feature for enhancing text representation, which is different from the vanilla sequence-learning model with GRU. After obtaining the word-embedding vector representation and context representation from biGRU, we concatenate them for final text representation.

For quantitatively understanding the contribution of final text representation by RTR, we demenstrate the comparison results in Table VI for 1000-image test set on MS COCO dataset and in Table VIII for 1000-image test set on Flickr30 K dataset, respectively.

We can observe that the performance can be improved from 64.6% to 69.9% at Recall@1 for image-to-sentence retrieval task and from 52.0% to 56.5% at Recall@1 for sentence-to-image retrieval task by employ average-pooling on word-embedding feature in Table VI, which is a large margin for cross-modal retrieval. Besides, the performance can be also promoted at Recall@10 for bi-directional retrieval. From Table VIII, we can observe the image-to-sentence retrieval and sentence-to-image retrieval yield results 55.7%/42.5% at Recall@1 and 89.2%/80.7% at Recall@10 by exploiting low-level word-embedding feature. Comparison with the baselines results, the proposed low-level representation approach significantly improve the performance. The above results denote that the utilize of local information from word-embedding feature can really learn a better text representation when training the image-sentence matching system.

A major issue with GRU is that it learns representation from previous time steps. Sometimes, we might have to learn representation from future time steps to better understand the context and eliminate ambiguity. So, we employ the bidirectional GRU to extract the context feature of the sentence. We can see that Recall@1 score is improved from 64.6% to 67.4% for image-to-sentence retrieval and 52.0% to 55.6% for sentence-to-image retrieval from Table VI. Moreover, Recall@10 score is promoted when exploiting both forward and backward information for context representation, too. But, the approach of biGRU context representation doesn't work well. We think that flickr30 K dataset has fewer data and biGRU has more learning parameters than unidirectional GRU, so it can not learn better representation on small dataset.

Finally, we integrate context feature of biGRU and low-level word-embedding feature to obtain text representation. The performance shown by *avg-4-BiGRU-emb-avg* get improved compared with both *avg-4-GRU-emb-avg* and *avg-4-BiGRU*, which indicate developing the low-level and high-level text representation is applicable for cross-modal retrieval. Furthermore, we adopt max-pooling on word-embedding feature to gain low-level feature for text representation. It can improve the performance of baseline, but the performance is not as good as average-pooling operation. Therefore, biGRU and average-pooling operation is eventually used for sentence representation in our framework.

TABLE IX
ABLATION STUDY OF RERANKING, BIDIRECTIONAL RETRIEVAL RESULTS ON MS COCO 1000-IMAGE TEST SET

MS COCO	*image − to − sentence*			*sentence − to − image*			*SUM*
	R@1	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*	
avg-4-max-1234-BiGRU-emb-avg	71.7	**93.6**	**97.2**	57.0	**87.9**	**94.4**	501.8
avg-4-max-1234-BiGRU-emb-avg-rerank	**72.8**	93.5	**97.2**	**57.5**	87.8	**94.4**	**503.2**

TABLE X
ABLATION STUDY OF RERANKING, BIDIRECTIONAL RETRIEVAL RESULTS ON FLICKR30 K 1000-IMAGE TEST SET

Flickr30K	*image − to − sentence*			*sentence − to − image*			*SUM*
	R@1	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*	
avg-4-GRU-emb-avg	55.7	82.6	89.2	42.5	72.1	80.7	422.8
avg-4-GRU-emb-avg-rerank	**56.1**	**82.7**	89.4	**42.5**	**72.3**	**80.9**	**423.9**

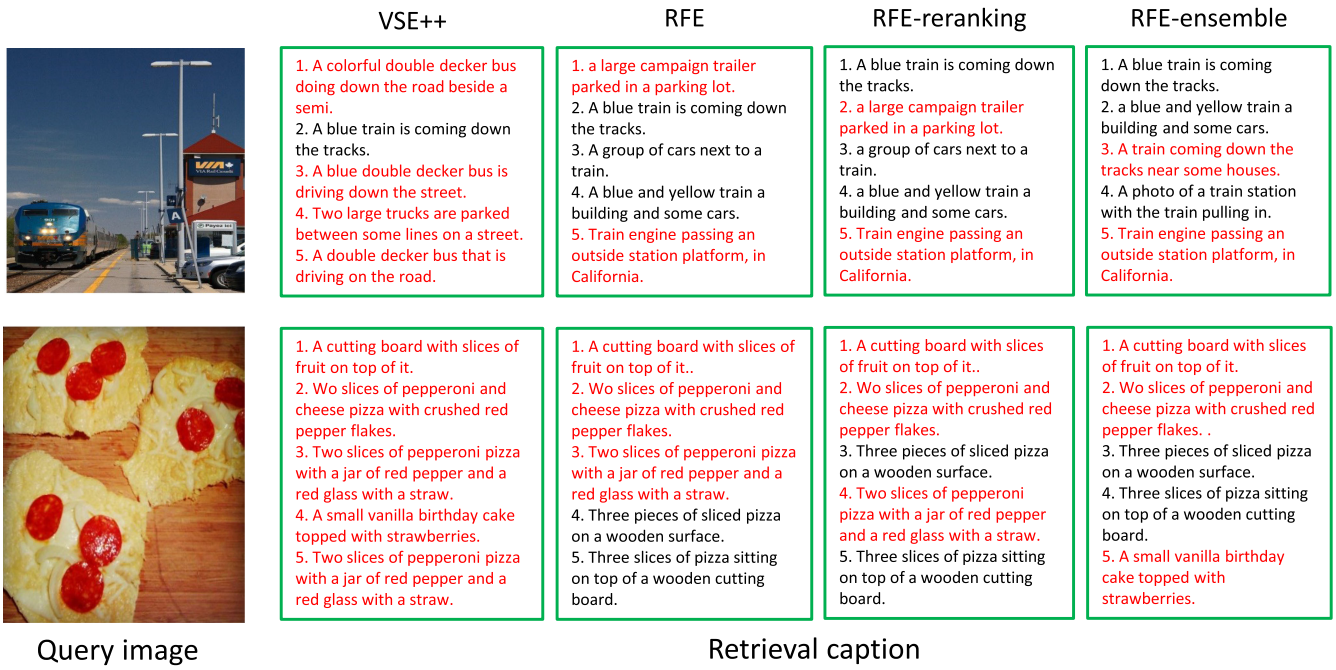

Fig. 6. Examples of image-text retrieval for ablation analysis. *RFE* denotes the model *avg-4-max-1234-BiGRU-emb-avg*. *RFE-reranking* is the model *avg-4-max-1234-BiGRU-emb-avg-rerank*. The ground-truth matching texts are in black. We observe that the proposed approach the improve the performance of the baselines model. In particular, it still works well even if the baselines model cannot search the true texts at the top-5 results.

2) RIR: We now proceed to evaluate the method of encoding image by RIR for learning cross-modal feature embedding. In this paper, we aim to learn the image feature that conveys both local and global information of the image, which can be obtained by max-pooling and average-pooling on Resnet blob, respectively. To compare with the baseline model, we handle max-pooling on Resnet-C1, Resnet-C2, Resnet-C3, and Resnet-C4 to extract local image feature. We present the ablation results on 1000-image test set in Table VI and on 5000-image test set in Table VII for MS COCO dataset. We select four kinds of ways to encode image local feature to validate the proposed RIR, for instance, *avg-4-max-4-BiGRU-emb-avg*, *avg-4-E-max-123-BiGRU-emb-avg*, *avg-4-E-max-34-BiGRU-emb-avg* and *avg-4-max-1234-BiGRU-emb-avg*. We can see that the performance of both average pooling and max pooling *avg-4-BiGRU-emb-avg* on the last layer of Resnet is improved from 70%/55.7% to 70.9%/57.1% at *recall 1* for cross-modal retrieval when comparison with only handling average pooling operation on the last layer *avg-4-max-4-BiGRU-emb-avg*. Furthermore, overall performance *SUM* is improved by 3.1%. It shows that we can get some local information by max pooling on the last layer, which can complement the global information by average pooling. The performance is improved from 70% to 71.7% at *recall 1* for image-to-sentence retrieval for four convolution layers with max pooling and the last layer with average pooling *avg-4-max-1234-BiGRU-emb-avg* compared with *avg-4-BiGRU-emb-avg*. Specially, the performance is improved by 3.7% for *SUM* metric. We think that there are different local characteristics for four convolutional layers, which can complement each other and global semantic information from texture level to semantic level. Therefore, we yield the best local features by fusing all four convolutional layers based on ablation study. We find that the enriched features are more helpful to retrieve complex images and long captions by visualize the magnitude of word-embedding vectors and the number of the image objects. In addition, we can observe that all the four models have better performance

than the baseline model *avg-4-GRU-emb-avg* in Table VI In addition, we find that it cannot drive up performance by concatenation the features from max-pooling on the first three Resnet blobs and average-pooling on Resnet C4 in Table VII. We hold the view that only max-pooling on Resnet C1, Resnet C2 and Resnet C3 ignores the semantic information in CNN. Anyway, we find out it is the best choice for image representation to concatenate four Resnet blob feature by max-pooling operation and average-pooling on Resnet C4 for MS COCO. Due to the limitation of data volume, RIR cannot work very well on Flickr30 K.

3) Re-Ranking and Ensemble: In order to further enhance the performance of cross-modal retrieval, we adopt two post-processing operations that is reranking [48] and ensemble. For a sentence (image) on test datasets, k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbors of image (sentence) on test datasets into a single vector, which is used for reranking inspired by [48] under the Jaccard distance. From the Table IX, we can know that the performance of the best single model *avg-4-max-1234-BiGRU-emb-avg* is improved from 501.8% to 503.2% after reranking the results at *SUM*. We also demonstrate the effectiveness on Flickr30 K shown on Table X. It's remarkable that the reranking method does not require training and can be directly used for testing, which is very practical for image-sentence matching owing to heavy time consuming when training. As we all know, it is the first time for reranking on cross-modal retrieval task in both MS COCO and Flickr30 K datasets.

As for ensemble, we combine features of the four different models in common space to obtain the image representation and text representation. From the Table II and Table VII, we can notice that the proposed ensemble method goes beyond the basic model on all the evaluation indicators, which validates the effectiveness of the proposed ensemble approach. Examples of cross-modal retrieval are shown in Fig. 6 for each component, which demonstrates the improvement when compared with baseline model.

V. Conclusion

We introduce a simple approach, i.e. RFE, to learn rich features embedding by mining the local and global information for both image and sentence. The RFE approach is effective to learn discriminative deep image and text feature for cross-modal retrieval. We achieve new state-of-the-art performance on the popular datasets for image-sentence matching task. This work paves a simple yet effective way to learn deep representation in embedding space for heterogeneous data, which makes a great contribution to the multi-modal learning community. In the future, we plan to develop more effective strategies for fusing the local and global feature, such as concatenating the key element by prototype selection. Moreover, we will develop object-level global and local feature by region proposal network, and fuse it with image-level global and local feature that is proposed in this paper.

References

[1] X. Lin and D. Parikh, "Leveraging visual question answering for image-caption ranking," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 261–277.

[2] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 299–307.

[3] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen, "Dual-path convolutional image-text embedding," 2017, *arXiv:1711.05535*.

[4] Y. Huang, Q. Wu, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6163–6171.

[5] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit*, 2018, pp. 7181–7189.

[6] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.

[7] D. Hu, C. Wang, F. Nie, and X. Li, "Dense multimodal fusion for hierarchically joint representation," in *Proc. IEEE Conf. Acoustics, Speech, Signal Process.*, 2019, pp. 3941–3945.

[8] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit*, 2015, pp. 53–61.

[9] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[10] L. Zhang, Y. Zhao, Z. Zhu, S. Wei, and X. Wu, "Mining semantically consistent patterns for cross-view data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 11, pp. 2745–2758, Nov. 2014.

[11] Y. Zhuang *et al.*, "Multimodal deep embedding via hierarchical grounded compositional semantics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 76–89, Jan. 2018.

[12] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[13] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[14] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multiordered discriminative structured subspace learning," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1220–1233, Jun. 2017.

[15] D. R. Hardoon, S. Szedmak, and J. Shawetaylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[16] A. Sharma, A. Kumar, H. Daume, and D. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit*, 2012, pp. 2160–2167.

[17] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[18] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[19] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4601–4611.

[20] Y. Yang *et al.*, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[21] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vision Conf.*, 2018.

[22] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2310–2318.

[23] J. Wehrmann and R. C. Barros, "Bidirectional retrieval made simple," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7718–7726.

[24] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vision*, 2018.

[25] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 2083–2090.

[26] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3864–3872.

[27] Q. Jiang and W. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3232–3240.

[28] K. Li, G.-J. Qi, J. Ye, and K. A. Hua, "Linear subspace ranking hashing for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1825–1838, Sep. 2017.

[29] R. Liu, S. Wei, Y. Zhao, Z. Zhu, and J. Wang, "Multi-view cross-media hashing with semantic consistency," *IEEE Multimedia*, vol. 25, no. 2, pp. 71–86, Apr.–Jun. 2018.

[30] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.
[31] C. Li *et al.*, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4242–4251.
[32] M. Hu *et al.*, "Collective reconstructive embeddings for cross-modal hashing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2770–2784, Jun. 2019.
[33] X. Zhang, H. Lai, and J. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 614–629.
[34] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4247–4255.
[35] Y. Wei *et al.*, "Modality-dependent cross-media retrieval," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, 2016, Art. no. 57.
[36] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 817–834.
[37] A. Fukui *et al.*, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.
[38] T. Liu, Y. Zhao, S. Wei, Y. Wei, and L. Liao, "Enhanced isomorphic semantic representation for cross-media retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 967–972.
[39] A. Jabri, A. Joulin, and L. van der Maaten, "Revisiting visual question answering baselines," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 727–739.
[40] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, "Joint global and co-attentive representation learning for image-sentence retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1398–1406.
[41] Y. Yang *et al.*, "Hierarchical multi-clue modelling for POI popularity prediction with heterogeneous tourist information," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 757–768, Apr. 2019.
[42] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, Nov. 2018.
[43] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.
[44] B. A. Plummer *et al.*, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2641–2649.
[45] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[47] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.
[48] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3652–3661.
[49] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
[50] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 3294–3302.
[51] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3128–3137.
[52] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 4437–4446.
[53] J. Mao *et al.*, "Deep captioning with multimodal recurrent neural networks (M-RNN)," in *Proc. Int. Conf. Learn. Representations*, 2015.
[54] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 595–603.
[55] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2623–2631.
[56] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learn. Representations*, 2016.
[57] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5005–5013.
[58] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc.IEEE Int. Conf. Comput. Vision*, 2017, pp. 4127–4136.
[59] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 686–701.

Xin Fu is currently working toward the Ph.D. degree with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include cross-media retrieval/completion, computer vision, and deep learning.

Yao Zhao received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. In October 2015, he visited the Swiss Federal Institute of Technology, Lausanne, Switzerland (EPFL). From December 2017 to March 2018, he visited the University of Southern California. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, video analysis and understanding, and artificial intelligence. He serves on the editorial boards of several international journals, including as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, a Senior Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, and an Area Editor for *Signal Processing: Image Communication*. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a Fellow of the IET.

Yunchao Wei received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2016, advised by Prof. Y. Zhao. He received the Winner prize of the object detection task (1a) in ILSVRC 2014, the runner-up prizes of all the video object detection tasks in ILSVRC 2017, the Winner Prizes of all human parsing tracks in the 2nd LIP challenge. His current research interests focus on weakly- and semi-supervised object recognition, multi-label image classification, video object detection, and multimodal analysis.

Yufeng Zhao received the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China. She is currently an Associate Professor with China Academy of Chinese Medical Sciences, Beijing, China. Her research interests include automatic image annotation, image retrieval, and multiple-instance learning.

Shikui Wei received the B.E. degree from Hebei University, Baoding, China, in 2003 and the Ph.D. degree in signal and information processing from Beijing Jiaotong University (BJTU), Beijing, China, in 2010. From 2010 to 2011, he was a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a Full Professor with the Institute of Information Science, BJTU. His research interests include computer vision, image/video analysis and retrieval, and machine learning.