

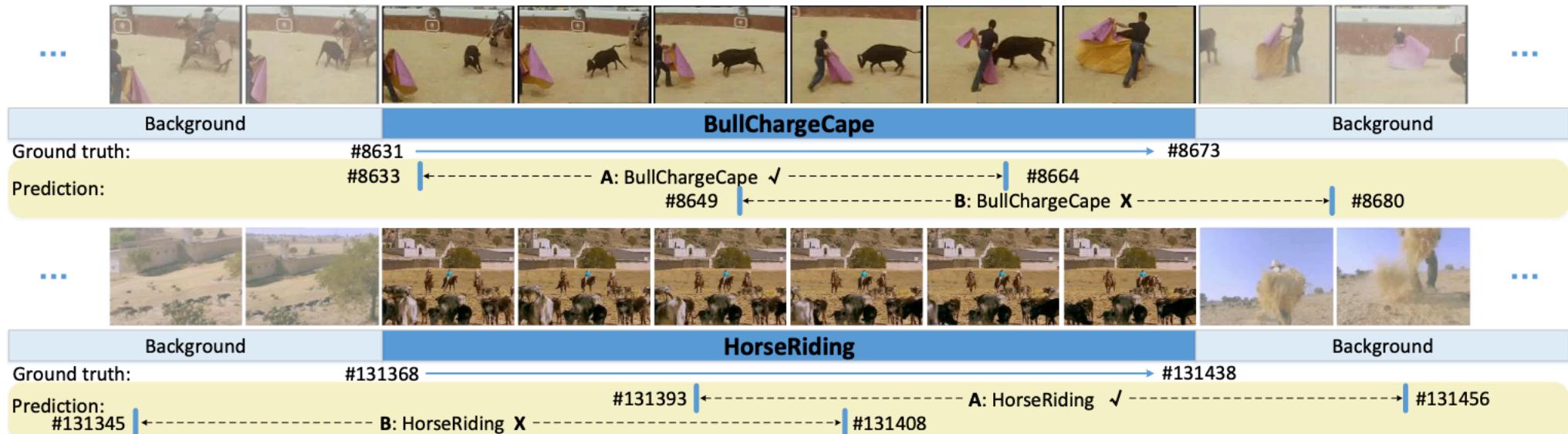
W-TALC: Weakly-supervised Temporal Activity Localization and Classification

Sujoy Paul, ECCV 2018

Introduction

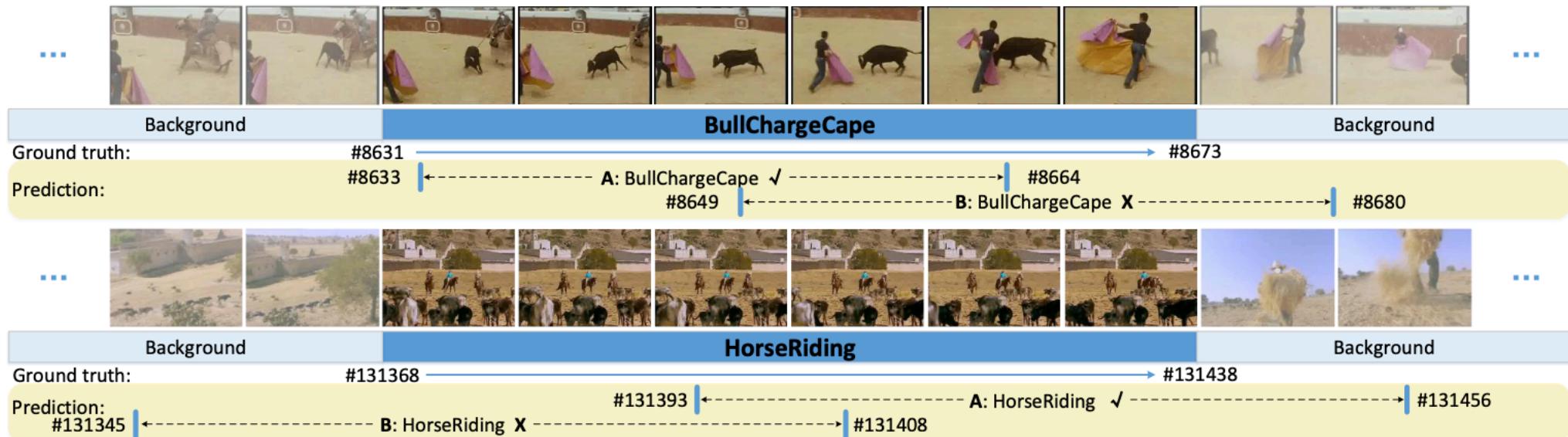
01. Temporal Action Detection

- Temporal Action Detection?
 - Untrimmed video에서 어디부터 어디까지 어떤 액션이다를 구하는 task
 - Proposal / Classification 두가지 관점에서 진행



01. Temporal Action Detection

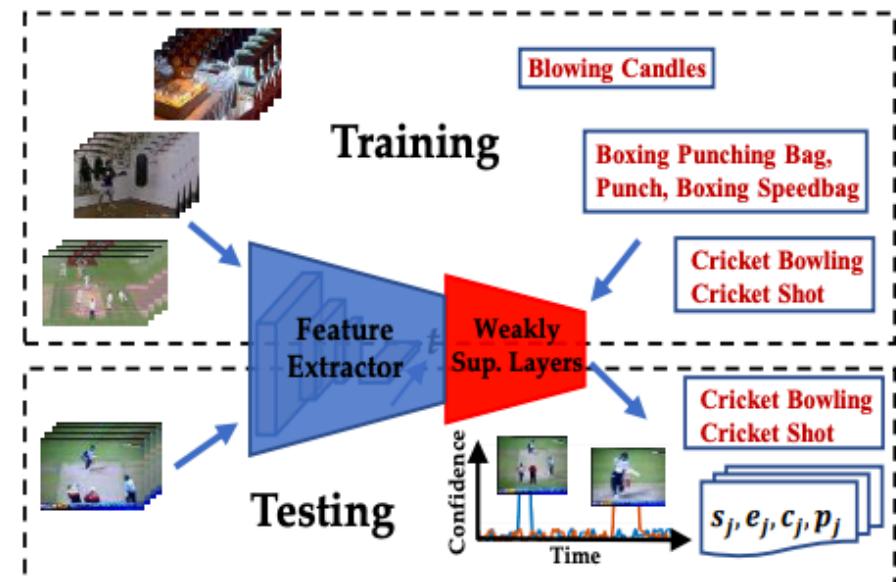
- Weak label
 - Video-label이 아닌 몇초부터 몇초까지 어떤 action이다 하는 정보를 얻는 것은 수작업 필요
 - Fully supervised setting으로부터 시작.
 - 인터넷에서 구할 수 있는 video에는 tag 존재 -> video, tag만 이용해서 action localization을 할 수 있지 않을까?
 - Weak-label을 사용한 **TALC(Temporal Activity Localization and Classification)** 제안



02. W-TALC

- Train, Test protocol
 - **Training set:** video label activity tag
 - video X : action A,A,B,C
 - **Output:** activity의 label만 측정하는 것이 아닌, activity의 시작 시간, 종료 시간, 카테고리, 신뢰도를 찾음

Fig. 1: This figure presents the train-test protocol of W-TALC. The training set consists of videos and the corresponding video-level activity tags. Whereas, while testing, the network not only estimates the labels of the activities in the video, but also temporally locates their occurrence representing the start (s_j) and end time (e_j), category (c_j) and confidence of recognition (p_j) of the j^{th} activity located by the model.



Related Works

01. Weakly supervised learning

- **Supervised learning**

- 목표값이 제시된 데이터가 학습 데이터로 주어진다. (data/label)
 - **Classification(분류)** : 데이터를 label에 따라 나누는 방법
 - Binary classification(예/아니오)
 - multi label classification(개/고양이/소..)
 - **Regression** : feature을 토대로 값을 측정. 결과는 연속된 그래프로 나온다.
- CNN, RNN을 주로 사용

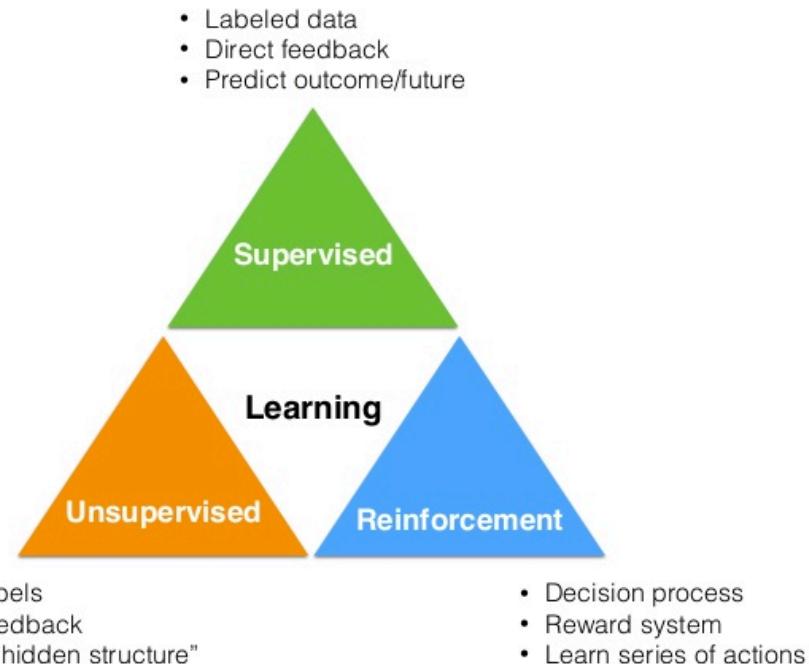
- **Unsupervised learning**

- 목표값이 제시되지 않은 데이터가 학습 데이터로 주어진다.
- 데이터의 숨겨진 특징(Hidden feature)이나 구조를 발견하는 데에 사용된다.
 - **Clustering(군집화)** : label이 주어지지 않은 데이터를, 일정한 개수의 cluster(군집)으로 모으는 방법
 - **Autoencoder** : Encoder(데이터->내부 표현), Decoder(내부 표현->데이터)로 구성된 모델.

01. Weakly supervised learning

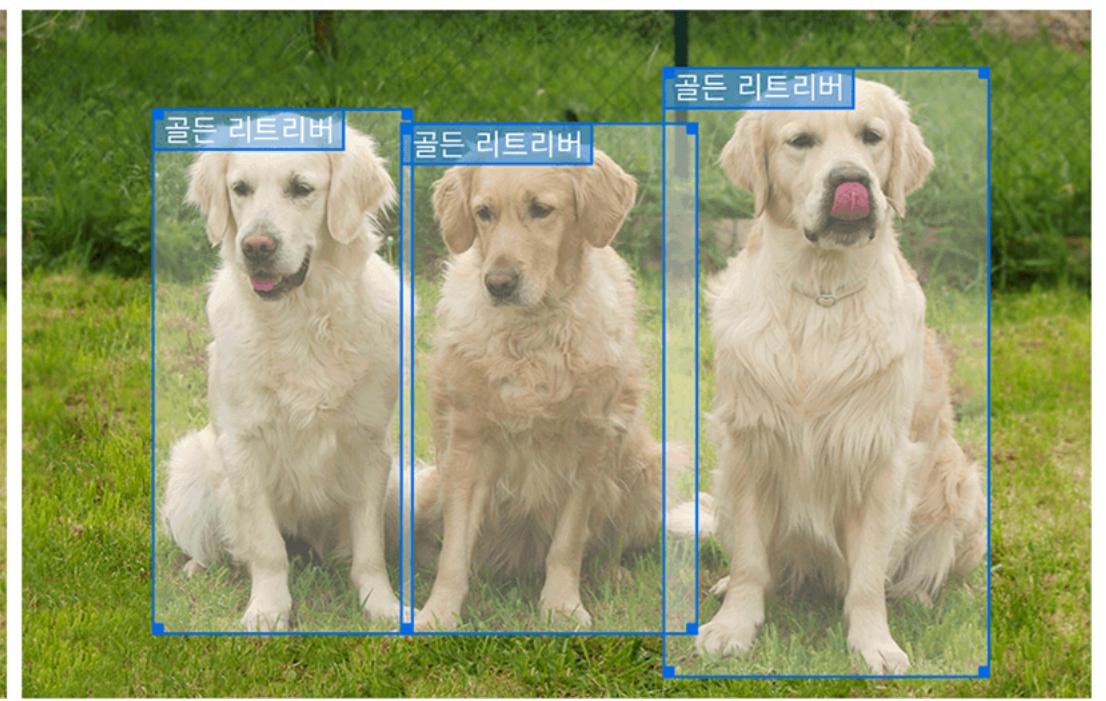
- **Reinforcement learning**

- Unsupervised learning의 일부.
- 어떤 state에서 action을 취하고 reward를 받아 점차 효율적인 방식으로 행동을 강화한다.
 - ex) Q-learning, DQN(Deep-Q-Network)



01. Weakly supervised learning

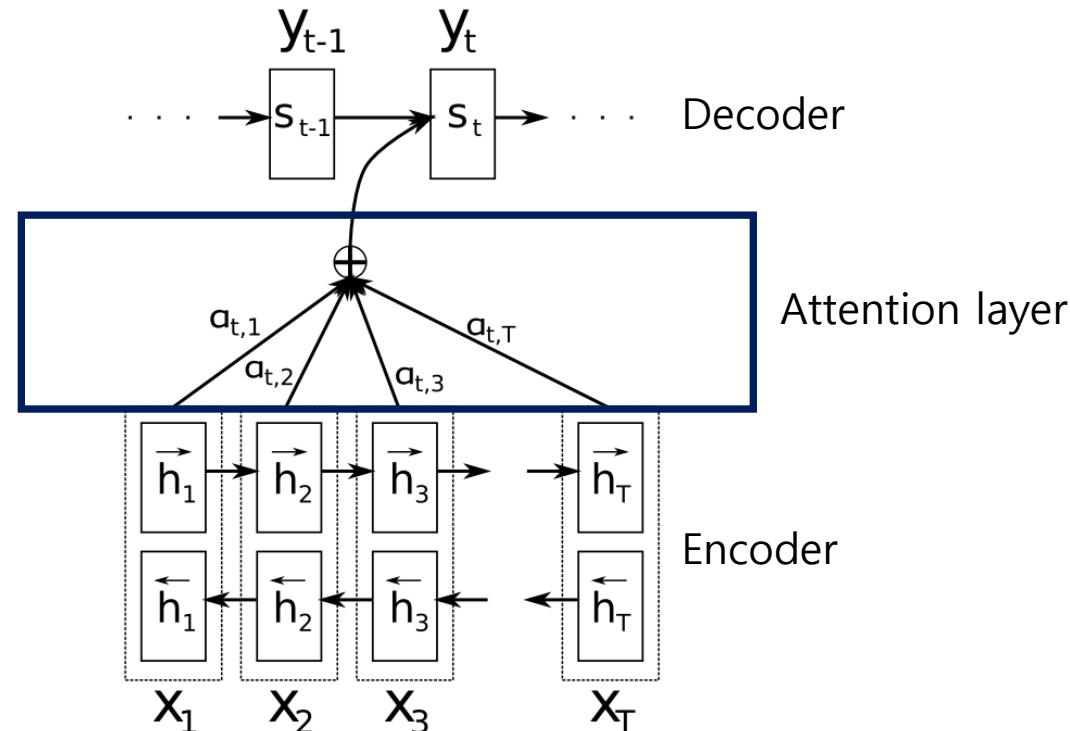
- Weakly supervised learning(약지도학습)
 - 반쪽짜리 정답set을 이용하여 supervised learning과 동일한 task를 수행
 - ex) weakly-supervised object localization
 - : 사물의 종류만 알려진 이미지 분류 dataset을 이용하여 **사물의 위치** 예측



02. Attention

- Attention

- Decoder에서 output을 예측하는 매 시점마다 encoder에서의 input을 다시 한번 참고하는 것
- 전체를 다 동일한 비율로 참고하는 것이 아닌, 해당 시점에서 연관이 있는 부분만 좀 더 attention해서 보는 것



03. Multiple Instance Learning

- MIL

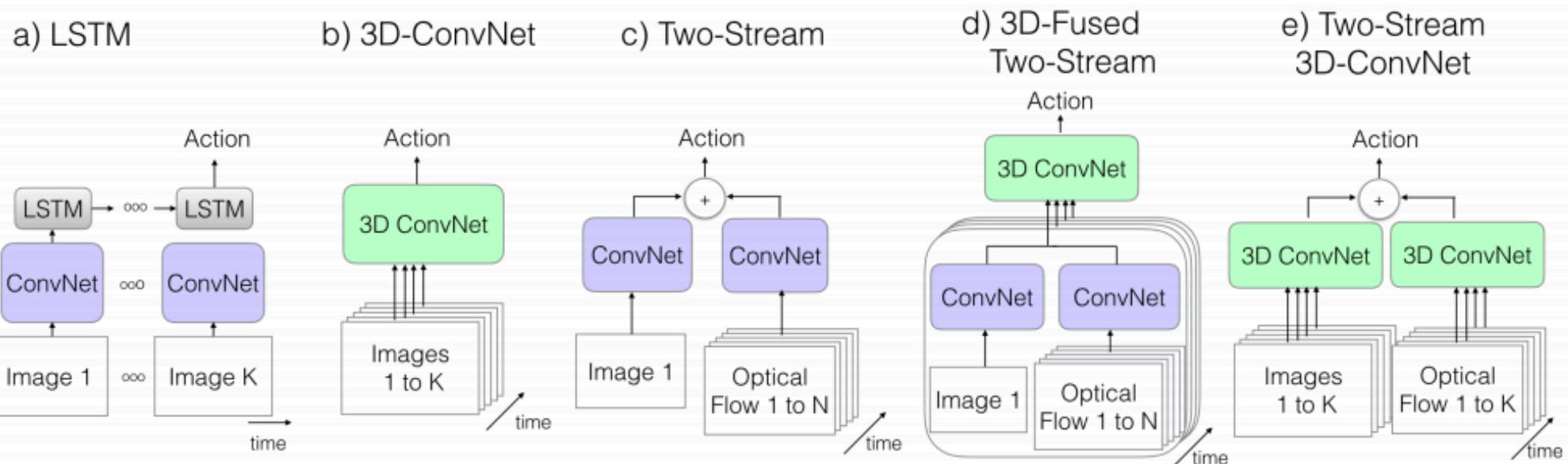
- Decoder에서 output을 예측하는 매 시점마다 encoder에서의 input을 다시 한번 참고하는 것
- 전체를 다 동일한 비율로 참고하는 것이 아닌, 해당 시점에서 연관이 있는 부분만 좀 더 attention해서 보는 것

Illustration of a MIL problem



04. Two stream network

- I3D
 - 3d convolution, inception v1 사용
 - Convolution filter를 time축으로 확장
 - weight를 시간축으로 n번 복제한 후, 나중에 1/n scaling

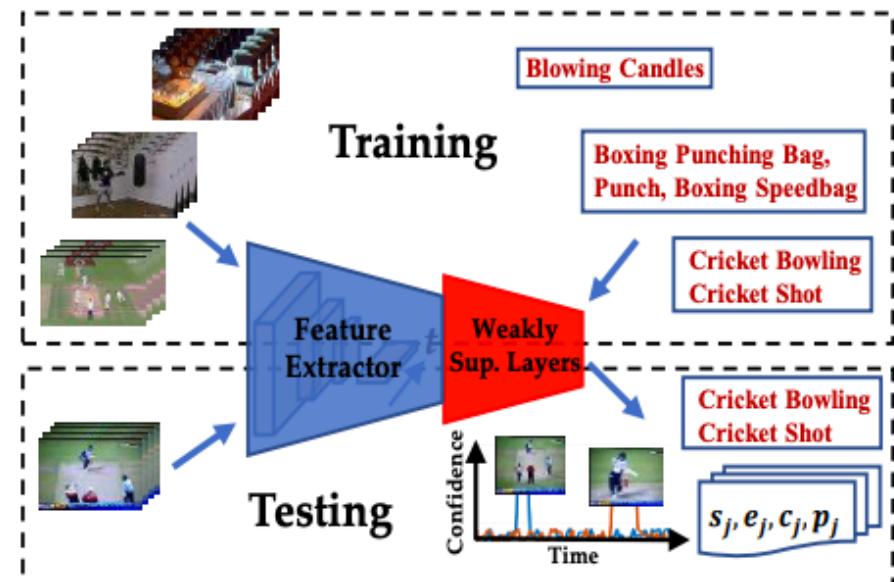


Method

01. W-TALC

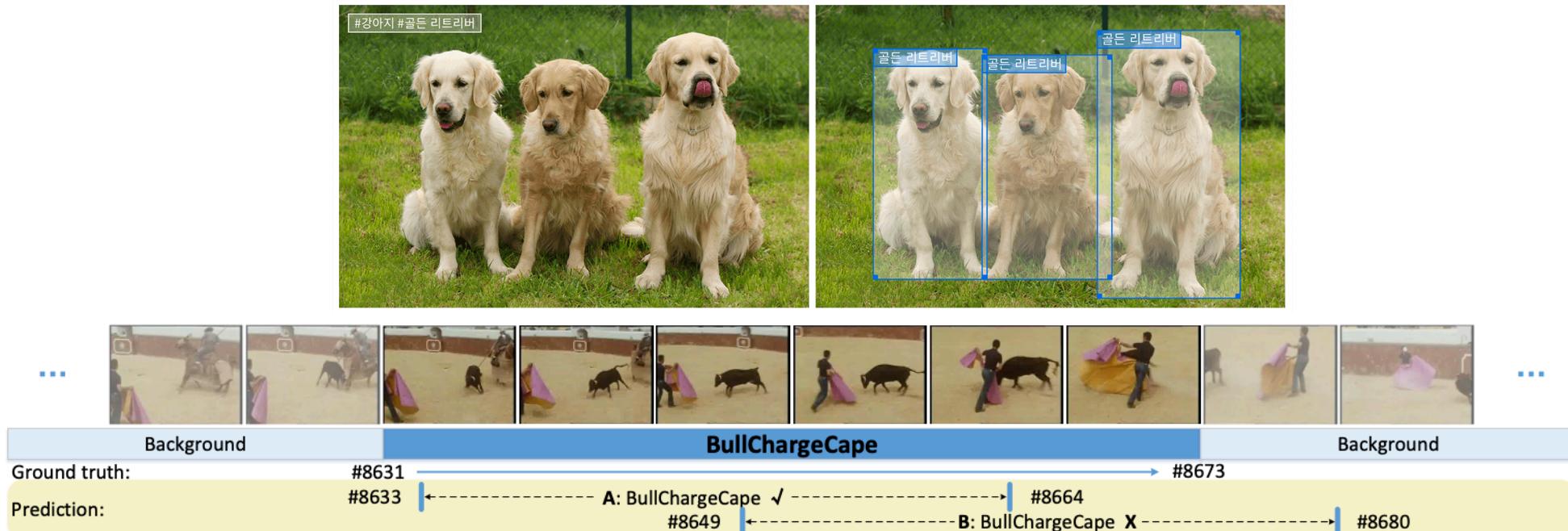
- Train, Test protocol
 - **Training set:** video label activity tag
 - video X : action A,A,B,C
 - **Output:** activity의 label만 측정하는 것이 아닌, activity의 시작 시간, 종료 시간, 카테고리, 신뢰도를 찾음

Fig. 1: This figure presents the train-test protocol of W-TALC. The training set consists of videos and the corresponding video-level activity tags. Whereas, while testing, the network not only estimates the labels of the activities in the video, but also temporally locates their occurrence representing the start (s_j) and end time (e_j), category (c_j) and confidence of recognition (p_j) of the j^{th} activity located by the model.



01. W-TALC

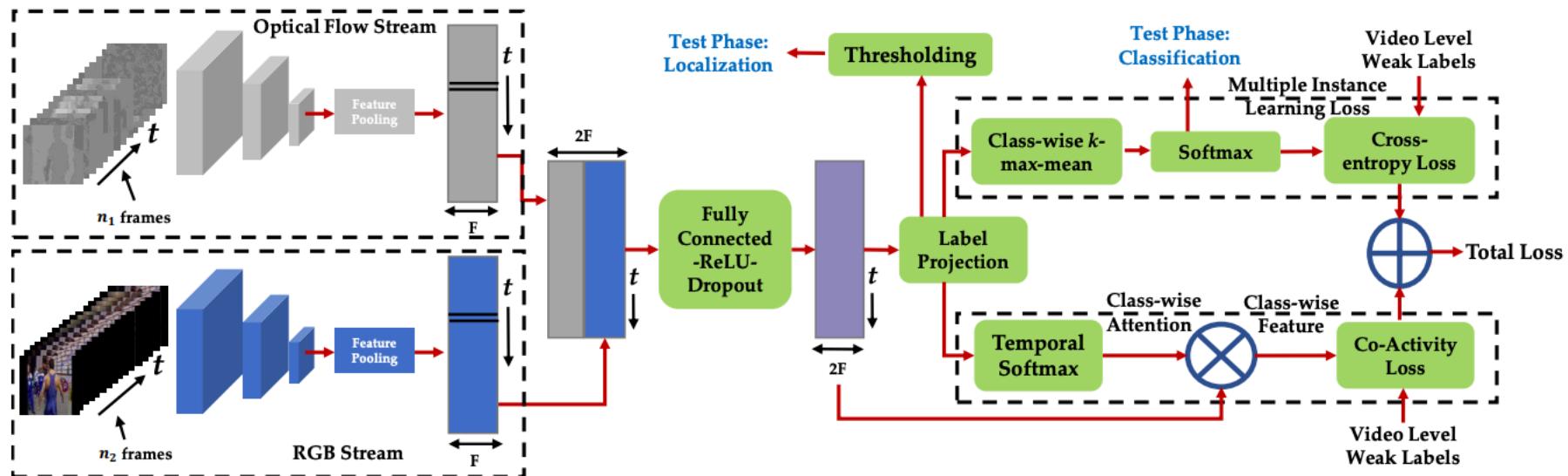
- Weak labeling
 - Weak TALC problem은 weak object localization과 유사
 - MIL(Multiple Instance Learning)을 주로 활용 : training에 이용할 수 있는 data 구조와 유사
 - Weak object localization보다 더 어렵다 : video의 시간 축에 따른 길이 등과 같은 문제
 - Activity localization에 weak labeling은 아직 탐구되지 않음



01. W-TALC

- Framework

- 1) RGB, Optical Flow stream의 feature vector를 concat
- 2) FC-ReLU-Dropout이 적용된 후 각 time마다 2048d의 feature를 얻음
- 3) feature들을 label projection module을 통해 카테고리에 대한 activation을 얻음
- 4) **Multiple Instance Learning Loss(MILL), Co-Activity Similarity Loss(CASL)** 2개의 loss function을 계산
-> network weight를 학습하기 위함



01. W-TALC

- Feature Extractor
 - Two-Stram network 사용 (Untrimmednets for weakly supervised action recognition and detection, CVPR 2017)
 - Network 통과하면 하나의 frame 당, 1차원의 feature vector
- Loss function
 - Video level label을 이용해 2개의 loss function을 계산하는 데 사용
 - (1) **Multiple Instance Learning Loss (MILL)**
 - Class-wise k-max-mean strategy 사용
 - (2) **Co-Activity Similarity Loss (CASL)**
 - 최소 하나의 activity category를 갖는 한 쌍의 비디오가 해당하는 시간 영역에서 비슷한 feature를 가져야 한다는 motivation에 기초
 - '자전거'에 대응하는 비디오 feature는 '자전거'에 대응하지 않는 다른 비디오 feature와 달라야 한다
 - 그러나 temporal label은 weakly-supervised data에 알려지지 않았으므로 CASL을 계산하기 위해 label space activation에서 얻은 attention을 weak temporal label로 사용

01. W-TALC

- Main Contribution

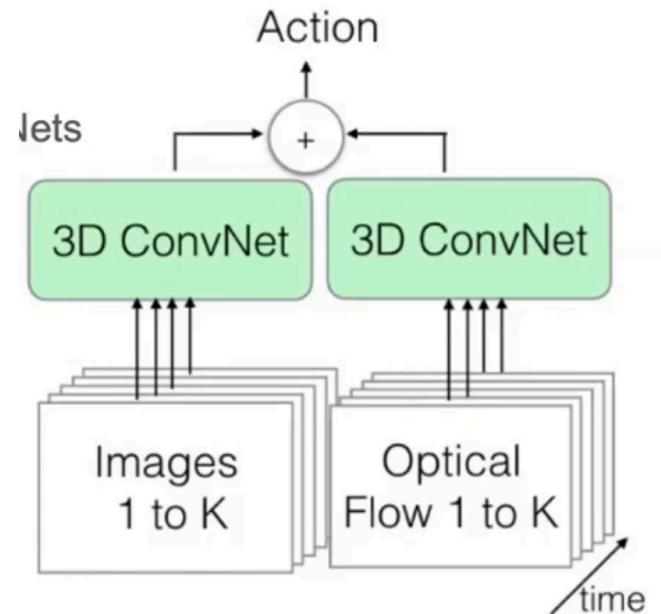
- feature extractor를 파인튜닝 하지 않고 task-specific parameter만 학습하는 W-TALC를 제안
 - 이 방법은 학습 중에 비디오의 라벨 순서를 고려하지 않으며, 같은 시간 동안 여러 activity 감지 가능
- **Co-Activity Similarity Loss(CASL)**를 소개하고, Multiple Instance Learning Loss(MILL)로 최적화
 - weakly-supervised task에 특정한 network weight를 학습
- 2개의 dataset으로 실험

02. Feature Extraction

- I3D Features

(Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR 2017)

- Kinetics가 pretrained된 I3D 사용
- output: 3D average pooling layer를 통과한 1024d feature



02. Feature Extraction

- Memory Constraints

- Untrimmed video : n분 이상의 긴 비디오
- 메모리 문제 발생
- Video를 시간 축에 따라 chunk로 분할, 시간 pooling을 통해 각 chunk를 vector로 줄임

- Long Video Sampling

- 비디오의 길이가 일정 이상 길면, 랜덤하게 T길이의 클립을 자르고 라벨 할당
- 일정 길이보다 길지 않으면 전체 비디오 처리

- Computational Budget and Fine-tuning

- Feature extractor를 fine-tuning 하지 않음
- Task-specific parameter만 학습

03. MIL Loss

- K-max Multiple Instance Learning
 - Multiple Instance Learning : positive bag / negative bag으로 분류
 - Positive bag: 하나 이상의 positive instance 포함
 - Negative bag: positive instance 포함 X
 - Bag을 training data로 학습해 분류
 - 전체 video를 Instance bag으로 간주
 - Loss를 계산하려면 각 video를 카테고리 별 **confidence score**로 표현해야 함
 - 각 비디오는 다수의 activity를 가지고 있고, **비디오에서 그 activity가 나타난 부분에 라벨 벡터를 나타냄**
 - 해당 카테고리에 해당하는 activation score를 해당 카테고리의 time dimension에 대해 k-max activation 평균 계산

$$s_i^j = \frac{1}{k_i} \max_{\substack{\mathcal{M} \subset \mathcal{A}_i[j,:], \\ |\mathcal{M}|=k_i}} \sum_{l=1}^{k_i} \mathcal{M}_l$$

$$\mathcal{L}_{MILL} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{n_c} -y_i^j \log(p_i^j)$$

s : i번째 video의 j번째 카테고리의 class-wise 신뢰 점수

P : S 에 대해 softmax (pmf)

M_l : set M 의 l번째 원소

04. CASL

- Co-Activity Similarity
 - 같은 activity A가 발생하는 부분의 feature는 서로 비슷해야 한다
 - 동일한 비디오 쌍에 대해, 한 비디오에서 action A가 발생하는 부분의 feature는 A가 등장하지 않는 부분과 달라야 함
 - 이러한 속성은 MILL에서는 적용되지 않음 → CASL 제안
 - Frame 단위의 label이 없으므로 class 단위의 activation 사용

$$\hat{\mathcal{A}}_i[j, t] = \frac{\exp(\mathcal{A}_i[j, t])}{\sum_{t'=1}^{l_i} \exp(\mathcal{A}_i[j, t'])}$$

A : attention. 특정 카테고리에 높은 attention을 갖는다는 것은 그 category에 높은 발생 가능성이 있다는 것
Video 별, class 별 activation score를 시간 축에 따라 정규화 (softmax의 비선형성을 이용)

04. CASL

- Feature similarity

- Feature vector : attention이 높고 낮은 영역의 클래스 별 feature vector 정의

$${}^H \mathbf{f}_i^j = \mathbf{X}_i \hat{\mathcal{A}}_i[j, :]^T$$

$${}^L \mathbf{f}_i^j = \frac{1}{l_i - 1} \mathbf{X}_i (1 - \hat{\mathcal{A}}_i[j, :]^T)$$

$$d[\mathbf{f}_i, \mathbf{f}_j] = 1 - \frac{\langle \mathbf{f}_i, \mathbf{f}_j \rangle}{\langle \mathbf{f}_i, \mathbf{f}_i \rangle^{\frac{1}{2}} \langle \mathbf{f}_j, \mathbf{f}_j \rangle^{\frac{1}{2}}}$$

- 두 feature vector의 유사성을 계산하기 위해 cosine similarity 사용

$$\begin{aligned} \mathcal{L}_j^{mn} &= \frac{1}{2} \left\{ \max \left(0, d[{}^H \mathbf{f}_m^j, {}^H \mathbf{f}_n^j] - d[{}^H \mathbf{f}_m^j, {}^L \mathbf{f}_n^j] + \delta \right) \right. \\ &\quad \left. + \max \left(0, d[{}^H \mathbf{f}_m^j, {}^H \mathbf{f}_n^j] - d[{}^L \mathbf{f}_m^j, {}^H \mathbf{f}_n^j] + \delta \right) \right\} \end{aligned}$$

(Ranking hinge loss : 한 쌍의 video x_m, x_n 이 주어졌을 때)

$$\mathcal{L}_{CASL} = \frac{1}{n_c} \sum_{j=1}^{n_c} \frac{1}{\binom{|\mathcal{S}_j|}{2}} \sum_{\mathbf{x}_m, \mathbf{x}_n \in \mathcal{S}_j} \mathcal{L}_j^{mn}$$

04. Optimization

- Total loss

$$\mathcal{L} = \lambda \mathcal{L}_{MILL} + (1 - \lambda) \mathcal{L}_{CASL} + \alpha \|\mathbf{W}\|_F^2$$

- Lambda = 0.5
- 각 batch가 공통적으로 1개 이상의 카테고리를 갖도록 최소 3쌍의 video를 갖게 batch 생성

- Classification and Localization

Video를 얻었을 때 class 별 confidence score 획득

Pmf를 threshold로 설정하여 1개 이상의 action을 갖도록 video 분류

- threshold보다 confidence score 낮은 카테고리 무시
- 나머지에 대해 시작축에 따라 activation에 따른 threshold를 사용해 localization 얻음

Experiments

01. Dataset

Frame 단위의 activity label이 있는 untrimmed video. Video에 관련된 activity tag만 사용

- ActivityNet v1.2
 - Data
 - Train data: 4819
 - Val data: 2383
 - Unlabeled: 2480
 - 100 class, video당 평균 1.5개의 action
- Thumos14
 - Data
 - Train data: 1010
 - Val data: 1574
 - 101 class, video당 평균 15.5개의 activity

02. Detection performance

Supervision	IoU →	0.1	0.2	0.3	0.4	0.5	0.7
Strong	Saliency-Pool [26]	04.6	03.4	02.1	01.4	00.9	00.1
	FV-DTF [36]	36.6	33.6	27.0	20.8	14.4	-
	SLM-mgram [39]	39.7	35.7	30.0	23.2	15.2	-
	S-CNN [44]	47.7	43.5	36.3	28.7	19.0	05.3
	Glimpse [64]	48.9	44.0	27.0	20.8	14.4	-
	PSDF [65]	51.4	42.6	33.6	26.1	18.8	-
	SMS [66]	51.0	45.2	36.5	27.8	17.8	-
	CDC [43]	-	-	40.1	29.4	23.3	07.9
	R-C3D [62]	54.5	51.5	44.8	35.6	28.9	-
Weak	SSN [68]	60.3	56.2	50.6	40.8	29.1	-
	HAS [47]	36.4	27.8	19.5	12.7	06.8	-
	UntrimmedNets [57]	44.4	37.7	28.2	21.1	13.7	-
	STPN (UNTF) [35] ↓	45.3	38.8	31.1	23.5	16.2	05.1
Weak (Ours)	STPN (I3DF) [35] ↓	52.0	44.7	35.5	25.8	16.9	04.3
	MILL+CASL+UNTF↓	49.0	42.8	32.0	26.0	18.8	06.2
	MILL+I3DF	46.5	39.9	31.2	24.0	16.9	04.4
	MILL+CASL+I3DF	53.7	48.5	39.2	29.9	22.0	07.3
Weak (Ours)	MILL+CASL+I3DF↓	55.2	49.6	40.1	31.1	22.8	07.6

UNTF: UntrimmedNet, I3DF: I3D

↓ : 20class만 사용

Thumos14

Supervision	IoU →	0.1	0.2	0.3	0.4	0.5	0.7	Avg.
Strong	SSN-SW [68]	-	-	-	-	-	-	24.8
	SSN-TAG [68]	-	-	-	-	-	-	25.9
Weak	W-TALC (Ours)	53.9	49.8	45.5	41.6	37.0	14.6	18.0

ActivityNet1.2

03. Classification Performance

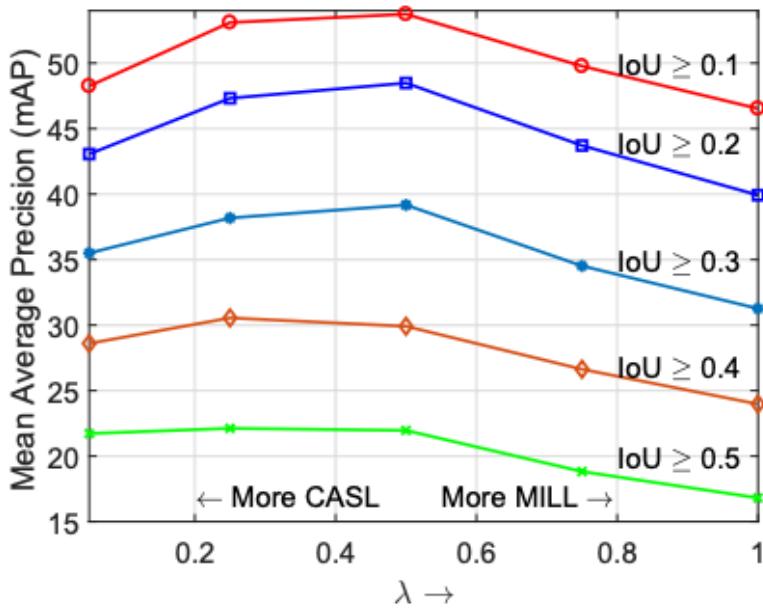
Methods	mAP	Supervision
EMV + RGB [67]	61.5	Strong ↑
iDT+FV [55]	63.1	Strong ↑
iDT+CNN [56]	62.0	Strong ↑
Objects + Motion [23]	71.6	Strong ↑
Feat. Agg. [22]	71.0	Strong ↑
Extreme LM [53]	63.2	Strong ↑
Temp. Seg. Net. (TSN) [58]	78.5	Strong ↑
Two Stream [45]	66.1	Strong ↑
Temp. Seg. Net. (TSN) [58]	67.7	Strong
UntrimmedNets [57]	74.2	Weak
UntrimmedNets [57]	82.2	Weak ↑
W-TALC (Ours w. I3D)	85.6	Weak

Thumos14

Algorithms	mAP	Supervision
C3D [51]	74.1	Strong ↑
iDT+FV [55]	66.5	Strong ↑
Depth2Action [23]	78.1	Strong ↑
Temp. Seg. Net. (TSN) [58]	88.8	Strong ↑
Two Stream [45]	71.9	Strong ↑
Temp. Seg. Net. (TSN) [58]	86.3	Strong
UntrimmedNets [57]	87.7	Weak
UntrimmedNets [57]	91.3	Weak ↑
W-TALC (Ours w. I3D)	93.2	Weak

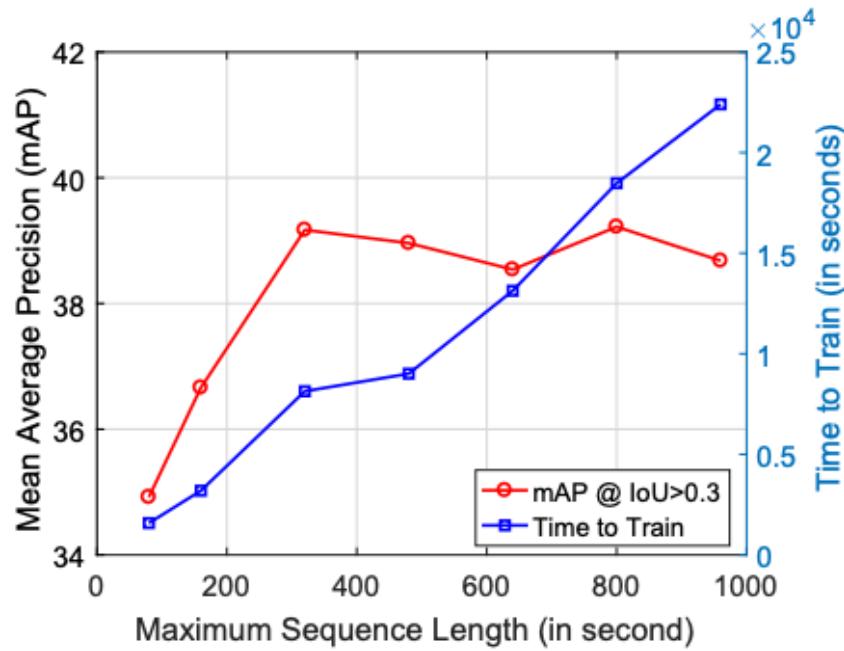
activityNet1.2

04. Results



(a)

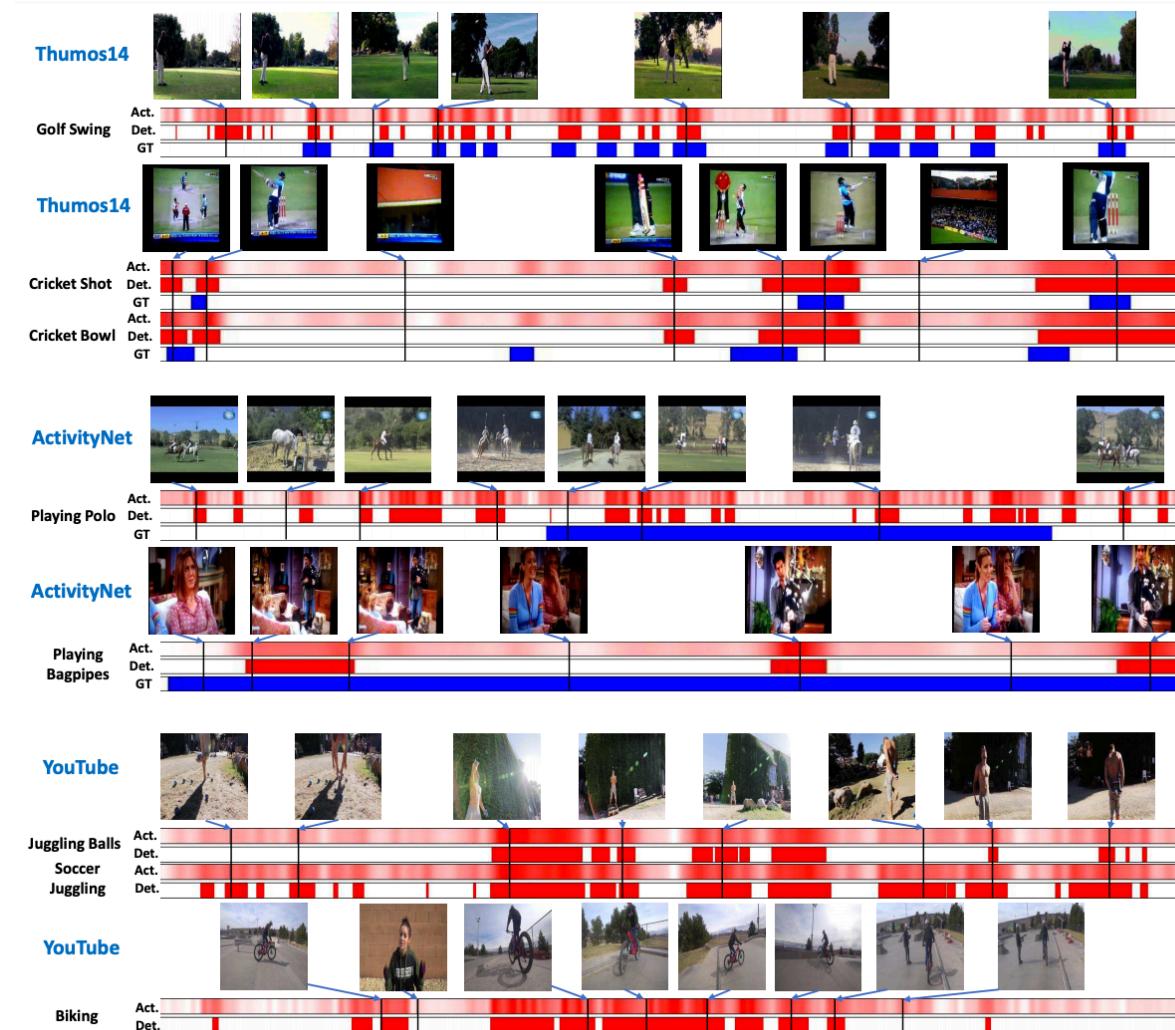
- Thumos14
- MILL, CASL 비율 바꾸면서 테스트
- Lambda = 0.5일 때가 가장 좋았다



(b)

- Thumos14
- $0.3 >= \text{IoU}$ 에서 video sequence 바꾸면서 테스트
- 320일 때가 가장 성능 좋았다

05. Results



Conclusion

01. Conclusion

- **Video level label**만으로 **weak supervision**을 사용해 temporal activity localization, video classification하는 모델 제시
- **Co-Activity Similiarty Loss(CASL)**를 제안하여 MILL을 보완
- weak-TALC에서 SOTA