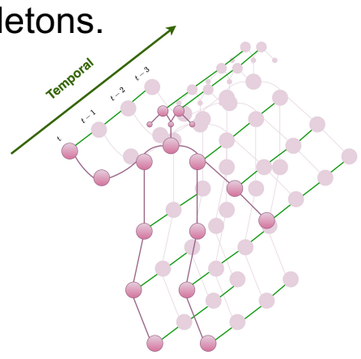
Temporal Attention-Augmented Graph Convolutional Network for Efficient Skeleton-based Human Action Recognition

Negar Heidari and Alexandros Iosifidis
Negar.heidari@eng.au.dk, Alexandros.iosifidis@eng.au.dk

Introduction: Skeleton-based Human Action Recognition

Recently, significant results have been obtained by employing Graph Convolutional Networks (GCNs) for skeleton-based human action recognition. In these methods, an action is represented as a sequence of body poses and each body pose is represented by a skeleton. The skeleton data is treated as a graph which models the spatial relationship between different body joints and the temporal dynamics in an action are expressed by a sequence of skeletons.



Most of these methods use deep feed forward networks to model the spatio-temporal features and process all the body skeletons in a sequence which is not efficient in terms of computational complexity. In this paper, we propose a temporal attention module (TAM) to extract the most informative skeletons in an action sequence, leading to increased computational efficiency in both the training phase and inference.

Method: Temporal Attention-Augmented GCN (TA-GCN)

In order to extract discriminative features in temporal dimension of data, we propose TAM which highlights the most informative skeletons in a sequence. Inspired by the baseline methods, ST-GCN and AGCN, at each layer of network, spatial and temporal convolutions are performed on data to extract the features in spatial and temporal dimensions, respectively. The TAM takes a tensor $\boldsymbol{H^{(l)}} \in R^{(C^{(l)} \times T \times N)}$ as input which can be the input data, i.e. $\boldsymbol{H^{(0)}} = \boldsymbol{X}$ or the output of a hidden layer of the network. First, two average pooling operations in both features and spatial dimensions are performed to produce $\boldsymbol{h^{(l)}} \in R^{(1 \times T \times 1)}$ which denotes the average value of each skeleton in the sequence. Then, the attention tensor a is produced by a fully connected layer as:

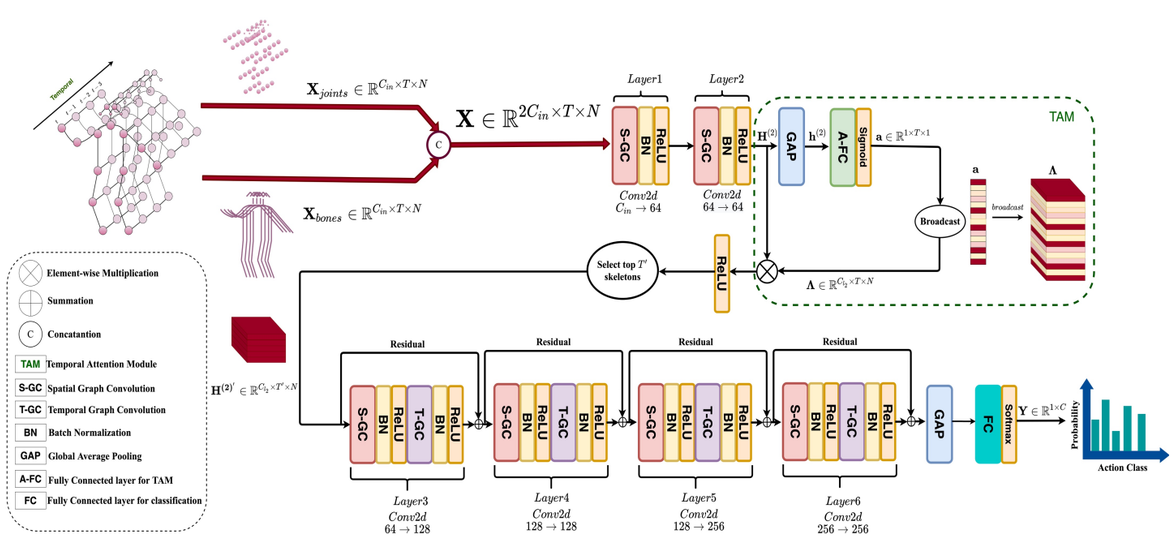
$$\boldsymbol{a} = Sigmoid\left(\boldsymbol{h^{(l)}}\boldsymbol{\Theta}\right),$$

Where $\boldsymbol{\Theta} \in R^{T \times T}$ denotes the learnable transformation matrix and the resulted attention tensor indicates the importance of each skeleton in the sequence. To highlight the most informative skeletons, the attention map is duplicated by $C^{(l)} \times N$ copies to produce $\boldsymbol{\Lambda} \in R^{(C^{(l)} \times T \times N)}$ which is subsequently dot multiplied to $\boldsymbol{H^{(l)}}$ as follows:

$$\widehat{\boldsymbol{H}}^{(\boldsymbol{l})} = ReLU(\boldsymbol{H^{(l)}} \otimes \boldsymbol{\Lambda})$$

To select a subset of T' skeletons from $H^{(l)}$, the values in the attention map are sorted in descending order and the skeletons corresponding to the highest attention values are selected to be introduced to the next layers of the network.

Figure. 1 Illustration of the proposed model diagram



Experiments

We conducted experiments on NTU-RGB+D and Kinetics-Skeleton datasets which are two widely adopted datasets for evaluating the performance of skeleton-based action recognition methods. We compare the performance of the proposed method with the of state-of-the-art methods in terms of classification accuracy and computational efficiency.

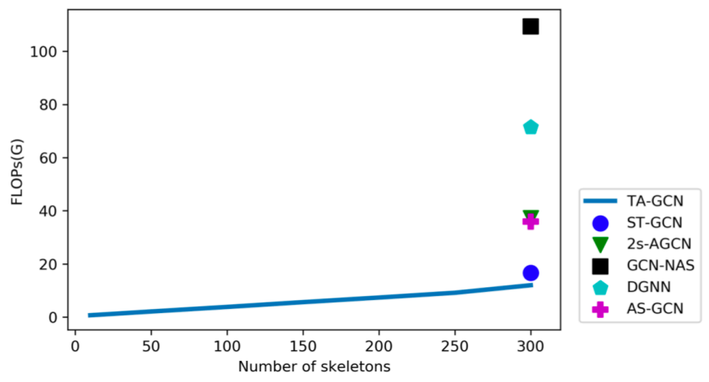
Table. 1 Comparisons of the classification accuracy with state-of-the-art methods on the test set of NTU-RGB+D dataset.

Method	CS(%)	CV(%)	#Streams	Skel.sel.
HBRNN (2015)	59.1	64.0	5	✗
Deep LSTM (2016)	60.7	67.3	1	✗
ST-LSTM (2016)	69.2	77.7	1	✗
STA-LSTM (2017)	73.4	81.2	1	✓
VA-LSTM (2017)	79.2	87.7	1	✗
ARRN-LSTM (2018)	80.7	88.8	2	✗
Two-Stream 3DCNN (2017)	66.8	72.6	2	✗
TCN (2017)	74.3	83.1	1	✗
Clips+CNN+MTLN (2017)	79.6	84.8	1	✗
Synthesized CNN (2017)	80.0	87.2	1	✗
3scale ResNet152 (2017)	85.0	92.3	1	✗
CNN+Motion+Trans (2017)	83.2	89.3	2	✗
ST-GCN (2018)	81.5	88.3	1	✗
DPRL+GCNN (2018)	83.5	89.8	1	✓
AS-GCN (2019)	86.8	94.2	2	✗
2s-AGCN (2019)	88.5	95.1	2	✗
GCN-NAS (2020)	89.4	95.7	2	✗
DGNN (2019)	89.9	96.1	4	✗
TA-GCN ($T' = 150$)	87.97	94.2	1	✓
2s-TA-GCN ($T' = 150$)	88.5	95.1	2	✓
4s-TA-GCN ($T' = 150$)	89.91	95.8	4	✓

Table. 2 Comparisons of the classification accuracy with state-of-the-art methods on the test set of Kinetics-Skeleton dataset.

Method	Top1(%)	Top5(%)	#Streams	Skel.sel.
Deep LSTM (2016)	16.4	35.3	1	✗
TCN (2017)	20.3	40.0	1	✗
ST-GCN (2018)	30.7	52.8	1	✗
AS-GCN (2019)	34.8	56.5	2	✗
2s-AGCN (2019)	36.1	58.7	2	✗
DGNN (2019)	36.9	59.6	4	✗
GCN-NAS (2020)	37.1	60.1	2	✗
1s-TA-GCN ($T' = 250$)	34.95	57.28	1	✓
2s-TA-GCN ($T' = 250$)	36.1	58.72	2	✓
4s-TA-GCN ($T' = 250$)	36.9	59.77	4	✓

Figure. 2 Computational complexity comparison between the proposed method when it selects different number of skeletons, and all the state-of-the-art methods which utilize all the skeletons of the input sequence.



Conclusion

- We proposed a temporal attention-augmented GCN to improve the computational efficiency in skeleton-based action recognition.
- Our method trains the attention mechanism to select the most informative skeletons for each action in an end-to-end manner.
- On two widely used benchmark datasets, the proposed method outperforms the baseline with a large margin and it has competitive performance with the state-of-the-art methods, while being up to 10 times less computationally complex.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871449