MSR Video to Language Challenge

An-An Liu[1], Ning Xu[1], Yurui Qiu[1], Xinhui Li[1], Mengjie Li[1], Yaoyao Liu[1],
Yongkang Wong[2], Weizhi Nie[1], Yuting Su[1], Mohan S. Kankanhalli[3]

[1]School of Electrical and Information Engineering, Tianjin University, China
[2]School of Computing, National University of Singapore, Singapore
[3]Smart Systems Institute, National University of Singapore, Singapore

Overview. In this paper, we propose a novel Two-Stream Fusion (denoted as "TSF") video caption model for the 2^{nd} Microsoft Research Video to Text (MSR-VTT) Challenge [3, 4]. The core contribution of this model is to jointly discover and integrate the dynamics of both visual (i.e., Resnet101 features [2]) and semantic (i.e, high-level attributes) streams for video captioning.

TSF. We present the overview of the proposed TSF model in Figure 1. Intuitively, a video can be complementarily represented by visual and semantic descriptors. The visual descriptor encodes the appearance information depicted in each frame, while the semantic descriptor encodes each video frame with high-level representation of semantic attributes (i.e., objects (nouns), motions (verbs) and properties (adjectives)). Given these two complementary information, our model first considers each modality as a unique stream, and uses a two-stream network to individually encode the dynamics of each modality. Particularly, we utilize the attention-LSTM unit [5] to enhance the individual feature learning. Then, a "combine unit" is deployed to linearly perform two-stream dynamic fusion for sentence generation. The softmax layer is deployed to get the probability distribution over the words.

Feature Extraction. We selected 50 equally-spaced frames out of each video. For the visual features, we used a pretrained Resnet101 model [2] to obtain 2,048 dimensional frame-wise visual features, which were extracted from the 'pool5' layer. For the semantic features, we used a set of attributes to represent the visual content in each frame. Particularly, we used MSCOCO as the extended dataset and selected 256 most frequent words from the training captions as the high-level attributes. Then, we associated each MSCOCO image with a set of attributes according to its captions. The attribute detectors were trained with binary SVM [1]. Finally, the SVM [1] predictions were aggregated as a 256-way vector and used as frame-wise semantic representation.

Implementation. We randomly select 9,000 videos as training set and 1,000 videos as validation set. Each word in the sentence was represented as a "one-hot" vector. The word embedding dimensionality was set to 468 and the dimension of hidden layers in attention-LSTM was 3,698. We optimized the hyperparameters with random search to maximize the log-probability on the validation set.

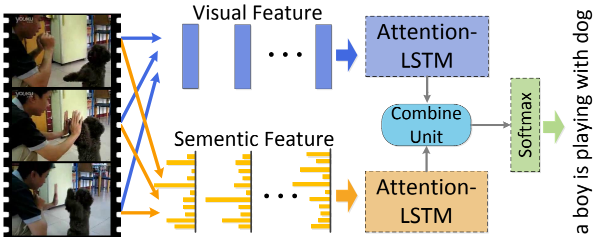
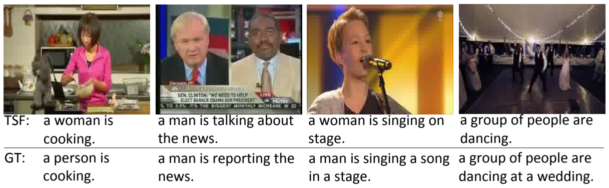


Figure 1: The structure of the proposed TSF model.



TSF:	a woman is cooking.	a man is talking about the news.	a woman is singing on stage.	a group of people are dancing.
GT:	a person is cooking.	a man is reporting the news.	a man is singing a song in a stage.	a group of people are dancing at a wedding.

Figure 2: Sentences generated by the TSF model on validation videos. GT represents the randomly selected ground truth sentence.

Qualitative Analysis. In Figure 2 , we show the four examples of generated descriptions by TSF. We can clearly see that the generated descriptions correspond well with the video clips. It illustrates that the dynamics of visual and semantic streams are complementary to boost video captioning. However, our model failed to generate the attribute "wedding" in the last prediction, which demonstrates that we should train more robust attribute detectors to capture semantic-stream of TSF for videos.

1. REFERENCES

[1] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3):27:1–27:27, 2011.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[3] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016.

[4] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[5] L. Yao, A. Torabi, K. Cho, N. Ballas, C. J. Pal, H. Larochelle, and A. C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015.

MM '17 Mountain View, CA USA
© 2017 ACM. ISBN 123-4567-24-567/08/06. . . $15.00
DOI: 10.475/123_4