2017 International Symposium on Intelligent Signal Processing and Communication Systems, November 6-9, 2017, Xiamen, China

WHEN DIARY MEETS LIFELOG VIDEO

Min Gao, Jiande Sun, En Yu, Xiao Dong

Shandong Normal University
School of Information Science and Engineering
Jinan 250014, Shandong province, China

Jing Li

Shandong Management University
School of Mechanical and Electrical Engineering
Jinan 250014, Shandong Province, China

ABSTRACT

As the increasing quantities of personal data is collected by individuals, the number of lifelog video is increasing. People make microblogging in the form of the text, later, in form of the text with pictures or videos. In this paper, a cross-media lifelog video retrieval approach is proposed to automatically match the corresponding lifelog video clip from a long lifelog video according to diary description(Fig.2). This model consists of a video captioning model and a text retrieval model. We train an encoder-decoder architecture to effectively learn video captioning by MSVD and MSR-VTT datasets. We use the similarity judgment to achieve the retrieval of the text. The similarity is measured by measuring the cosine distance between the two vectors. We experiment on some participants' lifelog videos and diaries. This approach is evaluated by investigating participants' satisfaction with results of lifelog video selected, the results show most of the testers were satisfied with the results.

***Index Terms*—** Diary, lifelog video, video captioning, L-STM, text retrieval

1. INTRODUCTION

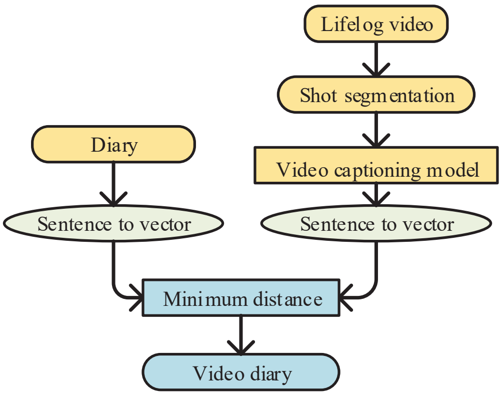


Fig. 1. An overall framework of cross-media lifelog video retrieval corresponding to diary model.

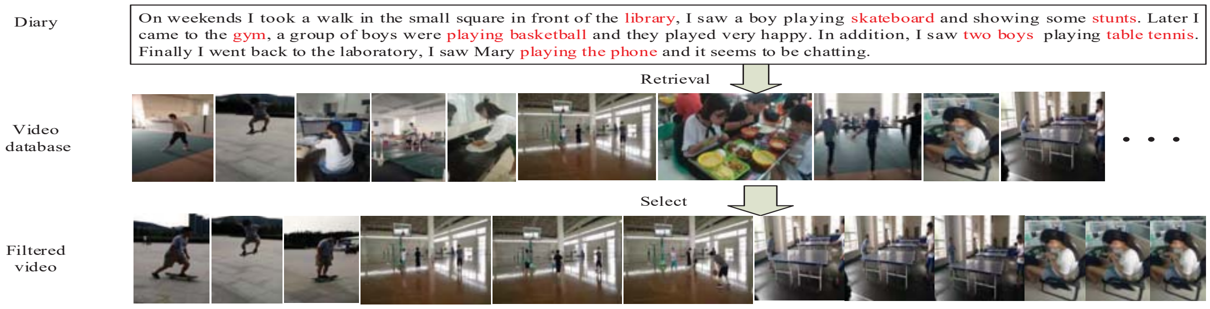
The rapid development of microblog is leading to a dramatic increase in video data sharing every minute. Once we make microblogging is only text-based gradually transformed into the diary description accompanied by pictures or videos. When we write a diary, we want to be able to automatically match the video clips. This can reduce the trouble and time to search for the corresponding videos. We keep uploading images and videos to the microblog, so the diary automatically matches lifelog videos is very necessary. The task of lifelog video retrieval by diary description (from the text to the visual) is presented. In this paper, we proposed an approach that the task of lifelog video retrieval corresponding to text description consists of a video captioning model and a text retrieval model (Fig.1).

Computer vision has been developed for more than 50 years. During this period, visual understanding of this field has made considerable progress. A few years ago, when we talk about images or video understanding, what we can do is to give an image or a video automatically labeled with some independent labels. Today, we have pushed visual understand forward a step by deep learning, that is, we can convert single labels into fluent natural language descriptions of the current video content. Visual and language is actually a cross field. For most people, it's easy to watch a short video and describe what's happening in the video. But for machines, it is complicated to extract the meaning from the video pixels and express it in a naturally smooth language. We aim to build a bridge of visual and natural language, not only need to understand the vision, but also know how to model the natural language. At the same time, the bridge is bi-directional, it can be from the visual to text (such as caption, sentiment, visual thinking answering, etc.), it can also be from the text to the visual (such as generation, search).

2. RELATED WORK

Image captioning comes from the basic idea of language translation, the process is generally first using CNN to encode the image to obtain visual features, and then use the RNN to decode this feature to generate image description. In the extraction of visual features, the computer visual field

978-1-5386-2159-2/17/$31.00 ©2017 IEEE



On weekends I took a walk in the small square in front of the library, I saw a boy playing skateboard and showing some stunts. Later I came to the gym, a group of boys were playing basketball and they played very happy. In addition, I saw two boys playing table tennis. Finally I went back to the laboratory, I saw Mary playing the phone and it seems to be chatting.

Fig. 2. Our algorithm retrieval lifelog videos by diary descriptions. Each sentence corresponds to a video clip with the most relevant content.The red words indicate the most relevant content of the video, which facilitate retrieval of the task.

commonly used methods are high-level semantic features and attention mechanisms, etc., it can also be used directly to the automatic encoder for processing. Image captioning work can be summarized as ”image captioning with X”, where X can be visual attention, visual attributes, entity recognition, dense caption and reinforcement learning modules. Video captioning and image captioning have different points, when we have to understand the video, we not only to understand the objects in each frame, but also to understand the movement of objects between multiple frames. Therefore, video understanding is often more complex than image comprehension. The main difference between a video description and an image description is that the video description adds a temporal sequences relationship. Video captioning task can be understood as the video image sequence to the text sequence of the sequence to sequence task. In recent years, most of the papers used LSTM to construct the encoder-decoder structure, which utilizes LSTM encoder to encode the features of the video image sequence and utilizes LSTM decoder to decode the text information.

In this paper, a cross-media retrieval approach is proposed to automatically match the corresponding video clips from a long lifelog videos corresponding to diary descriptions. The cross-media lifelog video retrieval model consists of a video captioning model and a text retrieval model. First, we need to segment this long lifelog video according to shot. Identifying shot [1] is vital. Usually, everyone’s lifelog video scene is complex. There is a serious lack of labeled training data. Unsupervised learning [2] becomes a focused issue. Then we need to preprocess video clips. Video captioning is the core of the entire video search model, and the video captioning has received more and more attention because of its many important applications such as human-computer interaction, video retrieval, description of movies for the blind. Recently, Wang et al. proposed dense trajectories [3] and later proposed improved dense trajectories [4] which is currently the state-of-the-art hand-crafted feature. With access to a large number of training and availability of powerful parallel machines (G-PUs, CPU clusters), convolution neural networks [5] have a substantial breakthrough in some artificial intelligence problems, related to speech recognition [6][7], text [8][9] and image based problems [10][11]. In this paper, we proposed to translate from video pixels to natural language with a single deep neural network. Deep neural networks can learn more powerful features (Donahue et al., 2013; Zeiler and Fergus, 2014). We utilized a model of the encoder-decoder structure to generate a word sequence. And we utilized a sentence similarity measurement model for text retrieval.

3. APPROACH

We introduce a method which takes a diary description as input and generates the corresponding diary video as output. Our goal is to find a set of sentences semantically relevant to the input diary description. In order to do this, first, we get a set of sentences generated by video captioning model. Then we learn a similarity function that sentences were generated by video captioning model and human diary description. Once the semantically relevant sentences which are generated by using video captioning model are selected, corresponding videos are determined.

3.1. Video segment

The primary task is to segment a long lifelog video, because the length of the video is 10 to 25 seconds in training datasets, so when the video is segmented, the length of videos is also controlled within this range. A lifelog video is segmented according to shot. Finally, a long lifelog video is divided into many 10 to 25 seconds video clips. Video shot boundary detection is under global concern as the first step of all kinds of video processing. In the shot boundary detection algorithm, the color histogram based on adjacent two images is the most commonly used class. The color histogram difference between successive frames is calculated and compared with a preset threshold, which is greater than the threshold, and the scene is considered to have changed.

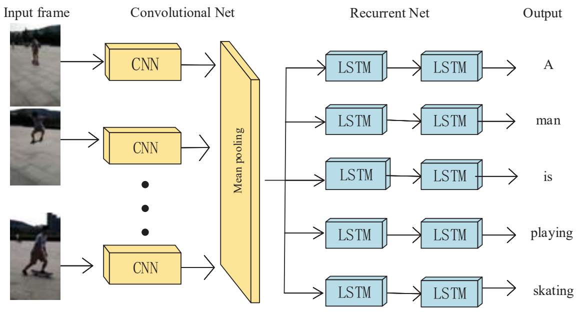


Fig. 3. The structure of our video captioning network. We extract fc$_7$ features for each frame, mean pool the features across the entire video, then, input feature vectors to LSTM network. The LSTM output a sentence.

3.2. Video captioning

Our framework is based on a sequence to sequence - video to text model [12], i.e., encoder-decoder framework(Fig.3), which is same as the previous works [12][13]. First, the input of the convolutional network is videos frame. Each frame is convoluted by a convolutional neural network to generate a fixed dimension vector representation. Then the output of the convolutional neural network is the input of the first layer LSTM. The second layer LSTM outputs a word sequence. The overall objective function we are optimizing is the log-likelihood over the whole training set,

$$\max_{\theta} \sum_{t=1}^{T} \log P\left(Z_t \left| v \right., y_{t-1}; \theta\right) \tag{1}$$

Where Z$_t$ is a one-hot vector (1-of-N coding, where N is the size of the word vocabulary) used to represent the word at the t-th time step, v is the feature vector output by the video encoder and θ represents the video captioning models parameters.

3.3. LSTMs for sequence generation

One of the key points of RNN is that they can be used to connect previous information to the current task, such as the use of past video clips to speculate on the understanding of the current clip. LSTM is a special type of RNN that can learn long-term dependency information. All RNNs have a chained architecture of a repeating neural network module. LSTM is also the same structure, but the duplicate module has a different structure(Fig.4). Unlike a single neural network layer, there are four repeating neural network modules in a very special way to interact. The key to LSTM is cell status. LSTM has the ability to remove or increase the information to the cell state by a well-designed structure called ”gate”. The gate is a way to get information through. They contain a sigmoid neural network layer and a pointwise multiplication operation.

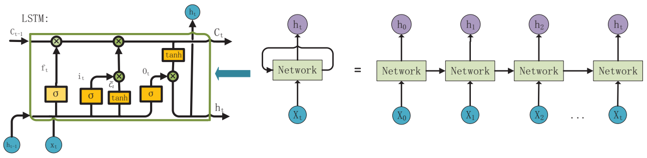


Fig. 4. The right part of the figure is the loop expansion of the RNN, the left part of the figure is the internal structure of the LSTM. In LSTM model, there are three gates to controlling information, i$_t$ is input gate, f$_t$ is forget gate, o$_t$ is output gate. The memory cell C$_t$ is at the core of the LSTM unit.

The first step in our LSTM is to decide what information we will discard from the cell state. This decision is done by a forget gate.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2}$$

The next step is to determine what kind of new information is stored in the cell state. It contains two parts. First, the sigmoid layer is called ”input gate” to determine what value we are going to update. Then, a tanh layer creates a new candidate value vector.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{3}$$

$$\tilde{C}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right) \tag{4}$$

Then the old cell state needs to be updated, so C$_{t-1}$ is updated to C$_t$. We multiply the old state with f$_t$ and discard the information we need to discard. Then add i$_t$ * C$_t$. This is the new candidate value.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

Ultimately, we need to determine what value to the output. This output will be based on our cell status, but also a filtered version. First, we run a sigmoid layer to determine which part of the cell state will be output. Next, we process the cell state by tanh and multiply it by the output of the sigmoid gate. Eventually, we only output the part of the output we determined.

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

Where σ is the sigmoidal non-linearity, and the weight matrices denoted by W$_X$ are the trained parameters.

3.4. Text retrieval

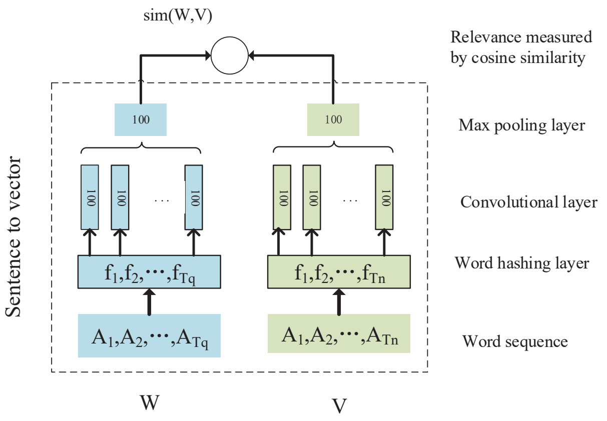


Fig. 5. A framework of similarity measurement between sentences. Word hashing layer uses sub-word unit as raw input to handle very large vocabulary. Convolutional and max-pooling layer used to identify key words/concepts in W and V. We utilize sentence to vector model to extract abstract semantic representations, which is a fix dimension vector. Last the similarity between two sentences is measured by cosine distance.

The features W(w_1,w_2,...,w_m) of sentences generated by the video captioning model are extracted by sentence to vector model, the features V(v_1,v_2,...,v_n)of human diary are also extracted by the sentence to vector model(Fig.5). All the features are represented by a fixed dimension vector. The similarity between them is measured by measuring the cosine distance between the two vectors. The cosine similarity is used as a measure of the magnitude of the difference between two individuals in the cosine of the two vectors in the vector space. The closer the cosine value is to 1, it means that the closer the angle is to 0 degrees, the more similar the two vectors are, the so-called ”cosine similarity”.

$$\begin{aligned} & sim_{\max} = \cos\left(v_j, w_j\right)_{\max} = \frac{v_j \cdot w_j}{\|v_j\| \times \|w_j\|} \\ & i = (1, 2, ..., m), j = (1, 2, ..., n) \end{aligned} \tag{8}$$

4. EXPERIMENTAL

4.1. Experimental datasets

Microsoft Research Video Description Corpus (MSVD): We perform experiments on the Microsoft Research Video Description Corpus (MSVD dataset). MSVD dataset is provided by Microsoft Research. The dataset consists of 1970 YouTube video clips. The duration of each clip is between 10 seconds to 25 seconds, and each video is annotated with approximately 40 English sentences.

MSR-VTT dataset: MSR-VTT dataset is also provided by Microsoft Research. MSR-VTT provides 10K web video clips with 41.2 hours and 200K clip-sentence pairs in total, covering the most comprehensive categories and diverse visual content, and representing the largest dataset in terms of sentence and vocabulary. Each clip is annotated with about 20 natural sentences by 1,327 AMT workers[14].In addition, the MSR-VTT also provides category information for each video (total of 20 categories). This category information is a priori and is known in the test set. At the same time, each video contains audio information.

4.2. Experimental settings

We train video captioning models with MSVD dataset and MSR-VTT dataset respectively. In our experiment, we extract a frame every two seconds for each video. Then we use GoogLeNet [15]and ResNet to extract the features of the frames. We use the feature as the input of the first layer LSTM. When we are training on MSVD and MSR-VTT, we use the following settings: All the LSTM units are set to 1,024, the visual feature size and the word size are set as 2048 empirically. We set the learning rate ρ=0.001, learning rate decay factor β= 0.99, encoder max sequence length is L_1=30, decode max sentence length is L_2=20, empirically. We train the model for 200 epochs. We utilize Tensorflow framework to conduct our experiments.

We use the PTBTokenizer in Stanford CoreNLP tools [16] to tokenize sentences and remove duplicate words and punctuation. This yields a vocabulary of 7063 in size for the MSVD dataset and a vocabulary of 14058 in size for the MSR-VTT dataset. In the training phase, we add a starting sentence tag (BOS) to begin each sentence and an ending tag (EOS) to end each sentence, so that our video captioning model can handle sentences of different lengths.

We segment a lifelog video into some clips according to the shots. The diary description may be quite different from the description of the machine’s understanding of the video, the reference diary descriptions (written by different individuals) for this record video are given. In consensus to the dataset, we also uniformly sample frames every 2 seconds and apply our proposed text retrieval video algorithm. The similarity between the generated sentence description by video captioning model and diary description is measured by measuring the cosine of the angle between the two vectors. The features of the generated sentence description by video captioning model and diary description are extracted by the sentence to vector model to obtain a vector representation of the fixed dimension. We set the dimension of this vector to 100. The maximum distance between the current and predicted word within a sentence is set to 5. Words with total frequency lower than minimum count will be dropped. We set minimum count to 5.

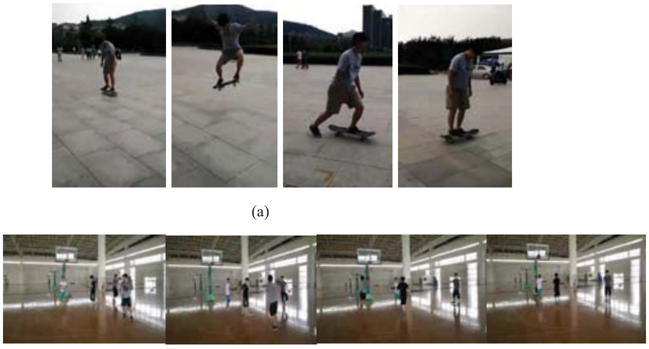
rank	A	B	C	D	E
proportion(%)	10%	50%	25%	10%	5%

Table 1. Evaluation of the Application of Text Retrieval Video Model by Testers. A represents very satisfied, B represents relatively satisfied, C denotes satisfied, D represents a relatively poor, E represents extremely poor.

4.3. Experimental results

The diary descriptions written by different individuals for their record video are given. Twenty people experimented with their diary descriptions, and then we surveyed the degree of satisfaction. We set five different levels of satisfaction: very satisfied, relatively satisfied, satisfied, relatively poor, extremely poor. Only a small number of people are not satisfied(Table 1).

Human description and machine description can not always be consistent, so when the description of the people and the machine is very different, the retrieval will be incorrect. There are some positive example(Fig.6) and negative examples(Fig.7). In correct example(Fig.6), there will be something relevant between video captioning generated sentence and diary description. However, in incorrect example(Fig.7), content is not related at all between video captioning generated sentence and diary description.



(b)

Fig. 6. (a) Video captioning generated sentence: a man is doing some skating stunts. Diary description: On weekends I took a walk in the small square in front of the library, I saw a boy was playing skateboard and showed some stunts. (b) Video captioning generated sentence: a man is throwing a ball in the air. Diary description: Later I came to the gym, a group of boys were playing basketball, they played very happy.

5. CONCLUSION

In this paper, we presented a text retrieval video application for matching human diary with lifelog video. The text retrieval video model consists of two parts: video captioning and text retrieval text. Video captioning utilizes sequence to sequence framework and is trained on MSVD dataset and MSR-VTT dataset. We use the similarity judgment to achieve the retrieval of the text. The similarity between two sentences is measured by measuring the cosine distance between the two vectors. Our algorithm can retrieval lifelog videos with corresponding content corresponding to diary description. The results of retrieval show most of the videos by diary description selected are correct. In this task, we have some problems to be considered and to be solved urgently. In the diary, we often refer to the name or title, but making the machine according to the name or title determine the gender or interpersonal relationship is difficult. In addition, sometimes there may be misalignment in time, we may describe the morning thing, and the machine found the afternoon video, which is possible. Last the person's diary description may be very different from the machine-generated sentence, which requires experiments to verify these problems.

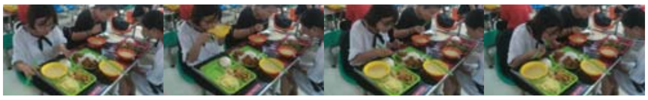


Fig. 7. Video captioning generated sentence: a man is seasoning some poultry. Diary description: At noon I went to the dining room for lunch with my long lost friends, we had a pleasant chat.

6. ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation of China (No.61572298), Key Research and Development Foundation of Shandong Province (2016GGX101009), Natural Science Foundation of Shandong Province (ZR2014FM012), and Scientific Research and Development Foundation of Shandong Provincial Education Department (J15LN60). And we gratefully acknowledge the support of NVIDIA Corporation with the donation of the TITAN X GPU used for this research. The contact author is Jiande Sun (jiandesun@hotmail.com).

7. REFERENCES

[1] Hehe Fan, Liang Zheng, Yi Yang, “Unsupervised Person Re-identification: Clustering and Fine-tuning,” in *arXiv preprint arXiv:1705.10444*. 2017.

[2] Hehe Fan, Xiaojun Chang, De Cheng, Yi Yang, Dong Xu, Alexander G. Hauptmann, “Complex Event Detection by Identifying Reliable Shots from Untrimmed Videos,” in *2017 IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2017.

[3] H.Wang, A. Klaser, C. Schmid, and C.-L. Liu, “Action recognition by dense trajectories,” in *CVPR*. 2011.

[4] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in *ICCV*. 2013.

[5] Y. Bengio and Y. LeCun, “Convolutional networks for images,speech, and time-series,” in *The Handbook of Brain Theory and Neural Networks*. 1995.

[6] A. Mohamed, T. Sainath, G. Dahl, B. Ramabhadran, G. Hinton,and M. Picheny, “Deep belief networks using discriminative features for phone recognition,” in *ICASSP*. 2011.

[7] A. Mohamed, D. Yu, and L. Deng, “Investigation of full sequence training of deep belief networks for speech recognition,” in *INTERSPEECH*. 2010.

[8] R. Collobert and J.Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *ICML*. 2008.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *arXiv preprint arXiv:1301.3781*. 2013.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*. 2012.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “ Rich feature hierarchies for accurate object detection and semantic segmentation,” in *arXiv preprint arXiv:1311.2524*. 2013.

[12] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney,Trevor Darrell, Kate Saenk, “ Sequence to Sequence – Video to Text,” in *ICCV* . 2015.

[13] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, “ Describing videos by exploiting temporal structure,” in *ICCV* . 2015.

[14] J Xu, T Mei, T Yao, Y Rui, “ MSR-VTT: A Large Video Description Dataset for Bridging Video and Language,” in *In Computer Vision and Pattern Recognition*. 2016.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed,D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “ Going deeper with convolutions,” in *CVPR*. 2015.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J.Bethard, and D. McClosky, “ The Stanford CoreNLP natural language processing toolkit,” in *ACL*. 2014.