Temporal-Relational CrossTransformers for Few-Shot Action Recognition Supplementary Material

Toby Perrett Alessandro Masullo Tilo Burghardt Majid Mirmehdi Dima Damen
<first>.<last>@bristol.ac.uk Department of Computer Science, University of Bristol, UK

X-Shot results

In the main paper, we introduced Temporal-Realational CrossTransformers (TRX) for few-shot action recognition. They are designed specifically for K-shot problems where $K > 1$, as TRX is able to match sub-sequences from the query against sub-sequences from multiple support set videos.

Table 1 in the main paper shows results on the standard 5-way 5-shot benchmarks on Kinetics [3], Something-Something V2 (SSv2) [4], HMDB51 [5] and UCF101 [6]. For completeness we also provide 1-, 2-, 3-, 4- and 5-shot results for TRX with Ω=$\{1\}$ (*i.e.* frame-to-frame comparisons) and Ω=$\{2, 3\}$ (*i.e.* pair and triplet comparisons) on the large-scale datasets Kinetics and SSv2. These are in Table 1 in this supplementary, where we also list results from all other works which provide these scores.

For 1-shot, in Kinetics, TRX performs similarly to recent few-shot action-recognition methods [8, 1, 7], but these are all outperformed by OTAM [2]. OTAM works by finding a strict alignment between the query and single support set video per class. It does not scale as well as TRX when $K > 1$, shown by TRX performing better on the 5-shot benchmark. This is because TRX is able to match query sub-sequences against similar sub-sequences in the support set, and importantly ignore sub-sequences (or whole videos) which are not as useful. Compared to the strict alignment in OTAM [2], where the full video is considered in the alignment, TRX can exploit several sub-sequences from the same video, ignoring any distractors. Despite not being as well suited to 1-shot problems, on SSv2 TRX performs similarly to OTAM. 2-shot TRX even outperforms 5-shot OTAM. Table 1 again highlights the importance of tuples, shown in the main paper, where TRX with Ω=$\{2, 3\}$ consistently outperforms Ω=$\{1\}$.

Figure 5 in the main paper shows how TRX scales on SSv2 compared to CMN [8, 9], which also provides X-shot results ($1 \leq X \leq 5$). The equivalent graph for Kinetics is shown in Fig. 1 here. This confirms TRX scales better as the shot increases. There is less of a difference between TRX with Ω=$\{1\}$ and Ω=$\{2, 3\}$, as Kinetics requires less temporal knowledge to discriminate between the classes than SSv2 (ablated in Sec. 4.3.1 and 4.3.2 in the main paper).

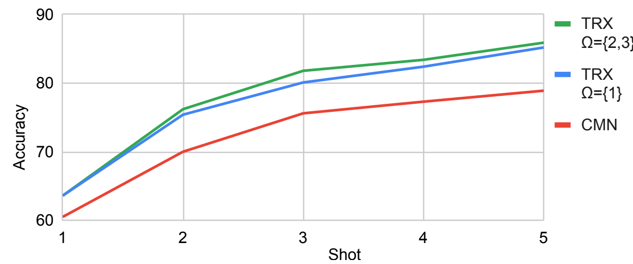

Figure 1: Comparing CMN [9] results to TRX for X-shot 5-way, for $1 \leq X \leq 5$ on Kinetics. TRX benefits from increasing the number of of videos in the support set, both for Ω=$\{1\}$ and Ω=$\{2, 3\}$.

The impact of positional encoding

TRX adds positional encodings to the individual frame representations before concatenating them into tuples. Table 2 shows that adding positional encodings improves SSv2 for both single frames and higher-order tuples (by +0.3% and +0.6% respectively). For Kinetics, performance stays the same as single frames and improves slightly with tuples (+0.4%) for the proposed model. Overall, positional encoding improves the results marginally for TRX.

References

[1] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition. In *British Machine Vision Conference*, 2019. 1, 2

[2] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-Shot Video Classification via Temporal Alignment. In *Computer Vision and Pattern Recognition*, 2020. 1, 2

[3] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Computer Vision and Pattern Recognition*, 2017. 1

[4] Raghav Goyal, Vincent Michalski, Joanna Materzy, Susanne Westphal, Heuna Kim, Valentin Haenel, Peter Yianilos, Moritz Mueller-freitag, Florian Hoppe, Christian Thurau,

Dataset	Method	Shot 1	2	3	4	5
Kinetics	CMN [8]	60.5	-	-	-	78.9
	CMN-J [9]	60.5	70.0	75.6	77.3	78.9
	TARN [1]	64.8	-	-	-	78.5
	ARN [7]	63.7	-	-	-	82.4
	OTAM [2]	**73.0**	-	-	-	85.8
	Ours - TRX Ω={1}	63.6	75.4	80.1	82.4	85.2
	Ours - TRX Ω={2, 3}	63.6	**76.2**	**81.8**	**83.4**	**85.9**
SSv2*	CMN-J [9]	**36.2**	42.1	44.6	47.0	48.8
	Ours - TRX Ω={1}	34.9	43.4	47.6	50.9	53.3
	Ours - TRX Ω={2, 3}	36.0	**46.0**	**51.9**	**54.9**	**59.1**
SSv2†	OTAM [2]	**42.8**	-	-	-	52.3
	Ours - TRX Ω={1}	38.8	49.7	54.4	58.0	60.6
	Ours - TRX Ω={2, 3}	42.0	**53.1**	**57.6**	**61.1**	**64.6**

Table 1: Comparison to few-shot video works on Kinetics (split from [9]) and Something-Something V2 (SSv2) (†: split from [9] *: split from [2]). Results are reported as the shot, *i.e.* number of support set videos per class, increases from 1 to 5. -: Results not available in published works.

Method	Positional Encoding	Kinetics	SSv2†
Ω={1}	×	85.2	53.0
Ω={1}	✓	85.2	53.3
Ω={2, 3}	×	85.5	58.5
Ω={2, 3}	✓	**85.9**	**59.1**

Table 2: The importance of incorporating positional encoding for single frames and the proposed model Ω={2, 3}.

Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *International Conference on Computer Vision*, 2017. 1

[5] H Kuehne, T Serre, H Jhuang, E Garrote, T Poggio, and T Serre. HMDB: A large video database for human motion recognition. In *International Conference on Computer Vision*, nov 2011. 1

[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv*, 2012. 1

[7] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H S Torr, and Piotr Koniusz. Few-shot Action Recognition with Permutation-invariant Attention. In *European Conference on Computer Vision*, 2020. 1, 2

[8] Linchao Zhu and Yi Yang. Compound Memory Networks for Few-Shot Video Classification. In *European Conference on Computer Vision*, 2018. 1, 2

[9] Linchao Zhu and Yi Yang. Label Independent Memory for Semi-Supervised Few-shot Video Classification. *Transactions on Pattern Analysis and Machine Intelligence*, 14(8), 2020. 1, 2