Du XY, Yang Y, Yang L *et al.* Captioning videos using large-scale image corpus. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(3): 480–493 May 2017. DOI 10.1007/s11390-017-1738-7

Captioning Videos Using Large-Scale Image Corpus

Xiao-Yu Du [1,2], *Member, CCF*, Yang Yang [3,4], *Member, CCF, ACM, IEEE*, Liu Yang [1,5]
Fu-Min Shen [3,4], *Member, CCF, ACM, IEEE*, Zhi-Guang Qin [1], *Senior Member, CCF, Member, ACM, IEEE* and Jin-Hui Tang [1,6,*], *Senior Member, CCF, IEEE, Member, ACM*

[1] *School of Information and Software Engineering, University of Electronic Science and Technology of China Chengdu 610054, China*

[2] *School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China*

[3] *Center for Future Media, University of Electronic Science and Technology of China, Chengdu 611731, China*

[4] *School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731, China*

[5] *Sichuan University West China Hospital of Stomatology, Chengdu 610041, China*

[6] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*

E-mail: duxiaoyu@cuit.edu.cn; dlyyang@gmail.com; yangliu1988322@163.com; fumin.shen@gmail.com qinzg@uestc.edu.cn; jinhuitang@njust.edu.cn

Revised December 26, 2016; revised April 7, 2017.

Abstract Video captioning is the task of assigning complex high-level semantic descriptions (e.g., sentences or paragraphs) to video data. Different from previous video analysis techniques such as video annotation, video event detection and action recognition, video captioning is much closer to human cognition with smaller semantic gap. However, the scarcity of captioned video data severely limits the development of video captioning. In this paper, we propose a novel video captioning approach to describe videos by leveraging freely-available image corpus with abundant literal knowledge. There are two key aspects of our approach: 1) effective integration strategy bridging videos and images, and 2) high efficiency in handling ever-increasing training data. To achieve these goals, we adopt sophisticated visual hashing techniques to efficiently index and search large-scale images for relevant captions, which is of high extensibility to evolving data and the corresponding semantics. Extensive experimental results on various real-world visual datasets show the effectiveness of our approach with different hashing techniques, e.g., LSH (locality-sensitive hashing), PCA-ITQ (principle component analysis iterative quantization) and supervised discrete hashing, as compared with the state-of-the-art methods. It is worth noting that the empirical computational cost of our approach is much lower than that of an existing method, i.e., it takes 1/256 of the memory requirement and 1/64 of the time cost of the method of Devlin *et al.*

Keywords video captioning, hashing, image captioning

1 Introduction

In recent years, driven by the rapid development of mobile devices, Internet and social sharing platforms, a tremendous amount of user-generated video clips have been emerging on the Web. As an important alternative to words and images, such UGC videos have significantly changed the way people live and communicate. Every minute, there are 300-hour videos uploaded to YouTube① and 39 million videos uploaded to Vine②.

Regular Paper
Special Section of CVM 2017
This work was partially supported by the National Basic Research 973 Program of China under Grant No. 2014CB347600, the National Natural Science Foundation of China under Grant Nos. 61522203, 61572108, 61632007, and 61502081, the National Ten-Thousand Talents Program of China (Young Top-Notch Talent), the National Thousand Young Talents Program of China, the Fundamental Research Funds for the Central Universities of China under Grant Nos. ZYGX2014Z007 and ZYGX2015J055, and the Natural Science Foundation of Jiangsu Province of China under Grant No. BK20140058.

*Corresponding Author

①http://www.youtube.com, Mar. 2017.

②https://www.vine.co, Mar. 2017.

Besides, averagely 1.5 million videos are uploaded to Miaopai③ every day. Such a large amount of videos are mostly lack of descriptions. How to semantically index these video data and make them easier to seek and comprehend is a significant problem[1].

As shown in Fig.1, video annotation[2] focuses on deriving simple semantic concepts from low-level features, i.e., describing videos in natural language words rather than numeric features. To comprehend more semantic information from video data, event detection[3], action recognition[1], refined frame tags[4], deduced tags[5] and many other semantic descriptions[6-7] were designed for learning more complex semantics[8]. Nevertheless, the approaches are based on classification and their results are limited in certain pre-defined concept/event set. The descriptive power of these results is very limited. Providing users comprehensive descriptions, video captioning, which describes videos with sentences or paragraphs, is in urgent need.

(a) (b)

Fig.1. Describing video clips using video annotation, action recognition, event detection and video captioning.

In existing video captioning approaches, convolutional neural network (CNN) and long-short term memory (LSTM) are widely used. The conventional way is first adapting CNN model to extract video features and then utilizing LSTM to generate video captions. Nonetheless, there are two disadvantages.

1) *Extremely Complex Neural Networks for Video Data.* The complex structure slows down the execution efficiency which limits the extensibility for variable large-scale datasets.

2) *Scarcity Problem in Existing Video Corpus.* For instance, one of the most popular video captioning datasets, YouTube2Text (Y2T), only contains 1 970 video clips with captions. Such scale of training data can hardly guarantee reliable models to generalize to new video data.

In order to address the above problems, we turn to seeking for auxiliary captioned data. Image datasets always contain various corpus (e.g., imsitu④ provides situation annotations[9]). They are always used as auxiliary information in video processing[10-11]. Moreover we notice that there are many image captioning datasets with sufficiently relevant captions. Integrating image captioning corpus may be a nice way to augment the scale of corpus for video captioning. There are two major image captioning directions. The first one is to extract image features (e.g., CNN feature, and the supervised and unsupervised features[12]) and then generate captions by LSTM. The other is to generate captions from the descriptions of the similar images of the query image. As mentioned before, the first direction is less extensible and does not perform well when we have the problem of data scarcity and data updating. Therefore, in this paper, we choose to follow the second direction and refer to it as the retrieval approach in the subsequent content.

With the frequently-updated training data, the retrieval approach only needs to process the new data. In contrast, the deep learning approach may have to retrain the complex model, which can hardly be afforded. Meanwhile, even if there are a very small amount of similar images for a query image, retrieval approach would perform well by accurately discovering the most relevant samples to guarantee the performance. It is

③http://www.miaopai.com, Mar. 2017.

④http://imsitu.org, Mar. 2017.

intuitive that when the size of training data is large enough, the retrieval approach is able to achieve good performance. However, the computational cost of k-nearest neighbor (kNN) search is extremely high as the dimensionality of visual feature is normally large, e.g., the fc7 layer (4 096-D) from Alexnet or VGG. Moreover, original visual features may contain redundancy and/or noise which degrades the performance.

In order to compensate the drawback of traditional kNN search, we propose to employ hashing technique to index visual samples. With the excellent ability of reducing storage and computational cost, hashing[13] has been extensively studied and applied to multimedia retrieval in large-scale databases. The main idea is to utilize a small set of binary bits to represent a sample instead of a large number of real-valued features[14]. The similarity between samples can be efficiently computed with the Hamming distance between the corresponding hash codes. Locality-sensitive hashing (LSH) is known as one of the most popular data-independent hashing methods. In recent years, data-dependent or learning-based hashing methods become more popular since they perform much better in accuracy and compressing rate. Iterative quantization (ITQ) is a typical unsupervised optimization algorithm. The popular unsupervised hashing method PCA-ITQ[15] compresses features by principle component analysis (PCA) and then optimizes the features by ITQ. The recently proposed supervised discrete hashing (SDH)[16] is a well-performed supervised hashing method.

In this paper, we propose a novel video captioning approach using hashing techniques. First of all, we utilize visual hashing methods to preprocess the visual elements including images and videos. We instantiate our method with SDH in this paper. Then we retrieve similar images to query video and collect the corresponding captions as candidate captions. At last, we re-rank the candidate captions and select the consensus one as our captioning result.

The main contributions of this paper are summarized as follows.

• We propose a novel video captioning approach by leveraging well-captioned image data. The sufficient semantic descriptions in large-scale image datasets help to avoid the deficiency of captioned videos.

• We devise a simple yet effective approach to align image and video data to alleviate the influence of domain difference, which bridges the images and videos by extracting representative video frames.

• By utilizing the hashing techniques, compared with the approach proposed in [17], the proposed retrieval captioning approach achieves 64 times speedup with only 1/256 times memory overhead. Hash codes are used for effective and efficiency feature compression, which makes retrieval approach easily extensible to emerging data.

The reminder of this paper is organized as follows. Section 2 presents related work. Section 3 elaborates the details of our proposed approach. Section 4 demonstrates the experimental results, followed by the conclusions in Section 5.

2 Related Work

The rapid development of deep neural network (DNN) motivates the progress of multimedia processing. Convolutional neural network (CNN)[18] is a kind of DNN to process image information[19]. Recurrent neural network (RNN)[20] is another kind of DNN widely used in natural language generation[21]. To reduce the complexity of DNN implementation, various DNN tools were developed. Caffe[22] is the famous one which can construct, train, test and improve the net in a simple way. Therefore many classical CNN models are trained on Caffe including Alexnet[18], GoogLeNet[23] and VGG[24]. Since the CNN models always perform well across domains[25], the trained models are frequently selected to extract semantic features.

The fully connection layers named "fc7" in Alexnet and VGG16 are always treated as semantic features of a given image. The performances in image classification and similarity judgment are better than the traditional features. While features are extracted, captioning approaches would generate sentences word by word. Some existing researches utilized probability models such as Markov chain[26], expectation maximization[27] and maximum entropy[28] models. Some directly utilized neural network models RNN[29], and LSTM[30], and some mixed the CNN and RNN[31-32] to train an integrative model. In the mentioned approaches, the descriptions are generated mechanically, and thus we call them generation approaches. Another kind of captioning approaches takes existing human captions as the captioning result. They are called retrieval approaches.

One of the retrieval approaches performs well in MSCOCO challenge[33]. It is a simple and even rude solution. Given a query image, several similar images are collected as candidate images. Then the candidate captions which refer to the candidate images are rearranged. The best caption is treated as the consensus caption. The retrieval approach was first pro-

posed in [34]. Ordonez *et al.*[34] constructed one million captioned images corpus and designed the captioning experiments. And the experiments demonstrated that the more images appended to dataset, the more effective the results are.

To retrieve similar images, Devlin *et al.*[17] used the k nearest neighbor (kNN) method. Hashing is another famous retrieving method especially in large-scale datasets. The main idea is effectively compressing features into a set of binary codes (hash codes) to efficiently calculate the similarity using Hamming distance[35]. Some hashing methods were accelerated[36] and some could extend the identifications[37]. To generalize the process, some hashing frameworks were proposed[38]. There are three kinds of hashing: 1) data-independent methods such as locality sensitive hashing (LSH)[39], 2) unsupervised methods such as RDSH[40], IMH[41] and PCA-ITQ[15], and 3) supervised methods such as LDA-HASH[42], CCA-ITQ[15], RDCM[43] and SDH[16]. PCA-ITQ and CCA-ITQ are dimensionality reduction methods with iterative quantization and perform well presented in [15]. Shen *et al.*[16] demonstrated that SDH performs much better than other supervised hashing methods. Generally the supervised methods perform better than the unsupervised ones, and the unsupervised ones perform better than data-independent ones. It is notable that the number of hash bits also significantly affects the performance. For example, LSH needs at least 512 bits to retrieve fairly similar images empirically.

Many image captioning approaches have been applied to video captioning. Improving the image generation approaches has been a widely-used way. Ballas *et al.*[44] arranged several CNNs in a CNN matrix to process the adjacent frames synchronously. Mazloom *et al.*[45] took the bag-of-words with the SIFT features[46], MFCC[47] and attention points as video features. Shetty and Laaksonen[48] incorporated frame features extracted by CNN, classification from CNN and video feature as the input of LSTM. Yao *et al.*[49] formed a 3D-CNN encoder-decoder frame, with which spatio-temporal vectors are divided to chunks as the input of LSTM. Pan *et al.*[50] proposed a hierarchical recurrent neural network to extract the video features. Sener *et al.*[51] put information from videos and subtitles into bag-of-words and took it as the input of LSTM. The researches demonstrate that the ideas from image captioning approaches benefit videos. Therefore, the retrieval approaches rarely used in video captioning are another captioning choice.

3 Proposed Approach

In this section, we describe the proposed video captioning approach. As shown in Fig.2, we first integrate video corpus and image corpus by mapping them into a common visual semantic space, and then compress them to binary codes. Given a query video, we retrieve the similar samples from the mapped corpus and collect the captions related to the similar elements as our candidate captions. Finally, candidate captions are rearranged, and the "best" caption also called "consensus caption" is selected as our caption for the query video.

3.1 Data Integration

Note that each video may consist of a sequence of frames, which makes it difficult to directly compare videos and images. In this part, we integrate videos and images in the same space where we could calculate their similarity. We design a simple yet effective approach to achieve this goal. Based on the fact that current UGC videos are very short (normally a few seconds), it is reasonable to assume that the semantics of a video is focused. In fact, we have analyzed the distribution of video length in the Y2T dataset, and the average length is 9 s. As illustrated in Fig.3, the representative videos are semantically focused. Meanwhile, if we use multiple frames to represent videos, the common part of these frames (e.g., background) may be over-emphasized. Therefore, it is possible for us to use only one frame to represent a video, which also provides a straightforward way to connect image representation and video representation. In this work, we adopt the fc7 layer of VGG16 as the visual feature for both the image and the video.

3.2 Visual Hashing

The visual features are usually in high-dimensional space (4 096-D for AlexNet CNN features). We compress the features to a set of binary codes with hashing methods. There are various hashing methods. We select several typical hashing methods in different kinds as our benchmark.

LSH. Locality sensitive hashing (LSH)[39] is a fundamental hashing method and it is data-independent. Codes are generated using random projection. Randomly constructing a weight matrix $\boldsymbol{W} \in \mathbb{R}^{d\times b}$, where d is the number of source feature dimensions and b is the number of hash bits, the codes can be calculated by $\boldsymbol{C} = (\boldsymbol{F} \cdot \boldsymbol{W} > 0)$. Here $\boldsymbol{F} \in \mathbb{R}^{n\times d}$ indicates n samples with d-dimensional features.

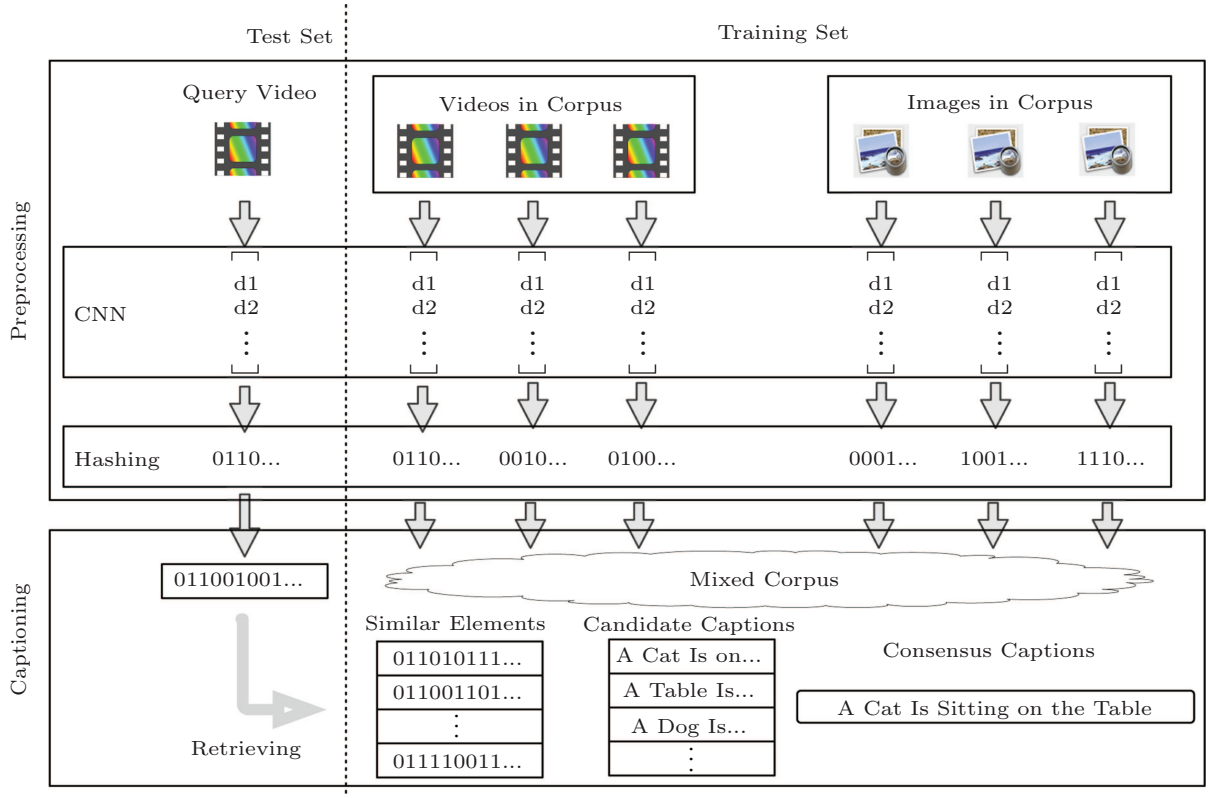


Fig.2. Video clip captioning flow.

Fig.3. Frames extracted from video clips. The main ideas to be explained are clear in the frames.

PCA-ITQ. PCA-ITQ[15] is one of the most popular unsupervised hashing methods. Different from LSH, it is optimized with data features to get better performance. PCA-ITQ contains two steps: 1) utilizing principle component analysis (PCA) to reduce feature dimensions, and 2) utilizing ITQ to optimize features by shifting and rotating. As shown in [15], we would get a transform matrix $\boldsymbol{T}$ to indicate the PCA reduction and a transform matrix $\boldsymbol{R}$ to indicate the ITQ operation. Therefore, the weight matrix of PCA-ITQ is $\boldsymbol{W} = \boldsymbol{T} \cdot \boldsymbol{R}$. The hash codes can be calculated by $\boldsymbol{C} = (\boldsymbol{F} \cdot \boldsymbol{W} > 0)$ as well.

SDH. The recently proposed supervised discrete hashing (SDH)[16] is a well-performed supervised hashing method. In this paper, we utilize SDH as our video hashing module. SDH learns binary codes and hash functions through the following problem.

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{W},\boldsymbol{F}} \quad & \sum_{i=1}^{n} ||y_i - \boldsymbol{W}^{\mathrm{T}} b_i||^2 + \lambda ||\boldsymbol{W}||^2 + \\ & \nu \sum_{i=1}^{n} ||b_i - H(x_i)||^2 \\ \text{s.t.} \quad & b_i \in \{-1, 1\}^L. \end{aligned}$$

That is,

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{W},\boldsymbol{F}} \quad & ||\boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}} \boldsymbol{B}||^2 + \\ & \lambda ||\boldsymbol{W}||^2 + \nu ||\boldsymbol{B} - H(\boldsymbol{X})||^2 \\ \text{s.t.} \quad & \boldsymbol{B} \in \{-1, 1\}^{L \times n}. \end{aligned}$$

Here $H(\cdot)$ is the hash function to be learned, λ and ν are penalty parameters, $\boldsymbol{X} = \{x_i\}_{i=1}^{n}$ is the data matrix,

$\boldsymbol{Y} = \{y_i\}_{i=1}^n \in \{0,1\}^{C\times n}$ is the ground truth label matrix, $\boldsymbol{W} = \{w_k\}_{k=1}^C \in \mathbb{R}^{L\times C}$ is the classification vector, and $\boldsymbol{B} = \{b_i\}_{i=1}^n \in \{-1,1\}^{L\times n}$ is the learned binary codes, where the i-th column b_i is the binary code for x_i.

The key of SDH is that binary code optimization is not solved with resort to continuous relaxations but directly with discrete optimization. In SDH, the optimization is processed with three alternating steps: W-step, F-step and B-step[16]. With the obtained hash function $H(\cdot)$, our hash codes could be generated as $C = (H(\boldsymbol{X}) > 0)$.

3.3 Video Captioning

In this part, we describe how to generate captions for videos using hashing and auxiliary images. The overall flowchart is illustrated in Fig.2, which is comprised of three major components. Preliminarily, we transform all training images and videos into hash codes using the learned hash functions. All the hash codes can be directly stored in the main memory. Given a target video v, we go through the following steps.

Step 1. We first export the representative frame and extract the fc7 layer of VGG as the visual feature. Then, we convert the visual feature to hash codes, denoted as q, using the learned hash functions.

Step 2. Given $\boldsymbol{q}$, we search the whole database of binary codes to obtain similar samples. Note that the similarity between two samples is calculated according to Hamming distance. Given a code of retrieved sample $\boldsymbol{p} \in \{0,1\}^L$, the Hamming distance of $\boldsymbol{q}$ and $\boldsymbol{p}$ is defined as

$$hd(\boldsymbol{q},\boldsymbol{p}) = \sum_{l=1}^{L} \mathrm{XOR}(q_l, p_l),$$

where q_l and p_l are the l-th elements of $\boldsymbol{q}$ and $\boldsymbol{p}$, respectively. The elements with the smallest Hamming distance are the similar samples.

Step 3. Finally, we collect the captions of all retrieved samples to form a candidate caption set, denoted as $\mathcal{C}$. To generate the desired caption for the target, we intend to first find a "maximum" subset of size m, denoted as $\mathcal{M} \subseteq \mathcal{C}$. Then, we decide the most representative caption c^* if c^* is the most similar one to the elements in $\mathcal{M}$. Following [17], we take BLEU-4[52] score as our caption similarity and our consensus caption is formally generated by:

$$c^* = \underset{c\in\mathcal{C}}{\mathrm{argmax}} \max_{\mathcal{M}\subseteq\mathcal{C}} \sum_{c'\in\mathcal{M}} Sim(c,c'),$$

where $Sim(\cdot,\cdot)$ computes the similarity of two captions.

4 Experiments

In this section, we discuss the performance of our proposed approach applied in video captioning. The experiments contain two steps. At the first step, we improve the image retrieval captioning approach, and compare the effects using different kinds of hashing methods and kNN method. At the second step, we apply our video captioning approach and illustrate the interesting results. We score the approaches in the experiments by BLEU[52], METEOR[53], CIDEr[54] and ROUGE[55], which are calculated with the MSCOCO evaluation server.

4.1 Datasets

Our experiments are designed on images and videos. Therefore the image and video datasets shown below are selected to illustrate the multiple aspects of our approach.

MSCOCO. It is an image captioning dataset provided by Microsoft Corporation. Each image has at least five captions. The training set, validation set and test set have been separated and an evaluation system is provided for results evaluating uniformly. There are totally 82 783 images in the training set (one of them is damaged), 40 504 images in the validation set and 40 775 in the test set. The test set is confidential for us; therefore we take the validation set as our test data.

YouTube2Text (*Y2T*). It is a user-generated video dataset with various clean and noisy captions. There are totally 1 970 videos and 122 665 captions. 80 839 of them are English captions and 33 855 are cleaned English captions. Statistically the average length of the videos is 9 s. Most of them can be described using one sentence. Following [44, 56-57] we take 1 300 videos as the training samples and the other 670 videos as the test data.

In our experiments, the official evaluation system of MSCOCO is used to score our results. The MSCOCO dataset is used to evaluate our image captioning approach and the Y2T dataset is used to evaluate our video captioning approach. MSCOCO is also used as an extended corpus for video captioning.

The numbers of n-grams in the mentioned corpus are shown in Table 1. Obviously the n-grams in image corpus are much larger than those in video corpus. Using image corpus may be a way to extend video corpus.

Table 1. Statistics of n-Grams in Y2T and MSCOCO

Dataset	1-Gram	2-Gram	3-Gram	4-Gram	Total
Y2T	13 091	81 133	159 958	200 092	454 274
MSCOCO Training	23 128	319 703	905 852	1 504 502	2 753 185
MSCOCO Validation	17 348	205 186	533 178	832 242	1 587 954
Total (unique)	32 134	456 186	1 336 145	2 261 555	4 086 020

4.2 Corpus Codes

At the first step of our experiments, the visual elements are referred to as binary codes using VGG16 and hashing methods. We explore the performances of typical hashing methods including data-independent hashing LSH, unsupervised hashing PCA-ITQ, and supervised hashing SDH. To evaluate the performances on large-scale corpus, we exploit the results using 512 bits hashing codes and 64 bits hashing codes.

Using 512 bits hashing codes costs only 64 bytes memory per image. Storing all the images costs less than 10 MB memory. Using 64-bit hashing codes costs much less. In contrast, taking the 4 096-dimensional fc7 layer of VGG16 as an image feature costs 16 KB per image and nearly 2 GB in total. That is about 256 times more than using 512-bit codes. The saved storage grows more obviously in larger scale datasets. Table 2 demonstrates the memory cost of our test datasets. 4 096-D indicates the features using 4 096 floats such as the original feature exported from the Alexnet. 512-bit and 64-bit indicate the features using 512 bits and 64 bits respectively. Obviously the compressed features are much smaller than the original features.

Table 2. Memory Requirements Using Different Sizes of Features for Visual Elements in Y2T and MSCOCO

Dataset	4 096-D	512-Bit	64-Bit
Y2T	31 MB	124.0 KB	16 KB
MSCOCO	1 894 MB	7.4 MB	1 MB

4.3 Similar Image Retrieval

Fig.4 demonstrates the similar image retrieving results using kNN, LSH, PCA-ITQ and SDH. From the five most similar images, we can hardly judge which method is better. The hashing methods could get effective candidate images as good as kNN. Therefore, hashing methods work well in retrieving similar images.

This change makes the execution cost much less time, as shown in Fig.5. With the increasing amount of data, kNN costs more and more time than the hashing methods. When there are about 10^6 images, using kNN costs about 10 minutes. In contrast, hashing methods cost several micro-seconds still. Obviously, after our improvements, the retrieval step is more appropriate for large-scale datasets.

4.4 Image Captioning Result

To explore the effects of image captioning using kNN, LSH, PCA-ITQ and SDH, we design several tests on the MSCOCO dataset. Following [17], we take half of the validation set (20 252 images) as tuning data and take the remains as the test data, and we score the results using the MSCOCO evaluation system.

There are two variables k and m for image captioning. We utilize tuning data to explore the best k and m for our approaches. Fig.6 shows the effect of k-m parameters. Generally the variances of different approaches are similar.

Table 3 presents that retrieval approaches using hashing methods perform as well as those using kNN method. The scores are based on BLEU[52], METEOR[53], CIDEr[54] and ROUGE[55]. This demonstrates our proposed image captioning approach is effective and efficient. Because of less memory and time cost, our approach could be used in a much larger dataset. The results also illustrate that although our approaches using different hashing methods have similar performances, there are still a fine distinction. The supervised method performs better than the unsupervised method, and the unsupervised method is better than the data-independent method, because of the descending order of scores.

Besides, we compare the codes using 512 bits with the codes using 64 bits. As shown in Table 3, using 64-bit, SDH still performs well, but LSH performs much worse. When resources are limited, using 64-bit SDH must be a better choice.

4.5 Video Captioning Result

Considering some frame in the video can denote the main contents of the video, we select one of the frames

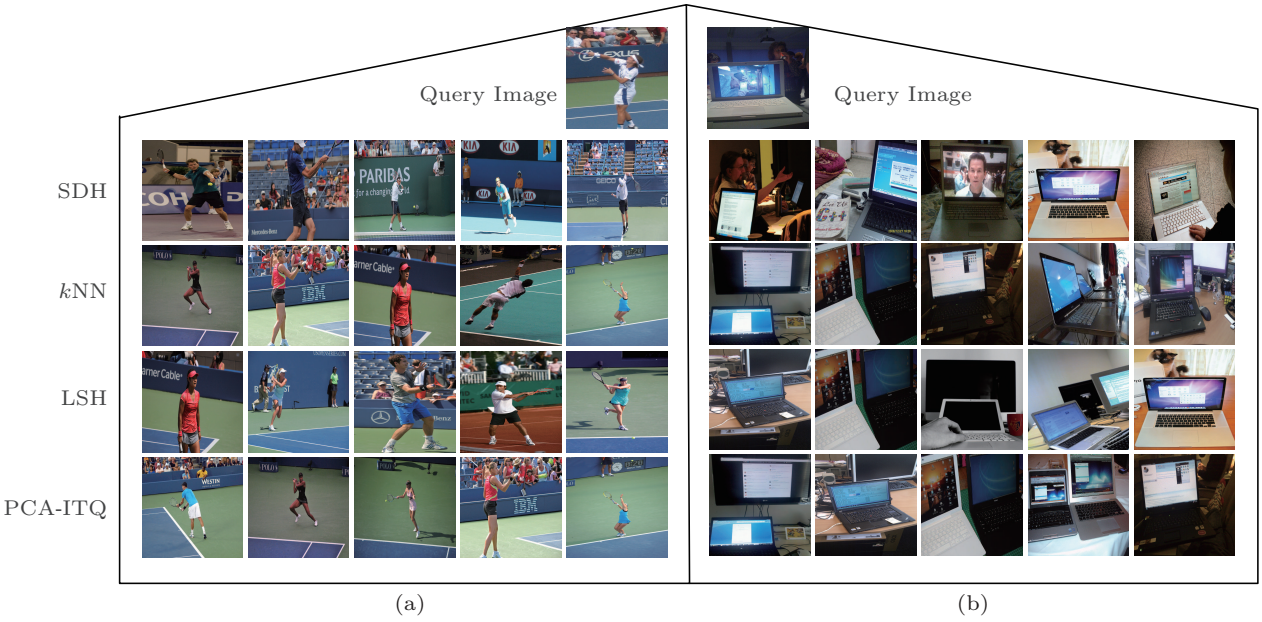


Fig.4. Similar images retrieved by SDH, kNN, LSH, and PCA-ITQ.

as the videos' indicator. Here we take the first frames and the middle frames of the videos for comparison. We take the test set of Y2T as ground truth, the entire MSCOCO dataset and the Y2T training set as our labeled corpus respectively. Table 4 presents the results using our SDH retrieval captioning approach. This simple and even rude method performs as well as the approach proposed in [57]. The result also demonstrates that the first frames and the middle frames perform similarly and both contain main video contents.

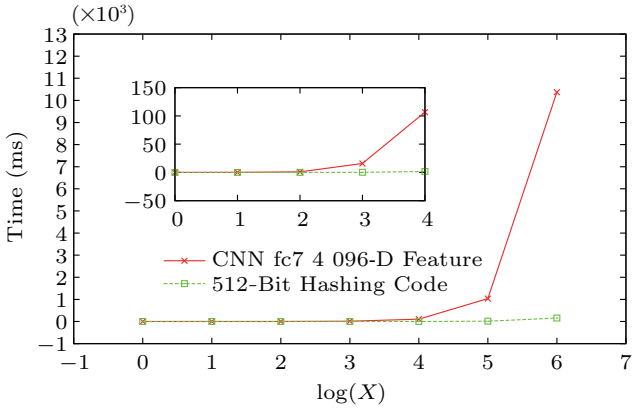


Fig.5. Time cost for corpus with X images.

For more details, we list some captioned videos in Fig.7 to present the captioning results. Obviously the generated captions are related to the query video. The inaccuracies are mainly caused by the deficiency of related corpus. The captions for the videos in Fig.7(a) use "puppy" and "dog" to describe the same animal. Even though both of them are correct, the "outer" corpus would obtain a lower score. Most situations similar to videos in Fig.7(b) stand for a party. Thus the caption based on MSCOCO describes some details of a party which causes a low score. In videos of Fig.7(c), the focuses of the two corpora are much different: Y2T says the process while MSCOCO says the status. Fig.7(d) demonstrates that only if there are enough phrases in MSCOCO, it may perform better than Y2T.

Generally, the Y2T training dataset is a better training dataset than the MSCOCO dataset in captioning Y2T test data, as demonstrated in Table 4, even though there are less descriptions in Y2T corpus. The BLEU-1 score presents that the results on MSCOCO contain similar information. But the much lower BLEU-4 score presents that the phrases and sentence structures in MSCOCO are much different. With the increasing size of corpus, the phrases and sentences would cover most of the grammars and this problem would be changed. Moreover, MSCOCO is not large enough so that the image types are deficient in our application. Fig.8 shows two unmatched samples. The first one is a girl putting some sticker on her face. There are 464 captions in MSCOCO related to "sticker", but only three captions are related to "woman", "women" or "girl". The second sample presents a similar problem. The captions including words "cat" and "turtle" are laid there. Besides, the results are limited by the extracted features. Although VGG16 is almost the best

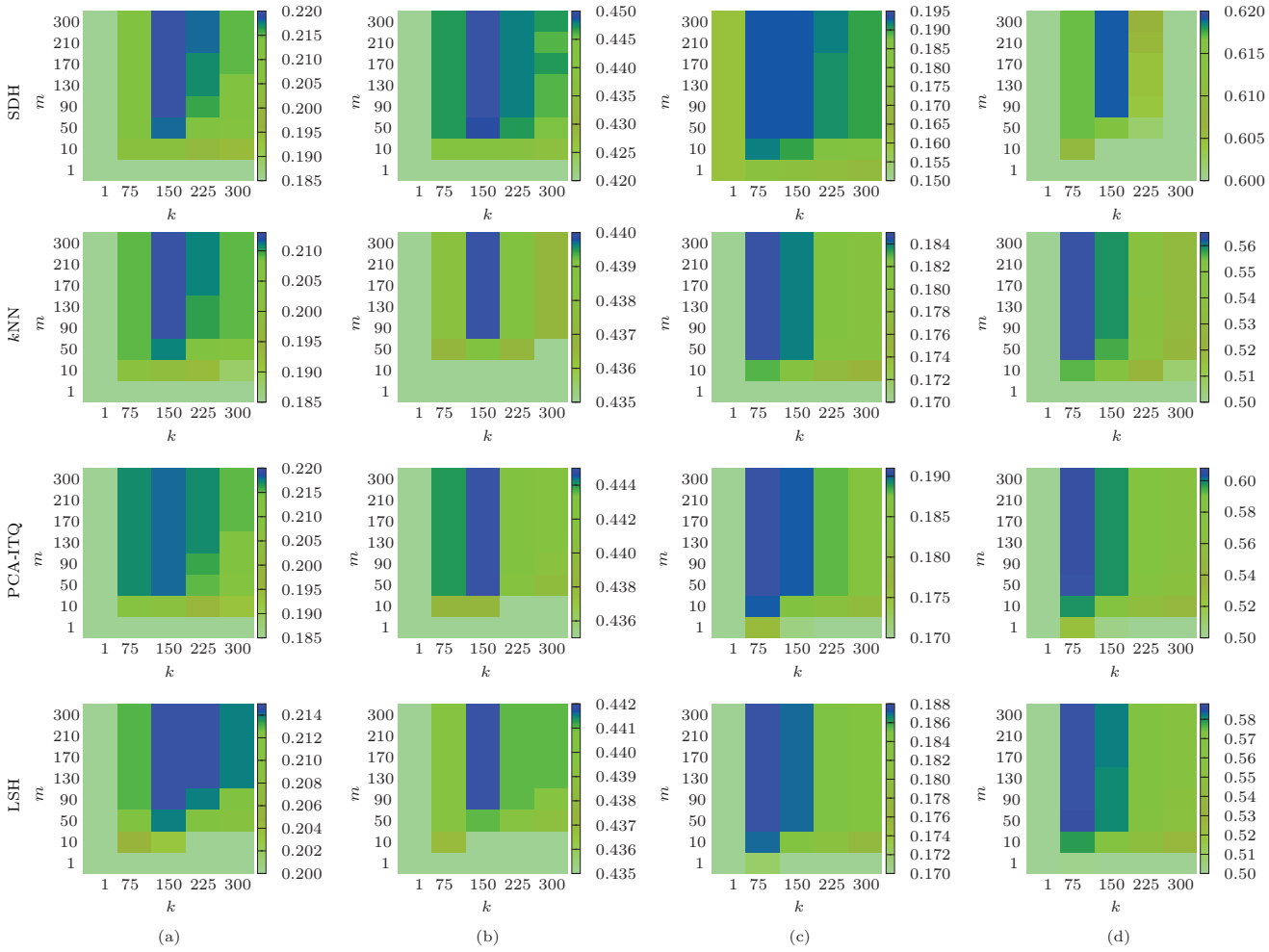


Fig.6. Captioning scores while selecting the consensus caption from m subsets of the captions for the k similar images. (a) BLEU-4. (b) ROUGE_L. (c) METEOR. (d) CIDEr.

Table 3. Scores of Our Approaches with Different Hashing Methods on MSCOCO Validation

Approach		k	m	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE
MS kNN[17]		130	150	58.8	41.2	29.2	21.4	56.0	18.4	44.0
512-bit	Ours-LSH	170	150	59.0	41.6	29.6	21.7	58.6	18.7	44.2
	Ours-PCA-ITQ	90	150	59.1	41.9	29.9	21.9	60.0	18.9	44.3
	Ours-SDH	90	150	59.5	41.9	29.7	21.6	61.2	19.2	44.6
64-bit	Ours-LSH	130	150	52.8	34.6	23.3	16.5	41.4	15.6	39.6
	Ours-PCA-ITQ	130	150	57.5	39.8	27.9	20.3	54.2	18.0	43.0
	Ours-SDH	130	150	58.6	40.7	28.6	20.7	58.1	18.5	43.8

Table 4. Results of Video Retrieval Captioning Approach on Y2T Using First Frame and Middle Frame

Frame	Corpus	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE
First	Y2T	61.0	42.4	30.6	20.6	18.80	22.0	54.1
	MSCOCO	51.4	28.3	13.8	7.4	9.10	15.8	41.0
Middle	Y2T	60.8	42.0	30.3	20.6	18.50	21.6	53.5
	MSCOCO	52.4	29.7	15.6	8.3	10.70	16.2	41.8
Thomason *et al.*[57]		-	-	-	13.68	-	23.90	-

(a)

(b)

(c)

(d)

Fig.7. Captions for four sample videos generated by our approaches with Y2T and MSCOCO corpus. (a) Y2T: a puppy plays with a ball. MSCOCO: a dog laying on the floor in a room. (b) Y2T: a man is playing a guitar onstage. MSCOCO: a group of people standing around and dancing. (c) Y2T: someone is slicing bread into slices. MSCOCO: a plate of food on a table. (d) Y2T: a man is riding a horse bareback. MSCOCO: a man riding on the back of an elephant.

features for image classification, it lacks salient point. While we are finding a "panda" appearing in a clip, there are only one panda image in 10-most similar images. Most of the retrieving results focus on the background and the rest are related to the color. To eliminate the effect from hashing methods, we retrieve the k-similar images using kNN. The results shown in Fig.9 confirm our discovery.

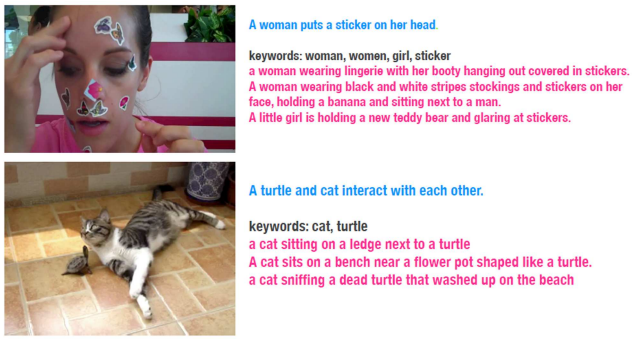


Fig.8. No appropriate captions for the query clip found. The blue caption is a user-generated caption for the clip. The red ones are all the captions related to the keywords.

Fig.9. Images similar to a "panda" image.

Finally we explore the relations between the corpus size and the captioning results. We integrate the training set and the validation set of MSCOCO and get totally 123 286 images. Then we design a 13-step experiment (s1∼s13): at the first step s1, we take the 10 000 images with their captions as our corpus. At each of the following steps we append 10 000 more images to the corpus. At the last step s13 we use the entire 123 286 images. Then we score the results at each step. Fig.10 presents that with the corpus size increased from s1 to s13, our approach gets a higher and higher score. Obviously large-scale corpus does help in video captioning.

5 Conclusions

In this paper, we proposed a novel video captioning approach to describe videos by leveraging freely-available image corpus with abundant literal knowledge. Hashing techniques are utilized to significantly improve both computation and storage efficiencies. The results demonstrated that our approach is exactly effective and efficient. Our simple captioning approach adapts to the large-scale datasets.

The volumes of multimedia data in the Internet are explosively increasing. Since the noisy data disturbs our approaches, how to clean the large-scale data, integrate them, and use other auxiliary features, may be the next topic.

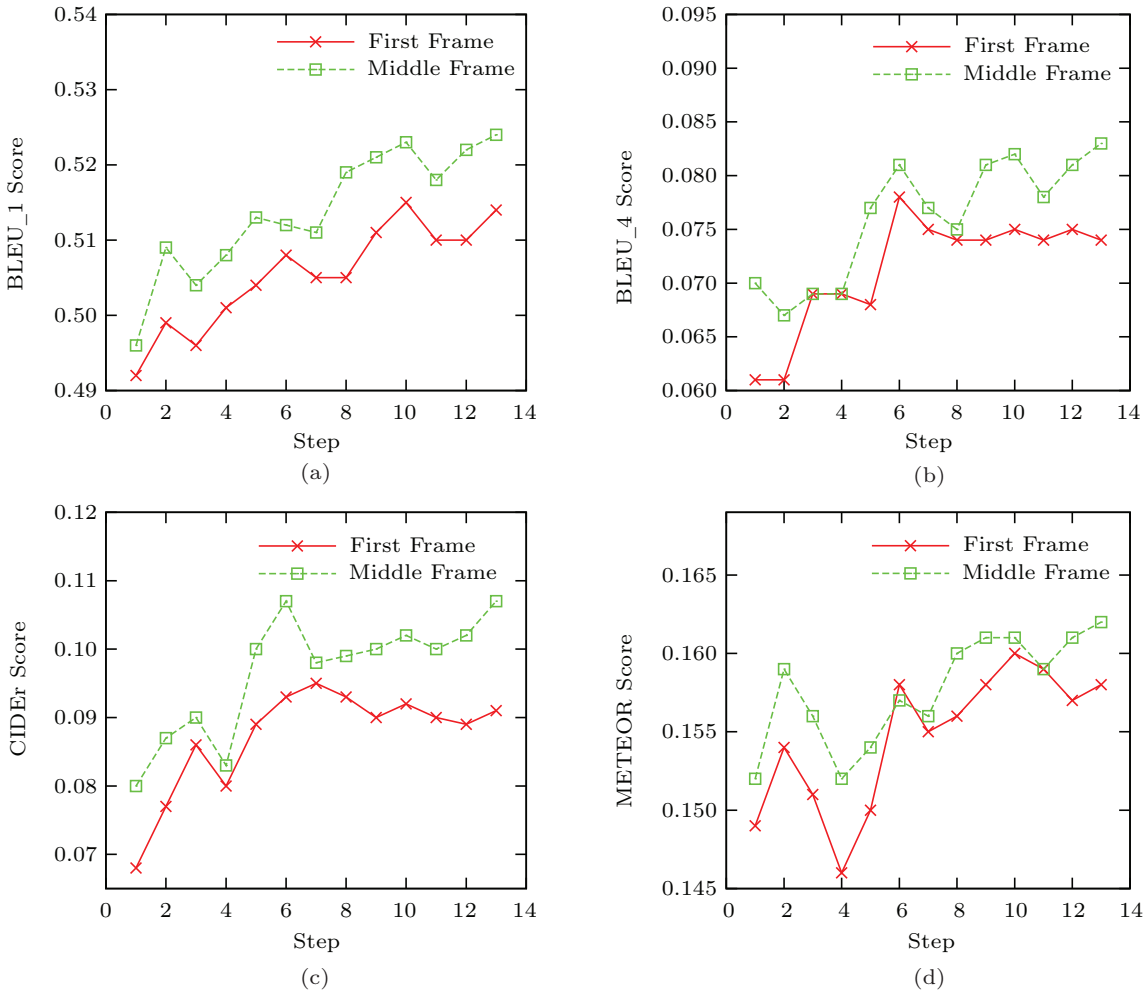


Fig.10. Scores of video captioning using different scales of MSCOCO corpus.

References

[1] Song Y, Tang J H, Liu F, Yan S C. Body surface context: A new robust feature for action recognition from depth videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, 24(6): 952-964.

[2] Qi G J, Hua X S, Rui Y, Tang J H, Mei T, Zhang H J. Correlative multi-label video annotation. In *Proc. the 15th*

ACM International Conference on Multimedia, Sept. 2007, pp.17-26.

[3] Chen J W, Cui Y, Ye G N, Liu D, Chang S F. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proc. International Conference on Multimedia Retrieval*, Apr. 2014.

[4] Tang J H, Shu X B, Qi G J, Li Z C, Wang M, Yan S C, Jain R. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, doi: 10.1109/TPAMI.2016.2608882.

[5] Tang J H, Shu X B, Li Z C, Qi G J, Wang J D. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2016, 12(4s): Article No. 68.

[6] Li Z C, Tang J H. Weakly supervised deep matrix factorization for social image understanding. *IEEE Transactions on Image Processing*, 2017, 26(1): 276-288.

[7] Li Z C, Liu J, Tang J H, Lu H Q. Robust structured subspace learning for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(10): 2085-2098.

[8] Yang Y, Zhang H W, Zhang M X, Shen F M, Li X L. Visual coding in a semantic hierarchy. In *Proc. the 23rd ACM International Conference on Multimedia*, Oct. 2015, pp.59-68.

[9] Yatskar M, Zettlemoyer L, Farhadi A. Situation recognition: Visual semantic role labeling for image understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016. pp.5534-5542.

[10] Yang Y, Zha Z J, Gao Y, Zhu X F, Chua T S. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Transactions on Multimedia*, 2014, 16(6): 1677-1689.

[11] Yang Y, Yang Y, Shen H T. Effective transfer tagging from image to video. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2013, 9(2): Article No. 14.

[12] Li Z C, Liu J, Yang Y, Zhou X F, Lu H Q. Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(9): 2138-2150.

[13] Wang J D, Shen H T, Song J K, Ji J Q. Hashing for similarity search: A survey. arXiv:1408.2927, 2014. https://arxiv.org/abs/1408.2927, Apr. 2017.

[14] Tang J H, Li Z C, Wang M, Zhao R Z. Neighborhood discriminant hashing for large-scale image retrieval. *IEEE Transactions on Image Processing*, 2015, 24(9): 2827-2840.

[15] Gong Y C, Lazebnik S. Iterative quantization: A procrustean approach to learning binary codes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2011 pp.817-824.

[16] Shen F M, Shen C H, Liu W, Shen H T. Supervised discrete hashing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.37-45.

[17] Devlin J, Gupta S, Girshick R, Mitchell M, Zitnick C L. Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467, 2015. https://arxiv.org/abs/1505.04467, Apr. 2017.

[18] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In *Proc. the 25th International Conference on Neural Information Processing Systems*, Dec. 2012, pp.1097-1105.

[19] Zhu Z, Liang D, Zhang S H, Huang X L, Li B, Hu S M. Traffic-sign detection and classification in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.2110-2118.

[20] Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S. Recurrent neural network based language model. In *Proc. the 11th Annual Conference of the International Speech Communication Association*, Sep. 2010, pp.1045-1048.

[21] Song J, Tang S L, Xiao J, Wu F, Zhang Z F. LSTM-in-LSTM for generating long descriptions of images. *Computational Visual Media*, 2016, 2(4): 379-388.

[22] Jia Y Q, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In *Proc. the 22nd ACM International Conference on Multimedia*, Nov. 2014, pp.675-678.

[23] Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp.1-9.

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. https://arxiv.org/abs/1409.1556, Apr. 2017.

[25] Razavian A S, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp.512-519.

[26] Norris J R. Markov Chains. Cambridge University Press, 1998.

[27] Moon T K. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 1996, 13(6): 47-60.

[28] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1996, 22(1): 39-71.

[29] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. In *Proc. the 31st International Conference on Machine Learning*, Jun. 2014, pp.595-603.

[30] Wu Q, Shen C H, van den Hengel A, Liu L Q, Dick A. Image captioning with an intermediate attributes layer. arXiv:1506.01144v1, 2015. https://www.arxiv.org/abs/1506.01144v1, Apr. 2017.

[31] Gao H Y, Mao J H, Zhou J, Huang Z H, Wang L, Xu W. Are you talking to a machine? Dataset and methods for multilingual image question answering. arXiv:1505.05612, 2015. https://arxiv.org/abs/1505.05612, Apr. 2017.

[32] Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. arXiv:1411.4389, 2014. https://www.arxiv.org/abs/1411.4389, Apr. 2017.

[33] Chen X L, Fang H, Lin T Y, Vedantam R, Gupta S, Dollar P, Zitnick C L. Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325, 2015. https://www.arxiv.org/abs/1504.00325, Apr. 2017.

[34] Ordonez V, Kulkarni G, Berg T L. Im2text: Describing images using 1 million captioned photographs. In *Proc. Neural Information Processing Systems*, Dec. 2011, pp.1143-1151.

[35] Charikar M S. Similarity estimation techniques from rounding algorithms. In *Proc. the 34th Annual ACM Symposium on Theory of Computing*, May 2002, pp.380-388.

[36] Shen F M, Zhou X, Yang Y, Song J K, Shen H T, Tao D C. A fast optimization method for general binary code learning. *IEEE Transactions on Image Processing*, 2016, 25(12): 5610-5621.

[37] Yang Y, Luo Y S, Chen W L, Shen F M, Shao J, Shen H T. Zero-shot hashing via transferring supervised knowledge. In *Proc. ACM Conference on Multimedia*, Oct. 2016, pp.1286-1295.

[38] Shen F M, Shen C H, Shi Q F, van den Hengel A, Tang Z M, Shen H T. Hashing on nonlinear manifolds. *IEEE Transactions on Image Processing*, 2015, 24(6): 1839-1851.

[39] Gionis A, Indyk P, Motwani R. Similarity search in high dimensions via hashing. In *Proc. the 25th International Conference on Very Large Data Bases*, Sept. 1999, pp.518-529.

[40] Yang Y, Shen F M, Shen H T, Li H X, Li X L. Robust discrete spectral hashing for large-scale image semantic indexing. *IEEE Transactions on Big Data*, 2015, 1(4): 162-171.

[41] Song J K, Yang Y, Yang Y, Huang Z, Shen H T. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proc. ACM SIGMOD International Conference on Management of Data*, Jun. 2013, pp.785-796.

[42] Strecha C, Bronstein A, Bronstein M, Fua P. Ldahash: Improved matching with smaller descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(1): 66-78.

[43] Luo Y D, Yang Y, Shen F M, Huang Z, Zhou P, Shen H T. Robust discrete code modeling for supervised hashing. *Pattern Recognition*, 2017, doi: 10.1016/j.patcog.2017.02.034.

[44] Ballas N, Yao L, Pal C, Courville A. Delving deeper into convolutional networks for learning video representations. arXiv:1511.06432, 2015. https://arxiv.org/abs/1511.06432, Apr. 2017.

[45] Mazloom M, Li X R, Snoek C G M. TagBook: A semantic video representation without supervision for event detection. arXiv:1510.02899v2, 2015. https://arxiv.org/abs/1510.02899v2, Apr. 2017.

[46] Lowe D G. Object recognition from local scale-invariant features. In *Proc. the 7th IEEE International Conference on Computer Vision*, Sep. 1999, pp.1150-1157.

[47] Xu M, Duan L Y, Cai J F, Chia L T, Xu C S, Tian Q. HMM-based audio keyword generation. In *Proc. Pacific-Rim Conference on Multimedia*, Dec. 2004. pp.566-574.

[48] Shetty R, Laaksonen J. Video captioning with recurrent networks based on frame- and video-level features and visual content classification. arXiv:1512.02949, 2015. https://arxiv.org/abs/1512.02949, Apr. 2017.

[49] Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A. Describing videos by exploiting temporal structure. In *Proc. IEEE International Conference on Computer Vision*, Dec. 2015, pp.4507-4515.

[50] Pan P B, Xu Z W, Yang Y, Wu F, Zhuang Y T. Hierarchical recurrent neural encoder for video representation with application to captioning. arXiv:1511.03476, 2015. https://arxiv.org/abs/1511.03476, Apr. 2017.

[51] Sener O, Zamir A R, Savarese S, Saxena A. Unsupervised semantic parsing of video collections. In *Proc. IEEE International Conference on Computer Vision*, Dec. 2015, pp.4480-4488.

[52] Papineni K, Roukos S, Ward T, Zhu W J. BLEU: A method for automatic evaluation of machine translation. In *Proc. the 40th Annual Meeting on Association for Computational Linguistics*, Jul. 2002, pp.311-318.

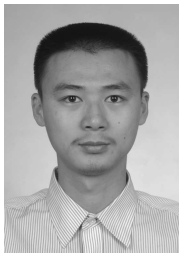
[53] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In *Proc. the 9th Workshop on Statistical Machine Translation*, Vol. 6, Apr. 2014, pp.376-380.

[54] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based image description evaluation. arXiv:1411.5726, 2014. https://arxiv.org/abs/1411.5726, Apr. 2017.

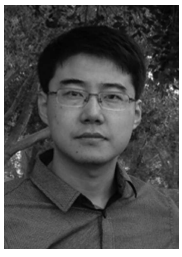
[55] Lin C Y. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL-04 Workshop on Text Summarization Branches Out*, Jul. 2004, pp.74-81.

[56] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Mooney R, Darrell T, Saenko K. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proc. IEEE International Conference on Computer Vision*, Dec. 2013, pp.2712-2719.

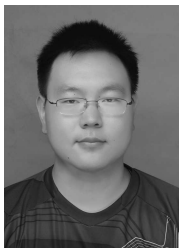
[57] Thomason J, Venugopalan S, Guadarrama S, Saenko K, Mooney R J. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proc. the 25th International Conference on Computational Linguistics*, Aug. 2014, pp.1218-1227.



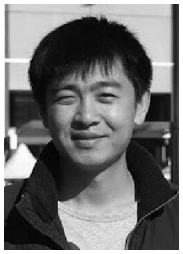
Xiao-Yu Du is currently a lecturer in the School of Software Engineering of Chengdu University of Information Technology, Chengdu, and a Ph.D. candidate of University of Electronic Science and Technology of China, Chengdu. He received his M.E. degree in computer software and theory in 2011 and B.S. degree in computer science and technology in 2008, both from Beijing Normal University, Beijing. His research interests include multimedia analysis and retrieval, computer vision, and machine learning.



Yang Yang is currently with University of Electronic Science and Technology of China, Chengdu. He joined National University of Singapore (NUS) as a research fellow, working with Prof. Tat-Seng Chua (2012∼2014). He was conferred his Ph.D. degree in information technology from the University of Queensland (UQ) in 2012. He obtained his Master's degree in computer science in 2009 from Peking University, Beijing, and Bachelor's degree in computer science in 2006 from Jilin University, Changchun, respectively. He research interests mainly focus on multimedia search, social media analysis, and machine learning.

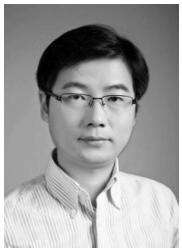


Liu Yang is a staff of Sichuan University West China Hospital of Stomatology, Chengdu. He is pursuing his Master's degree in School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu. His research interests in pattern recognition and computer vision.



Fu-Min Shen is currently a lecturer in School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu. He received his Ph.D. degree from School of Computer Science, Nanjing University of Science and Technology, Nanjing, in 2014. With the support of the China Scholarship Council from Apr. 2011 to Nov. 2012, he visited ACVT (Australian Centre for Visual Technology) and School of Computer Science, University of Adelaide, Australia. He received his Bachelor's degree in applied mathematics from Shandong University, Jinan, in 2007.

Zhi-Guang Qin received his Ph.D. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, in 1996. He is currently a full professor of the School of Information and Software Engineering with UESTC, Chengdu. His research interests include computer networking, information security, cryptography, and pattern analysis.



Jin-Hui Tang is a professor in School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing. He received his B.E. degree in communication engineering and Ph.D. degree in signal and information processing from the University of Science and Technology of China, Hefei, in 2003 and 2008 respectively. His current research interests include multimedia search and computer vision. He has authored over 160 journal and conference papers in these areas. He is a co-recipient of the Best Student Paper Award in MMM 2016, the Best Paper Finalist in ACM MM 2015, and Best Paper Awards in ACM MM 2007, PCM 2011 and ICIMCS 2011. He is a senior member of CCF and IEEE, and a member of ACM.