Supplement Material of Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition

Ke Cheng[1,2], Yifan Zhang[1,2], Congqi Cao[4], Lei Shi[1,2], Jian Cheng[1,2,3], and Hanqing Lu[1,2]

[1] NLPR & AIRIA, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
[4] School of Computer Science, Northwestern Polytechnical University
chengke2017@ia.ac.cn, {yfzhang, lei.shi, jcheng, luhq}@nlpr.ia.ac.cn,
congqi.cao@nwpu.edu.cn

1 The ablation study on K in DropGraph

We discuss the setting of K in spatial DropGraph and temproal DropGraph respectively.

In the spatial graph, the total number of nodes are typically less than 30 (25 nodes for NTU-RGBD, 18 nodes for Skeleton-Kinetics). Hence, we discuss $K = 0, 1, 2$ in Table 1. Note that when $K = 0$, the spatial DropGraph degenerates into dropping isolated joints, which is not effective. By expanding the drop area to 1^{st} neighbor nodes, spatial DropGraph provides efficient regularization. Expanding drop area to 2^{ed} neighbor may cause too strong regularization.

In the temporal graph, the total number of nodes are typically more than 100 (300 frames for NTU-RGBD, 150 frames for Skeleton-Kinetics). Hence, we discuss $K = 0, 5, 10, 20, 30$ in Table 2. When $K = 0$, the temporal DropGraph degenerated to drop isolated frames, which is not efficient. Temporal DropGraph can provide efficient regularization when $K = 5, 10, 20$. When $K = 30$, the regularization is too strong.

We use this setting for NTU-RGBD and NTU-RGBD-120. For NW-UCLA, we set temporal $K = 5$ due to the small number of frames in NW-UCLA action samples.

Model	DC-GCN	DC-GCN + spatial DG		
		$K = 0$	$K = 1$	$K = 2$
top-1	94.3	94.4	94.8	94.6
Δ	0	+0.1	+0.5	+0.3

Table 1. The different setting of K in spatial DropGraph. The accuracy (%) is evaluated on NTU-RGBD X-view task. Δ shows the improvement of accuracy.

Model	DC-GCN	DC-GCN + temporal DG ($K =$)				
		0	5	10	20	30
top-1	94.3	94.3	94.7	94.7	94.8	94.6
Δ	0	+0.0	+0.4	+0.4	+0.5	+0.3

Table 2. The different setting of K in temporal DropGraph. The accuracy (%) is evaluated on NTU-RGBD X-view task. Δ shows the improvement of accuracy.

2 Details for computing FLOPs[5].

In the main paper, we show the computational complexity of ST-GCN [2], Directed-GNN [1], and our proposed DC-GCN. Because the computational complexity was not explicitly discussed in some papers; we estimate them based on their description.

2.1 The FLOPs of ST-GCN [2].

ST-GCN [2] is composed of one input block and 9 residual blocks, as shown in Table 3. Each block contains a regular spatial convolution and a regular temporal convolution.

As introduced in our main paper, the regular spatial convolution in a block is a fusion of three graph convolutions, which operate on three different partitions respectively. Every graph convolution contains two matrix multiplications, whose computational complexity is $NCC' + N^2C'$ (For the blocks that $C = C'$, this formula can be simplified to $NC^2 + N^2C$). Every frame requires the same computation process. Therefore, the FLOPs of regular spatial convolution in a block is:

$$S_{FLOPs} = (3 \times (NCC' + N^2C')) \times T \tag{1}$$

The regular temporal convolution in a block is a 1D convolution on the temporal dimension whose kernel size is 9. Every body joint requires the same temporal convolution. Therefore, the FLOPs of a temporal convolution is:

$$T_{FLOPs} = (9 \times TCC') \times N \tag{2}$$

Besides, a fully-connected layer is used to get final scores for 60 classes, whose FLOPs is:

$$FC_{FLOPs} = C' \times classes = 256 \times 60 \tag{3}$$

ST-GCN [2] is composed of 10 spatial convolutions, 10 temporal convolutions, and one fully-connected layer. We compute the FLOPs of every part and get the total FLOPs: 8.1G.

Note that in the NTU RGB+D dataset, there are one or two people in each sample. For samples with only one person, the second person is padded with

[5] FLOPs: FLoating-number OPerations

zeros. The skeleton graphs of these two people are computed respectively. Thus, the total FLOPs for one action sample is 2 × 8.1G = 16.2G, including 4.0G on spatial graph convolution and 12.2G on temporal graph convolution.

Stage	ST-GCN
Block0	$C = 3, C' = 64, T = 300, N = 25$
Block1	$C = 64, C' = 64, T = 300, N = 25$
Block2	$C = 64, C' = 64, T = 300, N = 25$
Block3	$C = 64, C' = 64, T = 300, N = 25$
Block4	$C = 64, C' = 128, T = 150, N = 25$
Block5	$C = 128, C' = 128, T = 150, N = 25$
Block6	$C = 128, C' = 128, T = 150, N = 25$
Block7	$C = 128, C' = 256, T = 75, N = 25$
Block8	$C = 256, C' = 256, T = 75, N = 25$
Block9	$C = 256, C' = 256, T = 75, N = 25$
	global average pooling, FC, softmax

Table 3. The structure of ST-GCN, where C denotes the number of input channels, C' denotes the number of output channels. T denotes the number of temporal frames, N denotes the number of body joints on NTU RGB+D dataset.

2.2 The FLOPs of DC-GCN (Ours).

DC-GCN introduces zero extra FLOPs on ST-GCN backbone, so its FLOPs is 16.2G. As introduced in Section 4.3 in our main paper, multi-stream strategy is commonly employed in previous state-of-the-art approaches. We ensemble 4 stream DC-GCN [2], whose FLOPs is 64.8G.

2.3 The FLOPs of Directed-GNN [1].

Directed-GNN [1] fuses 4 streams: “joint stream”, “bone stream”, “joint motion stream”, and “bone motion stream”. All 4 streams use ST-GCN [2] as backbone. They divide them into two groups. One group contains the “joint stream” and “bone stream”. We call it “spatial group”. Another group contains the “joint motion stream” and “bone motion stream”. We call it “motion group”.

In both groups, they introduce two directed graph modules into every ST-GCN block to exchange information. Each directed graph module contains a fully-connected layer, whose input channel is $3C$ and output channel is C'. Therefore, the FLOPs of one directed graph module is $3CC'NT$, where N is the number of body joints and T is the number of frames.

In their paper, they mention two ST-GCN baselines. One is called 2s-ST-GCN which is the standard two-stream ST-GCN; the other is called 1s-ST-GCN whose number of channels is twice the original number. But they do not explicitly mention which baseline to build their Directed-GNN. We contact the author and confirm that: for spatial convolution, they use two-stream ST-GCN and introduce directed graph modules to exchange information; however, for temporal convolution, they concatenate the two-stream features on the channel dimension and double the number of channels. Notice that the FLOPs is *quadratic* in term of the number of channels.

With these analyses, we can compute the FLOPs of Directed-GNN [1]. The FLOPs of both "spatial group" and "motion group" are 63.4G, so the total FLOPs of Directed-GNN is 2 × 63.4G = 126.8G.

References

1. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
2. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)