Supplementary Material: Multi-modal Transformer for Video Retrieval

Valentin Gabeur[1,2], Chen Sun[2], Karteek Alahari[1], Cordelia Schmid[2]

[1] Inria*, karteek.alahari@inria.fr
[2] Google Research, {valgab,chensun,cordelias}@google.com

.1 Model complexity

Number of parameters. As shown below, using multiple modalities does not impact the number of parameters significantly. Interestingly, majority of the parameters correspond to the BERT caption encoding module. We also note that the difference in the video encoder comes from the projections. The number of parameters of a transformer encoder is independent of the number of input embeddings, as are the parameters of a CNN from the image size.

Our cross-modal architecture using 7 modalities has: 133.3M parameters, including caption encoder: 112.9M, video encoder: 20.4M (Projections: 3.3M, MMT: 17.1M). Our cross-modal architecture using 2 modalities has: 127.3M parameters, including caption encoder: 109.6M (decrease compared to 7 modalities due to using less gated embedding modules), video encoder: 17.7M (Projections: 0.6M, MMT: 17.1M).

Training and inference times. Training our full cross-modal architecture from scratch on MSRVTT takes about 4 hours on a single V100 16GB GPU.

If we replace our multi-modal transformer by collaborative gating [3], we reduce the number of parameters from 133.3M to 123.9M. However, the gain in inference time is minimal, from 1.1s to 0.8s, and is negligible compared to feature extraction, as detailed below.

Inference time for 1k videos and 1k text queries from MSRVTT on a single V100 GPU is as follows: approximately 3000s to extract features of 7 experts on 1k videos (480s just for S3D motion features), 1.1s to process videos with MMT, 0.9s to process 1k captions with BERT+gated embedding modules, 0.05s to compute similarities and rank the video candidates for the 1k queries.

.2 Results on additional metrics

Here, we report our results for the additional metrics R@1, R@10, R@50. Table 1 complements the results reported for the MSRVTT [8] dataset in Table **??** of the main paper. Similarly, Table 2 and Table 3 report the additional evaluations for Table **??** and Table **??** of the main paper on ActivityNet [2] and LSMDC [6] datasets respectively. We observe that the results on these additional metrics are in line with the conclusions of the main paper.

[*] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

Table 1: Retrieval performance on the MSRVTT dataset. 1k-A and 1k-B denote test sets of 1000 randomly sampled caption-video pairs used in [9] and [4] resp.

		Text $\longrightarrow$ *Video*					*Video* $\longrightarrow$ *Text*				
Method	Split	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Random baseline	1k-A	0.1	0.5	1.0	500.0	500.0	0.1	0.5	1.0	500.0	500.0
JSFusion [9]	1k-A	10.2	31.2	43.2	13	-	-	-	-	-	-
HT [5]	1k-A	12.1	35.0	48.0	12	-	-	-	-	-	-
CE [3]	1k-A	20.9$_{\pm 1.2}$	48.8$_{\pm 0.6}$	62.4$_{\pm 0.8}$	6.0$_{\pm 0.0}$	28.2$_{\pm 0.8}$	20.6$_{\pm 0.6}$	50.3$_{\pm 0.5}$	64.0$_{\pm 0.2}$	5.3$_{\pm 0.6}$	25.1$_{\pm 0.8}$
Ours	1k-A	**24.6**$_{\pm 0.4}$	**54.0**$_{\pm 0.2}$	**67.1**$_{\pm 0.5}$	**4.0**$_{\pm 0.0}$	**26.7**$_{\pm 0.9}$	**24.4**$_{\pm 0.5}$	**56.0**$_{\pm 0.9}$	**67.8**$_{\pm 0.3}$	**4.0**$_{\pm 0.0}$	**23.6**$_{\pm 1.0}$
HT-pretrained [5]	1k-A	14.9	40.2	52.8	9	-	-	-	-	-	-
Ours-pretrained	1k-A	**26.6**$_{\pm 1.0}$	**57.1**$_{\pm 1.0}$	**69.6**$_{\pm 0.2}$	**4.0**$_{\pm 0.0}$	**24.0**$_{\pm 0.8}$	**27.0**$_{\pm 0.6}$	**57.5**$_{\pm 0.6}$	**69.7**$_{\pm 0.8}$	**3.7**$_{\pm 0.5}$	**21.3**$_{\pm 0.6}$
Random baseline	1k-B	0.1	0.5	1.0	500.0	500.0	0.1	0.5	1.0	500.0	500.0
MEE [4]	1k-B	13.6	37.9	51.0	10.0	-	-	-	-	-	-
JPose [7]	1k-B	14.3	38.1	53.0	9	-	16.4	41.3	54.4	8.7	-
MEE-COCO [4]	1k-B	14.2	39.2	53.8	9.0	-	-	-	-	-	-
CE [3]	1k-B	18.2$_{\pm 0.7}$	46.0$_{\pm 0.4}$	60.7$_{\pm 0.2}$	7.0$_{\pm 0.0}$	35.3$_{\pm 1.1}$	18.0$_{\pm 0.8}$	46.0$_{\pm 0.5}$	60.3$_{\pm 0.5}$	6.5$_{\pm 0.5}$	30.6$_{\pm 1.2}$
Ours	1k-B	**20.3**$_{\pm 0.5}$	**49.1**$_{\pm 0.4}$	**63.9**$_{\pm 0.5}$	**6.0**$_{\pm 0.0}$	**29.5**$_{\pm 1.6}$	**21.1**$_{\pm 0.4}$	**49.4**$_{\pm 0.4}$	**63.2**$_{\pm 0.4}$	**6.0**$_{\pm 0.0}$	**24.5**$_{\pm 1.8}$

Table 2: Retrieval performance on the ActivityNet dataset.

	Text $\longrightarrow$ *Video*					*Video* $\longrightarrow$ *Text*				
Method	R@1↑	R@5↑	R@50↑	MdR↓	MnR↓	R@1↑	R@5↑	R@50↑	MdR↓	MnR↓
Random baseline	0.02	0.1	1.02	2458.5	2458.5	0.02	0.1	1.02	2458.5	2458.5
FSE [11]	18.2$_{\pm 0.2}$	44.8$_{\pm 0.4}$	89.1$_{\pm 0.3}$	7	-	16.7$_{\pm 0.8}$	43.1$_{\pm 1.1}$	88.4$_{\pm 0.3}$	7	-
CE [3]	18.2$_{\pm 0.3}$	47.7$_{\pm 0.6}$	91.4$_{\pm 0.4}$	6.0$_{\pm 0.0}$	23.1$_{\pm 0.5}$	17.7$_{\pm 0.6}$	46.6$_{\pm 0.7}$	90.9$_{\pm 0.2}$	6.0$_{\pm 0.0}$	24.4$_{\pm 0.5}$
HSE [11]	20.5	49.3	-	-	-	18.7	48.1	-	-	-
Ours	**22.7**$_{\pm 0.2}$	**54.2**$_{\pm 1.0}$	**93.2**$_{\pm 0.4}$	**5.0**$_{\pm 0.0}$	**20.8**$_{\pm 0.4}$	**22.9**$_{\pm 0.8}$	**54.8**$_{\pm 0.4}$	**93.1**$_{\pm 0.2}$	**4.3**$_{\pm 0.5}$	**21.2**$_{\pm 0.5}$
Ours-pretrained	**28.7**$_{\pm 0.2}$	**61.4**$_{\pm 0.2}$	**94.5**$_{\pm 0.0}$	**3.3**$_{\pm 0.5}$	**16.0**$_{\pm 0.4}$	**28.9**$_{\pm 0.2}$	**61.1**$_{\pm 0.2}$	**94.3**$_{\pm 0.4}$	**4.0**$_{\pm 0.0}$	**17.1**$_{\pm 0.5}$

Table 3: Retrieval performance on the LSMDC dataset.

	Text $\longrightarrow$ *Video*					*Video* $\longrightarrow$ *Text*				
Method	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Random baseline	0.1	0.5	1.0	500.0	500.0	0.1	0.5	1.0	500.0	500.0
CT-SAN [10]	5.1	16.3	25.2	46	-	-	-	-	-	-
JSFusion [9]	9.1	21.2	34.1	36	-	-	-	-	-	-
CCA [1] (rep. by [4])	7.5	21.7	31.0	33	-	-	-	-	-	-
MEE [4]	9.3	25.1	33.4	27	-	-	-	-	-	-
MEE-COCO [4]	10.1	25.6	34.6	27	-	-	-	-	-	-
CE [3]	11.2$_{\pm 0.4}$	26.9$_{\pm 1.1}$	34.8$_{\pm 2.0}$	25.3$_{\pm 3.1}$	-	-	-	-	-	-
Ours	**13.2**$_{\pm 0.4}$	**29.2**$_{\pm 0.8}$	**38.8**$_{\pm 0.9}$	**21.0**$_{\pm 1.4}$	**76.3**$_{\pm 1.9}$	**12.1**$_{\pm 0.1}$	**29.3**$_{\pm 1.1}$	**37.9**$_{\pm 1.1}$	**22.5**$_{\pm 0.4}$	**77.1**$_{\pm 2.6}$
Ours-pretrained	**12.9**$_{\pm 0.1}$	**29.9**$_{\pm 0.7}$	**40.1**$_{\pm 0.8}$	**19.3**$_{\pm 0.2}$	**75.0**$_{\pm 1.2}$	**12.3**$_{\pm 0.2}$	**28.6**$_{\pm 0.3}$	**38.9**$_{\pm 0.8}$	**20.0**$_{\pm 0.0}$	**76.0**$_{\pm 0.8}$

References

1. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: CVPR (2015)
2. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: ICCV (2017)
3. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. ArXiv **abs/1907.13487** (2019)
4. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. ArXiv **abs/1804.02516** (2018)
5. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)
6. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: CVPR (2015)

7. Wray, M., Larlus, D., Csurka, G., Damen, D.: Fine-grained action retrieval through multiple parts-of-speech embeddings. In: ICCV (2019)
8. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: CVPR (2016)
9. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: ECCV (2018)
10. Yu, Y., Ko, H., Choi, J., Kim, G.: End-to-end concept word detection for video captioning, retrieval, and question answering. CVPR (2017)
11. Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: ECCV (2018)