Video Description using Learning Multiple Features

Xin Xu[a], Chunping Liu[a,b,c**], Haibin Liu[a], Yi Ji[a], Zhaohui Wang[a]

[a]School of Computer Science and Technology, Soochow University, Suzhou 215006, China

[b]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China

[c]Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210046, China

***Abstract*—Generating descriptions for open-domain videos is a major challenge for computer vision due to the complex dynamics. In this paper, we propose a video description model based on multiple features. In the process of encoding, we exploit two complementary features. The spatial one is extracted from the raw frame by VGG-16 model. The temporal one is extracted from the SIFT flow image by a fine-tuned VGG-16 model. In the process of decoding, we further add the mean pooling feature which represents holistic feature of the video. For generating sentence of the video, we utilize two-layer LSTMs model to generate sentence about the video. We evaluate several variants of our model on the MSVD dataset for METEOR metrics. The experimental results show that our model can be beneficial for generating sequence about the video.**

Keywords—video description; SIFT flow; VGG-16; mean pooling; LSTM;

I. INTRODUCTION

With the development of artificial intelligence, people are increasingly interested in describing visual contents with natural language sentences, such as image caption or video description. For human being, it is easy to describe what happened in one video according to the visual information. While generating a sequence using visual information automatically is still a complicated and challenge task for the machine, video description also has a wide application prospect, such as video retrieval, video caption based on semantic content, describing videos for the blind and automated video surveillance.

Video description is a transform from sequence to sequence in a sense as it inputs a sequence of raw frames and outputs a sequence composed of meaningful words. General model for video description usually uses two-step pipeline [1,2,3]. Firstly, it extracts features of the video in the encoding stage and then generates sequence using these features in the decoding stage. Most methods identify a fixed tuple of sematic roles, such as subject, verb object and scene, which are used to generate sequence. Then these semantic contents are translated to a sentence using a sentence template. These methods simplified the process of video description, but the diversity of natural language structures is limited to a large extent. Venugopalan et al. [4] performed a mean pooling over frames to get the holistic feature about the video and use two-layer LSTM [5] to generate the sequence. Xu et al. [6] proposed a MM-VDN model which employs FCN [7] and MIL to learn features from different scales. These approaches only consider the spatial information of the video, and ignore the order of sequence which contains temporal information. Yao et al. [8] employed 3-D CNN [9] to extract spatio-temporal features and soft-attention [10] to select the most relevant temporal segments automatically. However, results show that 3-D CNN alone gives limited performance improvement without soft-attention mechanism.

In this paper, we extract features using a pre-trained VGG-16 [11] model on 1.2M images. Then the Long Short Term Memory [12] (LSTM), a special type of Recurrent Neural Network (RNN), is used to construct sentence sequence. Taking inspiration from video description models in [13], we propose an improved model for video description with spatial and temporal features in the encoding stage and the mean pooling feature in the decoding stage. Our model with a two-layer encoder-decoder LSTM is illustrated in Fig. 1. Firstly our model encodes the frames one by one and then decodes words one by one.

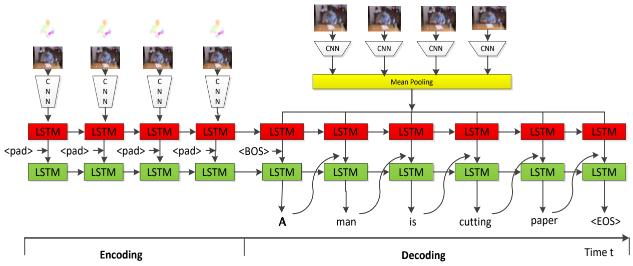


Fig. 1. Our model proposed which fuses multiple features during encoding stage and decoding stage.

II. OUR APPROACHES

Our model is a sequence-to-sequence process for video description, where the input is the sequence of frames $X(x_1,x_2\cdots,x_n)$, and the output is the sequence of words $Y(y_1,y_2\cdots,y_m)$. In our model, we estimate the conditional probability of an output sequence $Y(y_1,y_2\cdots,y_m)$ given an input sequence $X(x_1,x_2\cdots,x_n)$:

**Corresponding author: Chunping Liu, email: cpliu@suda.edu.cn

$$p(y_1,y_2\cdots,y_m|x_1,x_2\cdots,x_n) \tag{1}$$

Sutskever et al. [5] shown that it is effective to resolve such sequence-to-sequence problem with an LSTM such as machine translation. In this section, we describe our video description model based on CNN and LSTM in details by two parts.

A. LSTM for Sequence Generation

The main idea for video description is to first encode the input sequence of frames, representing the video using a latent vector representation, and then decode from that representation to a sentence.

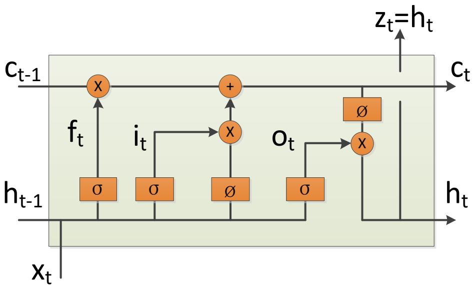


Fig. 2. LSTM structure with forget gate, input gate, output gate and memory cell.

Long Short Term Memory (LSTM) was proposed by Hochreiter and Schmidhuber in 1997 [12] and recently improved by Graves [14]. It is difficult for traditional RNN to learn long-range dependencies. However, LSTM which contains explicitly controllable memory units, is well known to be able to learn long-range temporal dependencies. Fig. 2 depicts the LSTM structure used in this paper.

The core of LSTM is the memory cell *c*, which updates its state through the controllable gate at every time step. The cell is modulated by gates which is a sigmoid function with a range of [0, 1]. These gates determine whether the LSTM keeps the value from the gate (if the layer evaluates to 1), or discards it (if it evaluates to 0). The forget gate f_t allows the LSTM to forget its previous memory c_{t-1}. The input gate i_t allows the LSTM to determine whether new information will be used for updating memory cell. Then a new candidate value g_t is created by a tanh function with input x_t and previous hidden state h_{t-1}. We multiply f_t with c_{t-1} to discard the information which should be discarded, then multiply g_t with i_t to keep the information needed. The updated memory cell c_t is computed by adding the result of the above two steps. The output gate o_t decides how much memory to transfer to the hidden state h_t. The final output of the result h_t is obtained by multiplying o_t with result which processes the memory cell c_t by a tanh function. The formula group for LSTM is defined as follows, where σ denotes sigmoid function, ϕ denotes hyperbolic tangent function tanh, $\odot$ denotes element-wise product with the gate value, and weight matrices denoted by W_{ij} and biases b_j are the trained parameters.

$$f_t = \sigma(W_{xf}\,x_t + W_{hf}\,h_{t-1} + b_f) \tag{2}$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{3}$$

$$g_t = \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{6}$$

$$h_t = o_t \odot \phi(c_t) \tag{7}$$

B. Multiple features fusion for Video Description

We exploit a similar two-layer encoder-decoder LSTM model proposed in [13] to describe video with sentence. However, our model as in Fig.1 inputs raw frames and SIFT flow images to train separate VGG-16 model for video description during encoding stage. After all frames are exhausted, the model inputs holistic feature of the video during decoding stage.

1) Encoder with spatial feature. In this work, we utilize the 16-layer VGG model (VGG-16) [10] pre-trained on the ImageNet dataset [16] to extract 4096 dimensional output of the fully connection layer (fc7) as the spatial feature for the raw frame. We then learn a new linear embedding of the features to a 500 dimensional space as the input to top layer LSTM. The weights of embedding are learned jointly with the LSTM layer during training.

2) Encoder with temporal feature. We follow the approach in [16] and first extract SIFT flow field between two consecutive frames on UCF-101 dataset. We then visualize SIFT flow field as SIFT flow images. These images are used with the label same as the category of activity to fine-tune a specific VGG-16 model for SIFT flow images. On MSVD dataset, we apply the same method to generate SIFT flow images. Then, we use the specific VGG-16 model to extract 4096 dimensional output of the fully connection layer (fc7) as the temporal features. We take the same approach to map the features to a 500 dimensional space.

3) Decoder with mean pooling feature. We extract features of all raw frames and then get only one holistic feature of the video after performing a mean pooling over these features extracted from raw frames. The Model in [13] inputs null pad as visual information after all frames are exhausted. In order to make full use of the holistic feature and establish more relationship between visual information and words, we take mean pooling feature as input at every time step during decoding stage.

4) Training and inference. Our model takes inspiration from the S2VT model in [13] which based on two-layer LSTMs. During encoding stage, the top LSTM layer inputs features from each frame while the bottom LSMT layer inputs the hidden representation h_t from the top layer and concatenates it with null pad. In this stage, there is no need to compute loss. When the encoding stage is finished, the bottom LSTM is fed with begin-of-sentence (<BOS>) tag. Then we start decoding for sequence generation. During decoding stage, the top LSTM layer inputs mean pooling feature of the video while the bottom LSMT layer inputs the hidden representation h_t from the top layer and concatenates it with the previous word. While training in the decoding stage, our model maximizes for the log-likelihood of the predicted output sentence given the hidden representation of the visual frame sequence and the previous words, formulated as in (8) where θ is model parameters and $Y(y_1,y_2,\cdots,y_m)$ is output sequence.

$$\theta^* = arg\,max_\theta \sum_{t=1}^{m} log\,p(y_t|h_{n+t-1},y_{t-1};\theta) \qquad (8)$$

In this stage, we employ stochastic gradient descent (SGD) to optimize this log-likelihood over the entire training dataset and the loss is propagated back in time. Then we take a softmax function to get the probability distribution over all words *y'* in the vocabulary *V* given the output of the bottom LSTM z_t at time *t*, as in (9):

$$p(y|z_t) = \frac{exp(W_y z_t)}{\sum_{y' \epsilon V} exp(W_{y'} z_t)} \qquad (9)$$

When second LSTM emits end-of-sentence (<EOS>) tag, the decoding stage is done. At train time, the previous word is ground truth during decoding stage. But at test time, the previous word is with the maximum probability via softmax function until it emits <EOS> token.

5) Fusion spatial feature with temporal feature. When models with raw frames and SIFT flow images have been trained, a shallow fusion technique is used to to integrate spatial and temporal features. At each time step of the decoding stage, the model computes probability of every candidate words. We then recomputed the probability of each new word, as in (10), where hyper-parameter α is tuned on the validation set.

$$\alpha \cdot p_{rgb}(y_t = y') + (1 - \alpha) \cdot p_{sift}(y_t = y') \qquad (10)$$

III. Experiments

A. DataSet

We evaluate our model on the Microsoft Video Description corpus [15] (MSVD), which is also known as the YouTube2Text dataset. MSVD is a video description dataset that contains 1970 short videos from YouTube. Each short video takes between 10s and 25s, describes a single behavior, and the dataset covers variable scenes. Each video corresponds to more than 100 text descriptions, including multiple languages. In our experiments, we take descriptions of English about 40 descriptions for each video. We pick 1200 videos for training, 100 videos for validation and 670 video for testing.

B. Evaluation Metrics

To evaluate the generated sentence, we use the METEOR [17] scores against all ground truth sentences. The METEOR score is computed based on the alignment between a give hypothesis sentence and a set of candidate reference sentences. Vedantam et al. [18] evaluated different metrics for image caption and results show that METEOR is always better that BLEU [19]. Therefore, we choose METEOR as the evaluation metrics. The performance is better with higher METEOR score. We employ the code released with Microsoft COCO Evaluation Server [20] for comprehensive comparison.

C. Results Analysis

1) Features during encoding stage. TABLE I shows the results of different features during encoding stage on the MSVD dataset. We found that taking SIFT flow images as input achieve better result than optical flow. Though our model which takes RGB images is not as good as result than S2VT with RGB images, fusing model with input of RGB images and SIFT flow images outperformed fusing model with input of RGB images and optical flow images. Results demonstrate that fusion of CNN feature and SIFT flow feature can improve the performance of video description.

TABLE I. Results of Features During Encoding Stage(%)

Model	*METEOR*
S2VT(RGB)[19]	29.2
Ours(RGB)	29.0
S2VT(optical flow)[19]	24.3
Ours(SIFT flow)	24.8
S2VT(RGB+optical flow)[19]	29.8
Ours(RGB+SIFT flow)	30.2

2) Features during decoding stage. TABLE II shows the results of mean pooling feature during decoding stage on the MSVD dataset. It justifies that adding mean pooling feature during decoding stage can improve the performance of video description.

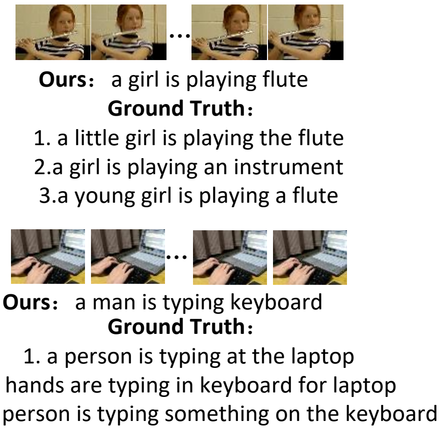
TABLE II. Results of Features During Decoding Stage(%)

Model	*METEOR*
Ours(RGB)	29.0
Ours(RGB+mean pooling)	29.4
Ours(SIFT flow)	24.8
Ours(SIFT flow+mean pooling)	27.6
Ours(RGB+SIFT flow)	30.2
Ours(RGB+SIFT flow+mean pooling)	30.6

3) Final Model. TABLE III shows the results of our final model compared with state-of-art works on the MSVD dataset. Our final model outperforms most of models. To show the efficiency of our visual-sentence translation, we give some visualized results of our final model as in Fig 3. and we find that generated sequences are very close to reference sequences.

TABLE III. RESULTS OF FINAL MODEL(%)

Model	*METEOR*
FGM[3]	23.9
Mean pool[4]	26.9
FCN+MIL[6]	29.0
3-D CNN+soft attention[8]	29.6
S2VT[13]	29.8
Ours(final)	**30.6**



Ours： a girl is playing flute
Ground Truth：
1. a little girl is playing the flute
2.a girl is playing an instrument
3.a young girl is playing a flute

Ours： a man is typing keyboard
Ground Truth：
1. a person is typing at the laptop
2. hands are typing in keyboard for laptop
3. the person is typing something on the keyboard

Fig. 3. Part of results of our final model against with ground truth

IV. CONCLUSION

In this paper, we propose a model for video description which fuses multiple features during encoding and decoding stage. The approach is evaluated on MSVD dataset for METEOR metric. The modification of original model can improve the capacity of video description. In the future, learning more text information underlying sentence from text corpora will be deserved to research.

ACKNOWLEDGMENT

This work was partially supported by National Natural Science Foundation of China (NSFC Grant No. 61170124, 61272258, 61301299, 61272005, 61572085), Provincial Natural Science Foundation of Jiangsu (Grant No. BK20151254, BK20151260), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (Grant No. 93K172016K08), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. Proceedings of the IEEE International Conference on Computer Vision, Sydney, pp. 2712-2719, December 2013.

[2] Krishnamoorthy N, Malkarnenkar G, Mooney R, Saenko K, Guadarranma S. Generating Natural-Language Video Descriptions Using Text-Mined Knowledge. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Washington, pp. 1-2, July 2013.

[3] Thomason J, Venugopalan S, Guadarrama S, Saenko K, Moony R. Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild. In the proceedings of the 25th International Conference on Computational Linguistics, Dublin, vol. 2, pp. 9, August 2014.

[4] Venugopalan S, Xu H, Donahue J, Rohrbach M, Mooney R, Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In: North American Chapter of the Association for Computational Linguistics – Human Language Technologies. Colorado, May 2015.

[5] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems vol. 4, pp. 3104-3112, 2014.

[6] Xu H, Venugopalan S, Ramanishka V, Rohrbach M, Saenko K. A Multi-scale Multiple Instance Video Description Network. Computer Science, vol. 6738, pp. 272-279, 2015.

[7] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence. vol. 10, pp. 1337-1342, 2014.

[8] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure. Proceedings of the IEEE International Conference on Computer Vision, Santiago, pp. 4507-4515, December 2015.

[9] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE International Conference on Computer Vision, Santiago, pp. 4489-4497. December 2015.

[10] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention. International Conference on Machine Learning. Lille, pp. 2048-2057, July 2015.

[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Proccedings of 3rd International Conference on Learning Representations. San Diego, May 2014.

[12] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. vol. 8, pp.1735-1780, 1997.

[13] Venugopalan S, Rohrbach M, Donahue J, Mooney R. Sequence to sequence-video to text. Proceedings of the IEEE International Conference on Computer Vision. Santiago, pp. 4534-4542, December 2015.

[14] Graves A. Supervised sequence labelling. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, pp. 5-13, 2012.

[15] D. L. Chen, W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, pp. 190-200, June 2011.

[16] Lowe D G. Distinctive image features from scale-invariant keypoint. International journal of computer vision, vol. 2, pp.91-110, 2004.

[17] Lavie, M D A. Meteor universal: Language specific translation evaluation for any target language. ACL, vol. 2014, pp.376, 2014.

[18] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, pp. 4566-4575, June 2015.

[19] Papineni K, Roukos S, Ward T, Zhu W J. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. Jeju Island, pp. 311-318, July 2012.

[20] Chen X, Fang H, Lin T Y, Vedantam R. Microsoft COCO captions: Data collection and evaluation server. CoRR. vol1504, pp. 325, 2015.