요약

- 제안 기법 (굵은 글씨가 확인된 기존 논문과 다른점)
    - **Joint Position과 joint별 다음프레임 position과의 차이 값을 concat**
    - learnable adjacency matrix 사용
    - **2s-AGCN의 AGC block에서 Dropout과 Convt 사이에 Temporal Channel-Wise Attention(TCA)모듈을 추가해서 block input feature와 시간축간 attention을 구현**
    - 구현 한계로 uni-labeling 사용(=kernel size 1)
- 실험 내용이 빈약하고 uni-labeling로만 실험했다는 것이 아쉬운 점

제안 기법

- Joint Position과 joint별 다음프레임 position과의 차이 값을 concat (그림 (2)의 Temporal difference 블록)
  입력 차원 [배치,프레임,관절, 3D position] => [배치,프레임,관절,6]
- learnable adjacency matrix 사용
- 2s-AGCN의 AGC block에서 Dropout과 Convt 사이에 Temporal Channel-Wise Attention(TCA)모듈을 추가 (그림 4, 6)
  - 2018년도 논문 'CBAM: Convolutional Block Attention Module'의 구조를 변경하여 사용
- 구현 한계로 uni-labeling 사용(=kernel size 1)

$$f_{out} = \sum_{i=1}^{n} W_i f_{in} \bar{A}_i$$

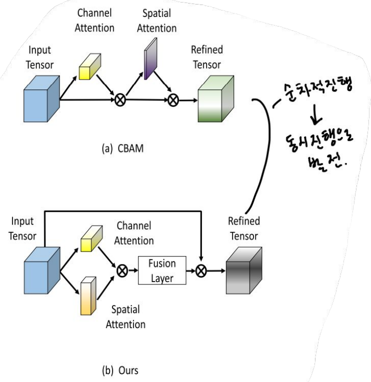
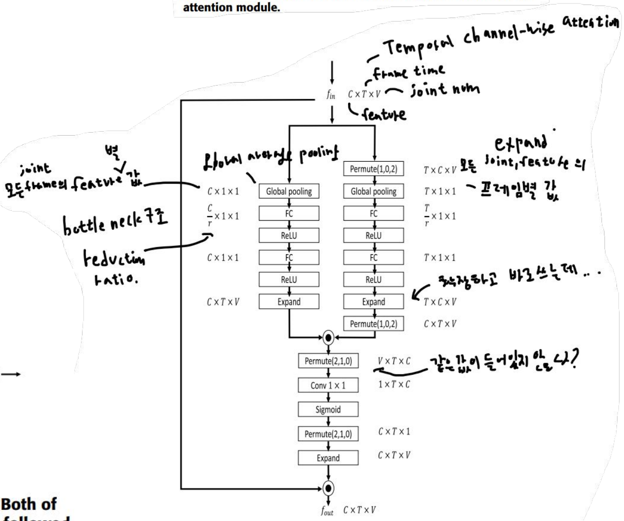


FIGURE 5. Illustration of the differences between CBAM and our attention module.



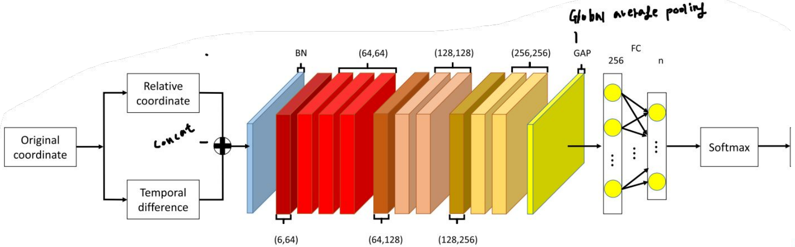


FIGURE 2. Illustration of the network architecture. The original coordinate is transformed into the relative coordinate and temporal difference. The relative coordinate and temporal difference are concatenated to form the network input. The network has ten pseud convolutional blocks. The pseudo graph convolutional block with different numbers of input and output channels uses a temporal s 2 to reduce the sequence length. A global average pooling layer is used after the last block. Fully connected layers and softmax are after the global average layer to compute scores of each class. Note that a and b in (a, b) mean the number of input channels of thi is a and the number of output channels of this block is b. ⊕ means that two tensors are concatenated.

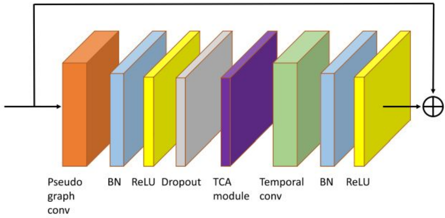


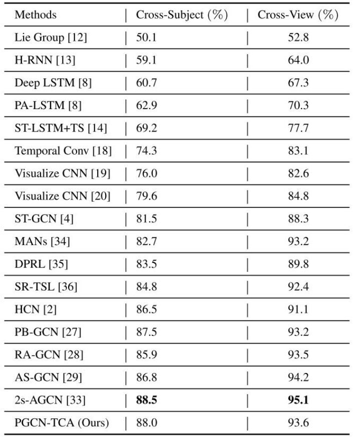
FIGURE 6. Illustration of our pseudo graph convolutional block. Both of the pseudo graph convolution and the temporal convolution are followed by a BN layer and a ReLU layer. The proposed mixed temporal and channel-wise attention module is applied between dropout layer and temporal convolution layer. In addition, a residual connection is added for each block except the first one.

FIGURE 4. Illustration of the proposed TCA module. The input f_{in} is a $C \times T \times V$ tensor, where C denotes the number of channels, T denotes the temporal length, and V denotes the number of joints. r is the reduction ratio. ⊙ denotes Hadamard product.

실험 결과

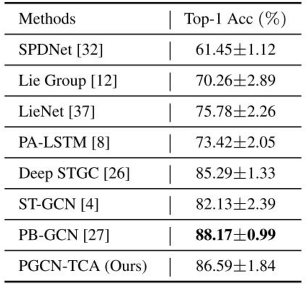
- 표 1) NTU-RGB+D 에서 타 논문 모델과 비교
- 표 2) HDM05 에서 타 논문 모델과 비교
- 표 3) 제안 기법을 제거했을때 정확도 비교
- 그림 8) Learnable Adjacency Matrix 사용에 대한 시각화
    - ○ 첫줄 3개 : ST-GCN의 Adjacency Matrix 시각화 (학습을 하지 않으므로 고정된 메트릭스값)
    - ○ 둘째줄~셋째줄 : PGCN-TCA 모델에서 10개의 PGC block 별 Adjacency Matrix 시각화

TABLE 1. Comparisons on the NTU-RGB+D dataset. The top-1 classification accuracies on both cross-subject and cross-view benchmarks are presented.



Methods	Cross-Subject (%)	Cross-View (%)
Lie Group [12]	50.1	52.8
H-RNN [13]	59.1	64.0
Deep LSTM [8]	60.7	67.3
PA-LSTM [8]	62.9	70.3
ST-LSTM+TS [14]	69.2	77.7
Temporal Conv [18]	74.3	83.1
Visualize CNN [19]	76.0	82.6
Visualize CNN [20]	79.6	84.8
ST-GCN [4]	81.5	88.3
MANs [34]	82.7	93.2
DPRL [35]	83.5	89.8
SR-TSL [36]	84.8	92.4
HCN [2]	86.5	91.1
PB-GCN [27]	87.5	93.2
RA-GCN [28]	85.9	93.5
AS-GCN [29]	86.8	94.2
2s-AGCN [33]	**88.5**	**95.1**
PGCN-TCA (Ours)	88.0	93.6

TABLE 2. Comparisons on the HDM05 datasets. The top-1 classification accuracy is presented.



Methods	Top-1 Acc (%)
SPDNet [32]	61.45±1.12
Lie Group [12]	70.26±2.89
LieNet [37]	75.78±2.26
PA-LSTM [8]	73.42±2.05
Deep STGC [26]	85.29±1.33
ST-GCN [4]	82.13±2.39
PB-GCN [27]	**88.17±0.99**
PGCN-TCA (Ours)	86.59±1.84

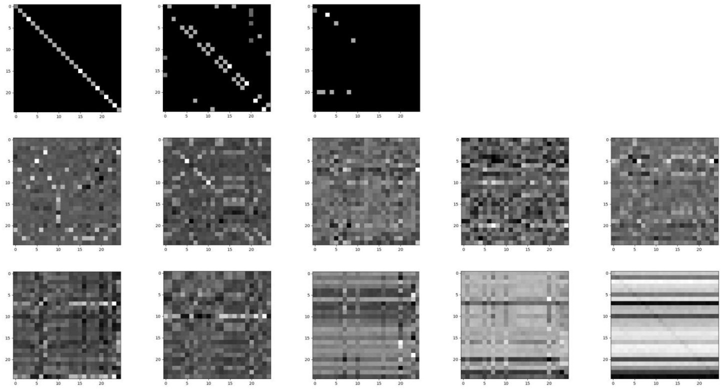
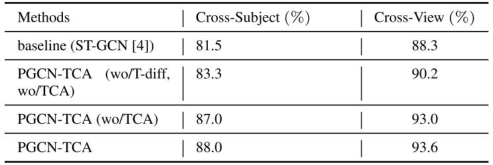


FIGURE 8. Visualization of the original adjacency matrices and our learned matrices on the NTU-RGB+D Cross-Subject benchmark. The figures in the first line are the original adjacency matrices used in ST-GCN [4] which are predefined and kept fixed through the training process. The figures in the second line are the learned adjacency matrices in our first five blocks and the others in the third line are the learned matrices in the last five blocks. The dark color means the value of the element is close to 0. The white color means the value of the element is close to 1.

TABLE 3. Comparisons of the top-1 accuracy when applying our model without X module. Note that wo/X means deleting the X module and T-diff means temporal difference.



Methods	Cross-Subject (%)	Cross-View (%)
baseline (ST-GCN [4])	81.5	88.3
PGCN-TCA (wo/T-diff, wo/TCA)	83.3	90.2
PGCN-TCA (wo/TCA)	87.0	93.0
PGCN-TCA	88.0	93.6