From Within to Between: Knowledge Distillation for Cross Modality Retrieval

Vinh Tran, Niranjan Balasubramanian, and Minh Hoai

Stony Brook University, Stony Brook, NY 11790, USA
{tquangvinh, niranjan, minhhoai}@cs.stonybrook.edu

1 Introduction

In this supplementary material, we would like to show the full results of our proposed method in comparison to the previous works. We add more metrics for comparison. We show conduct experiments using TEACHTEXT [4]. For each experiment, we report the mean and standard deviation of three randomly seeded runs. We highlight the best performances for each dataset in bold. In addition, we also show results in italic numbers where our proposed approach achieves second best results.

Table 1: **Comparison with the other methods on full split set of the MSRVTT dataset.** All models are trained without using the denoising trick [4]

Method	*Text* → *Video*					
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	Geom↑
VSE [10]	5.0	16.4	24.6	47.0	-	-
VSE++ [8]	5.7	17.1	24.8	65.0	-	-
W2VV [5]	6.1	18.7	27.5	45.0	-	-
M-Cues [14]	7.0	20.9	29.7	38.0	-	-
Dual [6]	7.7	22.0	31.8	32.0	-	-
HGR [3]	9.2	26.2	36.5	24.0	-	-
E2E [6]	9.9	24.0	32.4	29.5	-	-
MEE [13]	$11.1_{\pm 0.1}$	$30.7_{\pm 0.1}$	$42.9_{\pm 0.1}$	$15.0_{\pm 0.0}$	-	-
CE [4]	$11.0_{\pm 0.0}$	$30.8_{\pm 0.1}$	$43.3_{\pm 0.3}$	$15.0_{\pm 0.0}$	$81.8_{\pm 0.2}$	$24.4_{\pm 0.1}$
CE+ [4]	$13.8_{\pm 0.1}$	$36.5_{\pm 0.2}$	$49.4_{\pm 0.4}$	$11.0_{\pm 0.0}$	$69.4_{\pm 0.8}$	$29.2_{\pm 0.2}$
TT-CE [4]	$11.8_{\pm 0.1}$	$32.7_{\pm 0.1}$	$45.3_{\pm 0.1}$	$13.0_{\pm 0.0}$	$74.9_{\pm 0.4}$	$25.9_{\pm 0.1}$
TT-CE+ [4]	$14.6_{\pm 0.0}$	$37.9_{\pm 0.1}$	$50.9_{\pm 0.2}$	$10.0_{\pm 0.0}$	$63.1_{\pm 0.2}$	$30.4_{\pm 0.0}$
Ours+CE+	$\mathbf{14.7}_{\pm 0.1}$	$37.8_{\pm 0.1}$	$50.6_{\pm 0.1}$	$\mathbf{10.0}_{\pm 0.0}$	$65.4_{\pm 0.3}$	$30.4_{\pm 0.1}$
Ours+TT-CE+	$\mathbf{14.7}_{\pm 0.2}$	$\mathbf{38.1}_{\pm 0.1}$	$\mathbf{51.1}_{\pm 0.1}$	$\mathbf{10.0}_{\pm 0.0}$	$\mathbf{62.3}_{\pm 0.1}$	$\mathbf{30.6}_{\pm 0.1}$

To further improve the retrieval performance, we employ the recent Query-bank Normalization with Dynamic Inverted Softmax (DIS) [2]. The results are shown in Table 5.

Table 2: **Comparison with the other methods methods on the ActivityNet dataset**

Method	*Text* → *Video*					
	R@1↑	R@5↑	R@50↑	MdR↓	MnR↓	Geom↑
FSE [16]	18.2	44.8	89.1	-	-	-
MEE [13]	$19.7_{\pm 0.3}$	$50.0_{\pm 0.5}$	$92.0_{\pm 0.2}$	$5.3_{\pm 0.5}$	-	-
HSE [16]	20.5	49.3	-	-	-	-
MMT [9]	$22.7_{\pm 0.2}$	$54.2_{\pm 1.0}$	$93.2_{\pm 0.4}$	$5.0_{\pm 0.0}$	-	-
CE [4]	$19.9_{\pm 0.4}$	$50.1_{\pm 0.8}$	$92.2_{\pm 0.7}$	$5.3_{\pm 0.6}$	$21.3_{\pm 1.1}$	$40.4_{\pm 0.3}$
CE+ [4]	$19.4_{\pm 0.2}$	$49.3_{\pm 0.5}$	$65.4_{\pm 0.4}$	$6.0_{\pm 0.0}$	$22.5_{\pm 0.4}$	$39.7_{\pm 0.0}$
TT-CE [4]	$22.7_{\pm 0.8}$	$56.2_{\pm 0.1}$	$71.6_{\pm 0.8}$	$4.0_{\pm 0.0}$	$15.8_{\pm 0.1}$	$45.0_{\pm 0.6}$
TT-CE+ [4]	$23.5_{\pm 0.2}$	$57.2_{\pm 0.6}$	$\mathbf{73.6}_{\pm 0.2}$	$4.0_{\pm 0.0}$	$\mathbf{13.7}_{\pm 0.1}$	$46.3_{\pm 0.2}$
Ours+CE+	$20.6_{\pm 0.0}$	$50.6_{\pm 0.4}$	$66.9_{\pm 0.1}$	$5.0_{\pm 0.0}$	$19.5_{\pm 0.3}$	$41.1_{\pm 0.1}$
Ours+TT-CE+	$\mathbf{23.9}_{\pm 0.1}$	$\mathbf{57.3}_{\pm 0.4}$	$\mathit{73.5}_{\pm 0.1}$	$\mathbf{4.0}_{\pm 0.0}$	$\mathit{14.0}_{\pm 0.2}$	$\mathbf{46.5}_{\pm 0.0}$

Table 3: **Comparison with the other methods methods on the DiDeMo dataset**

Method	*Text* → *Video*					
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	Geom↑
S2VT [15]	11.9	33.6	-	13.0	-	-
FSE [16]	$13.9_{\pm 0.7}$	$36.0_{\pm 0.8}$	-	$11.0_{\pm 0.0}$	-	-
MEE [13]	$16.1_{\pm 1.0}$	$41.2_{\pm 1.6}$	$55.2_{\pm 1.6}$	$8.3_{\pm 0.5}$	-	-
CE [4]	$17.1_{\pm 0.9}$	$41.9_{\pm 0.2}$	$56.0_{\pm 0.5}$	$8.0_{\pm 0.0}$	$42.8_{\pm 2.8}$	$34.2_{\pm 0.4}$
CE+ [4]	$18.2_{\pm 0.3}$	$43.9_{\pm 1.1}$	$57.1_{\pm 0.9}$	$7.9_{\pm 0.1}$	$38.5_{\pm 3.4}$	$35.8_{\pm 0.4}$
TT-CE [4]	$21.0_{\pm 0.7}$	$47.5_{\pm 1.1}$	$61.9_{\pm 0.6}$	$6.0_{\pm 0.0}$	$35.1_{\pm 1.0}$	$39.5_{\pm 0.5}$
TT-CE+ [4]	$21.6_{\pm 0.8}$	$48.6_{\pm 0.5}$	$\mathbf{62.9}_{\pm 0.7}$	$6.0_{\pm 0.0}$	$\mathbf{31.5}_{\pm 0.8}$	$40.4_{\pm 0.4}$
Ours+CE+	$20.2_{\pm 0.7}$	$45.2_{\pm 0.7}$	$58.8_{\pm 0.8}$	$7.0_{\pm 0.0}$	$41.9_{\pm 2.0}$	$37.7_{\pm 0.4}$
Ours+TT-CE+	$\mathbf{21.7}_{\pm 1.1}$	$\mathbf{49.2}_{\pm 1.0}$	$\mathit{62.4}_{\pm 1.0}$	$\mathbf{5.7}_{\pm 0.6}$	$\mathit{32.4}_{\pm 1.2}$	$\mathbf{40.5}_{\pm 0.2}$

Table 4: **Comparison with the other methods methods on the MSVD dataset.** All models are trained without using the denoising trick [4]

Method	*Text* → *Video*					
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	Geom↑
VSE++ [8]	15.4	39.6	53.0	9.0	-	-
M-Cues [14]	20.3	47.8	61.1	6.0	-	-
MEE [13]	$21.1_{\pm 0.2}$	$52.0_{\pm 0.7}$	$66.7_{\pm 0.2}$	$5.0_{\pm 0.0}$	-	-
CE [4]	$21.5_{\pm 0.5}$	$52.3_{\pm 0.8}$	$67.5_{\pm 0.7}$	$5.0_{\pm 0.0}$	$20.4_{\pm 0.0}$	$42.3_{\pm 0.6}$
CE+ [4]	$25.1_{\pm 0.9}$	$56.5_{\pm 1.4}$	$70.9_{\pm 1.6}$	$4.0_{\pm 0.0}$	$17.8_{\pm 0.6}$	$46.5_{\pm 1.0}$
TT-CE [4]	$22.1_{\pm 0.4}$	$52.2_{\pm 0.5}$	$67.2_{\pm 0.6}$	$5.0_{\pm 0.0}$	$19.6_{\pm 0.5}$	$42.6_{\pm 0.4}$
TT-CE+ [4]	$25.1_{\pm 0.6}$	$56.8_{\pm 0.6}$	$71.2_{\pm 0.6}$	$4.0_{\pm 0.0}$	$16.8_{\pm 0.3}$	$46.6_{\pm 0.5}$
Ours+CE+	$\mathbf{26.0}_{\pm 0.1}$	$\mathbf{58.3}_{\pm 0.1}$	$\mathbf{72.9}_{\pm 0.3}$	$\mathbf{4.0}_{\pm 0.0}$	$\mathbf{16.1}_{\pm 0.1}$	$\mathbf{47.9}_{\pm 0.0}$
Ours+TT-CE+	$\mathit{25.5}_{\pm 0.2}$	$\mathit{57.1}_{\pm 0.1}$	$\mathit{71.7}_{\pm 0.2}$	$\mathbf{4.0}_{\pm 0.0}$	$\mathit{16.3}_{\pm 0.1}$	$\mathit{47.1}_{\pm 0.1}$

Table 5: **Comparison with other methods on the DiDeMo dataset.**

Method	R@1↑	R@5↑	R@10↑	MnR↓
S2VT [15]	11.9	33.6	-	-
FSE [16]	13.9	36.0	-	-
MEE [13]	16.1	41.2	55.2	43.7
CE [4]	17.1	41.9	56.0	-
TT-CE [4]	21.0	47.5	61.9	-
CE+ [4]	18.2	43.9	57.1	-
ClipBERT [11]	20.4	48.0	60.8	-
TT-CE+ [4]	21.6	48.6	62.9	-
Frozen [1]	34.6	65.0	74.7	-
MDMMT [7]	38.9	69.0	79.7	-
CLIP4Clip [12] (reported in [12])	43.4	70.2	**80.6**	**17.5**
CLIP4Clip-rerun (frozen layers + smaller batches)	42.0	69.1	78.1	18.8
CLIP4Clip-rerun + Caption Distillation (Ours)	43.2	69.7	79.2	**17.5**
CLIP4Clip-rerun + Video Distillation (Ours)	43.2	69.2	79.3	17.9
CLIP4Clip-rerun + Caption Distillation (Ours) + DIS [2]	**45.2**	69.8	79.5	**17.7**
CLIP4Clip-rerun + Video Distillation (Ours) +DIS [2]	45.0	**70.6**	80.0	17.8

2 Experimental results with frozen all encoder

We have performed the suggested experiments with CLIP4Clip on the Didemo dataset. The results are shown in Table 6. Our proposed distillation improves the retrieval performance in this setting.

Table 6: **Experimental results with frozen layers on Didemo dataset.**

Method	R@1↑	R@5↑	R@10↑	MnR↓
CLIP4Clip (frozen encoder)	38.5	**67.6**	76.4	21.8
CLIP4Clip (frozen encoder) + Caption Distillation (Ours)	40.1	67.3	**77.6**	20.6
CLIP4Clip (frozen encoder) + Video Distillation (Ours)	**40.5**	66.9	76.7	**20.5**

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the International Conference on Computer Vision (2021)
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. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (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 International Conference on Computer Vision (2021)
5. Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia **20**(12), 3377–3388 (2018)
6. Dong, J., Li, X., Xu, C., Ji, S., Wang, X.: Dual dense encoding for zero-example video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
7. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
8. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. In: Proceedings of the British Machine Vision Conference (2018)
9. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal Transformer for Video Retrieval. In: Proceedings of the European Conference on Computer Vision (2020)
10. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
11. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
12. 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)
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 video-text retrieval. In: ACM International Conference on Multimedia Retrieval (2018)
15. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)
16. Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: Proceedings of the European Conference on Computer Vision (2018)