Video Description Generation using Audio and Visual Cues

Qin Jin
Multimedia Computing Lab, School of Information
Key Lab of Data Engineering and Knowledge Engineering
Renmin University of China
qjin@ruc.edu.cn

Junwei Liang
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
junweil@andrew.cmu.edu

ABSTRACT
The recent advances in image captioning stimulate the research in generating natural language description for visual content, which can be widely applied in many applications such as assisting blind people. Video description generation is a more complex task than image caption. 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 using both audio and visual cues. We use 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.

Keywords
video description; image caption; audio analysis; deep neural networks.

1. INTRODUCTION
Describing visual content automatically in natural language sentences is a challenging task. It can be widely applied in many applications such as assisting blind people. With the recent success in generating natural language sentence descriptions for images which is also called image caption [1-3], generation of natural descriptions for videos has also attracted more and more attention in the research community. However, it is a more complex problem than image caption. 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 such as YouTube videos remains an open challenge.

Many of the recent works for video description generation rely on Long-Short Term Memory networks (LSTMs) [6], a type of recurrent neural networks (RNNs), from the visual point of view only. Visual information in videos has been captured by video-level representations [7-8], or frames-level representations [9], or sub-sampling on a fixed number of input frames [10]. However, human verbalization of video contents may rely not only on the visual information but also on other content-related information such as audio content that is not directly present in the visual source. In this paper, we study utilizing audio cues in the video together with visual cues for creating a better description system. Our approach is inspired by 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 audio and visual information. Section 4 presents the experiments and case studies on the MSVD corpus. Section 5 concludes the paper.

2. RELATED WORK
Image caption, generating a natural language description for images, has received a lot of attention and achieved some exciting results recently. Many works have been proposed [1-3, 11-14]. Most of them 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] etc. There are also studies to emphasize the novelty of generated descriptions [17].

The task of video description generation has also attracted more and more interest lately [4-5, 7-10, 18]. We can generally categorize video description approaches into two types. The first type normally relies on action recognition, object detection, attributes recognition, or event recognition to extract the information needed to render the linguistic entities for sentence generation [19]. The second type normally relies on the end-to-end sequence-to-sequence model. 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 [20] 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

ICMR'16, June 06-09, 2016, New York, NY, USA
 ISBN 978-1-4503-4359-6/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2911996.2912043.

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 TRECVID Multimedia Event Detection (MED), Multimedia Event Recounting (MER), Surveillance Event Detection (SED), and Semantic Indexing (SIN) Systems [21]. However, to the best of our knowledge, there hasn't been much work on video description based on deep models using acoustic information, although there are some works on video classification using deep models with both visual and acoustic cues [30]. Therefore, in this paper, we conduct a pilot study on utilizing acoustic information to improve the video description performance based on deep models using visual cues only.

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 representation 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 test 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 test phase, the trained LSTM-RNN is applied for sentence prediction. There are three key components: visual/acoustic feature representation, text sequence modeling.

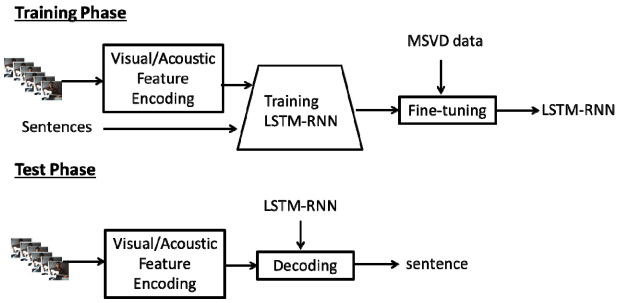


Figure 1: Illustration of video description system

3.1 Visual/Acoustic Feature Representation

A CNN is applied for visual feature representation. We use the pre-trained VGGNet [22] for visual feature extraction. A feature vector is extracted for each frame and mean pooling is applied to produce a video-level visual feature representation.

For acoustic feature representation, 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) [23] as our fundamental feature. The Fast Fourier Transformation (FFT) [24] 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) [25] 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 [26] is generated by applying an acoustic codebook to transform this set of MFCCs into a single fixed-dimension (4096) video-level feature representation.

3.2 Text Sequence Modeling

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:

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

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. We employ LSTM-RNN to encode the sentence description of the video, and decode a visual/acoustic feature representation of fixed length to generate natural language output. The encoding LSTM-RNN and decoding LSTM-RNN are shared.

4. EXPERIMENTS

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 MSCOCO 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.

4.2 Evaluation Metric

We use the METEOR [28] metric which was originally proposed to evaluate machine translation results for quantitative evaluation of the video description system. The METEOR score is computed based on the alignment between a given hypothesis sentence and a set of candidate reference sentences.

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 and the remaining 1729 videos are used in the following experiments. A METEOR score of 23.70% and 20.21% is achieved respectively if the 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 can provide complementary information that the visual-only system fails to capture. For example in Figure 2, the visual-only system predicts the description "A woman is exercising". 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 girl is dancing".

4.4 Fusion of Audio and Visual Cues

As shown in the above example, 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. 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. We can see that combining audio and visual cues together greatly improves the description performance over each single fine-tuned baseline.

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

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

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

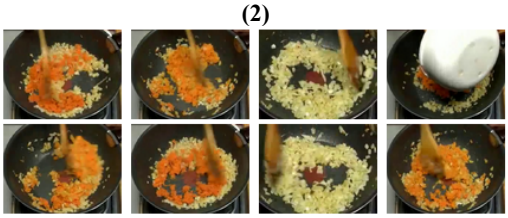
System	Audio	Visual	Combined
METEOR	20.21	23.70	**26.17**

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

(1)

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

(2)



(a) Visual-only system: A person is cooking (25.24%)
(b) Combined system: A woman is cooking (49.29%)

(3)

Figure 3. Example descriptions from visual-only system vs. from combined system

In Figure 3, 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, as shown in examples (1) and (3), or making correct object detection with the help of acoustic cues, as shown in example (2).

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. We will look at larger video databases such as [29] as well.

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, et al. "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition". *ICCV'13*.

[8] M. Rohrbach, et al. "Translating video content to natural language descriptions". *ICCV*'13.

[9] S. Venugopalan, et al. "Translating videos to natural language using deep recurrent neural networks". NAACL, 2015.

[10] L. Yao, et al. "Describing videos by exploiting temporal structure". *arXiv:1502.08029v4*, 2015.

[11] H. Fang, S. Gupta, et al. "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, et al. "Show, attend and tell: Neural image caption generation with visual attention". *arXiv:1502.03044*, 2015.

[15] P. Young, et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". *Trans. of ACL,* 2014.

[16] T.-Y. Lin, et al. "Microsoft coco: Common objects in context". *arXiv:1405.0312*, 2014.

[17] J. Devlin, et al. "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] A. Barbu, et al. "Video In Sentence Out". UAI 2012.

[20] Y. Pan, et al. "Jointly modeling embedding and translation to bridge video and language". *arXiv:1505.01861*, 2015.

[21] L. Brown, et al. "IBM Research and Columbia University TRECVID-2013 MED, MER, SED, and SIN Systems". TRECVID 2013.

[22] K. Simonyan, A. Zisserman. "Very deep convolutional networks for large-scale image recognition". *arXiv:1409.1556*, 2014.

[23] 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.

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

[25] N. Ahmed, et al. "Discrete Cosine Transform". *IEEE Transactions on Computers* 23(1): 90-93, 1974.

[26] S. Pancoast, M. Akbacak. "Softening quantization in bag-of-audio-words", *ICASSP 2014*, 1370 - 1374.

[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] T. Bart, B. Elizalde, D.A. Shamma, K. Ni, G. Friendland, D. Poland, D. Borth, L.J. Li. "YFCC100M: The new data in multimedia research". Communications of the ACM 59, No. 2 (2016): 64-73.

[30] Z. Wu, et al. "Exploring Inter-feature and inter-class relationships with deep neural networks for video classification". ACM Multimedia 2014.