Two-Stream Adaptive Graph Convolutional Networks for Skeleton-based Action Recognition

Yongwon Jo

School of Industrial and Management Engineering, Korea University

KOREA UNIVERSITY DMQA

Contents

Research Purpose

❖ **Two-Stream Adaptive Graph Convolutional Networks for Skeleton-based Action Recognition**

- 2019 Conference on Computer Vision and Pattern Recognition(CVPR)에서 발표된 논문
- 2022년 3월 25일 기준 548회 인용
- Spatial Temporal Graph Convolutional Networks (ST-GCN)의 문제점을 개선한 방법론

Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition

Lei Shi[1,2] Yifan Zhang[1,2*] Jian Cheng[1,2,3] Hanqing Lu[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{lei.shi, yfzhang, jcheng, luhq}@nlpr.ia.ac.cn

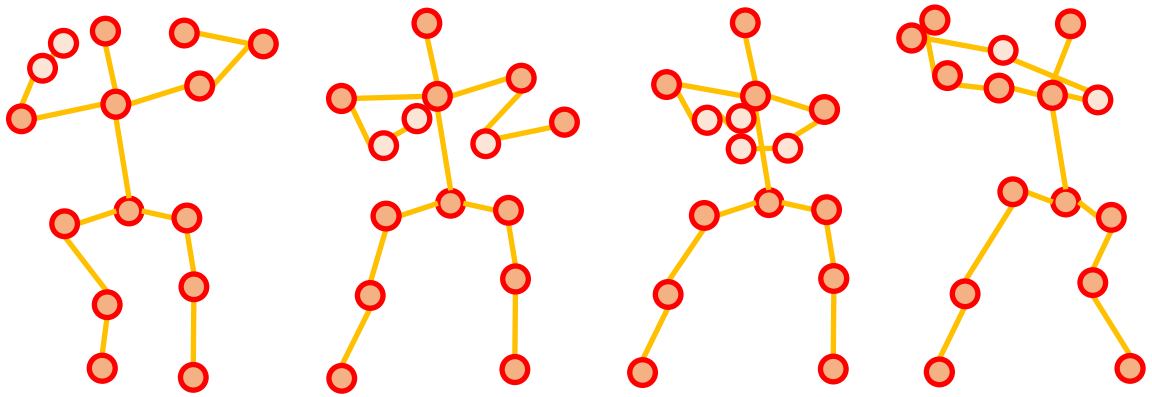
Research Purpose

❖ **Skeleton-based Human Action Recognition(HAR)**

- 인간 자세 추정(Human pose estimation) 결과를 입력 받아 행동을 인식하는 문제
- 인간 자세 추정 결과는 인간 주요 관절과 이들의 연결 정보로 구성
- 관절을 **Node(Vertex)**로 연결 정보를 **Edge**로 정의하고 이를 **Skeleton**이라 정의

영상 내 프레임

개별 프레임에 대한 HPE 추정 결과(입력 데이터)



Research Purpose

❖ **Spatial Temporal Graph Convolutional Networks (ST-GCN)**

- Skeleton을 Graph로 정의하고 이를 Graph Convolutional Networks로 특징 추출 및 행동 인식
- 인간 관절(Node)끼리 연결된 정보(Edge)는 동일하다는 가정으로 Adjacency matrix 정의
  - ➢ 육체적 연결 정보는 고정되어 있음
- 합성곱 연산 특징 상 먼 관절 사이 연산이 불가능하다는 단점
  - ➢ 예를 들어, 왼손과 오른 발 사이 관계 정보를 반영 불가

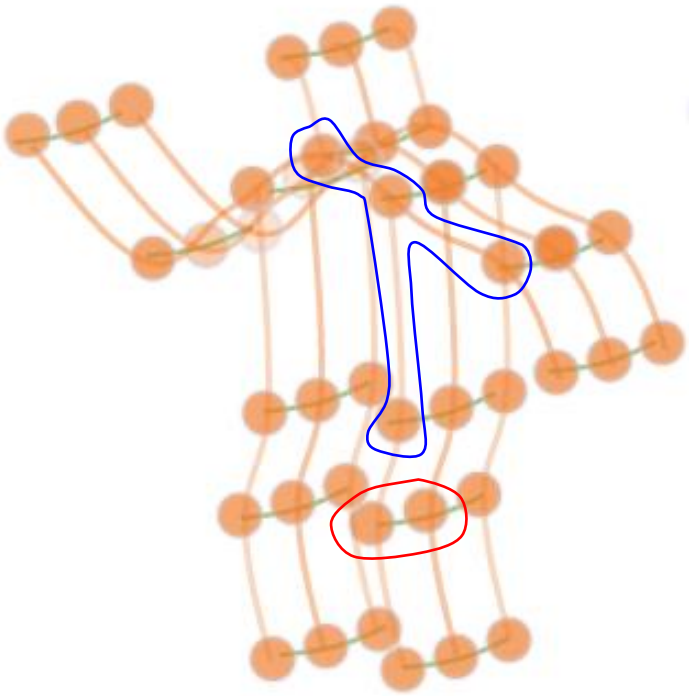
❖ **Two-Stream Adaptive Graph Convolution Networks (2S-AGCN)**

- 추가적인 Graph 를 정의해(Two-stream) 추가적인 관절 사이 연결 정보를 학습(Adaptive)
- 새롭게 정의한 Graph와 기존 Skeleton에서 추출한 Graph 를 학습(Graph)
- 행동 범주에 적합한 Graph학습을 통한 먼 관절 사이의 관계를 반영 가능(Data-driven)

Two-Stream Adaptive Graph Convolutional Networks

❖ **Two-Stream Adaptive Graph Convolutional Networks (2S-AGCN)**

- 기본적으로는 ST-GCN 과 동일한 형태를 가지는 Networks
- 단일 프레임 내 관절(Node)와 이들의 연결 정보(Edge) 를 추출하는 Spatial Convolutional Networks
- 프레임 사이 정보를 추출하는 Temporal Convolutional Networks



Spatial Convolutional Networks

Temporal Convolutional Networks

$$f_{out}(v_i) = \sum_{v_j \in B_i} \frac{1}{Z_{ij}} f_{in}(v_j) \cdot w(l_i(v_j))$$

- f: $feature\ map$
- v_i: $A\ vertex\ of\ the\ graph$
- B_i: $The\ sampling\ area\ of\ the\ convolution$
- w: $weights\ of\ the\ convolution$
- l_i: $a\ mapping\ function\ in\ ST-GCN$

Two-Stream Adaptive Graph Convolutional Networks

❖ Graph Convolution in the Spatial Dimension of ST-GCN

- 2S-AGCN 내 Spatial 차원에서 Graph Convolution 연산

$$f_{out} = \sum_{k}^{K_v} W_k (f_{in} A_k) \odot M_k$$

- K_v: $The\ kernel\ size\ of\ the\ spatial\ dimension (= 3)$
- $A_k = \Lambda_k{}^{-\frac{1}{2}} \overline{A_k} \Lambda_k{}^{\frac{1}{2}},\ where\ \overline{A_k}\ is\ the\ adjacency\ matirx\ of\ the\ subset\ graph\ in\ GCN$
- $\Lambda_k{}^{ii} = \sum_j \overline{A_k}^{ij} + \alpha\ , \alpha\ is\ 0.001$
- M_k: $Attention\ map$

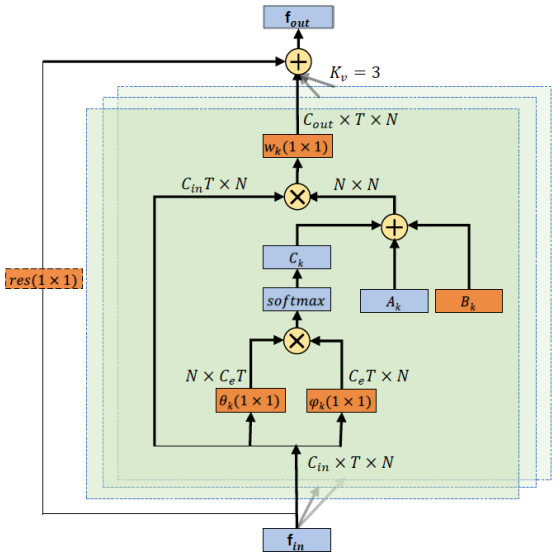
❖ Graph Convolution in the Temporal Dimension of ST-GCN

- Kernel size를 2로 고정하고 위 Graph Convolution 연산 진행

Two-Stream Adaptive Graph Convolutional Networks

❖ Adaptive Graph Convolution Layer

- 1st part(A_k): 기존 Adjacency matrix와 동일하며 관절 사이 물리적인 관계 반영
  - ➢ 머리와 목이 물리적으로 연결되어 있음을 의미하는 것
- **2nd part(B_k)**: Adjacency matrix와 동일하지만 matrix 내 원소들은 학습으로 결정(Data-driven features)
  - ➢ 물리적 연결 뿐만 아니라 연결의 강함 정도를 표현하며 이는 학습을 통해 값이 정해지는 것
- **3rd park(C_k)**: 데이터에 의해 결정되며 관절을 Embedding(θ_k, φ_k) 후, 관절간 유사도를 원소로 사용



Adaptive Graph Convolution layer

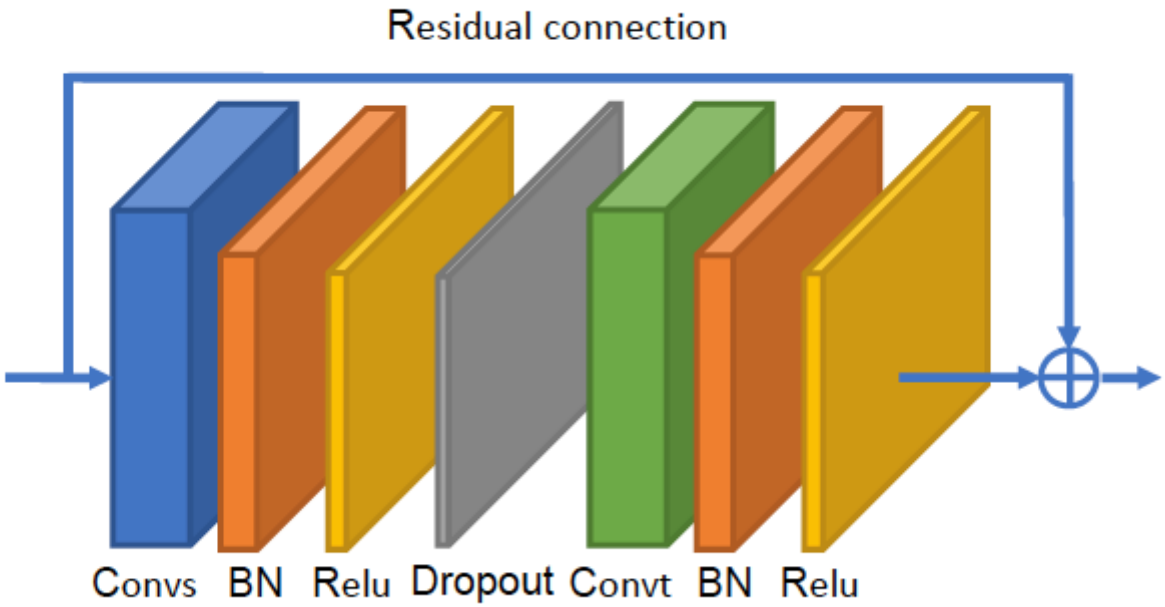
연산 방식

$$f_{out} = \sum_{k}^{K_v} W_k f_{in}(A_k + B_k + C_k)$$

Two-Stream Adaptive Graph Convolutional Networks

❖ **Adaptive Graph Convolution Block for the Spatial Dimension**

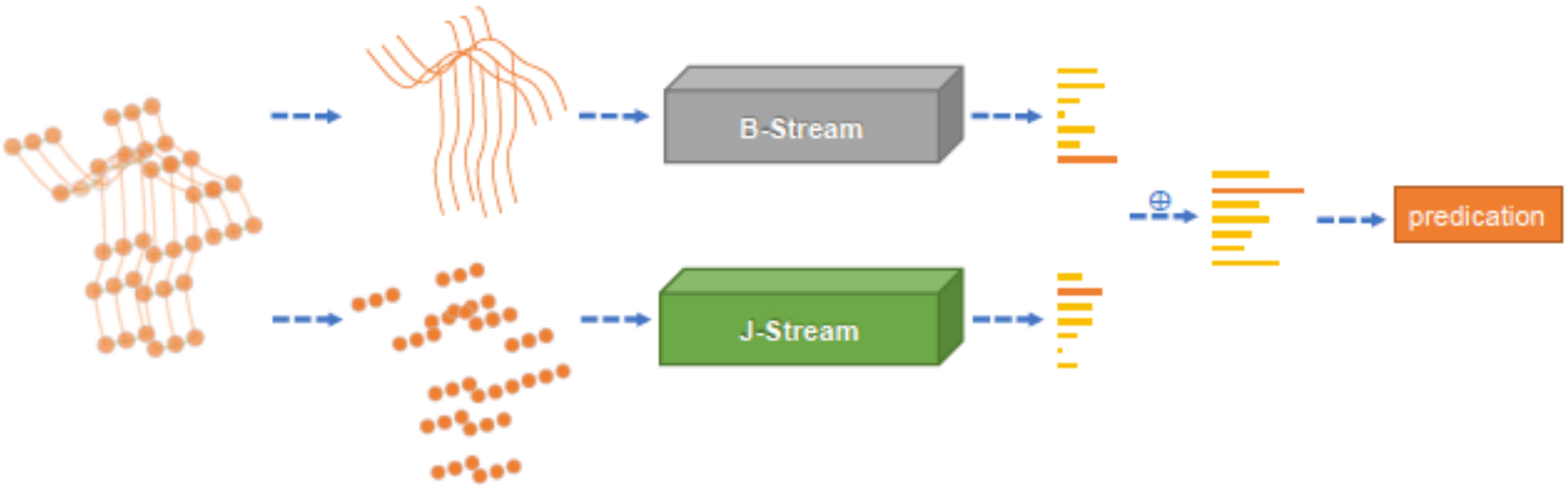
- Temporal dimension에 대한 Graph Convolution 연산은 ST-GCN과 동일
- 앞에서 설명한 Adaptive Graph Convolution layer로 구성된 Network Block
- (AGCN + Batch Normalization + ReLU) + Dropout + (AGCN + Batch Normalization + ReLU)
- Residual connection 기법을 사용해 안정적인 학습을 도모



Two-Stream Adaptive Graph Convolutional Networks

❖ **'2S'-AGCN 인가?**

- 전체 관절에 대한 중심 좌표를 중력 중심 좌표(The center of gravity)라 정의
- 중력 중심과 모든 좌표들을 연결하여 새로운 Skeleton 을 생성
- 기존 Skeleton(B-Stream)과 중력 중심 좌표를 포함하는 Skeleton(J-Stream)을 AGCN에 입력
- 두 AGCN에서 나온 Softmax 값을 더 해 최종 행동 범주 할당



Experiments

❖ 실험에 사용한 데이터 셋

- Kinetics: YouTube에서 수집한 동작 영상 30만 클립과 이에 대한 행동(400) 분류 정보 존재
  - ➢ OpenPose를 사용해 2차원 관절 좌표와 관절별 확률 값 산출
- NTU-RGB+D: 5만 6천 행동 영상과 행동 종류 60개, 3차원 관절 좌표 정보 존재
  - ➢ RGB 영상과 원근감 인식이 가능한 Depth 카메라로 인간 행동 영상을 촬영
- 2D Skeleton(Kinetics), 3D Skeleton(NTU-RGB+D) 각각을 입력 데이터로 하는 ST-GCN 학습

Kinetics

NTU-RGB+D

- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.
- Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010-1019).

KOREA UNIVERSITY DMQA

Experiments

❖ ST-GCN 및 연결 관계 정보(A, B, C) 존재 여부에 따른 성능 비교

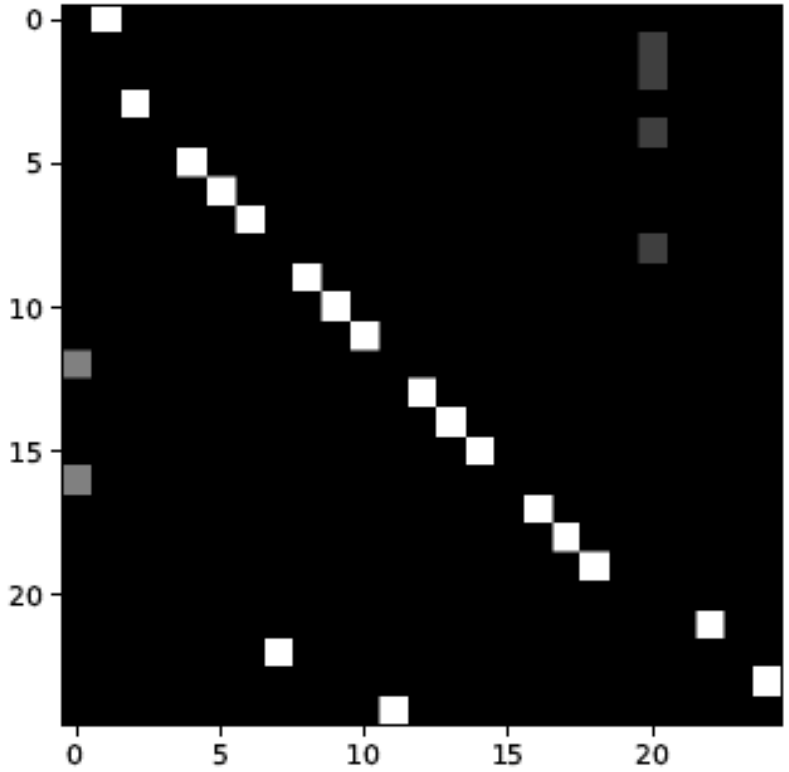
- Attention 정보를 포함하는 M 행렬이 Graph Convolution에 포함되어 성능 증가
- ST-GCN 대비 다양한 연결 관계 정보를 제공함으로써 성능 향상
- 연결 정보들이 조합되어 최종적인 성능 향상 성공

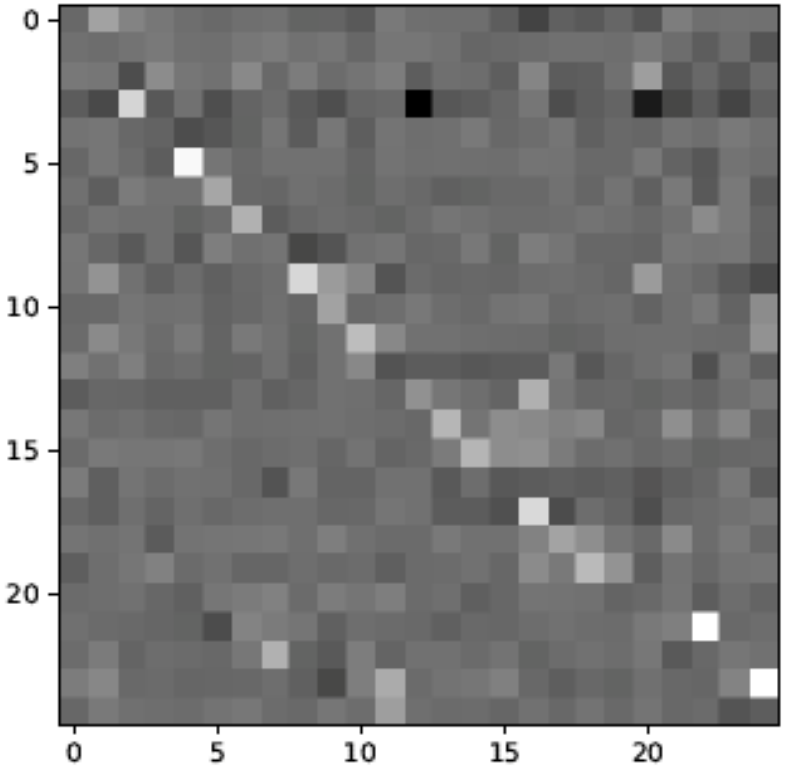
Methods	Accuracy (%)
ST-GCN	92.7
ST-GCN wo/M	91.1
AGCN wo/A	93.4
AGCN wo/B	93.3
AGCN wo/C	93.4
AGCN	93.7

Experiments

❖ 연결 관계 정보(B, C) 추가에 따른 Attention 정보 변경 발생

- (Left) 인간의 연결 정보(Adjacency matrix) vs (Right) 학습된 Adjacency matrix(B)
- 학습된 B Matrix는 관절 위치 상 먼 경우에도 이들의 특징을 반영할 수 있음을 확인 가능

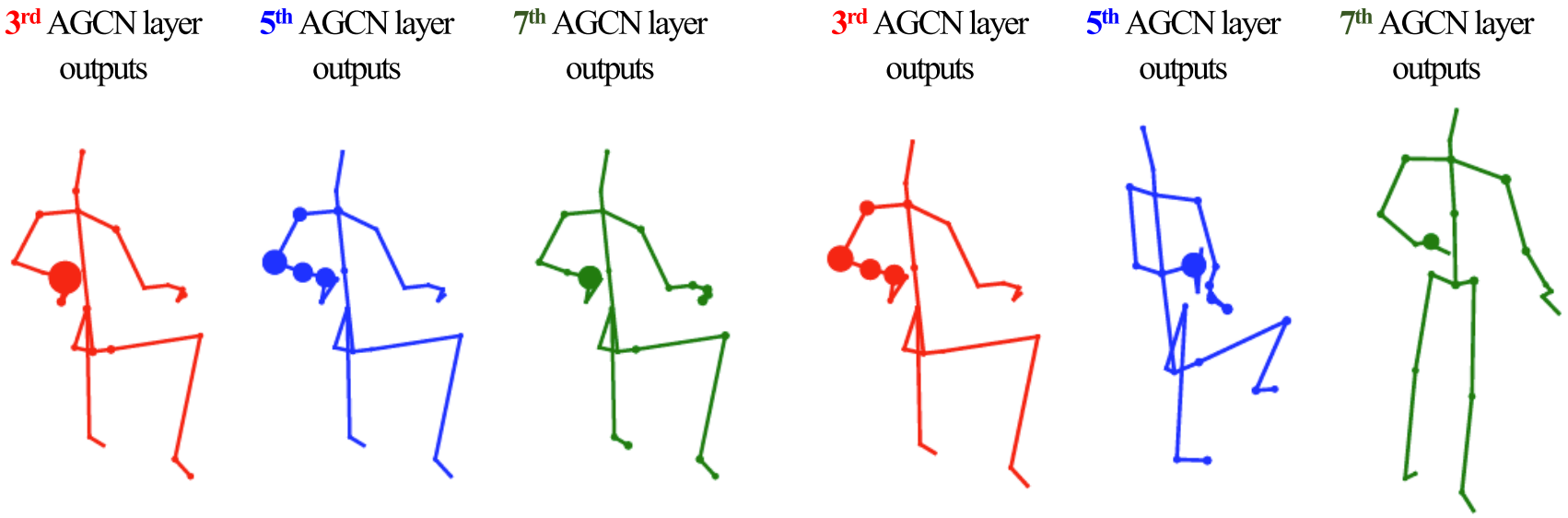




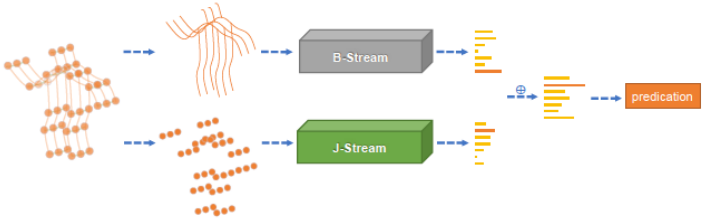
Experiments

❖ 연결 관계 정보(B, C) 추가에 따른 Attention 정보 변경 발생

- AGCN layer 출력 층 변경에 따른 관절이 행동에 미치는 정도를 시각화한 그림
- 2S-AGCN 초기에는 주변 관절 사이 연결 정도가 강함을 확인 가능(원의 크기↑)
- 하지만 여러 AGCN layer 통과 후 반대 쪽 손목에서도 연결 정도가 강해짐을 확인



Experiments

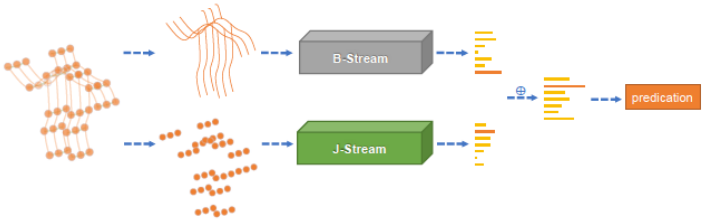


❖ Two-Stream 형태와 Single-Stream 형태의 AGCN 성능 비교

- 여러 형태의 Graph를 입력함에 따라 성능 증가 성공
- 하지만 J-Stream만 사용했을 때 성능이 좋다는 것은 의문
- 중력 중심(Center of the gravity)가 어떠한 영향을 미쳤는지 살펴보고 싶다는 생각

Methods	Accuracy (%)
Js-AGCN	93.7
Bs-AGCN	93.2
2s-AGCN	95.1

Experiments



❖ State-of-the-art 모델과의 성능 비교

- X-Sub: 학습 데이터 내 행동하고 있는 인간이 검증 데이터 내에는 존재하지 않는 경우
- X-View: 동일 행동을 두 카메라를 사용해 촬영해 카메라를 기준으로 학습/검증 데이터 구분
- 기존 순환 신경망(Recurrent neural networks, RNN) 계열 대비 뛰어난 성능
- 합성곱 신경망(Convolutional neural networks, CNN) 계열보다도 뛰어난 성능

NTU-RGBD Dataset

Methods	X-Sub (%)	X-View (%)
Lie Group [31]	50.1	82.8
HBRNN [6]	59.1	64.0
Deep LSTM [27]	60.7	67.3
ST-LSTM [22]	69.2	77.7
STA-LSTM [29]	73.4	81.2
VA-LSTM [33]	79.2	87.7
ARRN-LSTM [19]	80.7	88.8
Ind-RNN [20]	81.8	88.0
Two-Stream 3DCNN [21]	66.8	72.6
TCN [14]	74.3	83.1
Clips+CNN+MTLN [13]	79.6	84.8
Synthesized CNN [23]	80.0	87.2
CNN+Motion+Trans [18]	83.2	89.3
3scale ResNet152 [17]	85.0	92.3
ST-GCN [32]	81.5	88.3
DPRL+GCNN [30]	83.5	89.8
2s-AGCN (ours)	88.5	95.1

Kinetics-Skeleton Dataset

Methods	Top-1 (%)	Top-5 (%)
Feature Enc. [8]	14.9	25.8
Deep LSTM [27]	16.4	35.3
TCN [14]	20.3	40.0
ST-GCN [32]	30.7	52.8
Js-AGCN (ours)	35.1	57.1
Bs-AGCN (ours)	33.3	55.7
2s-AGCN (ours)	36.1	58.7

Conclusion

❖ Conclusion

- 2S-AGCN은 그래프 연결 정보를 Data-driven하게 정의하고 이를 예측 모델에 반영
- 기존에 정의된 Skeleton과 중력 중심을 포함한 새로운 Skeleton 정의
- 두 기여점이 정확히 반영되어 Skeleton-based HAR 문제의 State-of-the-art 달성

Thank you