Supplementary Materials: Active Learning for Video Description With Cluster-Regularized Ensemble Ranking

David M. Chan[1], Sudheendra Vijayanarasimhan[2], David A. Ross[2], and John Canny[1 2]

University of California at Berkeley, USA[1] {davidchan,canny}@berkeley.edu

Google Research, USA[2] {svnaras,dross}@google.com

1 Datasets

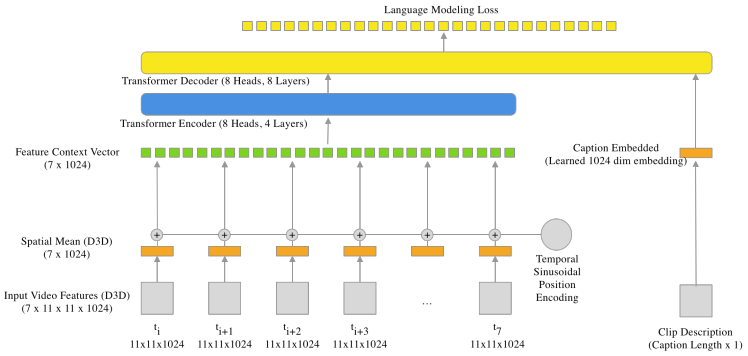
We demonstrate the performance of our model on two common video description datasets, MSR-VTT [1] and the LSMDC [2].

MSR-VTT: The MSR Video to Text Dataset (MSR-VTT) [1] is a large-scale benchmark for video description generation. The dataset was generated by collecting a set of 257 popular video queries, selecting 118 videos for each query. These videos were then annotated using Mechanical Turk with 20 natural language sentences. This provides 10K web video clips, with 41.2 hours of video, and 200K clip-description pairs. The clips have an average length of approximately 15 seconds.

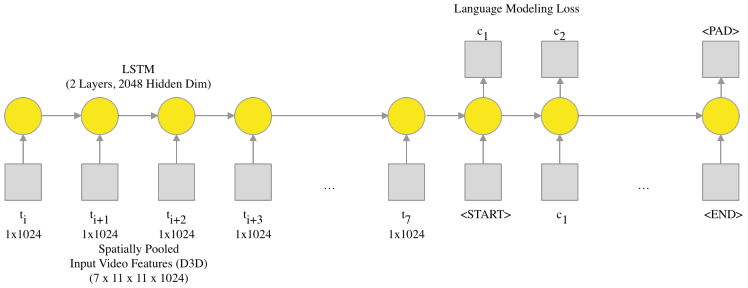
LSMDC: The Large Scale Movie Description Challenge (LSMDC) [2] combines two common bench-mark datasets: M-VAD [3] and MPII-MD [4]. The dataset consists of video descriptions extracted from professionally generated Descriptive Video Services tracks on popular movies. The dataset contains 118,081 clips from 202 unique films. Each clip has approximately one sentence of description, are 2-5 seconds each, and the names of most characters are replaced with a signifying "SOMEONE" tag. The LSMDC dataset has a very wide text coverage, with almost 23,000 unique vocabulary tokens.

2 Models

We use two models in the paper, specified by Figure 1 below. These models are relatively standard in the Video Description literature. Reference numbers refer to references from the main paper.



(a) Model architecture of our transformer-based model based on [5]. The only difference in this model and that from [5] is we drop the computation of the video masking which is unnecessary in our task. We use self-attention over spatially pooled input vectors to produce a context for a Transformer Decoder [6], which is a conditional cross-attention used to produce the output with a language modeling loss. The best-case performance of this model is similar to the stat-of-the-art method for vision only features presented in [7].



(b) Model architecture of the S2VT model [8]. An LSTM encoder is used to encode the spatially pooled video features. The hidden state of the encoder is used to initialize the hidden state of a decoder, which produces the output tokens. The final performance is slightly better than the performance reported in [8].

Fig. 1: Model diagrams for the two models used in this paper. (a), the transformer-based architecture. (b), the S2VT based architecture.

3 Qualitative Examples

Iteration 5
A man is talking to the camera.
A man is talking.
A man is talking.
A man is a man talking.

Iteration 10
A woman is sitting on a couch.
A man is talking to a man.
A man is talking.
A man is sitting on a couch.

Iteration 15
A girl is sitting on a couch.
A man is talking.
A woman is talking to a man.
A man is sitting on a couch.

Iteration 20
A man and a woman are sitting on a couch.
A man is sitting on a couch.
A man is sitting on a couch.
A man and a man is sitting on a couch.

Ground Truth: A man and a woman on a couch race to find the answer to a quesion on their phones

Ensemble-Divergence **Coreset** **Random** **ALISE**

Fig. 2: An example description (Selected randomly) traced during the learning process using multiple methods. While this sample does not have a high visual correlation with the ground truth, this is the case for many videos in the MSR-VTT dataset. As we can see, over the course of training the captions evolve, and the proposed method is able to quickly capture the information present in the scene.

References

1. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 5288–5296
2. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision **123** (2017) 94–120
3. Torabi, A., Pal, C., Larochelle, H., Courville, A.: Using descriptive video services to create a large data source for video annotation research. arXiv preprint arXiv:1503.01070 (2015)
4. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3202–3212
5. Zhou, L., Zhou, Y., Corso, J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. (2018)
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. (2017) 5998–6008
7. Aafaq, N., Mian, A., Liu, W., Gilani, S.Z., Shah, M.: Video description: A survey of methods, datasets, and evaluation metrics. ACM Computing Surveys (CSUR) **52** (2019) 1–37
8. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)