SDA 시니어 논문세션

Game State and Spatio-temporal Action Detection in Soccer using Graph Neural Networks and 3D Convolutional Networks

2026년 5월 23일(토)

Index

- 1. Introduction
- 2. Methodology
- 3. Evaluation
- 4. Critic
- 5. Future work

Introduction

• 축구 데이터 분석의 진화: 1950년대 수동 분석에서 현대의 카메라·AI 기반 자동 분석으로 발전

• 현재의 한계: 선수 추적(Tracking)은 자동화되었으나, 정밀한 이벤트 주석(Annotation)은 여전히 작업자의 수동 작업에 의존하여 비용과 시간이 많이 소요됨

• 기술적 난제: 기존 시공간 액션 탐지 모델은 개별 선수의 시각 정보에만 의존하여, 축구 특유의 복잡한 선수 간 상호작용 및 전술적 맥락을 파악하지 못함

Introduction

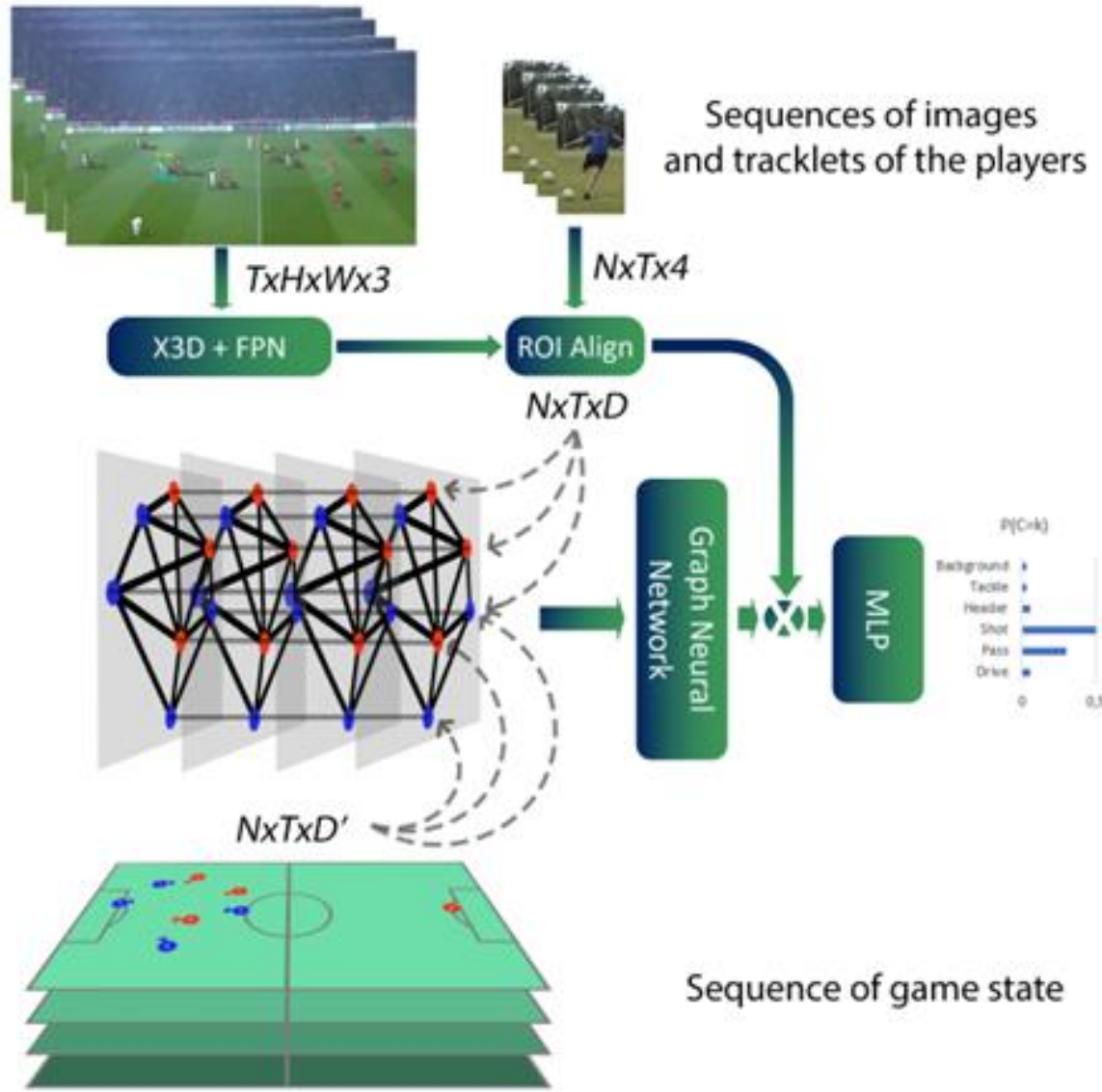
- Spatio-temporal Action Detection(시공간적 액션 탐지)

- 축구 선수의 행동은 서로 유기적으로 연결되어 있음. 따라서 주변 선수의 위치, 속도, 팀 정보를 추가하면 순수 시각 모델보다 더 정확한 예측이 가능할 것이라는 가설로 시작

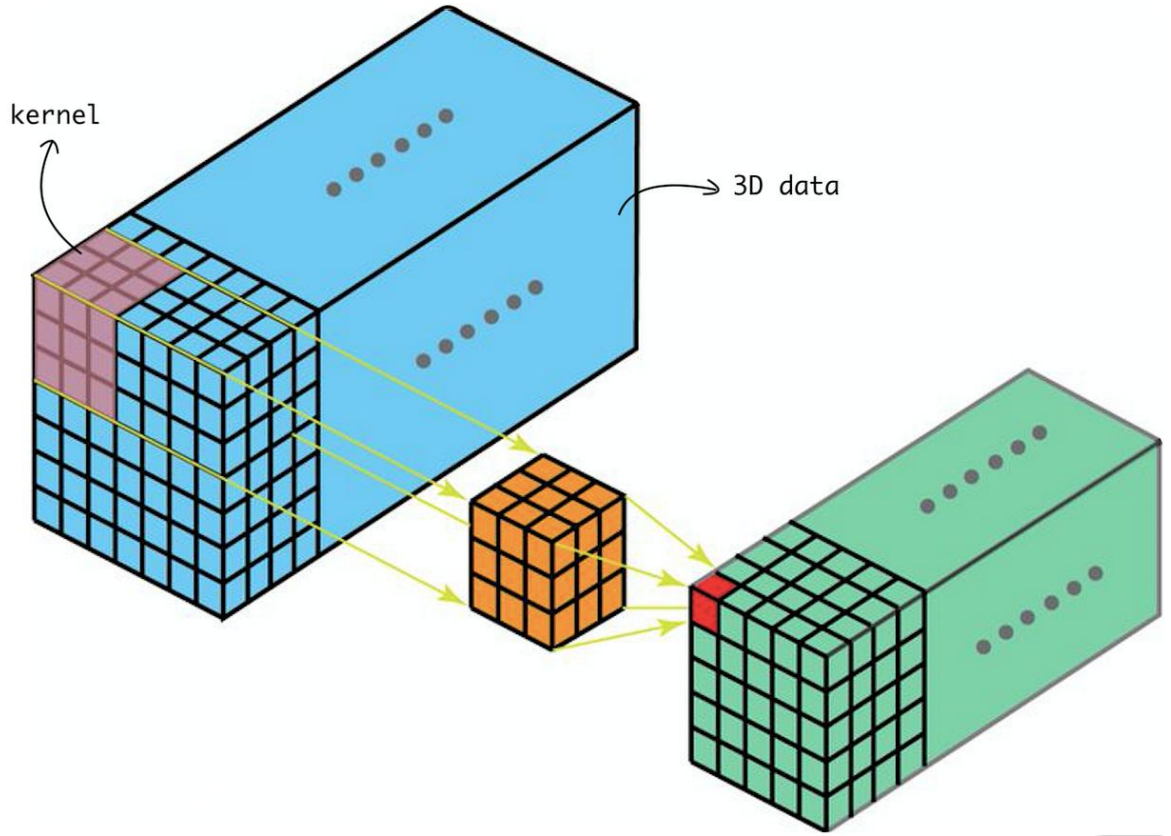
➔3D-CNN + GNN

- 시각적 특징과 게임 상태(Game state) 정보를 결합하는 하이브리드 모델 제안

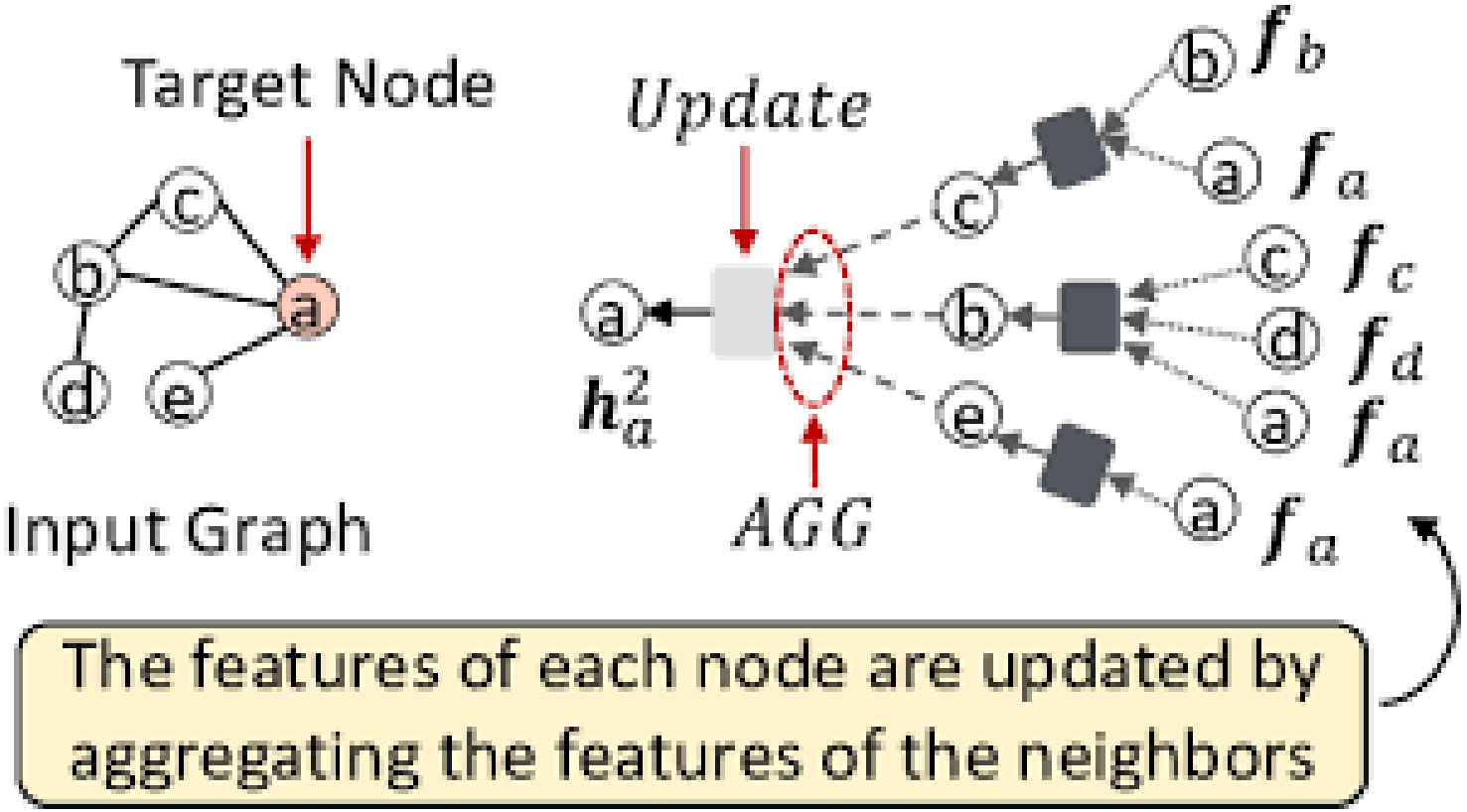
- 실제 현장에서 사용 가능하도록 단일 GPU(RTX A6000)에서 학습 가능한 효율적인 시스템 구축, computational cost(계산 비용) 제한

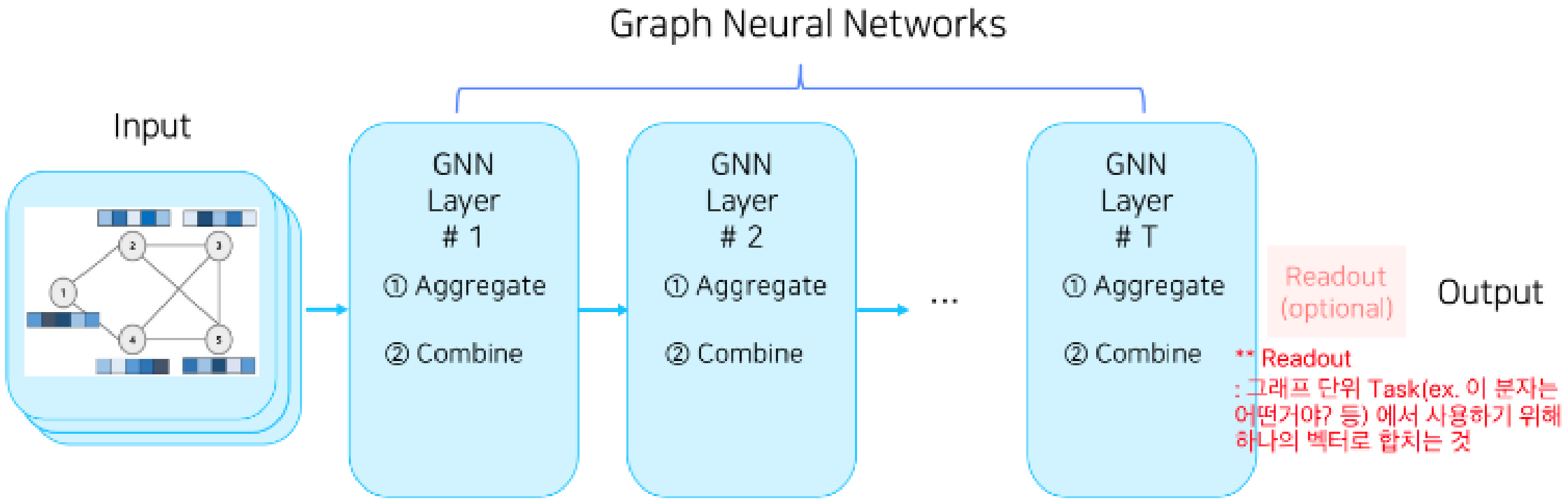

Sequences of images
and tracklets of the players
TxHxWx3
NxTx4
X3D + FPN
ROI Align
NxTxD
Graph Neural Network
MLP
P(C=k)
Background
Tackle
Header
Shot
Pass
Drive
0
0,5
NxTxD'
Sequence of game state

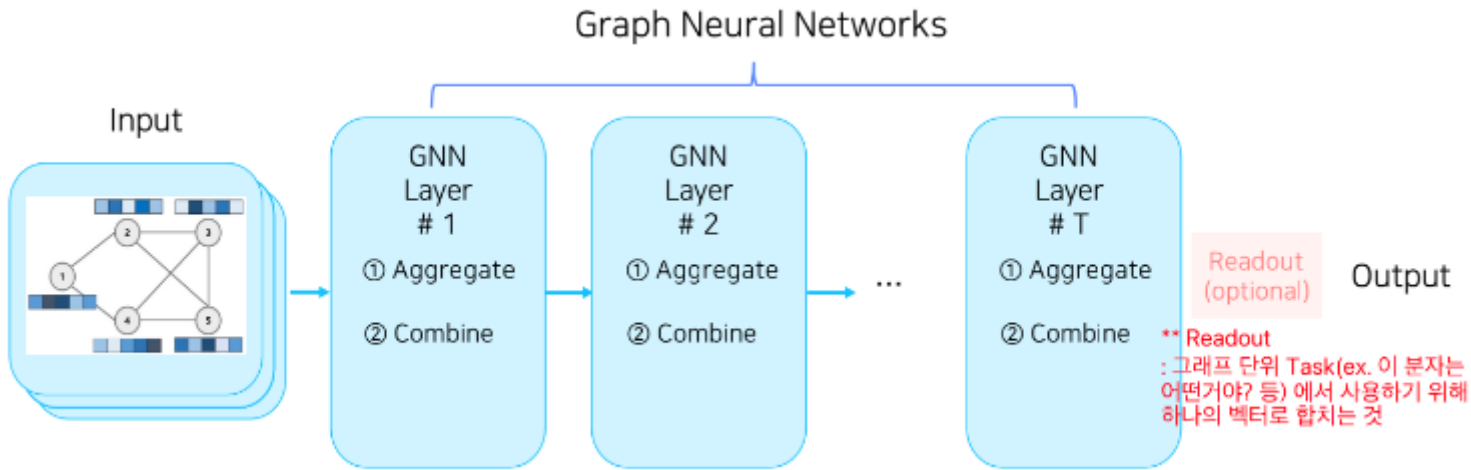
Video Understanding (3D CNN)

- 

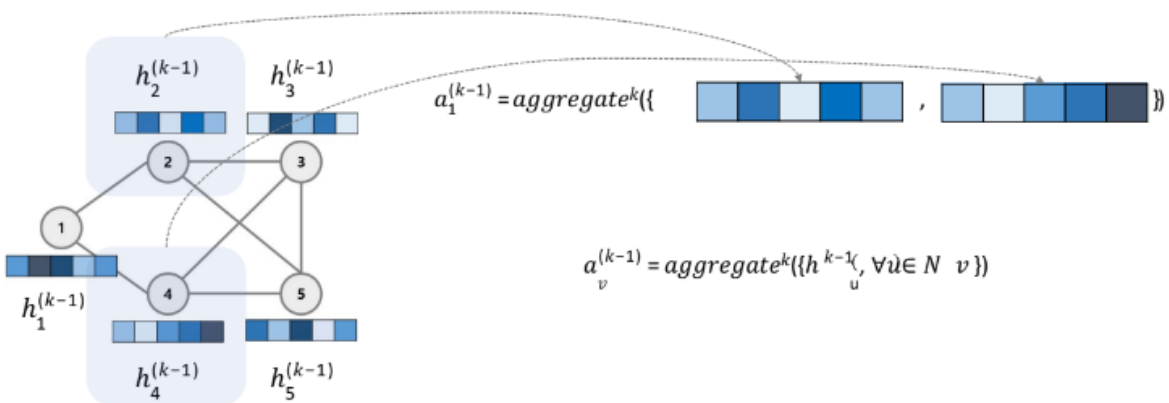
GNN (Graph Neural Network)



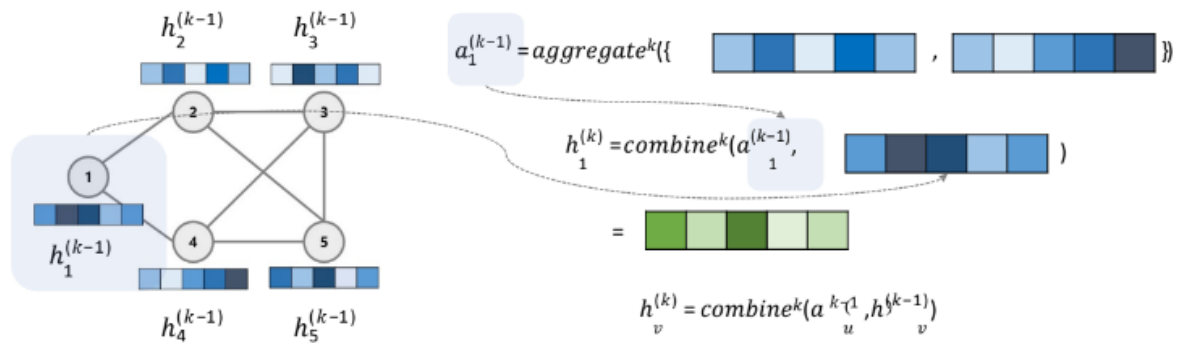

Graph Neural Networks
Input
GNN Layer # 1
① Aggregate
② Combine
GNN Layer # 2
① Aggregate
② Combine
...
GNN Layer # T
① Aggregate
② Combine
Readout (optional)
Output
** Readout
: 그래프 단위 Task(ex. 이 분자는 어떤거야? 등) 에서 사용하기 위해 하나의 벡터로 합치는 것



Aggregate(Message passing)



Combine(Update)



Methodology

- 기존 데이터셋(MultiSports)의 한계: 선수 추적 데이터 및 카메라 보정 모델 부재로 게임 상태 연구가 불가능함

		# classes	# inst.	Max #inst.	Min #inst.	# act. bbox	#non-act. bbox	Position & Velocity
Multisports	Aerobic gym.	21	8703	-	-	325k	0	X
Multisports	Volleyball	12	7645	-	-	139k	0	X
Multisports	Basketball	18	9099	-	-	213k	0	X
Multisports	Football	15	12254	-	-	225k	0	X
Multisports	Total	66	37701	3477	3	902k	0	X
Footovision	Football	8	19969	2500	2500	620k	19,9M	✓

Figure 2: Comparison of the characteristics of the MultiSports and Footovision datasets

Methodology

- 기존 데이터셋(MultiSports)의 한계: 선수 추적 데이터 및 카메라 보정 모델 부재로 게임 상태 연구가 불가능함

- Footovision 데이터셋 구축

    - 규모: 997개 경기, 20,000개 비디오 클립 (각 3초)

    - 8개 클래스: 드리블, 패스, 크로스, 헤더, 스로인, 슛, 태클, 블락

    - 포함 정보: 선수/심판 바운딩 박스, 트랙 ID, 팀 정보, 경기장 내 실제 위치(x,y) 및 속도

Methodology

After applying TAAD to our clip of T frames, with N players, we obtain a feature tensor $\Phi_{X3D} \in \mathbb{R}^{NxTx192}$, and we note $\Phi_{X3D}^{(i,t)}$ the visual feature vector of the player i and time t:

시각적 특징 추출

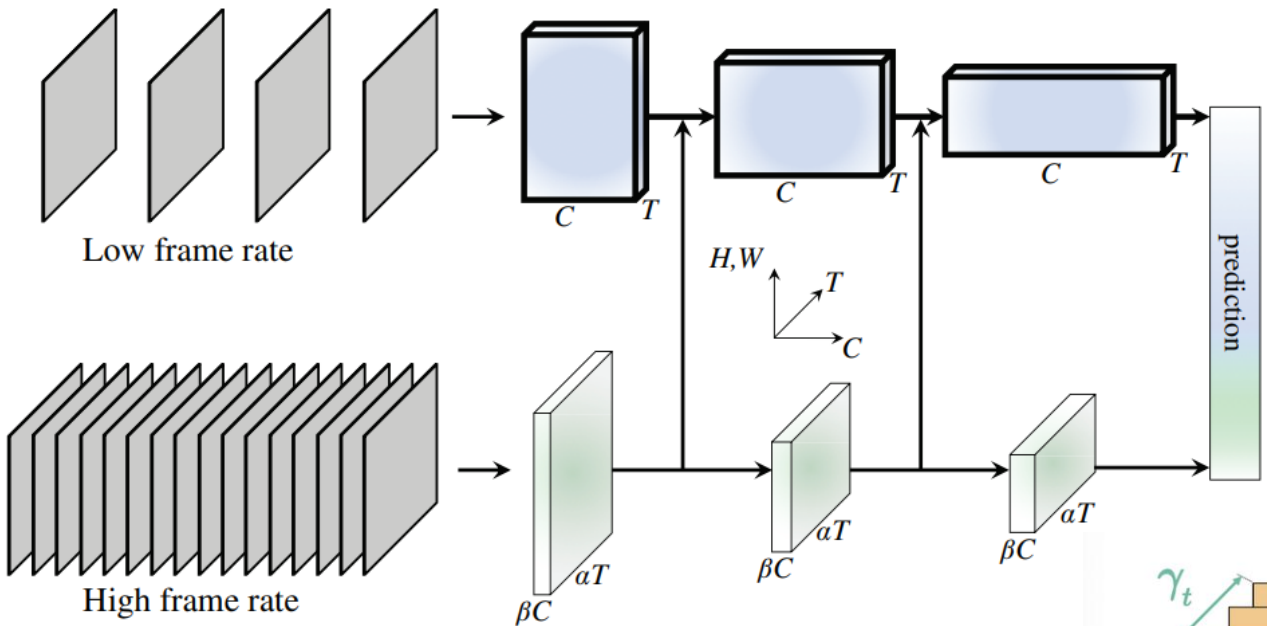
$$\Phi_{X3D}^{(i,t)} = \mathbf{ROI}(\mathbf{X3D}(sequence), bbox^{(i,t)}) \quad (1)$$

- 베이스라인 모델: 카메라 움직임에 강한 TAAD* 채택

- 백본 네트워크: 기존의 무거운 SlowFast 대신 경량화된 X3D 사용
  - 장점: 파라미터 수는 훨씬 적으면서 성능은 유사하여 과적합 방지 및 계산 효율성 높음
- 처리 방식: ROI Align**을 통해 선수의 경로를 따라 192차원의 특징 벡터 추출

** Track-aware action detector, 행위자를 먼저 찾고 그 경로(Track)를 따라 액션을 분석*

*** 샘플링 연산, 선수들의 바운딩 박스(Bounding Box) 데이터를 기반으로 X3D 백본의 결과물에서 각 선수의 시각적 특징 벡터(D=192)를 추출*



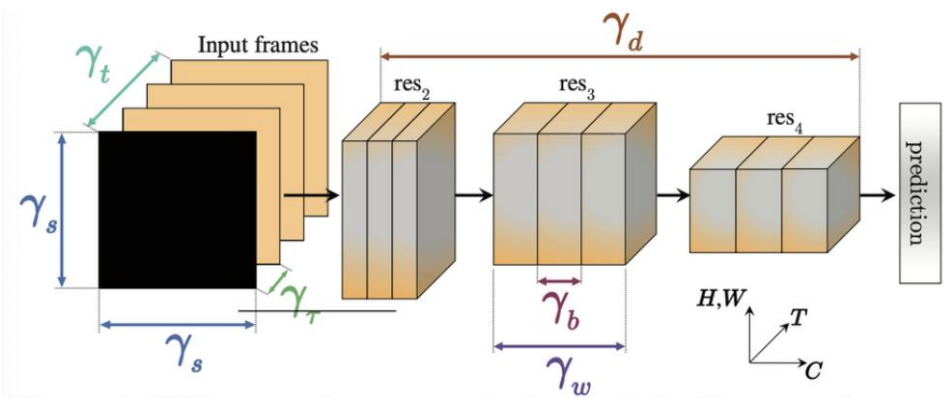


Figure 1. **X3D** networks progressively expand a 2D network across the following axes: Temporal duration γ_t, frame rate γ_τ, spatial resolution γ_s, width γ_w, bottleneck width γ_b, and depth γ_d.

Methodology

- its normalized position $(p_x^{(i,t)}, p_y^{(i,t)})$ at time t,
- its normalized velocity $(v_x^{(i,t)}, v_y^{(i,t)})$ at time t,
- its team membership, either 0 or 1,
- a projection of its visual feature vector $\Phi_{X3D}^{(i,t)}$ in a space of lower dimension (D'=64), noted $\varphi_{X3D}^{(i,t)}$.

$$h_u^{(k+1)} = MAX\left(MLP\left(h_u^{(k)} \middle\| h_v^{(k)} - h_u^{(k)}\right)\right) \quad (3)$$

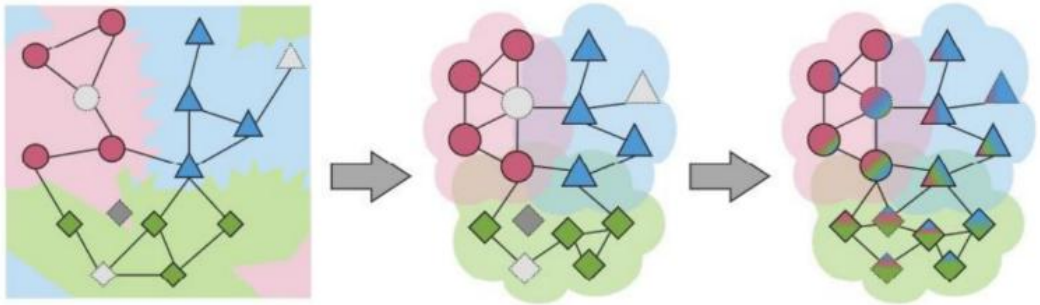
Where MLP is a shallow Multi-layer Perceptron and MAX a channel-wise max pooling operation.

로컬 게임 상태 인코딩(GNN)

$$\mathbf{x}_t^i = \left(p_x^{(i,t)}, p_y^{(i,t)}, v_x^{(i,t)}, v_y^{(i,t)}, 0|1, \varphi_{X3D}^{(i,t)}\right) \quad (2)$$

- 노드 구성: 각 타임스텝의 선수를 노드로 설정 (x_i^t = 위치 + 속도 + 팀 + 시각 특징 투영값)
- 그래프 구축: 동일 선수의 시간적 연결 + 가장 가까운 6명의 선수와 공간적 연결
- 에지 컨볼루션(Edge Conv): 노드의 전역 정보와 이웃과의 상대적 정보(상대 위치, 상대 속도 등)를 동시에 학습
- 3~4개 레이어만 사용하여 오버 스무딩(Over-smoothing) 방지

<주의> Over-Smoothing Issue



◦ GCN에서 Layer 개수별 정확도

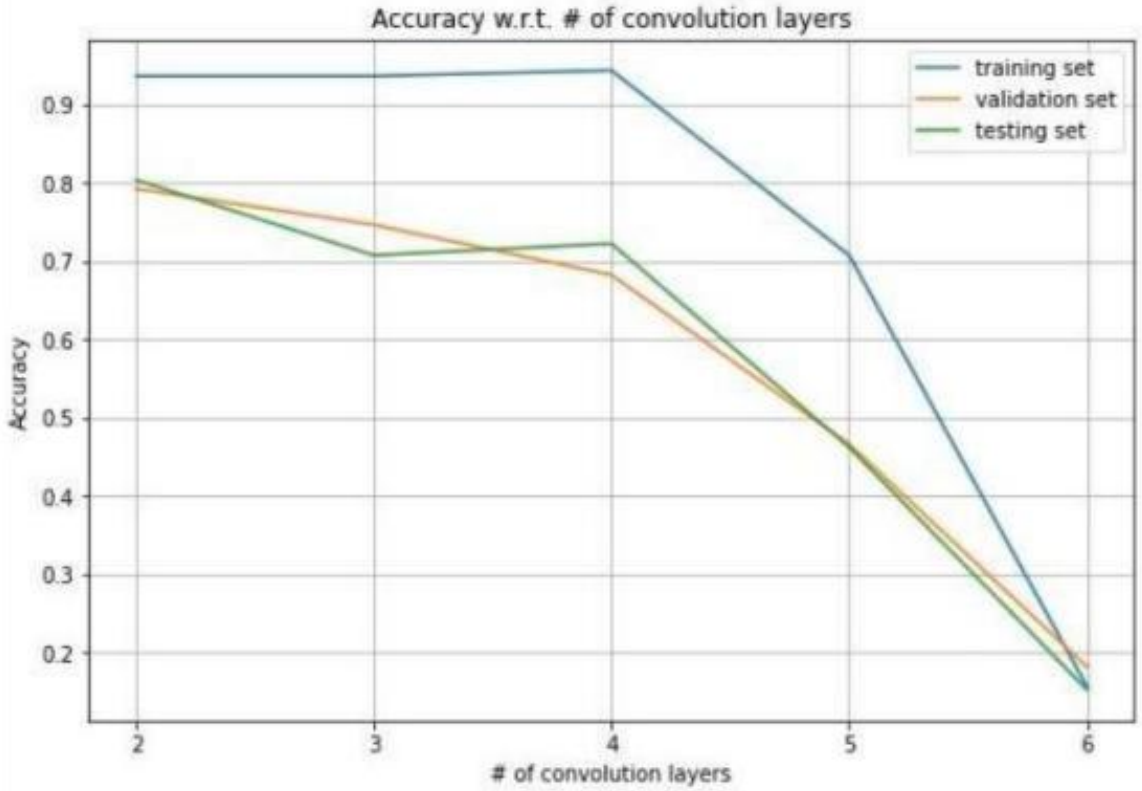
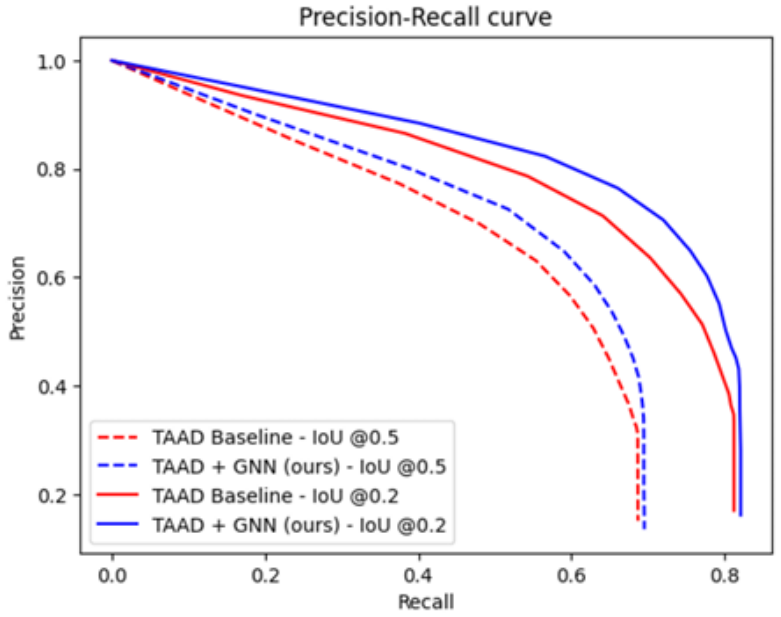


Figure 1: Accuracy of GCN on Cora w.r.t. # of convolution layers.

Evaluation

성능 비교 결과

Method	mAP	
	0,2	0,5
TAAD		
(X3D + FPN + TCN)	59,58	46,4
TAAD + GNN (ours)		
(X3D + FPN + Local Game State GNN)	**64,48**	**49,7**



- 평가 지표: mAP(IoU 임계값 0.2 및 0.5)

- TAAD 단독 모델보다 GNN 결합 모델의 성능이 모든 지표에서 우수함
  - mAP@0.2 59.58% -> 64.48%
  - mAP@0.5 46.4% -> 49.7%

- 구조화된 게임 상태 정보 통합이 탐지 정확도 향상에 기여함을 증명

Evaluation

Actions	Increase in number of True Positives	Decrease in number of False Positives
Ball-Drive	0,0%	23,0%
Pass	-0,5%	13,1%
Cross	-7,9%	39,9%
Throw	-0,4%	31,0%
Shot	3,9%	23,3%
Header	0,9%	29,0%
Tackle	13,8%	47,8%
Ball-Block	1,3%	36,0%
Overall	0,4%	29,0%

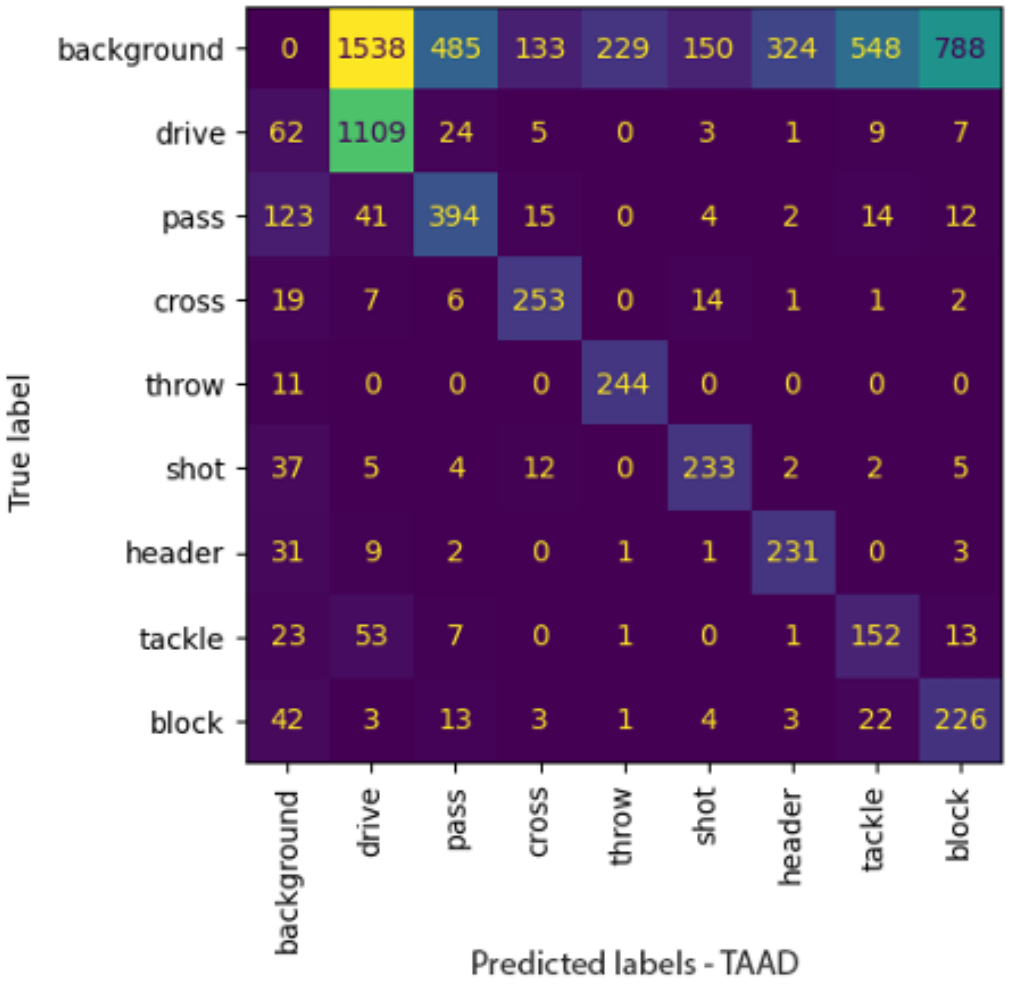
실무적 유용성 분석

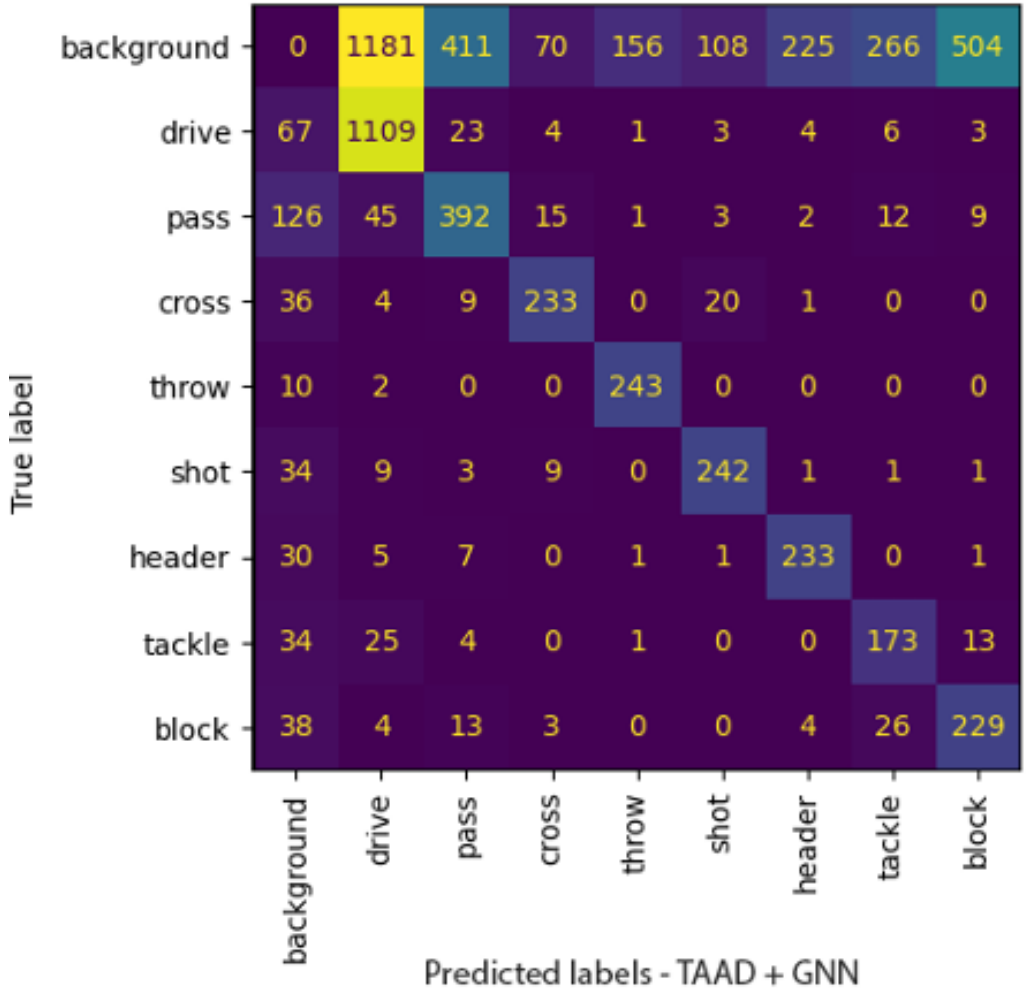
- 고재현율 모드: 데이터 분석 현장에서는 이벤트를 놓치지 않는 것이 중요함

- 오탐지(False Positives) 감소: 재현율 80% 수준에서 GNN 결합 모델은 기존 대비 오탐지를 약 29% 감소

- 태클: 오탐지 48% 감소, 정탐지 14% 증가 (가장 큰 개선)
- 크로스: 골문 근처에서 슛과 혼동되어 약 8% 하락

Evaluation


background 0 1538 485 133 229 150 324 548 788
drive 62 1109 24 5 0 3 1 9 7
pass 123 41 394 15 0 4 2 14 12
cross 19 7 6 253 0 14 1 1 2
throw 11 0 0 0 244 0 0 0 0
shot 37 5 4 12 0 233 2 2 5
header 31 9 2 0 1 1 231 0 3
tackle 23 53 7 0 1 0 1 152 13
block 42 3 13 3 1 4 3 22 226
True label
background drive pass cross throw shot header tackle block
Predicted labels - TAAD


background 0 1181 411 70 156 108 225 266 504
drive 67 1109 23 4 1 3 4 6 3
pass 126 45 392 15 1 3 2 12 9
cross 36 4 9 233 0 20 1 0 0
throw 10 2 0 0 243 0 0 0 0
shot 34 9 3 9 0 242 1 1 1
header 30 5 7 0 1 1 233 0 1
tackle 34 25 4 0 1 0 0 173 13
block 38 4 13 3 0 0 4 26 229
True label
background drive pass cross throw shot header tackle block
Predicted labels - TAAD + GNN

Critic

(+)

• 비논리적 탐지 제거: 시각 모델은 단순 제스처만으로 스로인을 오탐지할 수 있으나, GNN은 “이미 다른 선수가 공을 몰고 있다 " 는 맥락을 파악해 이를 차단함

• 실제 접촉 판단: 헛발질 상황에서 주변 선수들의 반응 (가속도 등)을 통해 실제 접촉 여부를 더 정교하게 판단함

Critic

(-)

• 밀집 상황 및 Occlusion: 선수들이 매우 가까이 붙어 있거나 겹쳐 있는 경우(공중볼 경합), 시각적 특징이 거의 동일해져 모델이 혼동함

• 의도 파악의 어려움: 크로스와 슛처럼 선수 위치와 동작이 유사한 액션은 여전히 구분하기 까다로움

Future work

향후 연구 방향

- 시간적 시야 확장: 현재의 2초(50프레임) 제한을 넘어선 장기 맥락 파악 필요
- Seq2Seq 모델 도입: 긴 이벤트 시퀀스를 통해 경기의 전체적인 논리와 역학을 학습
- 사후 정제 시스템: 전체 경기 지식을 바탕으로 탐지된 시퀀스의 노이즈를 제거하거나 사후 예측 수행
- 추가특징 결합: 미래 예측에 사용되는 다양한 게임 상태 특징들을 추가로 도입하여 모델 고도화