Supplementary Materials for Paper "Temporal-attentive Covariance Pooling Networks for Video Recognition"

Zilin Gao[†], **Qilong Wang**[‡], **Bingbing Zhang**[†], **Qinghua Hu**[‡], **Peihua Li**[†*]

[†]School of Information and Communication Engineering, Dalian University of Technology
[‡]College of Intelligence and Computing, Tianjin University
gzl@mail.dlut.edu.cn, qlwang@tju.edu.cn, icyzhang@mail.dlut.edu.cn
huqinghua@tju.edu.cn, peihuali@dlut.edu.cn

A Details of Kinetics-400 and Mini-Kinetics-200

Kinetics-400 [1] is a large-scale dataset containing 400 action categories, in which training set and validation set have 246K and 20K videos, respectively. The dataset is released by providing YouTube links. Because some links are broken, we use the dataset collected by [2], which has 234,643 and 19,761 videos for training and validation in total, respectively.

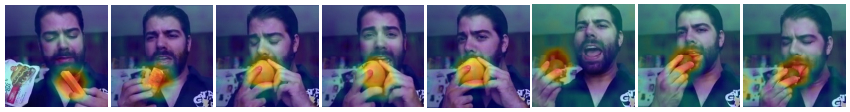
Mini-Kinetics-200 [3] dataset involving of 200 categories is a subset of Kinetics-400, where 400 and 25 videos are used for training and validation per category, respectively. We use the same categories and videos sampling strategy in [3]. The full dataset contains 80K training videos and 5K validation videos. Since broken links, we collect 77,161 and 4,988 videos for training and validation in total.

B Visualization of Attention Maps Learned by TCP

To give a quality analysis of TCP, we visualize results of the learnt temporal attention using some example videos on K-400. The qualitative results are shown in Figure S1, where we can see that the attention module in our TCP can effectively focus on key moving parts (e.g., hands and food in "tasting food" Figure S1(a)) for action recognition, while suppressing the remaining irrelevant regions.

*Corresponding author.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

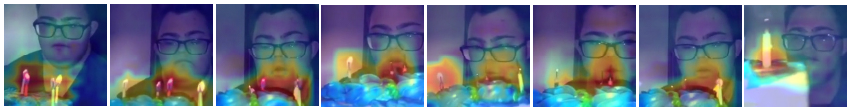


(a) Tasting food

(b) Snatch weight lifting

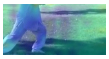
(c) Riding unicycle

(d) Clapping



(e) Blowing out candles

(f) Playing kickball



(g) Tai chi

Figure S1: Visualization of learned attention maps of some video examples using our TCP on K-400.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*, 2017.

[2] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[3] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018.