INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Generating Natural Video Descriptions via Multimodal Processing

Qin Jin [1,2], *Junwei Liang*[3], *Xiaozhu Lin*[1]

[1] Multimedia Computing Lab, School of Information, Renmin University of China
[2] Key Lab of Data Engineering and Knowledge Engineering of Ministry of Education, Renmin University of China, Beijing 100872
[3] Language Technologies Institute, Carnegie Mellon University, USA

Abstract

Generating natural language descriptions of visual content is an intriguing task which has wide applications such as assisting blind people. The recent advances in image captioning stimulate further study of this task in more depth including generating natural descriptions for videos. Most works of video description generation focus on visual information in the video. However, audio provides rich information for describing video contents as well. In this paper, we propose to generate video descriptions in natural sentences via multimodal processing, which refers to using both audio and visual cues via unified deep neural networks with both convolutional and recurrent structure. Experimental results on the Microsoft Research Video Description (MSVD) corpus prove that fusing audio information greatly improves the video description performance. We also investigate the impact of image amount vs caption amount on the image caption performance and see the trend that when limited amount of training is available, number of various captions is more important than number of various images. This will guide us to investigate in the future how to improve the video description system via increasing amount of training data.

Index Terms: Video Description, Multimodal Processing, Deep Neural Networks

1. Introduction

Describing visual content automatically in natural language sentences is an intriguing and challenging task. It has attracted a lot of research interest lately. With the recent success in describing images with a natural sentence [1-3], generating natural descriptions for videos has also attracted more and more attention in the research community. The task of automatically generating descriptions for videos is a very complex problem. Although there have been successful examples in specific domains with a limited set of known actions and objects [4-5], generating descriptions for open-domain videos or videos "in-the-wild" such as YouTube videos remains an open challenge.

Many of the recent works for video content description use Long-Short Term Memory Recurrent Neural Networks (LSTM-RNNs) [6] based on the visual information only. Visual information in videos has been captured by holistic video representations [7-8], or pooling over frames [9], or sub-sampling on a fixed number of input frames [10]. However, human description of video contents may rely not only on the visual information but also on other content-related information such as audio content which is not directly present in the visual source. In this paper, we study improving video description via multimodal processing which uses both audio and visual information in videos. Our approach follows the recent progress in image caption such as in [1]. We also utilize a LSTM-RNN to model sequence dynamics and connect it directly to a convolutional neural network (CNN) and an acoustic feature extraction module which process incoming video frames for visual and acoustic encoding.

The rest of the paper is organized as follows: section 2 summarizes related works. Section 3 describes the key components of our video description system using acoustic and visual information. Section 4 presents the experiments and case studies on the MSVD corpus and analysis of the impact of image vs. caption on performance when limited training data is available. Section 5 concludes the paper.

2. Related Work

Generating natural language description for images, the image caption task, has received a lot of attention and achieved some exciting results recently [1-3, 11-14]. Most of the work rely on two networks: CNN and RNN in particular with LSTM. CNN is used to provide image encoding and LSTM-RNN is used to translate from images to sentences of flexible length. Some public datasets have been accumulated in the community such as the Flickr30k corpus [15] and the Microsoft COCO (MSCOCO) corpus [16]. There are also studies to emphasize the novelty of generated descriptions [17].

With the success in image caption, video description task has attracted more and more interest lately [4-5, 7-10, 18]. Most of the works study the task of describing short video clips with a single sentence. Similar to image caption methods, they also rely on CNNs and LSTM-RNNs for video description. It has been shown that pre-training the LSTM-RNN network for image captioning and fine-tuning it to video description is beneficial [9]. Some work [19] also builds a 2-D and/or 3-D CNN for learning powerful video representation and the LSTM-RNN network for generating sentences and a joint embedding model for exploring the relationships between visual content and sentence semantics.

Most previous researches targeting the problem of generating natural language descriptions for videos rely on visual content only. However, acoustic information also plays an important role in explaining and understanding an event/action in videos. In semantic concept annotation and classification of videos, fusion of audio and visual cues has been demonstrated very helpful, such as in the TRECTVID Multimedia Event Detection (MED), Multimedia Event Recounting (MER), Surveillance Event Detection (SED), and Semantic Indexing (SIN) Systems [20]. In this paper, we conduct a pilot study on utilizing both acoustic and visual

http://dx.doi.org/10.21437/Interspeech.2016-380

information to improve the visual only video description baseline.

3. System Description

Our video description system relies on deep models such as CNN and LSTM-RNN, which is similar to what was proposed by Vinyals et al. in [1]. An illustration of our description system is shown in Figure 1. The visual-only description system shares the similar system structure with the audio-only system. The difference lies in the feature encoding component. The visual-only system uses CNN for feature encoding while the audio-only uses bag-of-acoustic-words for feature encoding. There are two phases in the system execution: training phase and testing phase. In the training phase, the LSTM-RNN model is trained using target domain training data or is pre-trained using related auxiliary data and fine-tuned on the target domain data. In the testing phase, the trained LSTM-RNN is applied for sentence prediction. There are three key components: visual/acoustic encoding, text encoding and text decoding.

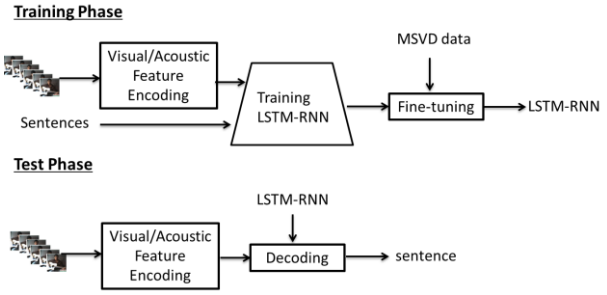


Figure 1: *Illustration of video description system*

3.1. Visual and Acoustic Encoding

A CNN is applied for visual encoding. We use the pre-trained VGGNet [21] for visual feature extraction. A feature vector is extracted for each frame of the video and mean pooling is applied to produce the video-level visual encoding.

For acoustic encoding, we first extract the single channel soundtrack from the video and re-sample it to 8kHz. We then apply feature extraction from the soundtrack. We use the Mel-frequency Cepstral Coefficients (MFCCs) [22] as our fundamental feature. The Fast Fourier Transformation (FFT) [23] is first applied over short-time window of 25ms with a 10ms shift. The spectrum of each window is warped to the Mel frequency scale, and the discrete cosine transform (DCT) [24] was applied over the log of these auditory spectra to compute MFCCs. Each video is then represented by a set of 39-dimensional MFCC feature vectors (13-dimensional MFCC + delta + delta delta). Finally, a bag-of-audio-words type of feature representation [25] is generated by applying an acoustic codebook to transform this set of MFCCs into a single fixed-dimension (4096) video-level feature encoding The 4096 acoustic codewords were trained using Kmeans clustering on the audio data that we collected from freesound.org as in [26].

3.2. LSTM-RNN for Text Encoding and Decoding

Standard RNNs learn to map a sequence of inputs $(X_1,...,X_N)$ to a sequence of outputs $(Z_1,...,Z_N)$ via a sequence of hidden states $(h_1,...,h_N)$. The memory cell in LSTM model encodes, at every time step, the knowledge of the inputs that have been observed up to that step. The cell is modulated by gates that are all sigmoidal. The gates decide whether the LSTM keeps or discards the value from them. The recurrences for the LSTM are defined as:

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1}) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1}\right) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1}) \\
h_t &= o_t \odot \phi(c_t)
\end{aligned}
\tag{1}
$$

where i, f, o, c, W represent the input gate, forget gate, output gate, memory cell and weight matrix respectively. σ is the sigmoidal non-linearity, ϕ is the tangent non-linearity, and $\odot$ is the product with the gate value.

After we extract the visual/acoustic features, we train and/or fine-tune a LSTM-RNN network as illustrated in Figure 1. We employ LSTM-RNN to encode the sentence description of the video, and decode a visual/acoustic feature encoding of fixed length to generate natural language output. The encoding LSTM-RNN and decoding LSTM-RNN are shared.

4. Experiments

This section presents our video description experiments on the MSVD corpus and analysis of the impact of image amount vs caption amount on the image caption performance.

4.1. Data Description

We conduct our experiments on the Microsoft Research Video Description (MSVD) Corpus [27]. The MSVD corpus contains 1970 YouTube clips with duration between 10 seconds to 25 seconds, mostly depicting a single activity. Each video was then used to elicit short sentence descriptions from annotators. There are multi-lingual human-generated descriptions for each video in the corpus. We only use the English descriptions which amount to about 40 sentences per video. We split the video dataset according to [9] into a training set, a validation set and a testing set which consists of 1200 videos for training, 100 for validation and 670 for testing. The training split contains about 48.7k text sentences; the validation split contains 4.3k text sentences and the testing split contains 27.7k text sentences. We apply simple preprocessing on the text data by converting all text to lower case, tokenizing the sentences and removing punctuation.

We also use the Microsoft COCO corpus [16], which is a new image recognition, segmentation, and captioning dataset, to pre-train the video description system based on visual information.

To pre-train the video description system based on audio information, we collect more audio data in-the-wild from freesound.org which contains user-collected recordings with descriptions and tags. In total, we collect over 10,000 audio files with a total duration of about 200 hours, covering a wide range of sound categories such as activities, locations, occasions, objects, scenes, and nature sounds etc. Each audio comes with tags and descriptions made by its uploader, but the format and quality of the descriptions differ greatly from each other. To ensure the quality of the data we use to train our model, we only keep the description sentences that match the tags to avoid unrelated descriptions. In the end, each audio comes with a description of one or two sentences.

(a) Visual-only : a woman is cooking
(b) Audio-only: a man is slicing a potato

(a) Visual-only: A woman is talking
(b) Audio-only: A woman is dancing

Figure 2. *An example description generated by the audio-only and visual-only system respectively*

4.2. Evaluation Metric

We use the METEOR [28] and BLEU [29] metrics which were originally proposed to evaluate machine translation results for quantitative evaluation of the video description system. The METEOR metric is designed to address some of the deficiencies inherent in the BLEU metric. The METEOR metric also includes some other features that are not considered in other metrics, such as synonymy matching, where instead of matching only on the exact word forms, the metric also matches on synonyms. The metric also includes a stemmer, which lemmatises words and matches on the lemmatised forms.

4.3. Baseline Results

For the visual-based video description system, we pre-train the LSTM-RNN on the MSCOCO dataset. We then fine-tune the model on the MSVD training set using a low learning rate. The system yields a METEOR score of 25.0% on the MSVD test set, which is comparable to the state-of-the-art description results as reported in [8-9]. The audio-based video description system achieves a METEOR score of 18.8% if we train the LSTM-RNN directly on MSVD training set. If the system is pre-trained on the freesound data as described in section 4.1 and then fine-tuned on the MSVD training data, the METEOR score is improved to 19.6%. The results show that visual-only system achieves better description performance than audio-only system.

As we look closely into the videos in the MSVD corpus, we find that some videos are post-edited with pure music. For the purpose of investigating how much additional information that audio content can contribute to the visual-only based description, we think such videos may not be useful for this purpose. We therefore filter out those videos that are post-edited with pure non-content-related music or with no soundtrack from the MSVD corpus. About 12% of the video data were filtered out, leaving 1729 videos in total in the following experiments. A METEOR score of 23.70% and 20.21% is achieved respectively if the fine-tuned visual-only description system and audio-only description system are evaluated on the filtered test set. The performance of the audio-only system improves a bit but that of the visual-only system drops slightly which is not surprising since the filtering is biased towards audio. Although the visual-only system outperforms the audio-only system, from some detailed case analysis, we find that the audio-only system provides complementary information that the visual-only system fails to capture. For example in Figure 2, in a cooking video with a man's voice in the background, the audio-only system detects "man" while the visual-only system confuses the gender of the cook. In the second example, the visual-only system predicts the description "A woman is talking". While the audio-only system detects acoustic characteristics – the music playing in the room - and predicts the correct dancing activity and generates a better description "A woman is dancing".

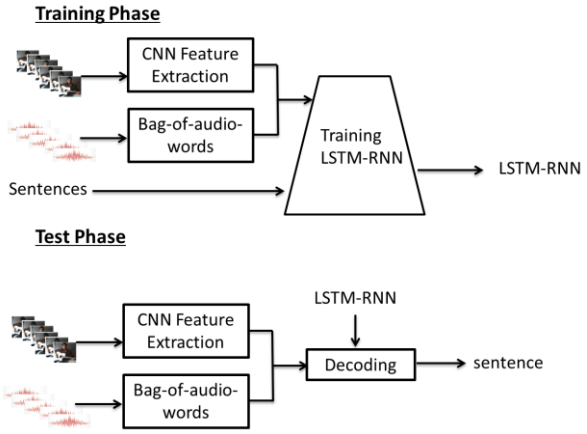


Figure 3: *Audio-visual combined video description system*

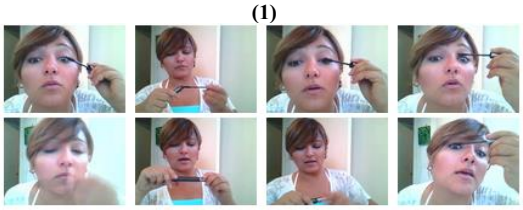
4.4. Fusion of Acoustic and Visual Cues

As shown in the above example that audio and visual information are complementary and should be combined for a better description system. We therefore construct a video description system using both visual and acoustic cues. In this paper, we achieve the combination at the feature representation level by simply concatenating the video-level CNN visual features and the bag-of-audio-words acoustic features. We then train the LSTM-RNN model on the MSVD training set. The system structure is illustrated in Figure 3. To reduce the dimension of the simply concatenated audio+visual feature representation, we apply PCA on the bag-of-audio-words feature to reduce its dimension to 400 before concatenation. The description performance comparison in METEOR among the audio-only, visual-only and audio+visual combined systems is presented in Table 1. As we can see from the table, combining audio and visual cues together greatly improves the description performance over each single fine-tuned baseline system. In Figure 4, we showcase some video description examples from visual-only system vs. audio+visual combined system. On these examples, the combined system achieves higher METEOR score than the visual-only system. The METEOR score is shown in the brackets following each sentence. We observe that the combined system can provide more accurate description such as identifying the correct gender of the person by taking the acoustic cues into account.

Table 1: *Comparison among the audio-only, visual-only and combined description systems (METEOR in %)*

System	Audio	Visual	Combined
METEOR	20.21	23.70	**26.17**

(a) Visual-only system: A woman is talking (35.06%)
(b) Combined system: A man is talking (100%)
(1)



(a) Visual-only system: A woman is talking (27.68%)
(b) Combined system: A woman is applying makeup (52.46%)
(2)

(a) Visual-only system: A car is running (9.04%)
(b) Combined system: A plane is flying (49.29%)
(3)

Figure 4. *Example descriptions from visual-only system vs. from combined system*

4.5. Impact on Description Performance

We notice that the amount of available data for training a video description system is quite small compared to the amount of image caption data because generating a video description is more complex than generating an image caption. It is often the case that there are multiple manual descriptions or captions from different human annotators for a single video or image. Naturally we would ask the question: for training a good image caption or video description system, is it more important to have more video/image data or is it more important to have more manual caption data. The finding will guide us to make different focus on increasing the amount of video description training data. We therefore conduct the following experiment to compare the image caption performance by changing the amount of training images and the amount of training captions to see which factor makes more impact on the caption performance.

We use the Flickr8K dataset [30] which contains 8092 images from the Flickr.com website as training dataset. There are five captions for each image that were generated by different annotators using a crowdsourcing service. We use 1000 images from the Microsoft COCO [16] train2014 dataset as testing dataset. We compare caption results with the following three training setups.

S1. Baseline setup: using all the 8092 images and all their captions to train the image caption system based on CNN+LSTM-RNN same as shown in Figure 1.

S2. Reduce number of images: using randomly selected 4000 images from the Flickr8K dataset and all their captions to train the image caption system.

S3. Reduce number of captions: using all the 8092 images and randomly selecting 3 out of 5 captions for each image to train the image caption system.

Table 2 compares the image caption performance in BLEU with the above three different training setups. We can see that reducing the number of captions leads to more performance degradation than reducing the number of images. Although the number of images*captions (8092*3=24,276) in S3 setup is larger than that (4000*5=20,000) in S2 setup, the image caption performance is worse. We therefore suspect that the number of various captions is more important for training an image caption system when limited amount of training data is available. In the next steps, we will consider generating more descriptions via automatic expansion approaches for the training videos in MSVD dataset. We will verify its impact on the video description performance.

Table 2: *Image caption performance comparison with different training setups (BLEU in %)*

Training Setup	BLEU1	BLEU2	BLEU3	BLEU4
S1: Baseline	54.6	36.20	22.7	14.5
S2: Reducing #images	53.3	34.4	21.3	13.6
S3:Reducing #captions	51.9	33.3	20.1	12.6

5. Conclusions

Generating natural language descriptions of video content is a challenging problem. Most works for video description generation focus only on visual information in the video. However, audio also provides rich information for describing video contents. In this paper, we investigate using acoustic information in addition to the visual information in the video for natural language description generation. Our system relies on CNN and LSTM-RNN two networks. We simply combine both acoustic and visual information at the video representation level. Experiments on the Microsoft Research Video Description corpus show that fusing audio information improves the description performance greatly. Case studies show that audio information can fix the acoustically related errors in the visual-only description output. Therefore, there is a lot of benefit to explore using acoustic information for video description prediction. In this work, the visual and acoustic information are both captured at holistic video representation level.

In the future work, we will explore more powerful representation by taking into account the sequential aspect and synchronizing visual and acoustic information for joint modeling. Through comparison experiments, we also observe that when limited amount of training is available, number of various captions is more important than number of various images. This will guide us to investigate in the future how to improve the video description system via increasing amount of training data in the future work.

6. Acknowledgement

This work was partially supported by the Beijing Natural Science Foundation (No. 4142029), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

7. References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. “Show and tell: A neural image caption generator”. arXiv:1411.4555, 2014.

[2] 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”. CVPR, 2015.

[3] J. Mao, W. Xu, Y. Yang, J.Wang, and A. L. Yuille. “Deep captioning with multimodal recurrent neural networks (m-rnn)”. arXiv:1412.6632, 2014.

[4] H. Yu and J. M. Siskind. “Grounded language learning from videos described with sentences”. ACL, 2013.

[5] P. Das, C. Xu, R. F. Doell, and J. J. Corso. “A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching”. CVPR, 2013.

[6] S. Hochreiter and J. Schmidhuber. “Long short-term memory”. Neural Computation, 1997.

[7] 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-shoot recognition”. ICCV’13.

[8] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. “Translating video content to natural language descriptions”. ICCV’13.

[9] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko. “Translating videos to natural language using deep recurrent neural networks”. NAACL, 2015.

[10] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. “Describing videos by exploiting temporal structure”. arXiv:1502.08029v4, 2015.

[11] H. Fang, S. Gupta, F.N. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J.C. Platt, C.L. Zitnick, G. Zweig. “From captions to visual concepts and back”. CVPR, 2015.

[12] A. Karpathy, L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. CVPR, 2015.

[13] R. Kiros, R. Salakhutdinov, R.S. Zemel. “Unifying visual-semantic embeddings with multimodal neural language models”. Trans. of ACL, 2015.

[14] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio. “Show, attend and tell: Neural image caption generation with visual attention”. arXiv:1502.03044, 2015.

[15] P. Young, A. Lai, M. Hodosh, J. Hockenmaier. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. Trans. of ACL, 2014.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. “Microsoft coco: Common objects in context”. arXiv:1405.0312, 2014.

[17] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, M. Mitchell. “Language models for image captioning: The quirks and what works”. arXiv:1505.01809, 2015.

[18] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. “Sequence to sequence - video to text”. arXiv: 1505.00487, 2015.

[19] Y. Pan, T. Mei, T. Yao, H. Li, Y. Rui. “Jointly modeling embedding and translation to bridge video and language”. arXiv:1505.01861, 2015.

[20] Brown, L. et al. “IBM research and Columbia University TRECVID-2013 Multimedia Event Detection (MED),Multimedia Event Recounting (MER), Surveillance Event Detection (SED), and Semantic Indexing (SIN) Systems”. TRECVID 2013.

[21] K. Simonyan, A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. arXiv:1409.1556, 2014.

[22] S.B. Davis, and P. Mermelstein . “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. IEEE Trans. on Acoustics, Speech, and Signal Processing, 28(4), 357–366, 1980.

[23] D.N. Rockmore, “The FFT: an algorithm the whole family can use”. Computing in Science Engineering 2(1): 60–64, 2000.

[24] N. Ahmed, T. Natarajan, K.R. Rao. “Discrete Cosine Transform”. IEEE Transactions on Computers 23(1): 90–93, 1974.

[25] S. Pancoast, M. Akbacak. “Softening quantization in bag-of-audio-words”, ICASSP 2014, 1370 - 1374.

[26] Q. Jin, J. Liang, X. He, G. Yang, J. Xu, X. Li. “Semantic concept annotation for user generated videos using soundtracks. International Conference on Multimedia Retrieval (ICMR) 2015.

[27] D. Chen, W. Dolan. “Collecting highly parallel data for paraphrase evaluation”. ACL, 2011.

[28] S. Banerjee and A. Lavie. “Meteor: An automatic metric for MT evaluation with improved correlation with human judgments”. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation/Summarization, 2005.

[29] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. “BLEU: a method for automatic evaluation of machine translation”. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[30] M. Hodosh, P. Young, J. Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics”. Journal of Artificial Intelligence Research 47, 853–899, 2013.