Describing Videos using Multi-modal Fusion

Qin Jin[†], Jia Chen[♮], Shizhe Chen[†], Yifan Xiong[†], Alexander Hauptmann[♮]

[†]School of Information, Renmin University of China, China
{qjin, cszhe1, xiongyf}@ruc.edu.cn
[♮]Language Technologies Institute, Carnegie Mellon University, USA
{jiac, alex}@cs.cmu.edu

ABSTRACT

Describing videos with natural language is one of the ultimate goals of video understanding. Video records multi-modal information including image, motion, aural, speech and so on. MSR Video to Language Challenge provides a good chance to study multi-modality fusion in caption task. In this paper, we propose the multi-modal fusion encoder and integrate it with text sequence decoder into an end-to-end video caption framework. Features from visual, aural, speech and meta modalities are fused together to represent the video contents. Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) are then used as the decoder to generate natural language sentences. Experimental results show the effectiveness of multi-modal fusion encoder trained in the end-to-end framework, which achieved top performance in both common metrics evaluation and human evaluation.

Keywords

video description generation; multi-modal fusion; end-to-end framework

1. INTRODUCTION

It is an intriguing challenge to automatically describe videos containing complex and diverse contents with natural language sentences. It has a wide range of applications such as assisting blind people or improving search quality for online videos. Inspired by the recent success of image captioning [1, 2], where natural language sentences are generated to describe image content, researchers have been paying more attention to the generation of video captions.

Different from image caption, generating video descriptions encounters two additional challenges: Firstly, video contains temporal information. Semantic concepts involved in the video may evolve over time. Secondly, video consists of multi-modalities. Besides visual information, videos also contain other contents such as aural and speech modalities, which provide additional information. The diversity of groundtruth captions also reflects information from multiple modalities.

Many research efforts have been made to address the first aforementioned challenge in video description generation. Venugopalan et al. [3] use the LSTM to encode the frames in the video to a fixed length vector and learn the encoder in the end-to-end framework. Yao et al. [4] exploit the local temporal structure underlying the video. They use the spatio-temporal convolution neural network (3-D CNN) as action features to encode local temporal structure and temporal attention mechanism based on soft-alignment method to exploit global temporal structure. Pan et al. [5] use a hierarchical recurrent encoder to encode frames in the video to a fixed length and learn the encoder in the end-to-end framework. There are relatively few works focusing on the second challenge of the multi-modality issue. Jin et al. [6] investigate to combine acoustic and visual information at the video representation level to generate video descriptions and achieve significant improvement over visual-only baseline.

As the decoder of the video caption generation, LSTM is widely used in previous works [3, 4, 5, 6, 7]. Pan et al. [7] improve the optimization goal for the decoder. The coherence loss is to locally maximize the next word given previous words and visual content and the relevance loss aims to enforce the relationship between the semantic of the sentence and visual content by creating a visual-semantic embedding space. They joint optimize the two losses in a unified model. Yu et al. [8] exploit a hierarchical-RNN framework including a sentence generator and a paragraph generator as the decoder. The framework models inter-sentence dependency which enables it to generate a paragraph for a long video.

In this paper, we tackle the video description generation problem in the context of MSR Video to Language Challenge. We mainly focus on utilizing multi-modal features from visual, audio, speech and meta-data information to improve the description performance. We propose a multi-modal fusion encoder and combine it with the text sequence decoder in an end-to-end framework. We address the following two questions in the challenge:

1. How much performance gain can additional modalities bring to visual-only systems?
2. What specific categories can benefit from different modalities?

We examine these questions empirically by evaluating our framework on different feature combinations. Experimental results show that combining multi-modal cues can significantly improve the description performance and generate more semantically accurate and comprehensive sentences.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2984065

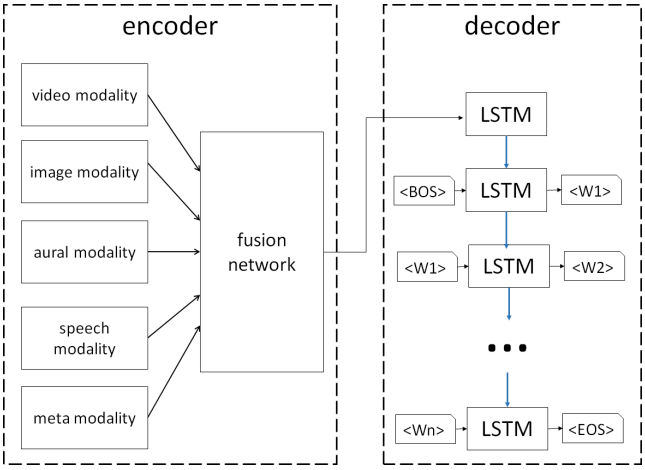


Figure 1: The key components of our system

Our system is the winner in the common metrics evaluation and ranks the second place in the human evaluation.

The rest of the paper is organized as follows: Section 2 describes our solutions for the video to text challenge. Section 3 presents the experiments on the challenge dataset. Section 4 draws the conclusions and future work.

2. PROPOSED SOLUTION

Our solution is based on the typical encoder-decoder framework of video caption task. We design a multi-modal fusion encoder and use the vanilla RNN decoder in our solution. The multi-modal fusion encoder encodes multi-modal features to a fixed length output vector. The initial hidden state of the vanilla RNN decoder is set to be the output of multi-modal fusion encoder and the initial word is set to be <BOS>. The dimension of multi-modal fusion encoder output is aligned to the dimension of decoder hidden state. An illustration of our proposed system is shown in Figure 1.

2.1 Multi-modal Features

We list all the modalities that are used in our solution, including image, video, aural, speech and meta-data. Different modalities cover different information in the video, corresponding to different parts in the video caption.

Image modality feature: The image modality reflects the static content of videos. DCNN is the state-of-the-art model in many visual tasks such as object detection, scene recognition etc. In this task, we use the fc7 layer of VGG-19 [9] pre-trained on ImageNet to extract image features every 16 frames. We perform mean pooling to get the final video-level feature.

Video modality feature: The video modality captures the temporal and motion content. We extract the fc6 layer of C3D [10] and Improved Dense Trajectory (iDT) [11]. Both of them are the state-of-the-art video features. The C3D model is pre-trained on the Sports-1M dataset. We segment the video to non-overlap shots of 16 frames length. The C3D feature is extracted on each shot and we apply mean pooling to aggregate them into one vector for a video. iDT feature is aggregated by fisher vector and we use kernel PCA to reduce dimensionality.

Aural modality feature: Aural modality is complementary to visual modalities, which is especially useful to distinguish different scene events. We extract the Mel-frequency Cepstral Coefficients (MFCCs) [12] over short-time window of 25ms with 10ms shift. We use two encoding strategies to aggregate MFCC frames into one video-level feature vector: Bag-of-Audio-Words (BoAW) [13] and Fisher Vector (FV) [14].

Speech modality feature: Speech modality provides semantic topics and details for what happened in the video. We use the IBM Watson API to extract the transcription of each video. To be specific, the speech modality feature is composed of two parts. The first part is the length of speech content. The length of speech content is usually long in a video about people talking while the length of speech content is usually short in a video about people exercising. We quantize the length of transcriptions into several bins and represent it as the one-hot vector. Another feature is related to the semantic meaning of the content. For example, "news" is a typical word in the outdoor reporters' videos and "car" often appears when people introduce their cars. We use word2vec pretrained on Google News Dataset[15] to generate a codebook with 100 codewords by clustering all the word vectors from training transcriptions. Each cluster represents a topic. Finally we form a bag-of-topics feature representation based on the codebook for each video.

Meta modality feature: Meta-data such as category provides a strong prior information about video content. It is also useful for dynamically weighting other modalities. For example, speech modality plays an important role in the howto category while video modality plays an important role in the sports category. For each video, we encode the category label in a one-hot feature vector.

2.2 Multimodal Fusion Encoder and Text Sequence Decoder

The multi-modal feature encoder is formed as a multi-layer feed-forward network. In the competition we only explore the network structure of one full connect layer without activation function. Other network structures are worth exploring and may bring additional performance boost.

For text sequence decoder, we use the vanilla LSTM [16] which is able to learn complex long-range dependences. The memory cell of LSTM encodes the history information up to the current time step and update with the current input under the control of three gates. The recurrences for the LSTM are defined as:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1}) \quad (1)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1}) \quad (2)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1}) \quad (3)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1}) \quad (4)$$
$$h_t = o_t \odot c_t \quad (5)$$

where i, f, o, c, W represent the input gate, forget gate, output gate, memory cell and weight matrix respectively. σ and ϕ is the sigmoid and hyperbolic tangent function. $\odot$ is the element-wise product.

We learn the encoder and decoder jointly in the end-to-end framework. In training, the groundtruth word is used as the input of each time step for text sequence decoder. In testing, the estimated output word is feed to the next step input for text sequence decoder.

Table 1: Performance of Multi-Modality Fusion

modality	model	BLEU@4	METEOR	ROUGE	CIDEr
video	`c3d`	36.94	27.27	58.40	41.85
	`c3d+iDT`	34.96	26.59	57.58	36.51
video+aural	`c3d+mfccbow`	39.80	27.89	60.02	41.07
	`c3d+mfccfv`	40.07	27.76	60.02	39.60
	`c3d+mfccbow+mfccfv`	41.32	28.21	60.45	43.66
video+aural+image	`c3d+mfccbow+mfccfv+vgg19`	41.81	28.67	60.41	43.35
video+aural+speech	`c3d+mfccbow+mfccfv+asr`	40.55	27.98	60.10	42.28
video+aural+meta	`c3d+mfccbow+mfccfv+category`	43.70	28.95	61.35	45.74

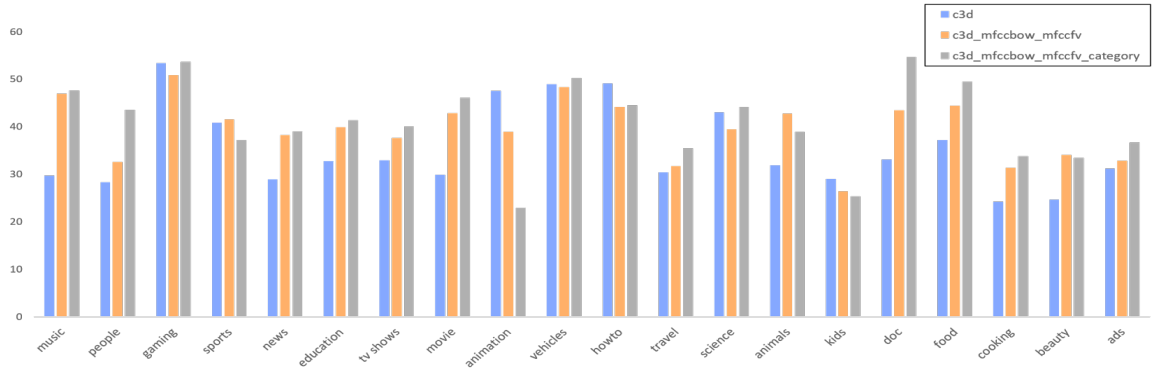


Figure 2: BLEU@4 Performance Comparison among Multi-modal Fusion

3. EXPERIMENTS

3.1 Experimental Setup

The challenge dataset is the MSR-VTT corpus [17], which is composed of 10,000 clips from 20 categories. Each video clip is annotated with 20 sentences by different workers in Amazon Mechanical Turk. In the challenge, we are provided with 6513 training videos and 497 videos for tuning parameters. Following standard data split, we randomly select 200 videos out of the 6513 videos for local validation and report experiment results on the 497 videos as local test set. That is, we use 6313 videos for training, 200 videos for local validation and 497 videos for local test. We evaluate the results comprehensively on all major metrics, including BLEU [18], METEOR [19], ROUGE-L [20] and CIDEr [21].

3.2 Implementation Details

We preserve words which appear more than three times, resulting in a vocabulary size of 10,866. We add begin-of-sentence tag <BOS> and end-of-sentence tag <EOS> to our vocabulary. The max generated caption length is set to 30. For videos without soundtrack or speech transcripts, we simply pad zeros as the audio features or speech features. The dimensionality of the fusion network output, word vector and the size of hidden layer of LSTM are all set to 512 empirically. We apply dropout with rate of 0.5 on the input and output of LSTM to prevent overfitting and use ADAM algorithm to minimize the log-likelihood loss with learning rate of 1×10^{-4}. Beam search with beam width of 5 is used to generate sentences during test process.

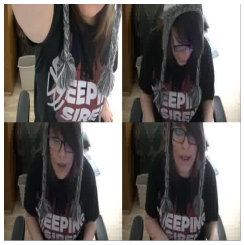
3.3 Overall Evaluation of Multi-modal Fusion

To study multi-modality fusion, we start with video modality feature `c3d`, which achieves a decent performance as shown in Table 1. Combining state-of-the-art video feature `iDT` doesn't improve the performance. Then we add aural modality feature `mfccbow` and `mfccfv`. Adding `mfccbow` alone improves the performance by 2.94% on BLEU@4 and 0.62% on METEOR. Adding `mfccfv` alone also improves the performance by 3.07% on BLEU4 and 0.49% on METEOR. Adding `mfccbow` and `mfccbow` together further improves the performance on all metrics consistently. It indicates that the two aural modality features, `mfccbow` and `mfccfv` are complementary.

With video and aural modality fusion `c3d+mfccfv+mfccbow`, we reach the performance plateau. It is very difficult to improve the performance by simply adding more modality features such as image modality feature `vgg19` or speech modality feature `asr` as shown in Table 1. However, adding meta modality feature `category` overcomes the performance plateau by a significant improvement on all metrics, 2.68% on BLEU@4, 0.74% on METEOR, 0.9% on ROUGE, 2.08% on CIDEr. There are two potential reasons for such significant boost. First, the model learns to use `category` to limit the sentence space in generation. Second, the model learns to dynamically weight multi-modal features with `category` as different modalities excel at different categories on caption task. It is intriguing to distinguish between these two potential reasons and we will investigate it in our future work.

3.4 Detailed Study of Multi-modal Fusion

We further conduct category-wise evaluation on multi-modal fusion performance. As shown in Figure 2, adding aural modality significantly improves the performance in categories such as music and movie, but deteriorates the per-



(1) A woman is talking to a camera.
(2) **A woman is singing a song.**
GT: A girl is sitting at a piano playing and singing.

(a)

(1) A man is talking to a woman.
(2) There is a suit man is talking with a man.
GT: A man in a white shirt and dark suit jacket is talking about merging living and retail space.

(b)

Figure 3: Examples of sentence generation results from different modalities and groundtruth. (1) c3d, (2) c3d+mfccbow+mfccfv. GT is a randomly selected groundtruth human description. Sentences in bold highlight the best generated sentences.

Table 2: Performance on Challenge Test Set

	common metrics evaluation				human evaluation		
run	BLEU@4	METEOR	ROUGE	CIDEr	C1	C2	C3
run1(p)	0.408	0.282	0.448	0.609	3.261	3.091	3.154
run2	0.426	0.288	0.467	0.617	-	-	-

formance in animation. For some categories like animation, video modality performs best. It seems that audio modality brings little information to these categories. Furthermore, adding meta modality `category` consistently improves the performance on most categories.

We present some examples of generated sentences from `c3d` model and `c3d+mfccfv+mfccbow` model in Figure 3. In Figure 3(a), audio modality helps to distinguish talking and singing successfully especially when the behaviour is not obvious from visual information. Figure 3(b) shows that acoustic information helps to identify the gender of the speaker.

3.5 Performance on Test Set

The performance on the challenge test set is presented in Table 2. For human evaluation, the subset of testing sentences generated from the primary run is scored according to C1 (coherence), C2 (relevance) and C3 (helpful for blind) on a scale of 1-5. The generated sentences in our primary run are selected from the `c3d`, `c3d+mfccbow+mfccfv`, `c3d+mfccbow+mfccfv+asr`, `c3d+mfccbow+mfccfv+category` systems by category according to their categories performance on the validation set. The results in the second run are simply generated from the `c3d+mfccbow+mfccfv+category` system.

The primary run might overfit the validation set so do not generalize well on the test set. The multi-modal fusion system alone achieves the best performance in the challenge.

4. CONCLUSION AND FUTURE WORK

This paper proposes a multi-modal fusion encoder to fuse image, video, aural, speech and meta modalities. We integrate the multi-modal fusion encoder with the text sequence decoder in an end-to-end framerwork to generate video descriptions. Adding audio and meta modalities can significantly improve the overall performance of the visual modality description system. In the future, we will explore different structure of the fusion network to improve the performance. We will also explore adding attention to expose each modality feature directly to the text sequence decoder for dynamic fusion in each prediction step.

5. ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001202.

6. REFERENCES

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057, 2015.

[3] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence - video to text. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4534–4542, 2015.

[4] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4507–4515, 2015.

[5] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. *CoRR*, abs/1511.03476, 2015.

[6] Qin Jin and Junwei Liang. Video description generation using audio and visual cues. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 239–242. ACM, 2016.

[7] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and

Yong Rui. Jointly modeling embedding and translation to bridge video and language. *CoRR*, abs/1505.01861, 2015.

[8] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. *CoRR*, abs/1510.07712, 2015.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: generic features for video analysis. *CoRR, abs/1412.0767*, 2:7, 2014.

[11] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[12] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[13] Stephanie Pancoast and Murat Akbacak. Softening quantization in bag-of-audio-words. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1370–1374. IEEE, 2014.

[14] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

[17] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[19] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *In Proceedings of the Ninth Workshop on Statistical Machine Translation*. Citeseer, 2014.

[20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

[21] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.