Supplementary Material

Learning Linguistic Association Towards Efficient Text-Video Retrieval

Sheng Fang et al.

In this supplementary material, more experimental results are given first in Appendix A. Then we will introduce the detailed experimental settings in Appendix B. Afterward, we provide detailed analysis about our proposed Adaptive Distillation strategy in Appendix C. Finally, some examples of the support sets of our demo dataset and the corresponding attention weights are visualized in Appendix D.

A More Experimental Results

A.1 Complete results on MSVD and VATEX

To save limited space, we do not give a complete version of some experimental results in the main paper. The complete tables on MSVD and VATEX are given in this section. Note that, CE+ is an improved version of CE with higher quality multi-modal features, which is proposed by Croitoru *et al.* [4]. The authors only provide the high-quality features of MSR-VTT and MSVD datasets, so we can only apply our LINAS to CE+ on these two datasets. Although 'TeachText - CE+' achieves the best performance on VATEX dataset, we can not compare LINAS with TeachText on VATEX.

Table 1. Comparison on MSVD dataset.

Method	Text2Video				Video2Text				SumR
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
VSE++[7]	15.4	39.6	53.0	9	-	-	-	-	-
M-Cues[14]	20.3	47.8	61.1	6	-	-	-	-	-
MoEE[13]	21.1	52.0	66.7	5	-	-	-	-	-
CE[11]	21.5	52.3	67.5	5	-	-	-	-	-
Support Set[15]	23.0	52.8	65.8	5	27.3	50.7	60.8	5	280.4
Dual Encoding[5]	11.2	32.6	45.3	13	14.6	29.4	38.5	21	171.5
LINAS - Dual Encoding	12.2	33.9	47.2	12	17.0	33.7	43.4	17	187.5
CE+[4]	25.1	56.5	70.9	**4**	26.3	54.3	66.8	5	299.9
TeachText - CE+[4]	25.1	56.9	71.2	**4**	26.2	55.0	65.6	**4**	300.0
LINAS - CE+	**25.7**	**57.6**	**72.5**	**4**	**27.0**	**55.7**	**68.2**	**4**	**306.7**

Table 2. Comparison on VATEX dataset.

Method	Text2Video				Video2Text				SumR
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
CE[11]	31.1	68.7	80.2	-	41.3	71.0	82.3	-	374.6
W2VV++[10]	32.0	68.2	78.8	-	41.8	75.1	84.3	-	380.2
HGR[3]	35.1	73.5	83.5	2	-	-	-	-	-
TeachText - CE+[4]	**53.2**	**87.4**	**93.3**	**1**	-	-	-	-	-
VSE++[7]	30.4	65.8	76.8	3	42.9	75.1	84.3	**2**	375.2
LINAS - VSE++	33.7	70.1	80.6	3	47.3	78.9	86.1	**2**	396.6
Dual Encoding[5]	34.6	71.3	80.7	2	45.1	75.7	84.6	**2**	392.0
LINAS - Dual Encoding	36.7	72.7	82.2	2	48.0	79.1	**88.0**	**2**	406.6
Hybrid Space[6]	36.8	73.6	83.7	-	46.8	75.7	85.1	-	401.7
LINAS - Hybrid Space	37.7	74.4	83.3	2	**48.6**	**79.4**	86.8	**2**	**410.2**

A.2 Experiments on Text Embeddings

To demonstrate the effectiveness of our LINAS framework, we further do some visualization and statistical analysis on the caption embeddings. Figure 1 and Figure 2 are the visualizations of the text features in the learned common space with t-SNE [12] of 'Dual Encoding' and 'LINAS - Dual Encoding' respectively. We randomly select 10 videos from the test set of MSR-VTT dataset. The points sharing a same color are the captions corresponding to a same video. Different colors in the figure represent different videos. We can observe that points of the same color are closer in Figure 2 compared with Figure 1, which means that our LINAS can make embeddings of relevant captions closer. We also make some statistical analysis to prove this conclusion. After applying LINAS, the average variance of relevant caption embeddings is reduced from 0.66 to 0.6. Since LINAS takes advantage of the complementarity between captions to learn the ability of association, it is not surprising that it can narrow the distance between relevant captions. This result explains why LINAS can improve retrieval performance to a certain extent.

B Detailed Experimental Settings

B.1 Datasets

MSR-VTT is a commonly used text-video retrieval dataset that contains 10,000 videos of about 10 seconds, along with 20 captions per video. The official partition scheme of this dataset [18] has 6,513 videos for training, 497 videos for validation, and 2,990 videos for testing. '1k-A' is another partition edition of

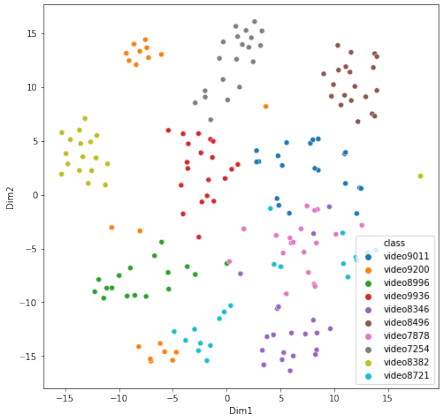

Fig. 1. Dual Encoding

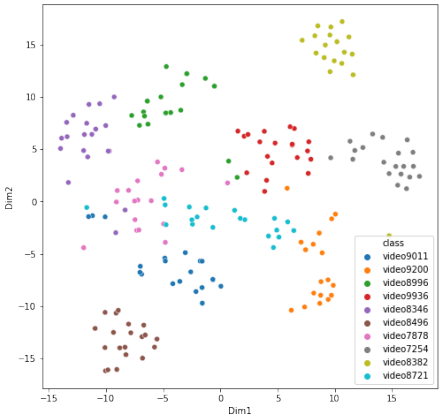

Fig. 2. LINAS - Dual Encoding

MSR-VTT provided by Yu *et al.* [19], which uses 9,000 videos for training and 1,000 for testing. VATEX is a larger multilingual dataset collected from YouTube, which includes 25,991 videos for training, 3,000 for validation, and 6,000 for testing. Each video in VATEX has 10 English captions and 10 Chinese captions. In this work, only English captions are used. Since the annotations of the testing set of VATEX are private, we adopt the partition provided by [3], where the original validation set is split into two equal parts for validation and testing separately. MSVD [2] has a smaller number of videos, but each one has more reference captions. It contains 80,000 English descriptions for a total of 1,970 videos. The standard split of MSVD has 1,200 videos for training, 100 for validation, and 670 for testing.

B.2 Metrics

To measure the performance of retrieval models, we employ commonly used metrics including Recall at K (R@K), Median Rank (MedR) and mean Average Precision (mAP). R@K is the proportion of test queries for which there is at least one desired item in the top-K of the retrieval ranking list. We take K as 1, 5, 10 in our experiments generally. MedR measures the median rank of the first ground truth item in the retrieval. Higher R@K, mAP, and lower MedR indicate a better model. The sum of all R@K scores (SumR) is utilized for overall performance comparison.

B.3 Implementation Details

For fair comparison, we adopt 4096-d ResNeXt-ResNet [9, 17] video features of MSR-VTT and MSVD provided by Dong *et al.*[6], 1024-d I3D [1] video features

of VATEX provided by Wang *et al.*[16]. Moreover, our training strategy is the same with the chosen baseline, *i.e.* learning rate, optimizer, batch size, learning rate decay, etc. The hyperparameters for balancing distillation losses α and β are set to be 0.2 and 1. The δ in Huber loss is set to be 1. Each caption has a support set consists of 8 corresponding descriptions in our experiments. The text encoding, video encoding, and metric learning modules in our student model keep the same with corresponding teacher model, unless otherwise specified.

B.4 Model-agnostic

In the process of applying LINAS to various baseline methods, some practical adaptation is made due to the differences. We will introduce the settings detailedly in this section.

Hybrid Space. Based on Dual Encoding, Hybrid Space [6] further utilizes a concept common space. Since the concept embeddings of both modalities have the frequency-based soft labels as the supervision, our feature-level distillation is only conducted on the latent space. Since the combined similarity matrix is used during inference, the relational distillation is conducted on the hybrid space. The whole distillation loss is

$$\mathcal{L}_D = \alpha * (\mathcal{L}_{D_{text,latent}} + \mathcal{L}_{D_{video,latent}}) + \beta * \mathcal{L}_{D_{rel,hybrid}}. \tag{1}$$

VSE++. We use the 'globalmatch' version of HGR [3] as the 'VSE++' model. The officially released code and features by Chen *et al.* [3] are used in our experiments. The matching mechanism of 'VSE++' is similar to that of 'Dual Encoding', thus the whole distillation loss remains unchanged, that is

$$\mathcal{L}_D = \alpha * (\mathcal{L}_{D_{text}} + \mathcal{L}_{D_{video}}) + \beta * \mathcal{L}_{D_{rel}}. \tag{2}$$

MMT. In MMT [8] method, caption embeddings are first obtained through Bert. Then caption representations of different modalities are obtained through Gated Embedding Modules. For the video branch, the embeddings of different modalities are directly obtained through transformer. The similarities for ranking are the weighted sum of similarities of each modality. We conduct the feature-level distillation on the caption embeddings and the relational distillation on the combined similarity matrix. The whole distillation loss is

$$\mathcal{L}_D = \alpha * \mathcal{L}_{D_{text}} + \beta * \mathcal{L}_{D_{rel,combined}} \tag{3}$$

CLIP4Clip. We adopt the Mean Pooling version of CLIP4Clip as our baseline, which aggregates the frame-level features in an average way for video encoding. Though CLIP4Clip utilizes CLIP for achiving video and text embeddings, it still can be divided into text encoding, video encoding, and metric learning three modules. So the distillation loss on this baseline is the classical one, that is

$$\mathcal{L}_D = \alpha * (\mathcal{L}_{D_{text}} + \mathcal{L}_{D_{video}}) + \beta * \mathcal{L}_{D_{rel}}. \tag{4}$$

B.5 Efficiency

The 'Base' model is actually a pruned version of 'Dual Encoding' [5]. Dual Encoding performs multi-level encoding that represents the rich content of both modalities in a coarse-to-fine fashion for achieving powerful dense representations. In detail, it utilizes mean pooling, biGRU, biGRU-CNN, three encoders and concatenates their outputs for both videos and captions. In the pruned 'Base' model, the video representations are only calculated by the mean pooling and the caption representations are only achieved by biGRU. Afterward, the improved triplet loss [7] is employed for training, which focuses on the hardest negative samples in a minibatch.

C Adaptive Distillation

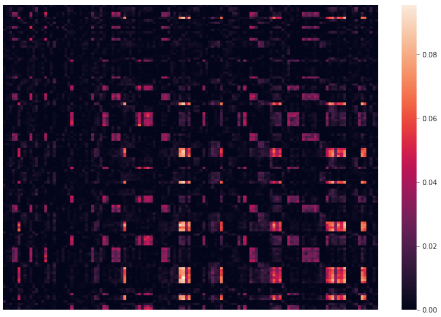

Fig. 3. Distance before reweight

Fig. 4. Distance after reweight

In the early stage of Adaptive Distillation process, the teacher model is well trained, while the student model is randomly initialized. We visualize the distance $L_\delta(S^t(i,j), S^s(i,j))$ in the early stage of training mask in Figure 3. We can see that the elements in the distance matrix are numerically different and the mean value of diagonal elements is higher than that of non-diagonal elements. To balance the influence of numerical value in the process of training mask, we propose a reweight strategy. The distance after reweight $\frac{1}{S^t(i,j)} L_\delta(S^t(i,j), S^s(i,j))$ is shown in Figure 4. It can be found that, the distance matrix is nearly average after reweight. Then our model can learn useful knowledge for distillation without interference.

Moreover, we observe that in the process of mask learning, the numerical difference between diagonal elements and non-diagonal elements is becoming larger and larger. That is because when mask m pays more attention to diagonal elements, the model parameters θ will be optimized in the direction of reducing the diagonal elements in the distance matrix according to $\mathcal{L}_{train}(\theta, m)$. Once the diagonal elements of the distance matrix are reduced, m will further assign

larger weights to the diagonal elements according to $\mathcal{L}_{val}(\theta, m)$. In other words, the Adaptive Distillation for learing the mask m is a positive feedback process. Nevertheless, the information learned in the early stage is reliable, that is, the diagonal elements in the similarity matrix contain positive knowledge that is helpful for distillation.

D Support Sets on Demo Dataset

For validating generalized LINAS, we make a demo dataset based on MSR-VTT as introduced in the main paper. In this part, we visualize some examples of the support sets of the demo dataset constructed with the alternative scheme. Meanwhile, the attention weights of each caption in the support sets are given as well in the aggregation process.

The results are shown in Figure 5. In each piece, the sentence in red and the sequence of pictures represent the query caption and the corresponding video respectively. The captions in the red panel form the support set of the query caption. The darker the block following the caption, the greater its attention weight in the aggregation process.

We can see that the alternative scheme for constructing the support sets does collect some semantically relevant captions. Moreover, with these support set captions, LINAS still captures the complementary information for enriching the semantics of the text representations.

Fig. 5. Visualizations on demo dataset.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)
3. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained text-video retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10638–10647 (2020)
4. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11583–11593 (2021)
5. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9346–9355 (2019)
6. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
7. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. p. 12. BMVA Press (2018), http://bmvc2018.org/contents/papers/0344.pdf
8. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision (ECCV). vol. 12349, pp. 214–229. Springer (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2vv++ fully deep learning for ad-hoc video search. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1786–1794 (2019)
11. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. In: 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019. p. 279. BMVA Press (2019), https://bmvc2019.org/wp-content/uploads/papers/0363-paper.pdf
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
13. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516 (2018)
14. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal text-video retrieval. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 19–27 (2018)
15. Patrick, M., Huang, P., Asano, Y.M., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=EqoXe2zmhrh

16. 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)
17. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
18. 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)
19. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 471–487 (2018)