Video Action Transformer Network: Appendix

Rohit Girdhar[1]* João Carreira[2] Carl Doersch[2] Andrew Zisserman[2,3]
[1]Carnegie Mellon University [2]DeepMind [3]University of Oxford
http://rohitgirdhar.github.io/ActionTransformer

#layers↓ #heads→	2	3	6
2	27.4	28.7	27.6
3	28.5	28.8	27.7
6	29.1	28.3	26.5

Table 1: **Ablating the number of heads and layers.** We find fewer heads and more layers tends to give slightly better performance. All performance reported with Action Transformer head, when using GT boxes as proposals.

Trunk	Head	QPr	GT Boxes	Params (M)	Val mAP
I3D	I3D	-		16.2	21.3
I3D	I3D	-	✓	16.2	23.4
I3D	Tx	LowRes	✓	13.9	28.5
R3D [2]	Tx	LowRes	✓	17.7	26.6
R3D + NL [2]	Tx	LowRes	✓	25.1	27.2

Table 2: **Different trunk architectures.** Our model is compatible with different trunk architectures, such as R3D or Non-Local network proposed in [2]. We observed best performance with I3D, so use it for all experiments in the paper.

1. Additional Ablations

Number of heads/layers in Action Transformer: Our Action Transformer architecture is designed to be easily stacked into multiple heads per layer, and multiple layers, similar to the original unit [1]. We evaluate the effect of changing the number of heads and layers in Table 1. We find the performance to be largely similar, though tends to get slightly better with more layers and fewer heads. Hence, we stick with our default 2-head 3-layer model for all experiments reported in the paper.

Swapping out the trunk architecture: As we observe in Table 2, our model is compatible with different trunk architectures. We use I3D for all experiments in the paper given its speed and strong performance.

2. Visualizations

Video: We visualize the embeddings, attention maps and predictions in the attached video (`combined.mp4`).

Per-class top predictions: We visualize the top predictions on the validation set for each class, sorted by confidence, in the attached PDF (`pred.pdf`).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 1

[2] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 1

*Work done during an internship at DeepMind