Are Spatial-Temporal Graph Convolution Networks for Human Action Recognition Over-Parameterized?

Jianyang Xie[1], Yitian Zhao[2], Yanda Meng[3], He Zhao[1], Anh Nguyen[1], Yalin Zheng[1*]
[1] University of Liverpool, UK. [2]Ningbo Institute of Materials Technology and Engineering, CAS, China. [3]University of Exeter, UK.
{Jianyang.Xie, yzheng}@liverpool.ac.uk

1. Appendix / supplemental material

1.1. Implementations Details

All experiments are conducted on one A100 GPU with the PyTorch deep learning framework. All models are trained for 100 epochs with the Cosine Annealing learning rate scheduler by using SGD with momentum 0.9, weight decay $5e^{-4}$. The initial learning rate was set to 0.1. The batch size was set to 128. To accelerate the training process, the input of temporal length was set to 64 in the ablation study. For a fair comparison, the input of temporal length was set to 100 when comparing the stare-of-the-arts. The pre-processing approach follows the setting in [3].

1.2. Datasets

NTU RGB+D 60 [6]. The action samples are performed by 40 volunteers and categorized into 60 classes. Each sample contains an action and is guaranteed to have at most 2 subjects, which are captured by three Microsoft Kinect v2 cameras from different views concurrently. Two benchmarksre were commend : (1) cross-subject (NTU60-Xsub): training data comes from 20 subjects, and testing data comes from the other 20 subjects. (2) cross-view (NTU60–view): training data comes from camera views 2 and 3, and testing data comes from camera view 1.

NTU RGB+D 120 [5]. NTU RGB+D 120 is currently the largest dataset with 3D joint annotations for HAR, which extends NTU RGB+D 60 with additional 57,367 skeleton sequences over 60 extra action classes. Totally 113,945 samples over 120 classes are performed by 106 volunteers, captured with three camera views. This dataset contains 32 setups, each denoting a specific location and background. The authors of this dataset recommend two benchmarks: (1) cross-subject (NTU120-Xsub): training data comes from 53 subjects, and testing data comes from the other 53 subjects. (2) cross-setup (NTU120-Xset): training data comes from samples with even setup IDs, and testing data comes from samples with odd setup IDs.

Kinetics-400 [2]. Kinetics-400 is a large-scale action recognition dataset with 400 actions. The skeletons were provided by [3], where the Openose algorithm [1] was applied for joint estimation. The box threshold of human detection is set as 0.5. After the validation, there are a total of 236,489 skeleton sequences for training and 19,505 skeleton sequences for testing.

FineGYM [7]. FineGYM is a fine-grained action recognition dataset with 29,000 videos of 99 fine-grained gymnastic action classes. Skeletons are extracted with ground-truth human bounding boxes as described in [4].

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[3] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition, 2022. 1

[4] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 1

[5] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2684–2701, 2019. 1

[6] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1

[7] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understand-

*Corresponding Author

ing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020. 1