Temporal Shift and Attention Modules for Graphical Skeleton Action Recognition *Supplementary Material*

Haidong Zhu, Zhaoheng Zheng and Ram Nevatia
Department of Computer Science
University of Southern California
Los Angeles, California 90089
{haidongz,zhaoheng.zheng,nevatia}@usc.edu

In this supplementary document, we discuss the ablation studies for module analysis and fusion analysis. We first analyze how much improvement with the two modules we have on accuracy for skeleton action recognition compared with baseline methods, then we compare using different self-attention generation methods for generating the video feature and prediction, followed the ablation of how the hyperparameter k, the percentage of shifted features in G-TSM, is chosen.

A. Module analysis

We show the results for module analysis in Table I. To better illustrate the performance of the two modules respectively, we choose ST-GCN on NTU-RGB+D 60 dataset [1] *xview* setting, since the boost of our method on ST-GCN is more significant than MS-G3D. We use the top-1 accuracy on NTU-RGB+D 60 dataset *xview* setting to assess the models. We notice that both modules boost the performance compared with the original ST-GCN [2], while the performance is the best when you use both of these modules in the network. Also, note that when we remove the *tanh* layer from the ST-G-TSA network, the accuracy drops from 90.2 to 89.5. This indicates that if the network only focuses on a few temporal segments of the whole video, it overfits the data and fails to give the best prediction.

TABLE I
RESULTS FOR ACTION PREDICTION ACCURACY WITH DIFFERENT MODULE COMBINATIONS ON *xview* SETTING OF NTU-RGB+D 60 DATASET. ST-G-TSM AND ST-G-TAM ARE ST-GCN WITH GRAPH TEMPORAL SHIFT AND GRAPH TEMPORAL ATTENTION MODULE ONLY RESPECTIVELY.

Method	Accuracy
ST-GCN	88.4
ST-G-TSM	88.9
ST-G-TAM	88.9
ST-G-TSA	90.2

B. Temporal Fusion method

We show the results of using different temporal fusions for the video feature in Table II compared with G-TSM. Results are also reported on the *xview* setting of the NTU-RGB+D 60 dataset. Besides average fusion ST-G-TSM (the original fusion of MS-G3D and ST-GCN) and G-TSA, we also compared with RNN fusion (ST-G-TSM + RNN) and max-pooling fusion (ST-G-TSM + Maxpool). This is to see if naive RNN network or max-pooling results for the highest response from all frames can help improve the performance of action prediction.

TABLE II
RESULTS FOR ACTION PREDICTION ACCURACY FOR DIFFERENT TEMPORAL FUSION METHODS ON NTU-RGB+D 60 *xview* DATASET.

Method	Accuracy
ST-G-TSM	88.9
ST-G-TSM + RNN	87.3
ST-G-TSM + MaxPool	88.4
ST-G-TSA w/o *tanh*	89.5
ST-G-TSA	90.2

From the table, we note that the RNN fusion generates the video feature via fusing the recurrent frame features and shows some minor improvements. Still, the improvement is not comparable with ST-G-TSA. This shows that the RNN network fuses all the temporal features but fails to understand which frames are more important for the final action prediction. In contrast, the max-pooling fusion selects the highest response of the frame features as the video feature and yields a lower accuracy compared with average fusion. The highest response might not be the action prediction we want for action recognition and might overfit the frames in the training set instead of showing good generalization.

C. Percentage of Feature Shift

We test different percentages of shifted features in G-TSM for deciding the choice of hyperparameter k in Table III. Experiments are conducted on the NTU-RGB+D 60 dataset, where top-1 accuracies are reported as results. We compared four different ratios of information exchanging, 40%, 33.3%, 25% and 12.5% on both *xview* and *xsub* settings, which is the value of $\frac{2}{5}$, $\frac{1}{3}$, $\frac{1}{4}$, and $\frac{1}{8}$.

TABLE III
ACCURACY FOR DIFFERENT PERCENTAGE OF SHIFTED FEATURES k ON NTU-RGB+D 60 DATASET.

Percentage of shifted features	NTU-RGB+D 60	
	xview	*xsub*
0%	88.9	82.2
12.5%	89.6	82.5
25.0%	90.2	83.3
33.3%	88.9	83.2
40.0%	88.6	82.7

We notice that when we choose to exchange 25% of the features with previous and coming features, the performance is the best. If we shift more features, i.e., when k increases to 40%, the accuracy for action prediction drops. At this exchange rate, only 20% features are used to predict the current temporal segments, indicating if features for the current frame are not enough to make a precise prediction for the current temporal frame or segment. When we have the same ratio of features used for the temporal shift as well as the current frames, the model can gain the best performance.

REFERENCES

[1] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016, pp. 1010–1019.

[2] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *ECCV*. Springer, 2016, pp. 816–833.