CVPR2020 Oral

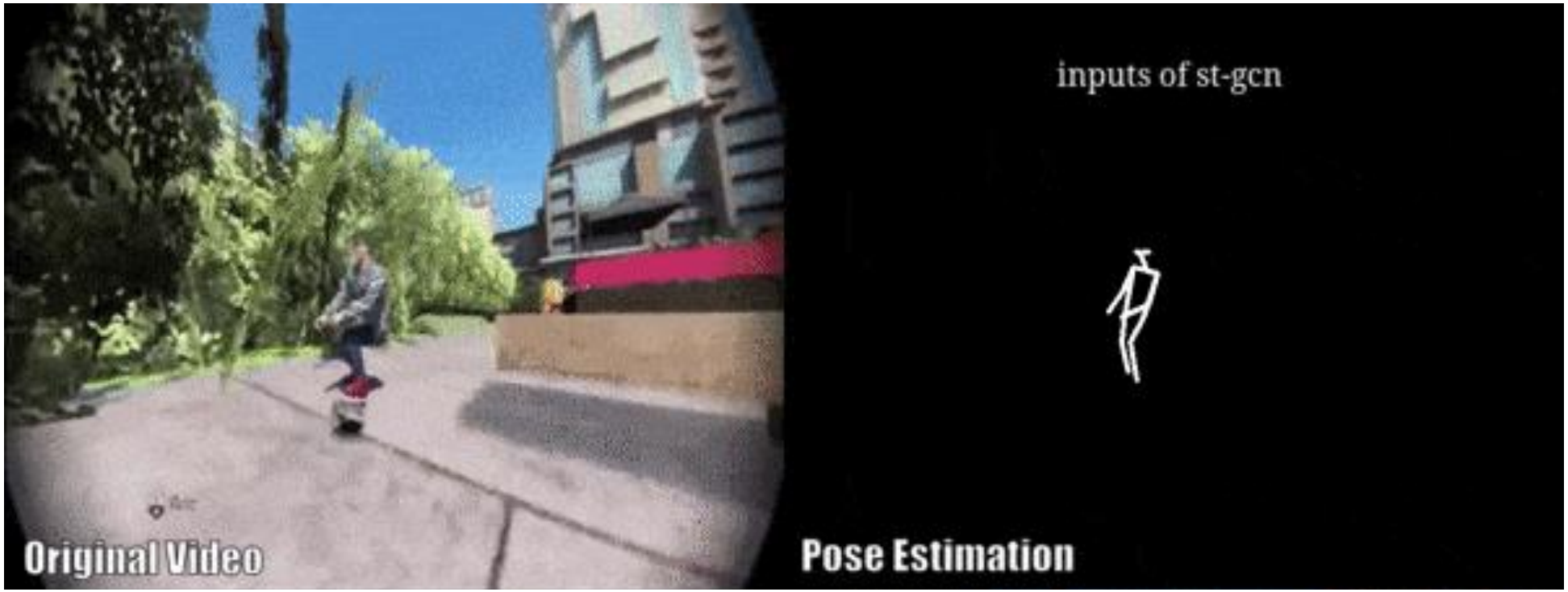
Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition

Dong-Hwan Jang

Seoul National University

Introduction

Skeleton-Based Action Recognition

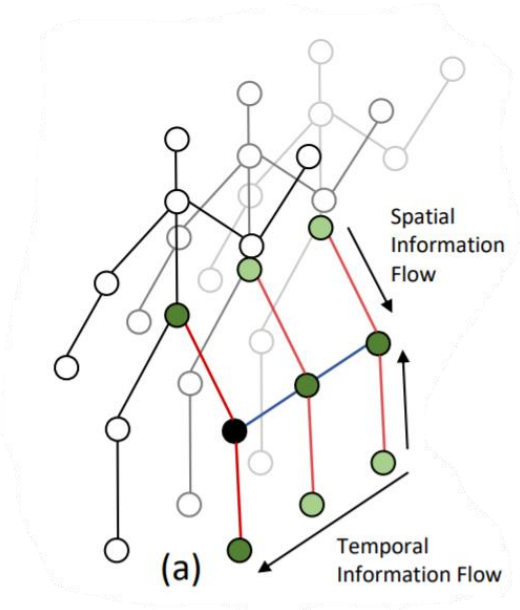


- Predicting actions from skeleton representations of human bodies instead of raw RGB videos
  - In recent work, the significant results have proven its merits

Introduction

Skeleton-Based Action Recognition

- Existing Approaches
  - GCN with higher-order polynomials of the skeleton adjacency matrix



$$\mathbf{X}_t^{(l+1)} = \sigma\left(\sum_{k=0}^{K} \hat{\mathbf{A}}^k \mathbf{X}_t^{(l)} \Theta_{(k)}^{(l)}\right),$$

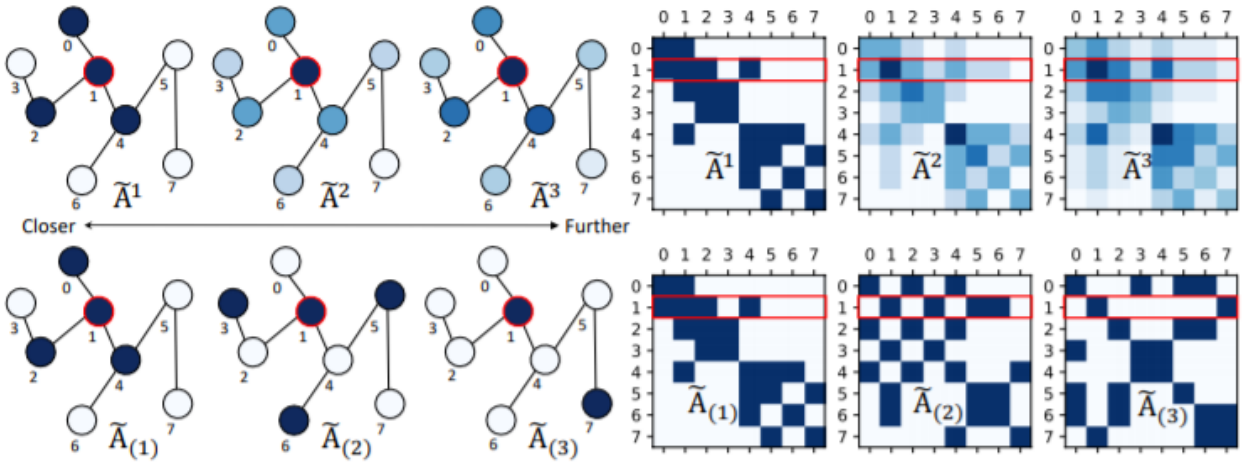
Introduction

Conditions for robust action recognition from skeleton

1) Extract **multi-scale structural features & long-range dependencies**

- Joints that are structurally apart can also have strong correlations
- Many existing approaches achieve this with higher-order polynomials of the adjacency matrix, A
  - **Biased weight problem**

Disentangling local dependencies

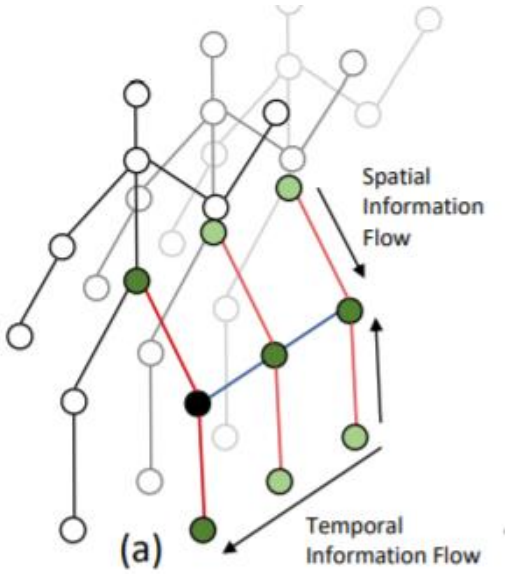


Introduction

Conditions for robust action recognition from skeleton

2) Leverage the complex **cross-spacetime joint relationships**

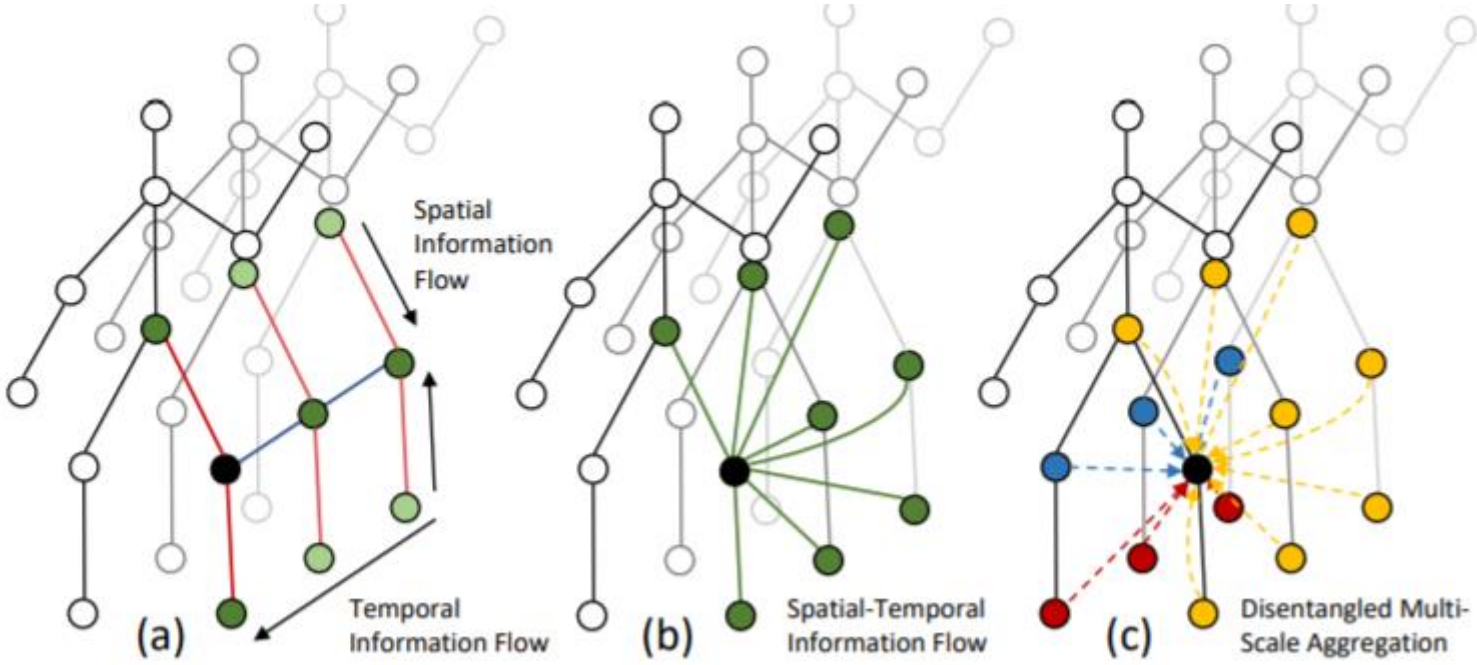
- Most existing approaches deploy interleaving spatial-only & temporal-only modules (similar to factorized C3D)
  - GCN – RNN or Conv1D
- E.g. “Standing Up” – past (*Upper body* “leaning forward) + future (*Lower body* “standing up”)



Skip Connection

Method

Multi-Scale G3D: MS-G3D



2) cross-spacetime 1) long-range dep.

Method

Multi-Scale G3D: MS-G3D

1) Cross Space-time relation

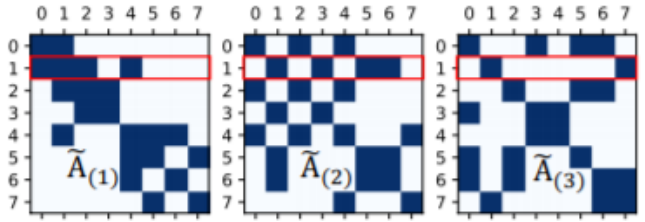
$$\tilde{\mathbf{A}}_{(\tau)} = \begin{bmatrix} \tilde{\mathbf{A}} & \cdots & \tilde{\mathbf{A}} \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{A}} & \cdots & \tilde{\mathbf{A}} \end{bmatrix} \in \mathbb{R}^{\tau N \times \tau N}$$

$$[\mathbf{X}_{(\tau)}^{(l+1)}]_t = \sigma\left(\tilde{\mathbf{D}}_{(\tau)}^{-\frac{1}{2}} \tilde{\mathbf{A}}_{(\tau)} \tilde{\mathbf{D}}_{(\tau)}^{-\frac{1}{2}} [\mathbf{X}_{(\tau)}^{(l)}]_t \Theta^{(l)}\right)$$

Method

Multi-Scale G3D: MS-G3D

$$[\mathbf{X}_{(\tau)}^{(l+1)}]_t = \sigma\left(\sum_{k=0}^{K} \tilde{\mathbf{D}}_{(\tau,k)}^{-\frac{1}{2}} \tilde{\mathbf{A}}_{(\tau,k)} \tilde{\mathbf{D}}_{(\tau,k)}^{-\frac{1}{2}} [\mathbf{X}_{(\tau)}^{(l)}]_t \Theta_{(k)}^{(l)}\right)$$



1) $\tilde{\mathbf{A}}_{(\tau,k)}$ is disentangled

2) Skip connections are considered in $\tilde{\mathbf{A}}_{(\tau,k)}$ $\tilde{\mathbf{D}}_{(\tau,k)}^{-\frac{1}{2}}$

Discussion

C3D-like Operation

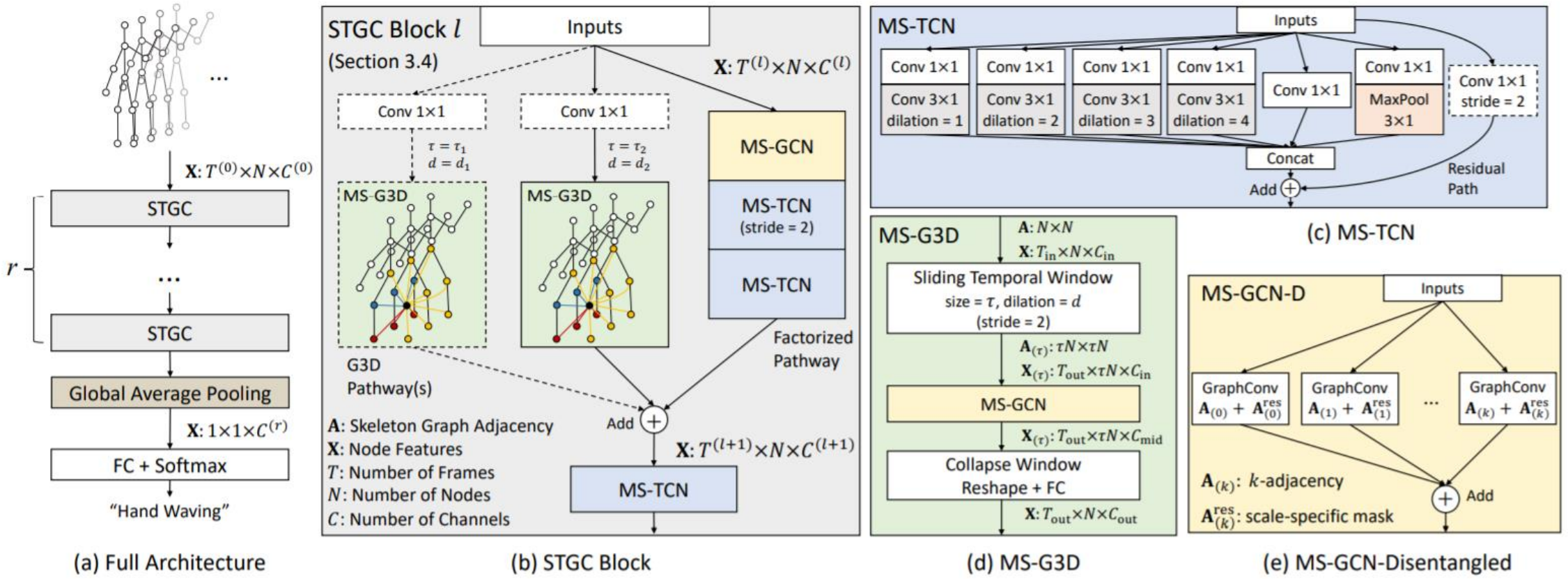
- Analogous to classical 3D convolutional blocks
- Θ is independent of τ, making less prone to overfitting with large τ

Dilated Convolution

- Dilated convolution picking a frame every d frames
- τ & d must be balanced

Method

Multi-Scale G3D: MS-G3D



Results

Methods	Number of Scales			
	$K = 1$	$K = 4$	$K = 8$	$K = 12$
GCN-E	85.1	85.6	86.5	86.6
GCN-D	85.1	87.0	86.9	86.8
GCN-E + Mask	86.1	87.0	87.5	87.7
GCN-D + Mask	86.1	86.9	87.9	87.8
G3D-E	85.1	85.5	85.4	85.5
G3D-D	85.1	86.4	86.5	86.4
G3D-E + Mask	86.6	87.0	86.5	86.2
G3D-D + Mask	86.6	87.4	87.1	87.0

Table 1: Accuracy (%) with multi-scale aggregation on individual pathways of STGC blocks with different K. "Mask" refers to the residual masks $\mathbf{A}^{\text{res}}$. If $K>1$, GCN/G3D is **Multi-Scale** (**MS-**).

Results

Model Configurations	Params	Acc (%)
Baseline (Js-AGCN [33])	3.5M	86.0
Baseline + MS-TCN	1.6M	86.7
MS-GCN (Factorized Pathway) Only	1.4M	87.8
with 2.5× Capacity	3.5M	88.5
with Dual Pathway	2.8M	88.6
MS-GCN (Factorized Pathway)		
with MS-G3D ($\tau = 3, d = 1$)	2.7M	89.0
with MS-G3D ($\tau = 3, d = 2$)	2.7M	89.1
with MS-G3D ($\tau = 3, d = 3$)	2.7M	89.1
with MS-G3D ($\tau = 5, d = 1$)	3.2M	89.2
with MS-G3D ($\tau = 5, d = 2$)	3.2M	89.2
with MS-G3D ($\tau = 7, d = 1$)†	3.0M	89.0
with 2 MS-G3D Pathways† $\tau = (3, 3), d = (1, 2)$	2.8M	89.3
with 2 MS-G3D Pathways† $\tau = (3, 5), d = (1, 1)$	3.2M	89.4

Results

Methods	NTU RGB+D 120	
	X-Sub (%)	X-Set (%)
ST-LSTM [26]	55.7	57.9
GCA-LSTM [27]	61.2	63.3
RotClips + MTCNN [16]	62.2	61.8
Body Pose Evolution Map [28]	64.6	66.9
2s-AGCN [33]	82.9	84.9
MS-G3D Net	**86.9**	**88.4**

Methods	NTU RGB+D 60	
	X-Sub (%)	X-View (%)
IndRNN [23]	81.8	88.0
HCN [20]	86.5	91.1
ST-GR [18]	86.9	92.3
AS-GCN [21]	86.8	94.2
2s-AGCN [33]	88.5	95.1
AGC-LSTM [34]	89.2	95.0
DGNN [32]	89.9	96.1
GR-GCN [8]	87.5	94.3
MS-G3D Net (Joint Only)	89.4	95.0
MS-G3D Net (Bone Only)	90.1	95.3
MS-G3D Net	**91.5**	**96.2**

Methods	Kinetics Skeleton 400	
	Top-1 (%)	Top-5 (%)
ST-GCN [50]	30.7	52.8
AS-GCN [21]	34.8	56.5
ST-GR [18]	33.6	56.1
2s-AGCN [33]	36.1	58.7
DGNN [32]	36.9	59.6
MS-G3D Net	**38.0**	**60.9**

Implementation

- **Source:** https://github.com/DongHwanJang/MS-G3D
- **Dataset**: Kinetics
- **Code in detail**
  - Main Code

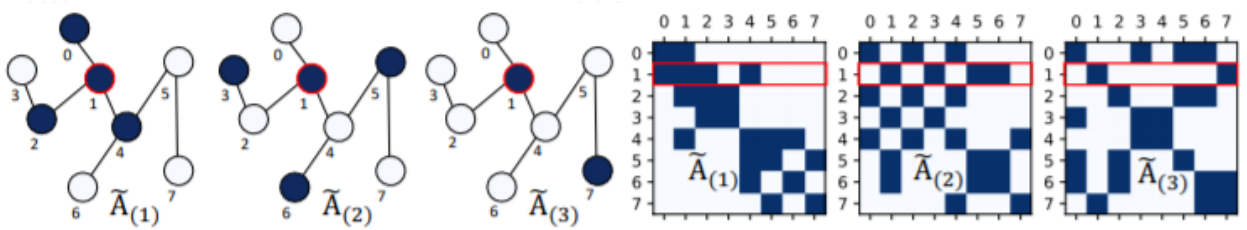
1) Disentanglement

```
A_scales = [k_adjacency(A, k, with_self=True) for k in range(num_scales)]
A_scales = np.concatenate([normalize_adjacency_matrix(g) for g in A_scales])
```

2) Skip-connection

1) Disentanglement

- A_k must be a adjacency matrix between nodes with k distance



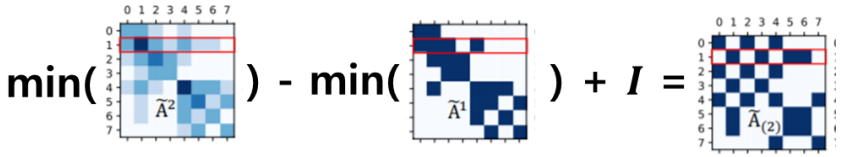
Implementation

- **Code in detail**
  1) Disentanglement (cont'd)
     - A_k must be a adjacency matrix between nodes with k distance

```
def k_adjacency(A, k, with_self=False, self_factor=1):
    assert isinstance(A, np.ndarray)
    I = np.eye(len(A), dtype=A.dtype)
    if k == 0:
        return I
    Ak = np.minimum(np.linalg.matrix_power(A + I, k), 1) \
       - np.minimum(np.linalg.matrix_power(A + I, k - 1), 1)
    if with_self:
        Ak += (self_factor * I)
    return Ak
```

Authors used a clever way to implement A_k It does NOT use any graph retrieval algorithms (including BFS) to find k-distance nodes, but apply power multiplication using the below equation.

$$min(A^k, 1) - min(A^{k-1}, 1) + I = \widetilde{A_k}$$



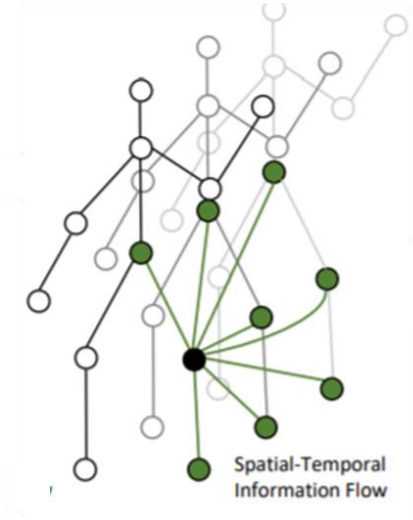
Implementation

- **Code in detail**

  2) Skip Connection

  - As mentioned earlier, the skip connection is applied by spanning the adjacency matrix

  - It connects the spatially adjacency nodes disregarding temporal relations



$$\Leftrightarrow \quad \tilde{\mathbf{A}}_{(\tau)} = \begin{bmatrix} \tilde{\mathbf{A}} & \cdots & \tilde{\mathbf{A}} \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{A}} & \cdots & \tilde{\mathbf{A}} \end{bmatrix} \in \mathbb{R}^{\tau N \times \tau N}$$

Implementation

- **Code in detail**

  2) Skip Connection (Cont'd)

  - After normalization of the disentangled $\widetilde{A_k}$ matrix, concatenates the matrix to form the final adjacency matrix for GCN operation

```
A_scales = [k_adjacency(A, k, with_self=True) for k in range(num_scales)]
A_scales = np.concatenate([normalize_adjacency_matrix(g) for g in A_scales])
```

2) Skip-connection

$$[\mathbf{X}_{(\tau)}^{(l+1)}]_t = \sigma\left(\tilde{\mathbf{D}}_{(\tau)}^{-\frac{1}{2}}\tilde{\mathbf{A}}_{(\tau)}\tilde{\mathbf{D}}_{(\tau)}^{-\frac{1}{2}}[\mathbf{X}_{(\tau)}^{(l)}]_t\Theta^{(l)}\right)$$

Implementation

- **Result**
  - Kinetics Validation dataset (Joint)
    - Similar to the paper's result

```
100%|██████████| 619/619 [06:21<00:00,  1.62it/s]
Accuracy:  0.35800161648817 94  model:  pretrain_eval/kinetics/joint
[ Fri Jun 19 23:49:58 2020 ]    Mean test loss of 619 batches: 3.0841087549684123.
[ Fri Jun 19 23:49:59 2020 ]    Top 1: 35.80%
[ Fri Jun 19 23:49:59 2020 ]    Top 5: 58.57%
[ Fri Jun 19 23:49:59 2020 ] Done.
```

Methods	Kinetics Skeleton 400	
	Top-1 (%)	Top-5 (%)
ST-GCN [50]	30.7	52.8
AS-GCN [21]	34.8	56.5
ST-GR [18]	33.6	56.1
2s-AGCN [33]	36.1	58.7
DGNN [32]	36.9	59.6
MS-G3D Net	**38.0**	**60.9**

Discussion

- **논문에서 밝히듯 MS-G3D 모델 하나만으로는 기존 GCN의 성능을 넘지 못함**
  - 하지만 local indenpency를 부여한 것이 augmentation으로 작용함으로 인해서, GCN의 용량만 키운 것보다 GCN+G3D가 압도적으로 좋은 성능을 보여줌
  - 다양한 접근법이 시도되는 Skeleton-based action recognition 분야에서 다양한 dataset에서 SOTA 성능을 보여준 것은 놀라운 기록
- **일반 GCN model로의 확장가능성**
  1) Disentangle은 먼 거리에 대한 penalt를 0로 만들어 주고, 가까운 거리와 먼거리를 모두 동일한 선상에 두게 만들기 때문에 human skeleton domain에만 적용할 수 있는 방법론이라 생각됨
  2) Skip Connection 역시도 temporal window 내에 만 있다면 공간상으로 연결된 node(t는 달라도)는 모두 skip connection으로 이어주기 때문에 일반적인 GCN dataset에 적용시키는데는 한계가 있을 수 있음

  - 하지만 그만큼 human skelenton task에 대해 최적화를 잘 시킨 방법론으로도 해석 가능