Video Caption Generation with Vision-Audio Guidance

UMich-COG team*
University of Michigan
Ann Arbor, MI, 48105, USA

1. INTRODUCTION

Our goal is to design a neural network-based model that can automatically describe the content of a video using a complete and natural English sentence. We adopt an encoder-decoder model structure that uses a deep convolutional neural network (CNN) as the encoder and a long short-term memory (LSTM) [1] as the decoder. Unlike existing video captioning methods, we combine visual features with audio features to yield a multi-modal video representation. The schematic illustration of our system is shown in Fig. 1.

2. METHODS

We preprocess the videos by uniformly sampling 15 frames from each video and resizing the sampled frames to 224×224. To extract visual features from those frames, we pass them separately through the VGG-16 [2] architecture, the weights of which were pretrained on the ILSVRC-2012 and MSCOCO datasets [3], and mean pool the outputs of the first fully connected layer (fc-6). Meanwhile, we extract Mel Frequency Cepstral Coefficients (MFCCs) [4] from the audio stream of each video and construct a single feature vector using a bag-of-words model. Codewords in MFCC space are found by performing k-means on the set of MFCC vectors extracted from the training and validation datasets. Each MFCC vector for a video is assigned the nearest codeword and feature vectors are constructed as histograms of this assignment. We find that 100 codewords (and thus, 100-dimensional feature vectors) work reasonably well. These feature vectors are passed through a fully connected layer and the 512-dimensional activation is then concatenated with the aforementioned VGG-16 visual features.

We then input the vision-audio feature to the guiding LSTM model [5], which "guides" the LSTM by feeding the visual-audio semantic feature as an extra input to the LSTM model at each time step. We adopt a two-layer guiding LSTM model. Words are represented by 512-dimensional vectors, learned alongside the rest of the decoder, and the training procedure is identical to that of [6] with regard to propagation of loss.

3. IMPLEMENTATION DETAILS

Our code is based on the open-source image caption generator, neuraltalk2 [7]. We used an open-source MFCC extraction Python script and the SciPy package to perform the audio feature construction. The encoding vocabulary for words is determined by both the MSR-VTT set and the MSCOCO set, with size 14239.

*Group members: Luowei Zhou, Parker Koch, Chenliang Xu and Jason J. Corso.
Corresponding author: Luowei Zhou (luozhou@umich.edu).

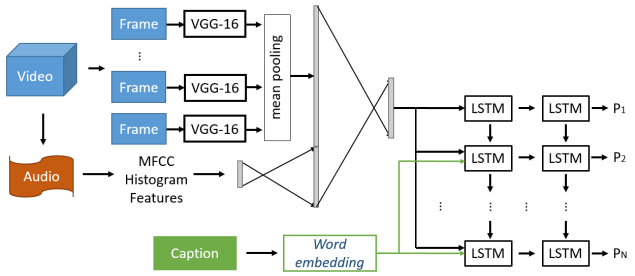

Figure 1: Schematic illustration of the proposed model.

We first pretrain our model on MSCOCO dataset for image captioning with only one-frame input and no audio input. We update our model in an end-to-end manner by enabling the CNN fine-tuning. This allows our model to learn text-related visual features. Then, we fine-tune the model on the MSR-VTT [8] training set with both vision and audio inputs. We first disable the CNN fine-tuning and the audio embedding updating for 100,000 iterations to make the learning process stable. After that, we enable the updating of the encoder weights, which allows the model to learn text-related vision-audio features. The updating strategies for both encoder and decoder are ADAM with batch size equal to 1. During training, we evaluate the intermediate models on the validation set, and save the one with the highest CIDEr score as our final model for the testing set evaluation.

4. REFERENCES

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

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[4] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[5] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding long-short term memory for image caption generation. *arXiv preprint arXiv:1509.04942*, 2015.

[6] 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.

[7] https://github.com/karpathy/neuraltalk2.

[8] 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.