MSR Video to Language Challenge

Ning Xu[1], Junnan Li[2,3], Yang Li[1], An-An Liu[1], Yongkang Wong[2], Weizhi Nie[1], Yuting Su[1], Mohan S. Kankanhalli[3]

[1]School of Electronic Information Engineering, Tianjin University, China
[3]Interactive & Digital Media Institute, National University of Singapore, Singapore
[2]School of Computing, National University of Singapore, Singapore

Overview In this paper, we propose a novel multimodal-sequence-to-sequence (denoted as "M-S2S") video caption model for the Microsoft Research Video to Text (MSR-VTT) Challenge [3, 6]. The core contribution of this model is to leverage the effectiveness of semantic knowledge (i.e., frame level image caption) together with the conventional visual feature on the Sequence-to-Sequence (S2S) [5] model for video caption generation.

M-S2S Our model uses a stack of three LSTMs with 1,000 hidden units each. Figure 1 depicts M-S2S model unrolled over time. The input is a variable number of single visual frames and their corresponding semantic descriptions. The stacked LSTM first encodes the frame level features in a sequential manner, where the top layer (marked as orange) encodes the visual knowledge and the middle layer (marked as pink) encodes the semantic knowledge. Once all frames are encoded, the model generates a video sentence word by word from the bottom layer (marked as blue) during the decoding stage. The encoding and decoding stage of M-S2S are jointly learned from a parallel corpus. For the treatment of frame level and word features, they are respectively embedded to a 500 dimensional space by applying a linear transformation. The weights of the embedding are jointly learned with the LSTM layers during training stage. We refer readers to [5] for more details of the S2S model.

Feature Extraction We employed the output of the fc7 layer (after applying the ReLU non-linearity) from the 16-layer VGG model [4] as frame level visual descriptor. Specifically, we periodically sample frames from the video (1 in every 20 frames) and extract a single 4,096 dimension descriptor for each frame. For the semantic knowledge, we employ a state-of-the-art image caption algorithm, namely Neuraltalk [2], to extract frame level sentence description from raw RGB data. Given an extracted caption, we use Sent2vec [1] to map each sentence to a single 1,024 dimension descriptor. Meanwhile, we follow the original S2S protocol [5] to encode the target output sequence of words.

Qualitative Result Four examples of generated captions are shown in Figure 2. In the first case, the prediction *A group of people are walking in a large room* does not appear in the training set. However, fragment of the prediction, such as *a group of people* (803 instances), *are walking* (388 instances), and *in a large room* (17 instances), can be found in different training videos. This suggests M-S2S can composes information from various data segment to describe a video. The prediction in the third case is sensible but the subject is mistranslated as *woman*. The last prediction demonstrate a model overfitting scenario due to dominant training sample, where the phrase *basketball* has over 1,000 occurrences in the training set. Lower learning rate or using more examples to fine tune the model are likely to prevent this problem.

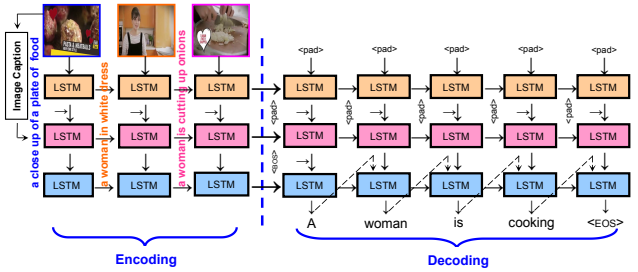

Figure 1: The structure of the proposed M-S2S model.

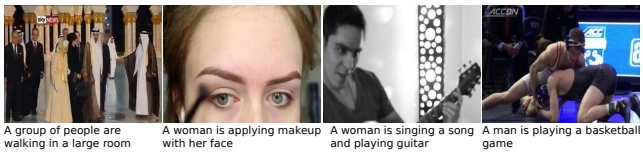

Figure 2: Sentences generated by the proposed M-S2S model for test videos.

1. REFERENCES

[1] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. P. Heck. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *CIKM*, pages 2333–2338, 2013.

[2] A. Karpathy and F. Li. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, pages 3128–3137, 2015.

[3] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language. In *CVPR*, 2016.

[4] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014.

[5] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to Sequence - Video to Text. In *ICCV*, pages 4534–4542, 2015.

[6] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, 2016.

MM '16 Amsterdam, The Netherlands

DOI: 10.475/123_4