Supplementary Materials for Multi-Query Video Retrieval

Zeyu Wang, Yu Wu, Karthik Narasimhan, and Olga Russakovsky

Princeton University
`{zeyuwang,yuwu,karthikn,olgarus}@cs.princeton.edu`

A Additional experiment results

A.1 Five-query experiment for Frozen

Due to space limit, only MSR-VTT [6] results are shown in Table 1 for Frozen [1] model. Here we include the results on MSVD [2] and VATEX [5] in Table 3. Similar findings can be drawn compared to the results on CLIP4Clip [3] model: the *similarity aggregation* outperforms *rank aggregation*. Dedicated multi-query training outperforms post-hoc inference methods. And that *weighted feature* training introduces additional improvement over *mean feature* training.

Table 3: Performance of different multi-query retrieval methods on MSVD and VATEX datasets with Frozen [1] backbone. The baseline is trained and evaluated with one query. Others are evaluated with five-query input. RA, SA are trained with one query. MF, TS-WF, LG-WF, and CG-WF are trained with five-query input. All numbers are the average over 100 evaluations with different query samples. Recall numbers are reported in percent.

	MSVD [6] (Frozen)					Vatex [5] (Frozen)				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
Baseline	39.6	70.5	80.7	2.0	13.9	26.7	56.9	70.4	4.0	26.4
RA	42.5	75.5	85.9	2.0	6.2	37.1	69.8	82.1	2.0	11.7
SA	53.2	85.3	93.0	1.0	3.7	40.6	72.2	83.8	2.0	10.0
MF	55.8	86.5	93.8	1.0	3.7	49.0	79.2	88.3	2.0	8.1
TS-WF	56.5	**87.1**	**93.9**	1.0	3.7	**49.7**	**79.7**	**88.7**	1.8	7.9
LG-WF	56.4	86.3	93.4	1.0	3.9	49.5	79.4	88.5	1.9	8.2
CG-WF	**56.7**	86.8	93.8	1.0	3.8	49.4	79.3	88.5	1.9	8.6

A.2 Concatenation method for MQVR

Another straightforward method for handling multiple queries is to concatenate them together into a long sentence. The benefit of this approach compared to *mean feature* and *weighted feature* methods discussed in section 4 is that it can fully utilize the ability of (pretrained) language models, *i.e.,* information contained in multi-query is extracted directly by the language model, instead of using language model to extract information in each query separately and then combine them together. However, there are two downsides. First, this approach is limited by the maximum sequence length the language model can handle and concatenating multiple queries can quickly exceed that limit. For example, CLIP [4] is pretrained with max sequence length of 77. So models built on it, like CLIP4clip [3], can't handle longer sentence[1]. Second, as computation cost for transformer-based language model scales quadratically with increase of sequence length, *concatenation* method is less computation efficient.

Figure 7 shows the comparison of *concatenation* with *mean feature* and *contextualized weight generation* for CLIP4clip [3] model trained with five-query [6][2]. As expected, *concatenation* outperforms the other two at low-query region. However, the performance stops increasing when more queries are available, at which point the concatenated sentence is longer than the max length the language model can handle.

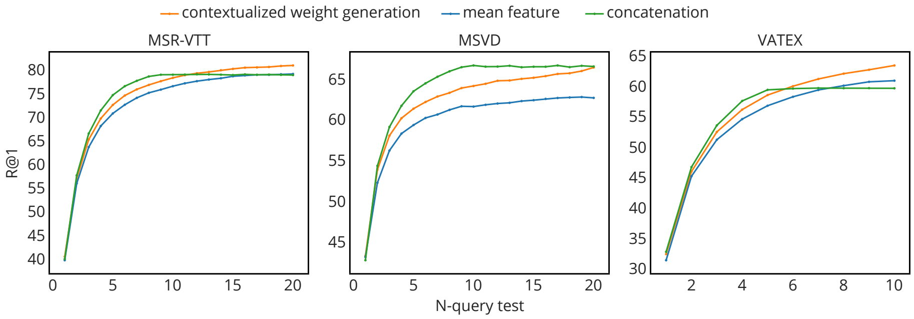


Fig. 7: R@1 performance for *mean feature*, *contextualized weight generation*, and *concatenation* when evaluated with varying number of queries.

B More qualitative examples

Additional qualitative examples of generated weights for different queries are shown in Figure 8. The weights correctly captures the relative quality among queries by giving higher weights to those containing more information.

---

[1] We tried to extend the max length the model can handle by extending the pretrained positional embedding, but that leads to worse results.

[2] When the input sentence is longer than the max sequence length which can be processed by the language model, the sentence is cut off at max length and only the leading tokens are input to the model.

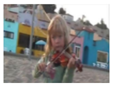
Video	**Query**	**Weight**	**Single-query rank**	**Multi-query rank**
	A person slicing up vegetables in their kitchen and preparing food	0.44	27	4
	In a kitchen a person is showing how to chop a large onion	**0.56**	1	
	A man is singing a song and playing guitar	0.39	20	2
	A man is singing and playing guitar with people dancing around him followed by two women handing out flowers	**0.61**	2	
	A dog and toddler are playing	0.48	1	1
	A baby climbs on top of a sleeping dog	**0.52**	1	
	A man is doing motor bike stunts	0.26	4	1
	A guy is riding a motocycle	0.32	3	
	A man is riding a motorcycle while lying back	**0.42**	1	
	The man mixed oxi clean with water	0.20	2	1
	A person fills a sink with water and soap	0.34	2	
	A person is pouring oxi clean powder into a sink full of water	**0.46**	1	
	A boy jumps around on some large rocks outdoors	0.27	1	1
	A young boy jumps on a some rocks in a park	0.31	1	
	A little boy wearing green tank top is walking on top of big rocks	**0.42**	1	
	A girl are singing a song	0.13	13	1
	A girl is playing the violin	0.23	1	
	A small girl standing on sand is playing the violin	0.30	1	
	A young girl is playing a violin on the beach	**0.34**	1	
	There is a man cooking a dish in the kitchen	0.08	34	4
	The person adds the liquid to the pot and stir with the leveler	0.29	4	
	A person adds broth to a pot and mixes it together	0.29	2	
	Broth is being added to a soup pot and stirred with a rubber spatula	**0.34**	4	
	A woman is cooking food	0.11	7	1
	Someone is making food	0.12	16	
	A woman is preparing and serving food in a kitchen	0.15	3	
	A woman shows how to prepare a pastry	0.28	1	
	A woman is demonstrate how to make cinnamon twists	**0.34**	1	
	A man cooking his kitchen	0.07	71	3
	A woman is sprinkling something on food	0.17	2	
	The woman is seasoning the shrimp and vegetables	0.25	8	
	A woman spices some seafood	0.25	7	
	The woman is seasoning the seafood and vegetables	**0.26**	3	

Fig. 8: Qualitative examples of a *contextualized weight generation* model trained with five queries.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. arXiv:2104.00650 (2021)
2. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: ACL (2011)
3. 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:2104.08860 (2021)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv:2103.00020 (2021)
5. 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: ICCV (2019)
6. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: CVPR (2016)