

Supplementary Material

Yuqi Liu^{1,2*}, Pengfei Xiong², Luhui Xu²,
Shengming Cao², and Qin Jin¹ ()

¹ School of Information, Renmin University of China

² Tencent

{yuqi657,qjin}@ruc.edu.cn,

xiongpengfei2019@gmail.com, {lukenxu,devancao}@tencent.com

1 Inverted Softmax.


The hubness phenomenon[8] is that a data point occurs among the k nearest neighbors of other data points. Dual softmax loss (DSL) was mentioned in CAMoE[3], which adopts a inverted softmax[11]. QB-Norm[2] proposes a query-bank normalization with dynamic inverted softmax (DIS) to deal with hubness problem. CLIP2TV[6] also reports its results with inverted softmax. We compare their results with basic inverted softmax during inference in Tab.1. Our results again surpass all other methods with significant improvement.

2 Evaluation Summary on Different Benchmarks

We compared our model to other state-of-the-art methods on different video retrieval benchmark datasets in the main paper. Table 2 summarizes the performance comparison between our proposed model TS2-Net and the previous best model on five different benchmark datasets. Among all these datasets, VATEX[12] and MSR-VTT[13] contain standard captions with average length of 15 words and 8 words respectively. LSMDC[10] contains videos in the movie domain. There is no movie overlap between the training and test set. So it can verify the generalization ability of a model. ActivityNet-Caption[4,7] and DiDeMo[1] offer paragraph-video retrieval, which means that the query text involves multi sentences and the video duration is long. Therefore, the query text in these two datasets contains more semantic information. We show some examples of these datasets in Fig.1. Our model TS2-Net consistently maintains the state-of-the-art performance on all benchmark datasets with very different characteristics, which demonstrates that our model TS2-Net has decent generalization ability.

3 More Qualitative Results

We provide more qualitative results in Fig.2. In the top left example, our model is able to differentiate the horses that are ‘having fun’ from the horses that are

⁰  Corresponding author

* This work is done when Yuqi is an intern at Tencent

Table 1. Text-to-Video retrieval results with Inverted Softmax

MSRVTT-1kA						DiDeMo					
Method	R@1	R@5	R@10	MdR	MeanR	Method	R@1	R@5	R@10	MdR	MeanR
QB-Norm[2]	47.2	73.0	83.0	2.0	-	QB-Norm[2]	43.5	71.4	80.9	2.0	-
CAMoE[3]	47.3	74.2	84.5	2.0	11.9	CAMoE[3]	43.8	71.4	79.9	2.0	16.3
TS2-Net	51.1	76.9	85.6	1.0	11.7	TS2-Net	47.4	74.1	82.4	2.0	12.9
CLIP2TV[6]	52.9	78.5	86.5	1.0	12.8						
TS2-Net(ViT16)	54.0	79.3	87.4	1.0	11.7						

Table 2. Text-to-Video retrieval results on five benchmarks. We select the previous best performance on each dataset for comparison.

Dataset	Method	R@1	R@5	R@10	MdR	rsum
MSR-VTT[13]	CLIP2Video[5]	45.6	72.6	81.7	2.0	199.9
	TS2-Net(Ours)	47.0	74.5	83.8	2.0	205.3
VATEX[12]	CLIP2Video[5]	57.3	90.0	95.5	1.0	242.8
	TS2-Net(Ours)	59.1	90.0	95.2	1.0	244.3
LSMDC[10]	CAMoE[3]	22.5	42.6	50.9	-	116.0
	TS2-Net(Ours)	23.4	42.3	50.9	9.0	116.6
DiDeMo[1]	CLIP4Clip[9]	42.5	70.2	80.6	2.0	193.3
	TS2-Net(Ours)	41.8	71.6	82.0	2.0	195.4
ActivityNet[4,7]	CLIP4Clip[9]	40.5	73.4	-	2.0	-
	TS2-Net(Ours)	41.0	73.6	84.5	2.0	199.1

stationary. In the top right example, our model can capture the small object and correctly retrieve the video with ‘thought bubbles’. Surprisingly, with token shift and token selection module, our model is able to distinguish some *adjective* words. For example, our model correctly retrieves the video with ‘mental bowl’ rather than ‘glass bowl’ (and ‘overweight people’ rather than normal people) in the bottom examples.

We show some failure cases as well in Fig.3, where our model fails to rank the groundtruth video at the top. However, we could argue for these failure cases and consider that our model may actually retrieve the more relevant video. For example, in the left case, the video retrieved by our model (in the red box) seems to be more relevant to the query text, since both ‘cup’ and ‘talking’ can be seen in our results, while the ‘talking’ can not be seen in the ground truth.

Based on further analysis, we consider that there are also many vague and general annotations in the datasets, such as the example shown in the right case in Fig.3. Such query annotations account for 1-2% of the dataset. We believe our model has a potential to gain in all metrics if such cases get fixed with more discriminative annotations.

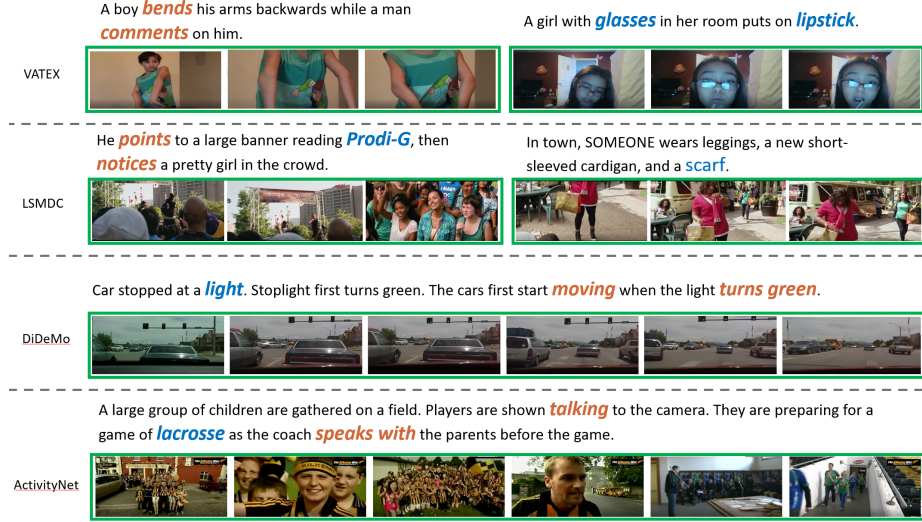


Fig. 1. Examples of Text-Video retrieval pairs from different benchmark datasets



Fig. 2. Visualization of more text-video retrieval examples. We rank the retrieval results based on their similarity scores. Green boxed: the correctly retrieved groundtruth video; Red boxed: incorrectly retrieved videos

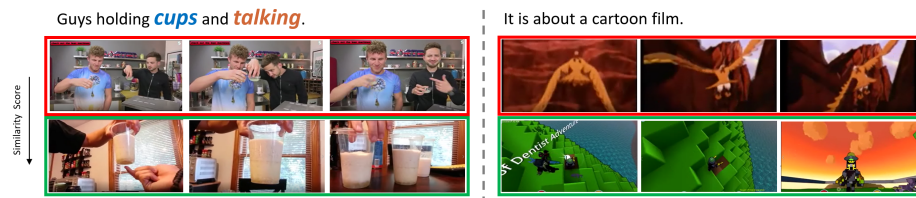


Fig. 3. Visualization of some *failure* text-video retrieval examples. We rank the retrieval results based on their similarity scores. Green boxed: the correctly retrieved groundtruth video; Red boxed: the *incorrectly* retrieved video by our model

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)
2. Bogolin, S.V., Croitoru, I., Jin, H., Liu, Y., Albanie, S.: Cross modal retrieval with querybank normalisation. arXiv preprint arXiv:2112.12777 (2021)
3. Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv preprint arXiv:2109.04290 (2021)
4. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
5. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
6. Gao, Z., Liu, J., Chen, S., Chang, D., Zhang, H., Yuan, J.: Clip2tv: An empirical study on transformer-based methods for video-text retrieval. arXiv preprint arXiv:2111.05610 (2021)
7. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
8. Liu, F., Ye, R.: A strong and robust baseline for text-image matching. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 169–176 (2019)
9. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
10. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision pp. 94–120 (2017)
11. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: Proceedings of the 5th International Conference on Learning Representations (2017)
12. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4581–4591 (2019)
13. 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. pp. 5288–5296 (2016)