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
Aarhus University, Department of Electrical and Computer Engineering

AARHUS UNIVERSITY

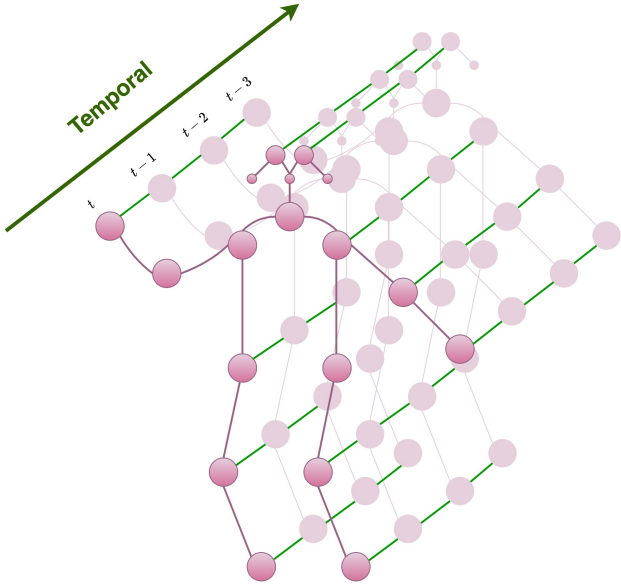
ICPR 2020

SKELETON-BASED HUMAN ACTION RECOGNITION

Graph Convolutional Networks (GCNs) have been successfully used for Skeleton-based human action recognition.

Spatio-temporal graph:

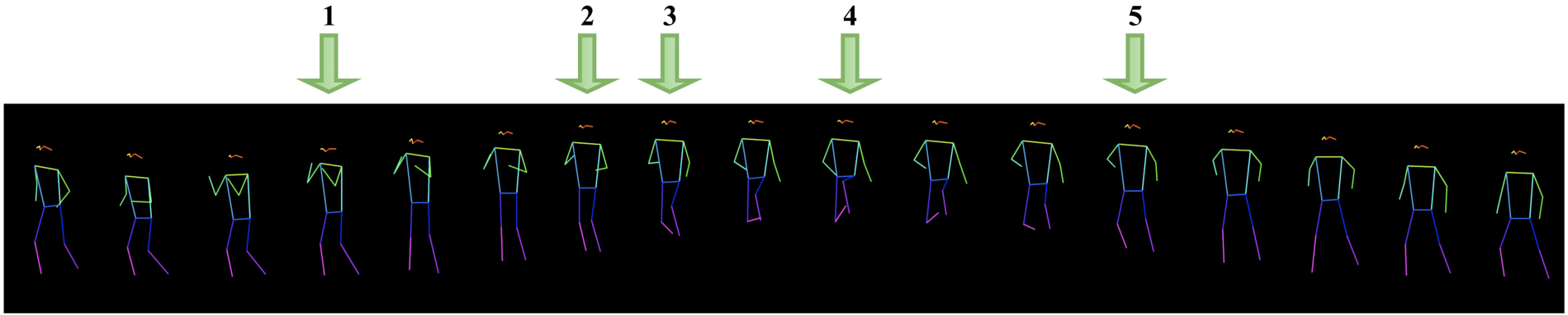
- Nodes: 2D or 3D human body joint coordinates
- Edges: spatial and temporal connections



HOW TO MAKE ST-GCN MORE EFFICIENT?

Solution:

pay different levels of attention to different temporal stages

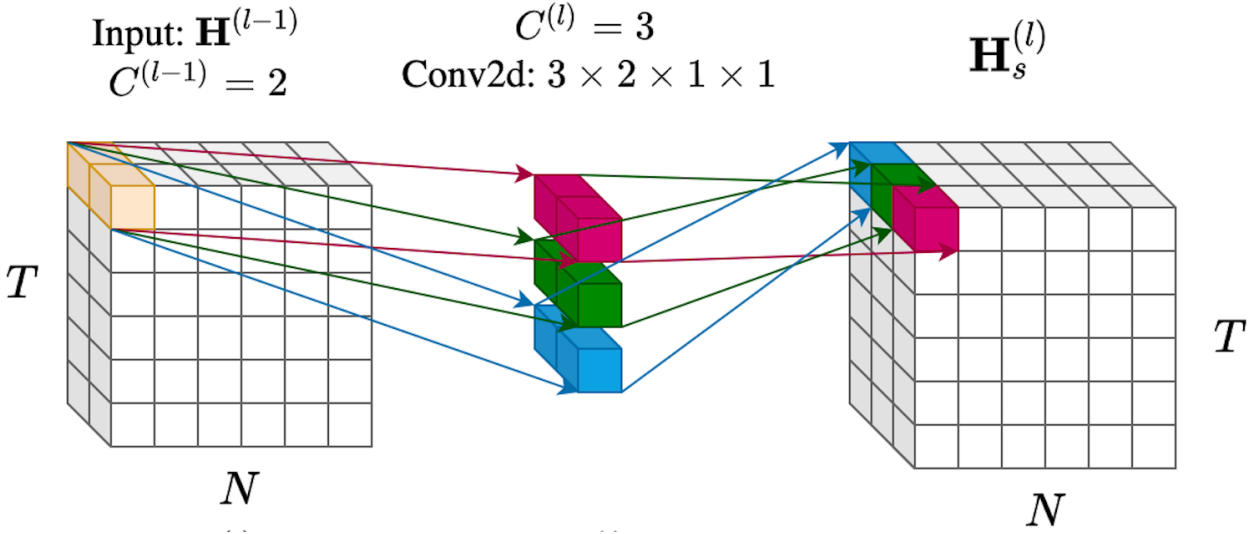


SPATIAL ATTENTION

$$H^{(l-1)} \in R^{C^{(l-1)} \times T \times N}$$

$$\hat{A}_p = D_p^{-\frac{1}{2}} A_p D_p^{-\frac{1}{2}}$$

$$M_p \in R^{N \times N}$$



- S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in AAAI conference onartificial intelligence, 2018.

SPATIAL ATTENTION

$$H^{(l-1)} \in R^{C^{(l-1)} \times T \times N}$$

$$\hat{A}_p = D_p^{-\frac{1}{2}} A_p D_p^{-\frac{1}{2}}$$

$$M_p \in R^{N \times N}$$

Spatial Convolution:

$$H_s^{(l)} = ReLU\left(\sum_p (\hat{A}_p + M_p) H^{(l-1)} W_p\right)$$

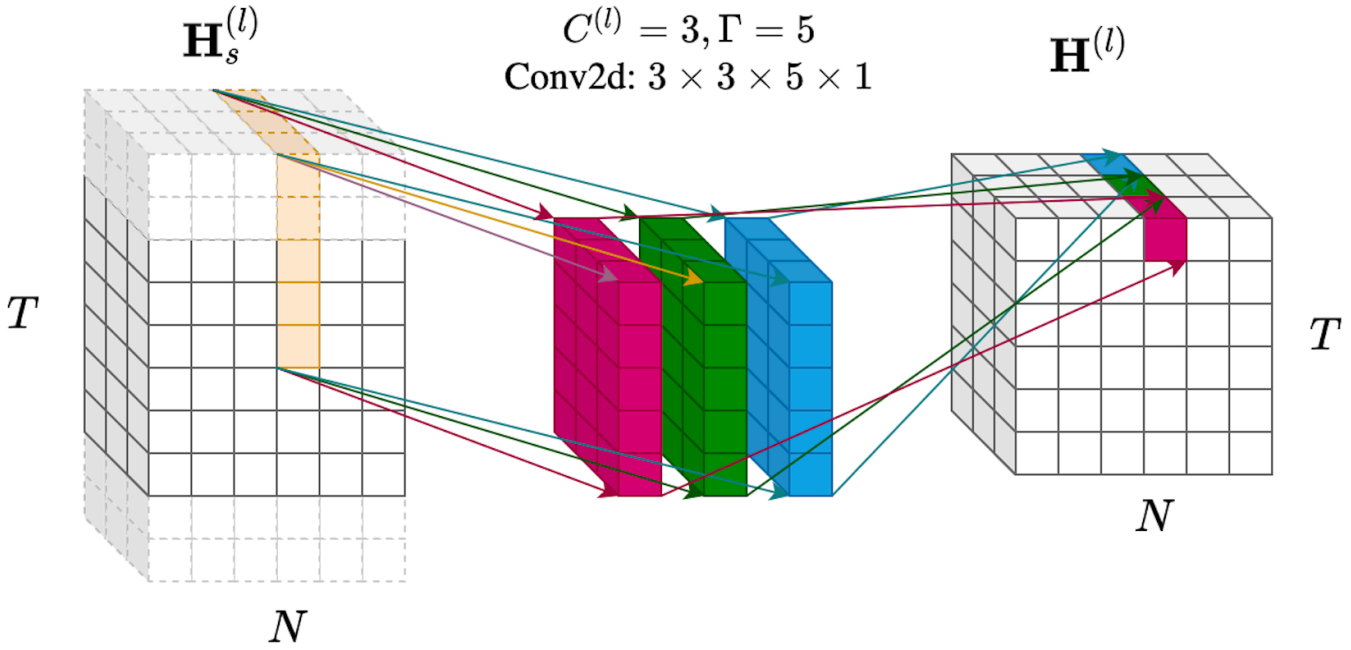
2D Convolution

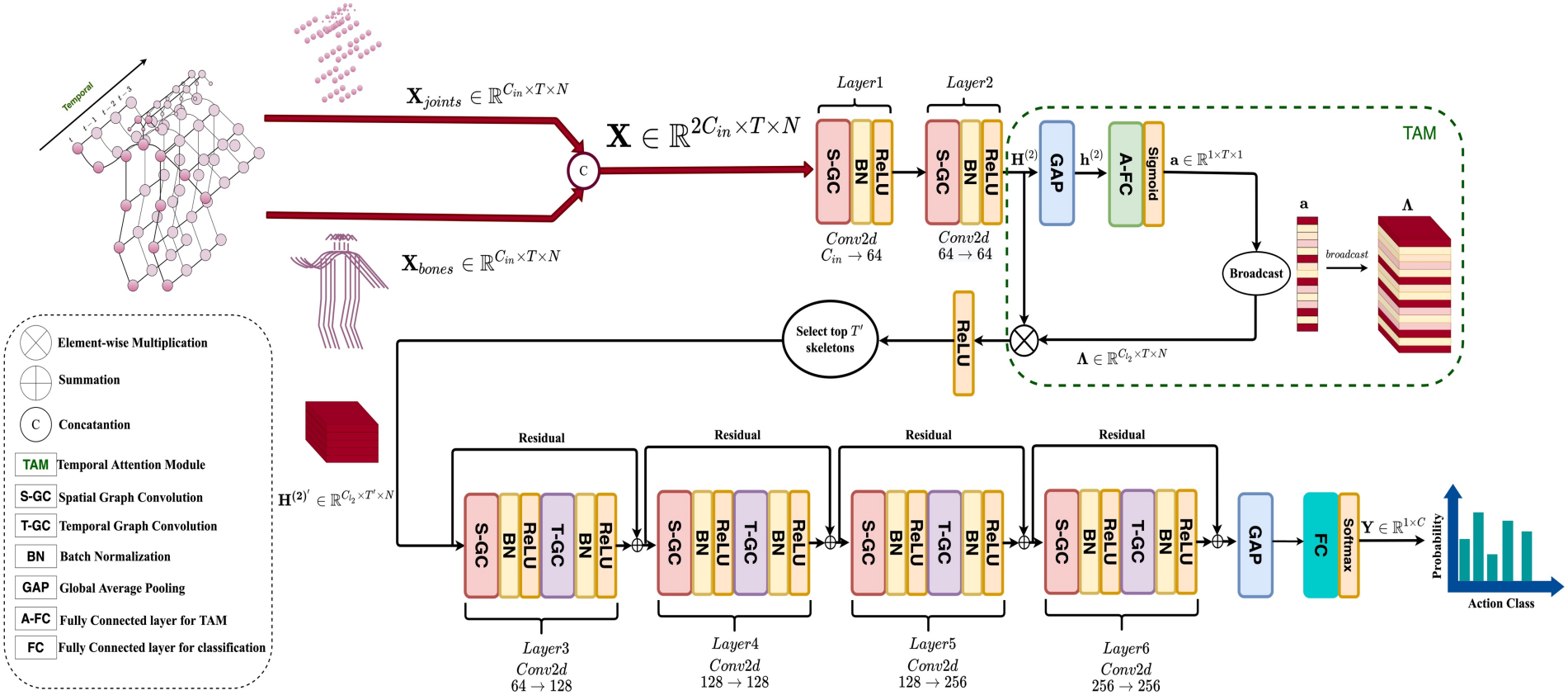
Spatial Attention

$$H_s^{(l)} \in R^{C^{(l)} \times T \times N}$$

- S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutionalnetworks for skeleton-based action recognition," inAAAI conference onartificial intelligence, 2018.

TEMPORAL CONVOLUTION




Temporal
$\mathbf{X}_{joints} \in \mathbb{R}^{C_{in} \times T \times N}$
$\mathbf{X}_{bones} \in \mathbb{R}^{C_{in} \times T \times N}$
C
$\mathbf{X} \in \mathbb{R}^{2C_{in} \times T \times N}$
Layer1
Layer2
S-GC
BN
ReLU
Conv2d
$C_{in} \rightarrow 64$
Conv2d
$64 \rightarrow 64$
$\mathbf{H}^{(2)}$
GAP
$\mathbf{h}^{(2)}$
A-FC
Sigmoid
$\mathbf{a} \in \mathbb{R}^{1 \times T \times 1}$
TAM
a
broadcast
Λ
Broadcast
$\mathbf{\Lambda} \in \mathbb{R}^{C_{l_2} \times T \times N}$
ReLU
Select top T' skeletons
$\mathbf{H}^{(2)'} \in \mathbb{R}^{C_{l_2} \times T' \times N}$
Residual
T-GC
Layer3
Conv2d
$64 \rightarrow 128$
Layer4
Conv2d
$128 \rightarrow 128$
Layer5
Conv2d
$128 \rightarrow 256$
Layer6
Conv2d
$256 \rightarrow 256$
FC
Softmax
$\mathbf{Y} \in \mathbb{R}^{1 \times C}$
Probability
Action Class
Element-wise Multiplication
Summation
Concatantion
TAM Temporal Attention Module
S-GC Spatial Graph Convolution
T-GC Temporal Graph Convolution
BN Batch Normalization
GAP Global Average Pooling
A-FC Fully Connected layer for TAM
FC Fully Connected layer for classification

RESULTS

Datasets:

1. NTU-RGB+D
2. Kinnetics-Skeleton

Table. 1: Classification Accuracy on 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	✓

RESULTS

Datasets:

1. NTU-RGB+D
2. Kinnetics-Skeleton

Table. 2: Classification Accuracy on 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	✓

COMPUTATIONAL COMPLEXITY

Table. 3: Computational Complexity on NTU-RGB+D Dataset

Method	FLOPs	# Params
1s-TA-GCN ($T' = 150$)	$\times 1$	$\times 1$
2s-TA-GCN ($T' = 150$)	$\times 2$	$\times 2$
4s-TA-GCN ($T' = 150$)	$\times 4$	$\times 4$
ST-GCN	$\times 2.9$	$\times 1.3$
AS-GCN	$\times 6.3$	$\times 3.2$
2s-AGCN	$\times 6.6$	$\times 3$
DGNN	$\times 12.6$	$\times 3.6$
GCN-NAS	$\times 19.3$	$\times 8.9$

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

THANK YOU

AARHUS
UNIVERSITET

open Deep Robotics