Supplementary Materials for Paper ID 17 "Temporal-attentive Covariance Pooling Networks for Video Recognition"

Anonymous Author(s)
Affiliation
Address
email

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.

κ	1	3	5	7
Top-1 Acc (%)	75.6	76.0	76.1	76.0
Top-5 Acc (%)	92.5	92.7	92.7	92.6

Table A1: Results (8 frames) of TCPNet based on 2D ResNet50 with various κ on Mini-K200.

κ	1	3	5	7	9	11	13	15
Top-1 Acc (%)	77.0	77.0	77.4	77.5	77.7	77.3	76.9	76.8
Top-5 Acc (%)	93.0	93.1	93.3	93.3	93.5	93.1	93.0	93.0

Table A2: Results (16 frames) of TCPNet based on 2D ResNet50 with various κ on Mini-K200.

B Effect of Kernel Size κ on TC-COV$_\kappa$

Our temporal covariance pooling TC-COV$_\kappa$ has a parameter, i.e., kernel size κ. Here we assess effect of kernel size κ on TC-COV$_\kappa$. Specifically, we conduct experiments on Mini-K200, where 8 frames or 16 frames of 112×112 images are used as inputs. The experimental settings are consistent with those in Section 4.1, and the results using 1 clip $\times$ 1 crop are reported for comparison. For 8 frames input, kernel size κ varies from 1 to 7. For 16 frames input, κ varies from 1 to 15. As listed in Table A1 and Table A2, TCPNet based on 2D ResNet50 with $\kappa = 5$ and $\kappa = 9$ respectively achieves the best results for 8 frames input and 16 frames input. They obtain 0.5% and 0.7% gains over those with $\kappa = 1$, respectively. These results show our temporal covariance pooling captures both intra-frame cross-correlation and inter-frame correlation, outperforming those only consider inter-frame

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

covariance pooling. The performance gradually decreases when $\kappa > 5$ and $\kappa > 9$, interaction among too large range will bring side effect on performance, while increasing computation cost.

References

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

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

[3] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018.