Actional-Structural Graph Convolutional Networks for Skeleton based Action Recognition (Supplementary Materials)

Maosen Li[1], Siheng Chen[2], Xu Chen[1], Ya Zhang[1], Yanfeng Wang[1], and Qi Tian[3]

[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2] Carnegie Mellon University
[3] Huawei Noah's Ark Lab

{maosen_li, xuchen2016, ya_zhang, wangyanfeng} @sjtu.edu.cn, sihengc@andrew.cmu.edu, tian.qi1@huawei.com

1. Theorem Proof

Theorem 1 *The actional-structural graph convolution is a valid linear operation; that is, when* $\mathbf{Y}_1 = \mathrm{ASGC}(\mathbf{X}_1)$ *and* $\mathbf{Y}_2 = \mathrm{ASGC}(\mathbf{X}_2)$. *Then,* $a\mathbf{Y}_1 + b\mathbf{Y}_2 = \mathrm{ASGC}(a\mathbf{X}_1 + b\mathbf{X}_2)$, $\forall a, b$.

Proof The operations in actional graph convolution (AGC) are all linear, as well as the structural graph convolution (SGC). The AGC satisfies

$$\begin{aligned} & \mathrm{AGC}(a\mathbf{X}_1 + b\mathbf{X}_2) \\ = & \sum_{c=1}^{C} \hat{\mathbf{A}}_{\mathrm{act}}^{(c)} (a\mathbf{X}_1 + b\mathbf{X}_2) \mathbf{W}_{\mathrm{act}}^{(c)} \\ = & a\left(\sum_{c=1}^{C} \hat{\mathbf{A}}_{\mathrm{act}}^{(c)} \mathbf{X}_1 \mathbf{W}_{\mathrm{act}}^{(c)}\right) + b\left(\sum_{c=1}^{C} \hat{\mathbf{A}}_{\mathrm{act}}^{(c)} \mathbf{X}_2 \mathbf{W}_{\mathrm{act}}^{(c)}\right) \\ = & a\mathrm{AGC}(\mathbf{X}_1) + b\mathrm{AGC}(\mathbf{X}_2). \end{aligned}$$

Similarly, SGC satisfies

$$\begin{aligned} & \mathrm{SGC}(a\mathbf{X}_1 + b\mathbf{X}_2) \\ = & \sum_{\ell=1}^{L} \sum_{p \in \mathcal{P}} \mathbf{M}_{\mathrm{struct}}^{(p,\ell)} \circ \hat{\mathbf{A}}^{(p)\ell} (a\mathbf{X}_1 + b\mathbf{X}_2) \mathbf{W}_{\mathrm{struc}}^{(p,\ell)} \\ = & a\mathrm{SGC}(\mathbf{X}_1) + b\mathrm{SGC}(\mathbf{X}_2). \end{aligned}$$

With both AGC and SGC operations, the actional-structural convolution (ASGC) is formulated as

$$\begin{aligned} \mathbf{Y}_1 &= \mathrm{ASGC}(\mathbf{X}_1) \\ &= (1-\lambda)\mathrm{SGC}(\mathbf{X}_1) + \lambda \mathrm{AGC}(\mathbf{X}_1), \end{aligned}$$

which is a linear summation of AGC and SGC. Therefore, we have

$$\begin{aligned} & \mathrm{ASGC}(a\mathbf{X}_1 + b\mathbf{X}_2) \\ = & (1-\lambda)\mathrm{SGC}(a\mathbf{X}_1 + b\mathbf{X}_2) + \lambda \mathrm{AGC}(a\mathbf{X}_1 + b\mathbf{X}_2), \\ = & (1-\lambda)(a\mathrm{SGC}(\mathbf{X}_1) + b\mathrm{SGC}(\mathbf{X}_2)) \\ & + \lambda(a\mathrm{AGC}(\mathbf{X}_1) + b\mathrm{AGC}(\mathbf{X}_2)), \\ = & a((1-\lambda)\mathrm{SGC}(\mathbf{X}_1) + \lambda\mathrm{AGC}(\mathbf{X}_1)) \\ & + b((1-\lambda)\mathrm{SGC}(\mathbf{X}_2) + \lambda\mathrm{AGC}(\mathbf{X}_2)) \\ = & a\mathrm{ASGC}(\mathbf{X}_1) + b\mathrm{ASGC}(\mathbf{X}_2) \\ = & a\mathbf{Y}_1 + b\mathbf{Y}_2. \end{aligned}$$

The ASGC is a linear operation for the input data. □

2. Model Architectures

In this section, we show the detailed architectures of the proposed AS-GCN model.

2.1. A-links Inference Module (AIM)

2.1.1 Encoder

Given the 3D joint positions of n joints across T frames, $\mathcal{X} \in \mathbb{R}^{n \times 3 \times T}$, we first downsample the videos to obtain 50 frames from the valid frames at regular intervals. If $T < 50$, we pad the sequences to be 50 frames with 0. For any joint v_i on the body, where $i \in \{1, 2, \ldots, n\}$, we represent the joint feature across 50 frames as $\mathbf{x}_i \in \mathbb{R}^{150}$. We set that there are four types of A-links for actional dependencies capturing. As for link feature aggregation, we use average operation for all links surrounding one joint. The operations in the encoder in AIM are presented in Table 1. The activation functions of MLPs in the encoder are exponential linear unit (elu) functions, and 'bn' denotes the batch normalization to the features. $\oplus$ is the concatenation operation.

Input	Operation	Output
$\mathbf{p}_i^{(0)} = \mathbf{x}_i$	$150 \xrightarrow{\text{elu}} 128 \xrightarrow{\text{elu}} 128$ (bn)	$\mathbf{p}_i^{(1)}$
$\mathbf{p}_i^{(1)}, \mathbf{p}_j^{(1)}$	$\mathbf{p}_i^{(1)} \oplus \mathbf{p}_j^{(1)}$	$\mathbf{Q}_{i,j}^{(2)}$
$\mathbf{Q}_{i,:}^{(2)}$	$(\frac{1}{n}\sum_{j=1}^{n}\mathbf{Q}_{i,j}^{(2)}) \oplus \mathbf{p}_i^{(1)}$ $384 \xrightarrow{\text{elu}} 128 \xrightarrow{\text{elu}} 128$ (bn)	$\mathbf{p}_i^{(2)}$
$\mathbf{p}_i^{(2)}, \mathbf{p}_j^{(2)}$	$\mathbf{p}_i^{(2)} \oplus \mathbf{p}_j^{(2)}$	$\mathbf{Q}_{i,j}^{(3)}$
$\mathbf{Q}_{i,j}^{(3)}$	$256 \xrightarrow{\text{elu}} 128 \xrightarrow{\text{elu}} 128$ (bn) $\rightarrow 4$	$\mathcal{A}_{i,j,:}$

Table 1: The architecture of the encoder in AIM

2.1.2 Decoder

We present the detailed configuration of the decoder of AIM. Given the position of joint v_i at time t, $\mathbf{x}_i^t$, the decoder aims to predict the future joint position $\mathbf{x}_i^{t+1}$ conditioned on the sourrounding A-links, $\mathcal{A}_{i,j,:}$. The architectures are presented in Table 2. $\mathrm{GRU}(\cdot)$ denotes a GRU unit, whose hidden feature dimension is 64. It predicts the future position of all joints conditioned on A-links and previous frames.

Input	Operation	Output
$\mathbf{x}_i^t, \mathbf{x}_j^t$	$3 \xrightarrow{\text{relu}} 64\ (c)$ $\oplus$ $\sum_{c=1}^{C}\mathcal{A}_{i,j,c} \cdot (128 \xrightarrow{\text{relu}} 64\ (c))$	$\mathbf{Q}_{i,j}^t$
$\mathbf{Q}_{i,j}^t$	$(\frac{1}{n}\sum_{j=1}^{n}\mathbf{Q}_{i,j}^t) \oplus \mathbf{p}_i^t$	$\mathbf{p}_i^t$
$\mathbf{p}_i^t, \mathbf{S}_i^t$	$\mathrm{GRU}(\mathbf{S}_i^t, \mathbf{p}_i^t)$, feature dimension: 64	$\mathbf{S}_i^t$
$\mathbf{S}_i^t$	$64 \rightarrow 3$	

Table 2: The architecture of the decoder in AIM

2.2. Backbone

The backbone network of the AS-GCN extracts the rich spatial and temporal feature of actions with the proposed ASGC and temporal CNN (T-CN). For example, we build AS-GCN on NTU-RGB+D dataset and Cross-Subject benchmark [1]. There are 25 joints, 3D spatial positions and 300 padded frames for each action. The architecture of the backbone is presented in Table 3. There are nine ASGC blocks consisting of the backbone of AS-GCN model. The input/output feature maps are 3D tensors, where the three axes represent the joint number, feature dimension and frame number, respectively. The shapes of operations have the consistent dimensions with input and output feature maps, where the first axis is the filter number or output feature dimension, and the other three correspond to the input shape. A and B are the types of A-links and S-links. For action recognition, we obtain the last feature map whose shape is [25,256,75] and apply a global average pooling operation on the time and joint axis, i.e. the 1st and 3rd axis. Thus we obtain a semantic feature vector of the action, whose dimension is 256.

In-Shape	Operation Shape	Out-Shape
[25,3,300]	ASGC:[64,1,64,1]×(n_A+n_S) T-CN:[64,1,64,7], stride=1	[25,64,300]
[25,64,300]	ASGC:[64,1,64,1]×(n_A+n_S) T-CN:[64,1,64,7], stride=1	[25,64,300]
[25,64,300]	ASGC:[64,1,64,1]×(n_A+n_S) T-CN:[64,1,64,7], stride=1	[25,64,300]
[25,64,300]	ASGC:[64,1,64,1]×(n_A+n_S) T-CN:[128,1,64,7], stride=2	[25,128,150]
[25,128,150]	ASGC:[128,1,128,1]×(n_A+n_S) T-CN:[128,1,128,7], stride=1	[25,128,150]
[25,128,150]	ASGC:[128,1,128,1]×(n_A+n_S) T-CN:[128,1,128,7], stride=1	[25,128,150]
[25,128,150]	ASGC:[128,1,128,1]×(n_A+n_S) T-CN:[256,1,128,7], stride=2	[25,256,75]
[25,256,75]	ASGC:[256,1,256,1]×(n_A+n_S) T-CN:[256,1,256,7], stride=1	[25,256,75]
[25,256,75]	ASGC:[256,1,256,1]×(n_A+n_S) T-CN:[256,1,256,7], stride=1	[25,256,75]

Table 3: The architecture of the backbone network of AS-GCN

2.3. Future Action Prediction Head

The architecture of the future action prediction head of AS-GCN model are presented in Table 4. We input the output feature map from the backbone network in to the prediction head. The input tensor are calculated by nine ASGC blocks. The first five blocks reduce the frame number to

In-Shape	Operation Shape	Out-Shape
[25,256,75]	ASGC:[128,1,256,1]×(n_A+n_S) T-CN:[128,1,128,7], stride=2	[25,128,39]
[25,128,39]	ASGC:[128,1,128,1]×(n_A+n_S) T-CN:[128,1,128,7], stride=2	[25,128,19]
[25,128,19]	ASGC:[128,1,128,1]×(n_A+n_S) T-CN:[128,1,128,7], stride=2	[25,128,10]
[25,128,10]	ASGC:[128,1,128,1]×(n_A+n_S) T-CN:[128,1,128,3], stride=2	[25,128,5]
[25,128,5]	ASGC:[128,1,128,1]×(n_A+n_S) T-CN:[128,1,128,5], stride=1	[25,128,1]
[25,131,1]	ASGC:[64,1,131,1]×(n_A+n_S) T-CN:[64,1,64,1], stride=1	[25,64,1]
[25,67,1]	ASGC:[32,1,67,1]×(n_A+n_S) T-CN:[32,1,32,1], stride=1	[25,32,1]
[25,35,1]	ASGC:[30,1,35,1]×(n_A+n_S) T-CN:[30,1,30,1], stride=1	[25,30,1]
[25,33,1]	FC:[30,1,33,1]	[25,30,1]

Table 4: The architecture of the prediction head of AS-GCN model

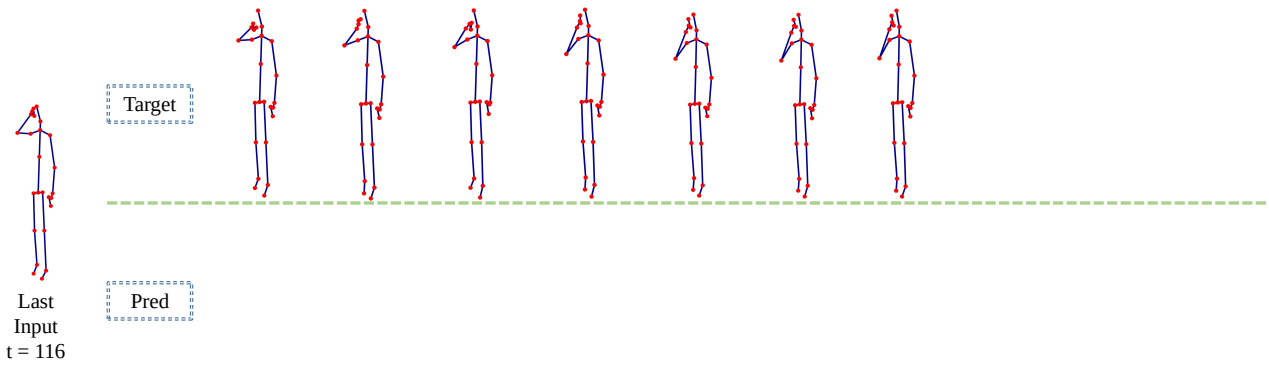
aggregate higher-level action features. The last four blocks work on action regeneration. For the last four blocks, we concatenate the last input frame to each feature map. Finally, with a residual connection, we obtain a tensor with shape [30,1,25] from a fully connected layer, which contains the joint position of the predicted 10 frames.

3. More Future Action Prediction Results

More future action prediction results of different actions are illustrated in Figure 1, which contains the action of 'wipe face', 'throw' and 'nausea or vomiting condition'. As we see, the actions are predicted with very low error.

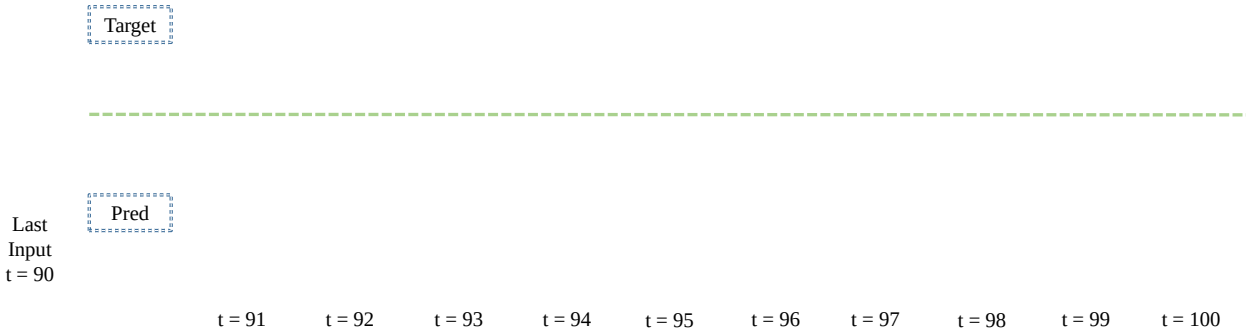
References

[1] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, June 2016. 4322



(a) One action of 'wipe face'. The 117th to 126th frames are predicted from the 116th frame.

(b) One action of 'throw'. The 69th to 78th frames are predicted from the 68th frame.



(c) One action of 'nausea or vomiting condition'. The 91th to 100th frames are predicted from the 90th frame.

Figure 1: Some predicted actions, including 'wipe face', 'throw' and 'nausea or vomiting codition' in NTU-RGB+D dataset.