MSR-VTT: A Large Video Description Dataset for Bridging Video and Language Supplementary Material

Jun Xu , Tao Mei , Ting Yao and Yong Rui
Microsoft Research, Beijing, China
{v-junfu, tmei, tiyao, yongrui}@microsoft.com

When organizing the Microsoft Research Video To Language challenge [1], we found that, in our previously released dataset [10], some sentences annotated by AMT workers are identical in one video clip or very similar in one category. Therefore, to control the quality of data and annotations, as well as the competitions, we removed those simple and duplicated sentences and replaced them with refined ones. We finally released the fixed dataset in our challenge website [1]. Due to these modifications of the dataset, the performance cannot be well matched with what we reported in our CVPR paper [10]. Here, we have reported the new performance in the following tables which also appeared in our CVPR paper (referred to as Table 1, 2, 3, 4, 5, 6, and 7, respectively). If you are trying to reproduce or compare the baselines conducted on our MSR-VTT dataset, please refer to this supplementary material and the updated performance reported in this material. However, please cite our CVPR paper [10] if you want to use the MSR-VTT as your dataset.

Feature	BLEU@4	METEOR
AlexNet	6.3	14.1
GoogleNet	8.1	15.2
VGG-16	8.7	15.5
VGG-19	7.3	14.5
C3D	7.5	14.5

Table 2. The performance of KNN baselines with different video representations and mean-pooling strategy.

References

[1] MSR-VTT Video to Language Challenge. http://ms-multimedia-challenge.com/. 1

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL*, pages 190–200, 2011. 2

[3] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of

Feature	BLEU@4	METEOR
AlexNet	32.3	23.4
GoogleNet	34.6	24.6
VGG-16	34.7	24.7
VGG-19	34.8	24.8
C3D	35.4	24.8

Table 3. BLEU@4 and METEOR for comparing the quality of sentence generation on different video representations. The experiments are all based on mean-pooling strategy, and the size of hidden layer in LSTM is set to 512. All values are reported as percentage (%).

Feature	BLEU@4	METEOR
Single frame	31.3	21.7
Mean pooling	34.8	24.8
Soft-Attention	35.6	25.4

Table 5. Performance comparison among different pooling methods (with VGG-19 feature and 512 hidden layers in LSTM).

Hidden layer size	BLEU@4	METEOR	Parameters
128	31.4	21.4	3.7M
256	33.9	23.3	7.6M
512	35.4	24.8	16.3M

Table 6. Performance comparison of different size of hidden layer in LSTM. The video representation here is the clip-based temporal representations by C3D and the pooling strategy is mean pooling.

Feature	Correctness	Grammar	Relevance
AlexNet	7.8	7.0	7.9
GoogleNet	6.2	6.8	6.4
VGG-16	5.3	6.9	5.4
VGG-19	5.4	6.7	5.2
C3D	5.1	6.4	5.3
C3D+VGG-16	5.1	6.1	5.0
C3D+VGG-19	4.9	6.1	5.1

Table 7. Human evaluation of different methods on MSR-VTT. Each method is evaluated by 5 persons (scale 1-10, lower is better).

Dataset	Context	Sentence Source	#Video	#Clip	#Sentence	#Word	Vocabulary	Duration (hrs)
YouCook [3]	cooking	labeled	88	–	2,668	42,457	2,711	2.3
TACos [4, 7]	cooking	AMT workers	123	7,206	18,227	–	–	–
TACos M-L [5]	cooking	AMT workers	185	14,105	52,593	–	–	–
M-VAD [8]	movie	DVS	92	48,986	55,905	519,933	18,269	84.6
MPII-MD [6]	movie	DVS+Script	94	68,337	68,375	653,467	24,549	73.6
MSVD [2]	multi-category	AMT workers	–	1,970	70,028	607,339	13,010	5.3
MSR-VTT-10K (ours)	20 categories	AMT workers	7,180	10,000	200,000	1,856,523	29,316	41.2

Table 1. Comparison of video description datasets. Please note that TACos M-L means TACos Multi-Level dataset. Although MSVD dataset has multiple video categories, the category information is not provided. In our MSR-VTT dataset, we provide the category information for each clip. Among all the above datasets, MPII-MD, M-VAD and MSR-VTT contain audio information.

Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR
MP-LSTM (AlexNet) [9]	77.0	59.2	44.6	32.3	23.4
MP-LSTM (GoogleNet)	78.9	60.7	45.8	34.6	24.6
MP-LSTM (VGG-16)	79.0	61.0	45.9	34.7	24.7
MP-LSTM (VGG-19)	79.2	61.3	46.0	34.8	24.8
MP-LSTM (C3D)	80.9	64.7	48.1	35.4	24.8
MP-LSTM (C3D+VGG-16)	81.5	65.2	48.4	35.7	25.1
MP-LSTM (C3D+VGG-19)	81.7	65.1	48.5	35.8	25.3
SA-LSTM (AlexNet)	77.8	60.8	45.8	34.8	23.8
SA-LSTM (GoogleNet) [11]	79.5	61.9	46.9	35.2	25.2
SA-LSTM (VGG-16)	79.9	63.1	48.7	35.6	25.4
SA-LSTM (VGG-19)	79.9	63.2	48.8	35.6	25.4
SA-LSTM (C3D)	81.2	65.1	49.2	36.1	25.7
SA-LSTM (C3D+VGG-16)	82.1	65.6	49.8	36.5	25.8
SA-LSTM (C3D+VGG-19)	82.3	65.7	49.7	36.6	25.9

Table 4. Performance comparison on our MSR-VTT dataset of seven video representations with mean pooling and soft attention method

videos through latent topics and sparse object stitching. In *Proceedings of CVPR*, pages 2634–2641, 2013. 2

[4] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 2

[5] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. *Pattern Recognition*, pages 184–195, 2014. 2

[6] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. *Proceedings of CVPR*, 2015. 2

[7] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *Proceedings of ICCV*, pages 433–440, 2013. 2

[8] A. Torabi, C. J. Pal, H. Larochelle, and A. C. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv:1503.01070*, 2015. 2

[9] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of ACL*, 2015. 2

[10] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[11] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of ICCV*, 2015. 2