

VIDEO CAPTIONING BASED ON JOINT IMAGE-AUDIO DEEP LEARNING TECHNIQUES

Chien-Yao Wang¹, Pei-Sin Liaw², Kai-Wen Liang², Jai-Ching Wang³, Pao-Chi Chang^{2,3}

¹Institute of Information Science, Academia Sinica, Taiwan

²Department of Communication Engineering, National Central University, Taoyuan, Taiwan

³Department of Computer Science and Information Engineering, National Central University, Taiwan

Abstract—With the advancement in technology, deep learning has been widely used for various multimedia applications. Herein, we utilized this technology to video captioning. The proposed system uses different neural networks to extract features from image, audio, and semantic signals. Image and audio features are concatenated before being fed into a long short-term memory (LSTM) for initialization. The joint audio-image features help the entire semantics to form a network with better performance.

A bilingual evaluation understudy algorithm (BLEU)—an automatic speech scoring mechanism—was used to score sentences. We considered the length of the word group (one word to four words); with the increase of all BLEU scores by more than 1%, the CIDEr-D score increased by 2.27%, and the METEOR and ROUGE-L scores increased by 0.2% and 0.7%, respectively. The improvement is highly significant.

Index Terms—Video captioning, sound event detection, acoustic scene classification, convolutional neural networks, long short-term memory, word embedding

I. INTRODUCTION

The image recognition technology involving deep learning has advanced, and video captioning technology has progressed. Videos now carry a description so that users can search and find videos on the basis of not only the title but also the video content. A video is a series of frames displayed one after another. Compared with an image, a video contains more contents with actions and movements. To convert the audio in a video to textual information, it is crucial that a network should learn the basic rules of the language.

Video captioning is a challenging research topic. There already exist a few research literatures utilizing video and audio signals to obtain the best results [1-4]. This work proposes an integrated video-audio captioning system. Feature normalization is utilized to achieve the balance of the input image and audio features. The importance of normalization must not be underestimated. The experimental results can confirm it.

To analyze the audio information in videos, in 2014, Venugopalan *et al.* proposed a device combining convolutional neural networks (CNN) and recurrent neural

networks [5]. In this work, feature extraction is performed on every frame in a video by using a CNN. Then, the recurrent neural network processes the sequential data of a video and converts it into text format.

In Section II, we present an integrated video-audio description technology that uses different CNN architectures to extract sound event detection features, acoustic scene classification features, and two-dimensional (2D) and three-dimensional (3D) features of an image; little data increment is required to obtain high quality video captions while using this approach. We experimentally demonstrate that our system outperforms the currently used systems. Section III provides the details of the results obtained using the Youtube2Text dataset. Finally, Section IV presents the conclusions.

II. PROPOSED SYSTEM

The proposed system consisting of three parts is shown as Fig. 1. The image, audio, and semantic descriptions are first processed in the preprocessing steps, and then the feature extraction process is performed using different CNN architectures. The extracted features are finally fed into the proposed semantic compositional network (SCN) architecture to generate the output for testing and evaluation.

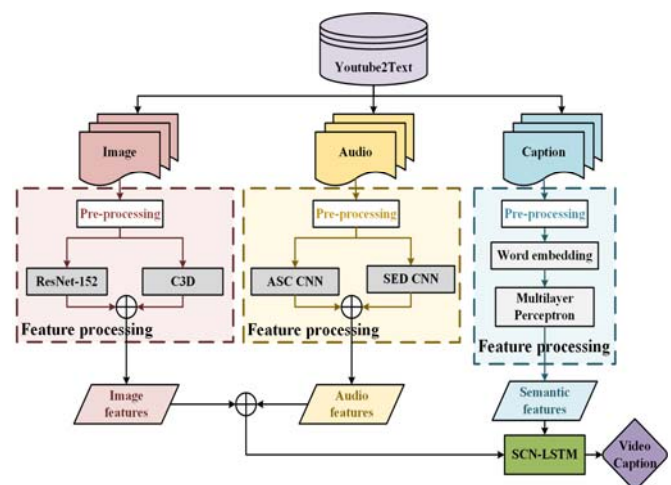


Fig. 1. Flowchart of the proposed video-captioning system

A. Data Preprocessing

For image description, only two frames per second are used as inputs to the system in order to reduce the amount of similar frames entering the system. This downsampling step considerably reduces data redundancy.

For audio description, we use the audio preprocessing method reported by Wu et al. [6]; this approach uses conventional audio signal processing methods. Original data is processed using short-time Fourier transform (STFT) and a Mel filter bank. The Hamming window length in the STFT is 40 ms per frame, and the moving distance is the length of the Hamming window. A log-mel spectrogram of size $40 \text{ mel} \times$ the number of frames is obtained finally.

For caption description, a large number of English sentences are acquired from a database. All words appearing in the sentences are encoded and stored in a “dictionary.” Then, the words are converted to numbers by using this dictionary. The database used contains 12,594 different words.

B. Feature Extraction

Before sending a video into SCN, features are extracted from the preprocessed image, audio, and caption contents of the video separately.

1) Image feature extraction

We use the ResNet-152 [7] and C3D [8] networks to extract 2D features from the images and 3D motion features from the frames in the video. The ResNet-152 network uses ImageNet [9] to form the pre-trained network. The images acquired from the Youtube2Text dataset with preprocessing were rescaled to a size of $224 \times 224 \times 3$. Then, these images were input into the pre-trained ResNet-152 network. The output of the fifth layer of the convolutional neural network is considered the 2D features of the video, and the feature size is 2048.

The C3D network uses the Sports-1M video dataset [10] for pretraining; then, the Youtube2Text dataset is preprocessed to adjust the image size to $112 \times 112 \times 3$. The length of the video was 16 frames, and eight overlapping frames were input in the C3D network. The output of fc7 is considered the 3D features of the video. The feature size of a frame is 512.

The experimental results in [11] evidence that averaging features of all frames of a video yields better results than using the features obtained from each frame separately. Thus, our approach averages the features from the ResNet-152 network ($2048 \times$ the number of frames) and the features from the C3D network ($512 \times$ the number of frames) separately to obtain the final 2D-CNN features of size 2048 and 3D-CNN features of size 512.

2) Audio feature extraction

The system performs sound feature extraction by using sound event detection (SED), which detects single or multiple sound events, and acoustic scene classification (ASC), which classifies the acoustic sound environment, to the preprocessed audio data, that is, a log-mel scale spectrum of size $40 \text{ mel} \times$ the number of frames.

For SED, the videos in the Youtube2Text database are marked and divided into eight categories. The training and verification datasets are further obtained by dividing those videos into two groups. Mini-batch is set to be 200, the learning rate is 0.00001, and the number of iterations is set to 1100 in the training process. Subsequently, we input all the video and audio content from the Youtube2Text database into the trained network and extracted the eight-dimensional sound event features that passed through max pooling.

For ASC, we used the TUT Acoustic Scenes 2016 [12] database provided by the 2016 Detection and Classification of Acoustic Scenes and Events for pretraining. For training, the Mini-batch is set to 256, the learning rate is 0.001, and the number of iterations is set to 200. Finally, the acoustic scenes are input into the Youtube2Text database to extract the 15-dimensional sound scene features that passed through max pooling.

We use the method presented in [6] to split the sound spectrum of the same film into several spectrograms of size 40×25 . Moreover, the two-layer asymmetric kernel convolutional neural network is utilized to train the convolution kernel of a size of 5×7 . The output of the last layer is used as a feature. Finally, the output characteristics of each spectrogram are averaged to obtain the audio feature representing the video

3) Semantic feature extraction

For semantic feature extraction, we used One-Hot Encode to process the encoded single word from the previous statistics. Then, the most commonly used 300 words, including commonly used nouns, verbs, and adjectives, in the training dataset are used to create the labels. Considering the entire problem as a multilabel classification problem, a support vector machine (SVM) and a sigmoid function was used to map 12594-dimensional text features into a 300-dimensional space.

C. SCN

The proposed architecture of the SCN, by referring to [13], is presented in Fig. 2. Initially, a concatenation of the previously obtained sound and image features is used for initializing the LSTM network.

After the weight matrix is generated using the SVM, the first output for a sentence is obtained from the SCN. The first output and the weight matrix are combined as the input for the next sentence.

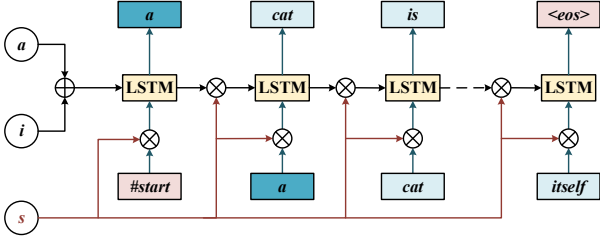


Fig. 2. SCN architecture combining image audio features: a is an audio feature, i is an image feature, and s is a semantic weight matrix

The Mini-batch is set to be 64, learning rate is 0.0002, number of maximum iterations is 20, the word embedding dimension is 300, dropout is 0.5, and the size of the LSTM hidden layers was 512. The algorithm Adam [14] is used to replace the traditional gradient descent algorithm for updating the weights.

III. EXPERIMENTAL RESULTS

A. Dataset

The video description database used is Youtube2Text dataset (or MSVD dataset) [15] created by Microsoft Research in 2010. This database contains 1970 YouTube videos. The length of each video is between 10 and 25 seconds. Youtube2Text dataset have provided a video description, and each video contains approximately 40 sentences in English.

However, some videos could not be downloaded due to regional problems or because they were removed from YouTube website. Finally, 1659 videos were obtained. From the 1659 videos, we removed 305 videos whose sounds were not related to the content of the video. Thus, 1354 videos were obtained; the videos were divided to training (835 clips), verification (69 clips), and testing (450 clips) sets.

B. Evaluation

For evaluating the accuracy of the text description, the result can be scored according to the fluency of the sentence and the relevance of the target sentence by manual judgement. However, the standards of subjective evaluation are hard to be consistent. A set of criteria were developed for evaluating the text descriptions. The algorithms mostly used for the evaluation are bilingual evaluation understudy BLEU [16], METEOR [17], ROUGE-L [18], and CIDEr-D [19].

C. SCN-LSTM Experiments

After feature extraction, we combined the 15-dimensional sound scene features and 8-dimensional sound event features to obtain a 23-dimensional sound feature. Moreover, image features were added for training. The experimental results on testing sets are presented in Table I.

By conducting experiments, the results reveal that the output of the semantic descriptions with considering the scenes and sound events at the same time accompanied by normalizing the sound features are all scored better than the semantic descriptions trained using only image features.

When the audio features were normalized to the $[-1, 1]$ interval, all scores increased by at least 1% for the word length from one to four words in the BLEU score, and the CIDEr-D score was as high as 2.27%. The METEOR and ROUGE-L scores also increased by approximately 0.2% and 0.7%, respectively.

We added different sound scenes and sound events features by using different normalized network. As presented in Table II, when the audio uses the same normalization, almost all indicators consider both sound scenes and sound events. Simultaneously, the semantic output of the network output was higher than other scores. Moreover, when the sound features were differently normalized, the highest scores were mostly concentrated in the normalization interval of $[-1, 1]$. The output is presented in Fig. 3.

TABLE I. EXPERIMENTAL RESULTS OF VIDEOS WITH ADDED ACOUSTIC SCENE AND SOUND EVENT FEATURES WITH DIFFERENT NORMALIZATION METHODS

Metrics	B 1	B 2	B 3	B 4	Meteor	Rouge-L	Cider-D
Image only (Base)	0.8224	0.7154	0.6260	0.5310	0.3490	0.7148	0.8160
Image+audio	0.8278	0.7227	0.6316	0.5365	0.3425	0.7144	0.7962
Improvement	0.54%	0.73%	0.56%	0.55%	-0.65%	-0.04%	-1.98%
Image+audio normalization $[-1\sim1]$	0.8334	0.7282	0.6401	0.5475	0.3513	0.7221	0.8387
Improvement	1.10%	1.29%	1.41%	1.66%	0.23%	0.73%	2.27%
Image+audio normalization $[0\sim1]$	0.8283	0.7244	0.6352	0.5398	0.3518	0.7218	0.8175
Improvement	0.59%	0.90%	0.92%	0.88%	0.27%	0.70%	0.15%


							
Ground truth: <i>a man is playing with his dog</i> Image only: a man is playing with a toy				Image+audio: <i>a monkey is playing</i> Image+audio [-1~1]: a man is playing with a dog Image+audio [0~1]: a man is playing with a dog			
Ground truth: <i>the man is playing basketball</i> Image only: a boy is playing				Image+audio: <i>a boy is playing basketball</i> Image+audio [-1~1]: a man is playing a basketball Image+audio [0~1]: a boy is playing football			

Fig. 3. Video semantics obtained using different input data

TABLE II. COMPARING DIFFERENT SOUND FEATURES AND DIFFERENT NORMALIZATIONS

Metrics	B 1	B 2	B 3	B 4	Meteor	Rouge-L	Cider-D
Original sound							
scene	0.8182	0.7100	0.6186	0.5196	0.3524	0.7157	0.7865
event	0.8209	0.7117	0.6185	0.5224	0.3460	0.7117	0.8043
scene + event	0.8278	0.7227	0.6316	0.5365	0.3425	0.7144	0.7962
Sound normalization: [-1~1]							
scene	0.8218	0.7118	0.6198	0.5217	0.3431	0.7131	0.8305
event	0.8161	0.7083	0.6195	0.5266	0.3469	0.7127	0.8064
scene + event	0.8334	0.7282	0.6401	0.5475	0.3513	0.7221	0.8387
Sound normalization: [0~1]							
scene	0.8230	0.7158	0.6237	0.5255	0.3434	0.7147	0.8373
event	0.8216	0.7141	0.6239	0.5298	0.3464	0.7124	0.7863
scene + event	0.8283	0.7244	0.6352	0.5398	0.3518	0.7218	0.8175

IV. CONCLUSIONS

We propose a new video-captioning system that can automatically describe the content of a video as a short description in text format. This would enable users to search and find videos easily and accurately. The proposed system uses video, image, and audio signals simultaneously and extracts the corresponding features by using different CNN network architectures; finally, the acquired features are all combined and input into the SCN-LSTM network to create a semantic description. In experiments, we used semantics to arrive at score. Using this mechanism, we found that adding sound features is helpful to output a better description of the proposed semantic network. The sound events and scene features were combined. All the scores in the set of evaluation criteria are greatly improved.

The architecture proposed in this study used for video description can be further investigated to achieve better performance. Audio feature extraction has scope for further improvement by pretraining more different audio sources to increase the accuracy. Furthermore, a multi-task CNN model [20] can be used for learning between sound scenes and sound event categories, sharing learning outcomes and acquiring audio features with both scenes and events, which

can improve our ability to extract the two characteristics of the sound instead of training two CNN networks separately, extracting features separately and then combining the two features.

REFERENCES

- [1] S. P. Chuang, C. H. Wan, P. C. Huang, C. Y. Yang, and H. Y. Lee, "Seeing and hearing too: Audio representation for video captioning," IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 381-388, 2017.
- [2] C. Hori, T. Hori, T. K. Marks, and J. R. Hershey, "Early and late integration of audio features for automatic video description," IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 430-436, 2017.
- [3] Liu, Sheng, Z. Ren, and J. Yuan, "SibNet: Sibling Convolutional Encoder for Video Captioning", ACM Multimedia Conference on Multimedia Conference, 2018.
- [4] Xu, Jun, *et al.* "Learning multimodal attention LSTM networks for video captioning", Proceedings of the 25th ACM international conference on Multimedia, 2017.
- [5] S. Venugopalan, H. Xu, and J. Donahue, "Translating videos to natural language using deep recurrent neural networks," arXiv preprint arXiv:14124729, 2014.

- [6] Y. C. Wu, P. C. Chang, C. Y. Wang, and J. C. Wang, "Asymmetric Kernel Convolutional Neural Network for acoustic scenes classification," IEEE International Symposium on Consumer Electronics (ISCE), May. 2018.
- [7] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," In CVPR, 2016.
- [8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. "Learning spatiotemporal features with 3d convolutional networks," In ICCV, 2015.
- [9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. "Imagenet large scale visual recognition challenge," IJCV, 2015.
- [10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-scale video classification with convolutional neural networks," In CVPR, 2014.
- [11] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. "Translating videos to natural language using deep recurrent neural networks," In NAACL, 2015.
- [12] M. Annamaria, T. Heittola, and T. Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection," IEEE 2016 24th European Signal Processing Conference, pp. 1128-1132, Aug. 2016.
- [13] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic Compositional Networks for Visual Captioning," CVPR, 2017.
- [14] D. Kingma and J. Ba. Adam: "A method for stochastic optimization." In ICLR, 2015.
- [15] 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," In ICCV, 2013.
- [16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a method for automatic evaluation of machine translation," In ACL, 2002.
- [17] S. Banerjee and A. Lavie. "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," In ACL workshop, 2005.
- [18] C.-Y. Lin. "Rouge: A package for automatic evaluation of summaries," In ACL workshop, 2004. 6
- [19] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. "Cider: Consensus-based image description evaluation," In CVPR, 2015.
- [20] A. H. Abdalnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN Model for Attribute Prediction," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 1949-1959, Nov. 2015.