# Summarization-based Video Caption via Deep Neural Networks

Guang Li[*1], Shubo Ma[*2], Yahong Han[1,3]
[1]School of Computer Science and Technology, Tianjin University, Tianjin, China
[2]School of Computer Software, Tianjin University, Tianjin, China
[3]Tianjin Key Laboratory of Cognitive Computing & Application, Tianjin University, Tianjin, China
lguang@live.cn, shuboma@tju.edu.cn, yahong@tju.edu.cn

## ABSTRACT

Generating appropriate descriptions for visual content draws increasing attention recently, where the promising progresses were obtained owing to the breakthroughs in deep neural networks. Different from the traditional SVO (subject, verb, object) based methods, in this paper, we propose a novel framework of video caption via deep neural networks. For each frame, we extract visual features by a fine-tuned deep Convulutional Neural Networks (CNN), which are then fed into a Recurrent Neural Networks (RNN) to generate novel sentences descriptions for each frame. In order to obtain the most representative and high-quality descriptions for target video, a well-devised automatic summarization process is incorporated to reduce the noises by ranking on the sentence-sequence graph. Moreover, our framework owns the merit of describing out-of-sample videos by transferring knowledge from pre-captioned images. Experiments on the benchmark datasets demonstrate our method has better performance than the state-of-the-art methods of video caption in language generation metrics as well as SVO accuracy.

## Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing

## Keywords

Video Caption, Summarization, Deep Learning, CNN, RNN

## 1. INTRODUCTION

Visual content understanding is an important subject in multimedia analysis and computer vision. The majority of previous works in this area has focused on labeling images or videos with a fixed set of visual categories, such as automatic image annotation, object recognition and action recognition. With great progresses made in related fields, some pioneering methods [7] [13] have been developed to address the challenge of generating natural language descriptions to visual content. In particular, a well description can not only caption the objects contained in visual content, but also narrate how these objects relate to each other as well as their attributes and the activities they are involved in [19]. In most cases, a well-devised language model is utilized to automatically generate sentences to the target visual content [15]. The typical practice is as follows: Firstly, a fixed tuple of role words, such as subject, verb, object and scene, are detected from the visual content. Then a pre-defined template is used to generate grammatical sentences. Their varieties, however, are always limited due to the fixed or pre-defined visual concepts and sentence template.

Inspirited by recent advances in machine translation, the Recurrent Neural Network (RNN), which is proved to be a powerful sequence model, shows the state-of-the-art performance in visual caption [4] [6]. Moreover, a fixed-length feature extracted by deep Convulutional Neural Networks (CNN) is fed into the RNN as input instead of the fixed word tuple. The combination of CNN and RNN has been shown to be a promising method in image caption [4] [6]. In this translation based model, an image ($I$) is treated as the source language and translated to the target language ($S$) by maximizing the likelihood $p(S|I)$ directly. However, because of the various temporal length of videos, it is difficult to obtain fix length CNN features for videos. Thus, the above translation based model can not be directly applied in video caption. As an alternative method, LSTM-YT [18] introduces a mean pooling operation to calculate a single mean vector of the extracted CNN features for video frames. Although LSTM-YT gains good performance, the mean vector of the CNN features may lose some key information in original extracted CNN features. Thus, the performance of LSTM-YT in video caption is restrained in situations where the scene changes fast and dramatically.

As each video clip can be taken as a spatial-temporal neighborhood, the characteristics of temporal consistency in videos may result in semantic consistency in the visual content of the sequential video frames. In this paper, we propose a novel framework of summarization-based video caption via deep neural networks. The proposed framework is illustrated in Fig. 1. Instead of trying to generate a fixed-length feature of video, our framework treats the target video clip as a a sequence of frames. We "translate" the frame sequence to the sentence sequence by the deep CNN and RNN. As is shown in Fig. 1, we develop a summarization process to generate the final description to the target video clip. In particular, we first construct an adjacency graph on the sentence sequence, in which each node denotes a sentence and the pairwise similarity is associated with each edge as the weight. After pruning the edges with small weights, we calculate the transitivity on the pruned graph until convergency. Finally, we output the top ranked sentence as the final video caption. Comparison results on the benchmark datasets demonstrate the better performance of our method.

[*]Equal contribution


*MM'15,* October 26-30, 2015, Brisbane, Australia

DOI: http://dx.doi.org/10.1145/2733373.2806314.

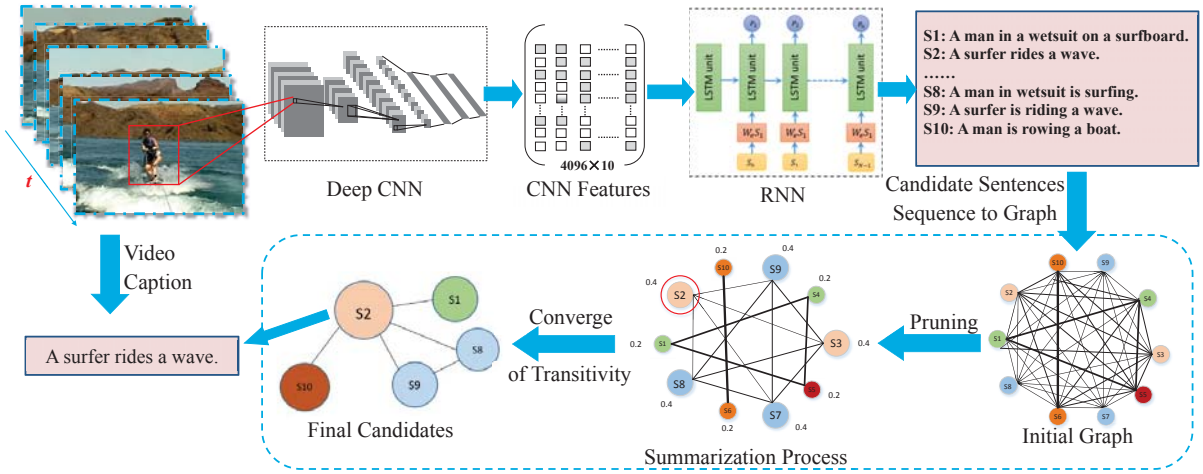


**Figure 1: Flowchart of the proposed framework.**

## 2. THE PROPOSED FRAMEWORK

Fig. 1 depicts our two-stage framework using neural and probabilistic approach and text analysis technique to generate descriptions from videos. For each video, we extract a number of frames by uniform sampling to form the sequence of representative video frames. In the first stage, a translation based natural language description generation process is constructed by deep CNN and RNN. Then, we develop an automatic summarization process to output the top ranked sentence as the final video caption.

### 2.1 Translation from Videos Frames to Sentences Sequence

Inspired by the principle of "translating" images [9], we maximize the log likelihood of the sentence $S$ given the corresponding frame $F$ and the model parameters $\theta$,

$$\theta^* = \arg\max_{\theta} \sum_{(F,S)} \log p(S|F;\theta), \tag{1}$$

where $S$ represents a sequence of words whose length is unbounded. By using the chain rule to model the joint probability over $S_{W_1}, ..., S_{W_n}$, Formula 1 can be expressed as:

$$p(S|F) = \sum_{t=0}^{N} \log p(S_{W_t}|F, S_{W_1}, ..., S_{W_{t-1}}), \tag{2}$$

where $N$ is the total number of words in sentence $S$ and $S_{W_t}$ represents the $i-th$ word in sentence $S$. And for the sake of convenience we have dropped $\theta$.

In order to get the conditional probabilities, RNN, parameterized by $\theta$, is selected to model $p(S_{W_t}|F, S_{W_1}, ..., S_{W_{t-1}})$. The RNN expresses the previously seen words as a hidden state or memory $h_{t-1}$. After seen a new input $x_t$, the memory $h_{t-1}$ is updated via a non-linear function $f$: $h_t = f_\theta(x_t, h_{t-1})$.

In our framework, we use a Long-Short Term Memory(LSTM) net [9], since it has shown state-of-the-art performance on sequence tasks such as machine translation and image captions [6][12][19].

For the encoding of images, we use a well-trained deep CNN model, which is pre-trained on ImageNet [3] and fine-tuned on the 200 classes of the ImageNet Detection Challenge [16]. In particular, we use Caffe [11] to extract CNN features.

### 2.2 LSTM-based Language Model

The basic unit of the LSTM is a memory cell $c$ encoding knowledge at every time step of what inputs have been observed up to this step (see Fig. 2). The cell is controlled by three sigmoid functions, the first one managing whether to consider the current input $x_t$ is named as input gate $i$, the second one allowing to forget its previous memory $c_{t-1}$ is named as forget gate $f$, and third one deciding how much of the memory will transfer to next hidden state $h_t$ is called output gate $o$. Such multiplicative gates makes it possible to train the LSTM robustly as these gates deal well with exploding and vanishing gradients. The detail of LSTM unit is described as follows:

$$i_t = \sigma\left(W_{ix}x_t + W_{im}m_{t-1}\right) \tag{3}$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fm}m_{t-1}\right) \tag{4}$$

$$o_t = \sigma\left(W_{ox}x_t + W_{om}m_{t-1}\right) \tag{5}$$

$$c_t = f_t \odot C_{t-1} + i_t \odot h\left(W_{cx}x_t + W_{cm}m_{t-1}\right) \tag{6}$$

$$m_t = o_t \odot c_t, \quad p_t = Softmax\left(m_t\right) \tag{7}$$

where the weight matrices $W$ are trained parameters, the non-linear functions ($\sigma$ and $h$) are sigmoid and hyperbolic tangent respectively, and the $\odot$ represents the and operation with a gate value. The last equation will produce a probability distribution $p_t$ over all words through a softmax function.

### 2.3 Summarization of Sentence Sequence

Inspired by the ranking based summarization method LexRank [5], we propose a summarization process to reduce outliers in sentences sequence that share little similarity with the others. The LexRank score is defined as follows:

$$LR(u) = (1-d)\sum_{v \in adj[u]} \frac{LR\left(v\right)}{deg\left(v\right)} + \frac{d}{N}, \tag{8}$$

where $LR(u)$ is the LexRank score of node $u$, $adj[u]$ is the set of nodes that are adjacent to $u$, and $deg(v)$ is the degree of the node $v$, $N$ is the total number of nodes in the graph, and $d$ is a "damping factor".

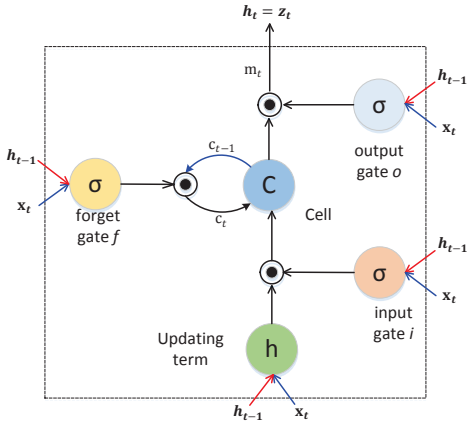


**Figure 2: LSTM unit:the memory block contains a cell $c$ controlled by three gates. The lines in blue show the recurrent connections - the output $m$ at time $t-1$ is fed back to the memory at time $t$ via the three gates.**

As shown in Fig. 1, a graph is generated from these sentences with each sentence as a vertex. The similarity between the sentences are represented as edges, and we delete the edge between the sentences whose similarity is too small. The similarity is defined by the cosine between two corresponding sentences. Here we set the initial LexRank score as $\frac{d}{N}$, and then each vertex transfer their LexRank score to their adjacent vertex according to the similarity of each other. According to the convergence property of Markov chains, the algorithm is guaranteed to get an stationary distribution.

## 2.4 Transfer Caption from Images to Videos

In our framework, we treat the video as a sequence of frames, so it is natural to extend the training data by using some captioned images as auxiliary for the lack of video descriptions. In our experiments, the $fc_7$ layer features of video frames is extracted from the CNN, as well as captioned auxiliary images. The initial model is trained roughly with a relatively high learn rate (about $10^{-3}$) on the mixed features of frames and images, then we fine-tune it on video frames with lower learn rate.

# 3. EXPERIMENTS

## 3.1 Datasets

We conduct our experiments on Microsoft Research Video Description Corpus [2], which consists of 1970 video clips collected from YouTube. Thus, we name the dataset by "YouTube" in the following. Each clip particularly contains a single action and lasts from 10 to 25 seconds. Each snippet has been aligned with a number of descriptions and we use roughly 40 available English descriptions per video. In our experiments, we follow the settings in [8] [17] to use 1200 videos as training data and 100 videos for validation and 670 videos for testing. For each video, we extract 10 frames by uniform sampling and select 5 sentences randomly for each frame.

In order to evaluate the performance of domain adaptation from images to video caption, we use a benchmark dataset Flickr8k [10] as the auxiliary training data, which is a subset of Flickr30k that consists of 8000 images. Each image is captioned with 5 sentences. We use 1000 images for validation and the rest images are combined with the video training set.

**Table 1: Comparison results in terms of BLEU at 4 (combined n-gram 1-4) and METEOR. All values are reported as percentage(%).**

| | BLEU | METEOR |
|---|---|---|
| FGM-YT | 13.68 | 23.9 |
| LSTM-YT | 31.19 | 26.87 |
| OUR-YT | 32.32 | 27.03 |
| LSTM-YT$_{flickr}$ | 32.03 | 27.87 |
| OUR-YT$_{flickr}$ | **35.09** | **29.26** |

**Table 2: Comparison results in terms of SVO accuracy. All values are reported as percentage(%).**

| | S | V | O |
|---|---|---|---|
| HVC-YT | **76.57** | 22.24 | 11.94 |
| FGM-YT | 76.42 | 21.34 | 12.39 |
| LSTM-YT | 71.19 | 19.4 | 9.7 |
| OUR-YT | 74.48 | 21.34 | 9.25 |
| LSTM-YT$_{flickr}$ | 75.37 | 21.94 | 10.74 |
| OUR-YT$_{flickr}$ | 74.33 | **23.13** | **13.43** |

## 3.2 Comparison Methods

We compare our framework (denote by OUR) with two state-of-the-art SVO-based methods HVC [17] and FGM [17]. In order to evaluate the effectiveness of the summarization process, we also compare with LSTM [18], which shows the best performance of video caption by the combination of deep CNN and RNN.

In the following, we let "-YT" denote the model is trained on the YouTube, for example OUR-YT and LSTM-YT, and "-YT$_{flickr}$" denote the model trained on YouTube and Flickr8k datasets, which means a transfer process of video caption from pre-captioned images, e.g., OUR-YT$_{flickr}$ and LSTM-YT$_{flickr}$.

## 3.3 Experimental Results

### 3.3.1 *Evaluation Metric*

To evaluate the generated sentences after summarization, we use the BLEU [14] and METEOR [1] as the criteria. As HVC [17] and FGM [17] used the SVO (subject, verb, object) accuracy, we also evaluate the results by the most frequent SVO accuracy.

### 3.3.2 *Comparison Results*

Firstly, we report the comparison results in Table 1 and 2. From the results we can see: (1) Comparing the results only trained on YouTube, our method obtains the best results in terms of BLEU and METEOR, see Table 1. (2) Comparing OUR-YT$_{flickr}$ with LSTM-YT$_{flickr}$, our method obtains better performance of transfer video caption. Moreover, by transferring knowledge from pre-captioned images, OUR-YT$_{flickr}$ obtains the best scores of BLEU and METEOR as well as the V and O scores of SVO, which further demonstrate the better performance of our method.

### 3.3.3 *Example Results of Transfer Caption*

In Fig. 3 we show some examples of video caption results. Four frames are extracted from a video clip. Note that there are some occlusions (dark noises) on the forth frame and thus the CNN-RNN generated sentence is not as accurate as the one obtained from a transfer caption process (see the sentence in green text). Furthermore, as there are some noises in the video frames, the output final video captions from FGM and LSTM are shown to be influenced by the noises, e.g., "A person playing the goal of the road" is not

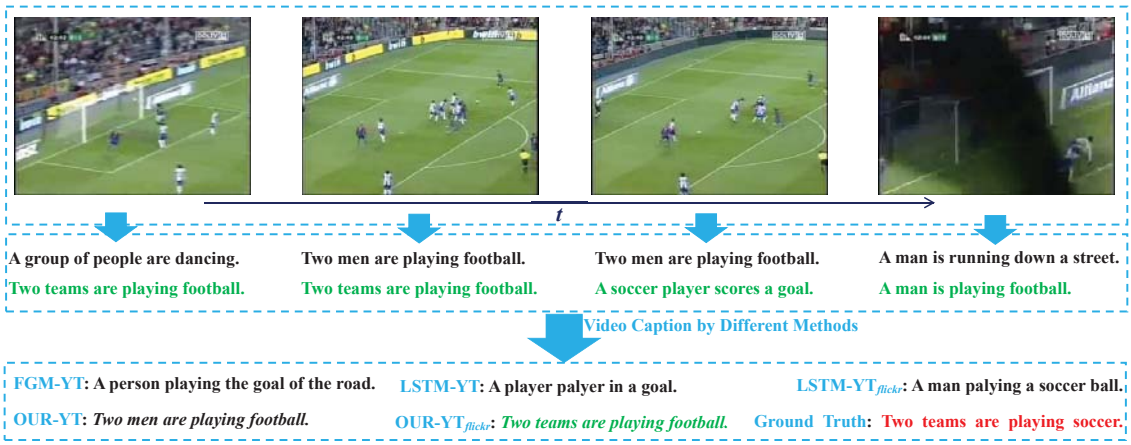


**Figure 3: Example of video caption results. Sentences in dark text under each frame are generated by our CNN-RNN framework, whereas the sentences in green are obtained by the help of transferring knowledge from pre-captioned images in Flickr8K.**

a correct sentence or some grammatical errors in "A player palyer in a goal". As our framework owns the merit of summarization and transferring caption, we generate sentences in better quality and are very close to the ground truth. Thus, these results further demonstrate the effectiveness of the proposed framework.

## 4. CONCLUSIONS

In this paper, we have proposed a framework of summarization-based video caption via deep neural networks. We translate the frame sequence to the sentence sequence by the deep CNN and RNN. To obtain the representative descriptions for video clip, we develop a summarization process to generate the final description to the target video clip. As shown in experimental results, our framework can generate better sentences than the state-of-the-art approaches, especially in noisy data. Moreover, the exploiting of pre-captioned images as auxiliary domain can help improve the performance of video caption. Future work includes utilizing the motion cue from videos to improve the prediction accuracy.

## 5. ACKNOWLEDGMENTS

This work was supported by NSFC (Grant 61202166, 61472276).

## 6. REFERENCES

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 190–200. Association for Computational Linguistics, 2011.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[5] G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479, 2004.

[6] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. *arXiv preprint arXiv:1411.4952*, 2014.

[7] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29. 2010.

[8] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, pages 2712–2719, 2013.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.

[13] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th annual meeting on association for computational linguistics*, pages 311–318, 2002.

[15] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, pages 433–440, 2013.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[17] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, 2014.

[18] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.