Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos

김민지

Video classification

문제:

이미지 프레임별로 비디오를 레이블하기 때문에 엄청난 비용이 든다.

Action Detection의 경우 이미지 프레임별로 bounding box까지 레이블 해야한다.

Action classification / Action detection

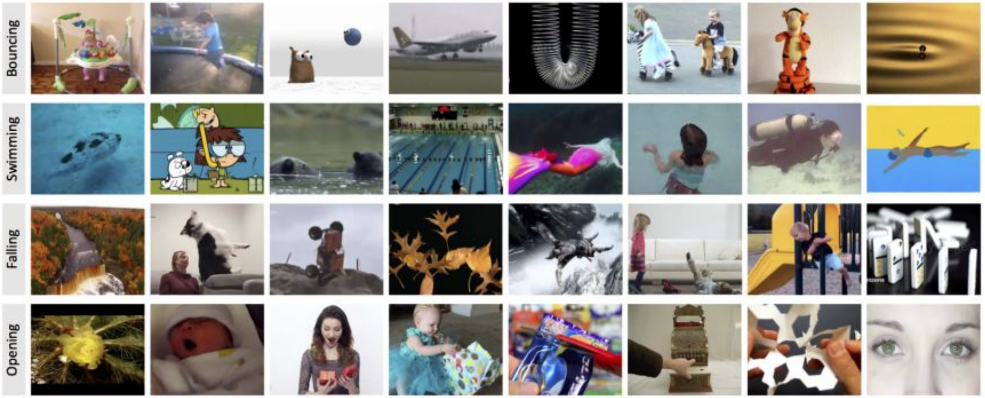
Action classification

: 짧은 클립에서 보여지는 액션을 분류

Action detection

: 주로 untrimmed video에서 한 사람 or 여러 사람의 행동을 감지

Kinetics / Moments in Time



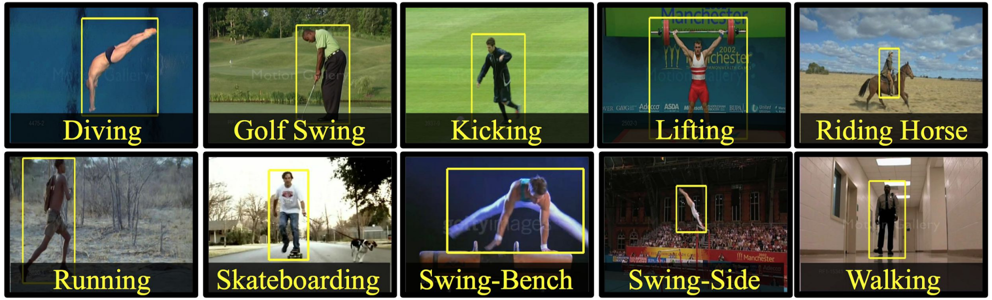
AVA / UCF101-24 dataset

Left: Sit, Talk to, Watch; Right: Crouch/Kneel, Listen to, Watch

Left: Stand, Carry/Hold, Listen to; Middle: Stand, Carry/Hold, Talk to; Right: Sit, Write

Left: Sit, Ride, Talk to; Right: Sit, Drive, Listen to

Left: Stand, Watch; Middle: Stand, Play instrument; Right: Sit, Play instrument



How to label?

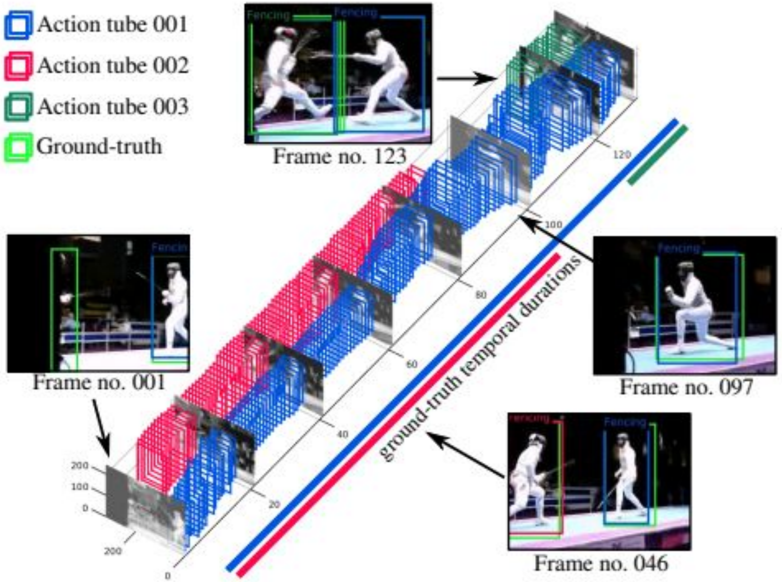
비디오에 나타나는 사람 중 몇 명까지 label 할 것인가?

액션의 시작과 끝은 어디인가?

만약 액션이 이어진다면, 다음 액션과의 경계는 정확히 어디인가?

Spatio-temporal action recognition

Action tublet



https://arxiv.org/pdf/1611.08563.pdf

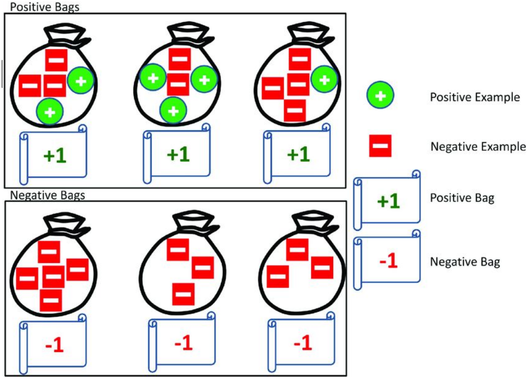
Proposed Approach

1. Multiple Instance Learning (MIL)
2. How to use it for spatio-temporal action recognition
3. When is MIL violated and how to reduce it using uncertainty

Multiple Instance Learning (MIL)

- Drug Activity Prediction 문제를 풀기 위해 고안.
- 약의 여러 분자들 중 하나라도 주어진 단백질에 반응을 하면 약이 적합하다, (positive)
  모든 분자가 주어진 단백질에 반응을 하지 않으면 약으로 적합하지 않다. (negative)

- 개별 Instance 단위가 아닌, 다중의 Instance가 모여있는 세트 (Bag) 단위로 학습하는 방법.
- <MIL의 가정>
  Positive Bag에는 주어진 조건을 만족하는 Instance가 적어도 하나 이상 있다



https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.etsmtl.ca%2FUnites-de-recherche%2FLIVIA%2FSeminars%2FIntroduction-to-Multiple-Instance-Learning.pdf&psig=AOvVaw3n4lRrwDv4TQPgbplpQPYS&ust=1597121371065000&source=images&cd=vfe&ved=0CA0QjhxqFwoTCIDinvLqj-sCFQAAAAAdAAAAABAR

Multiple Instance Learning (MIL)

Tasks that can be performed in MIL

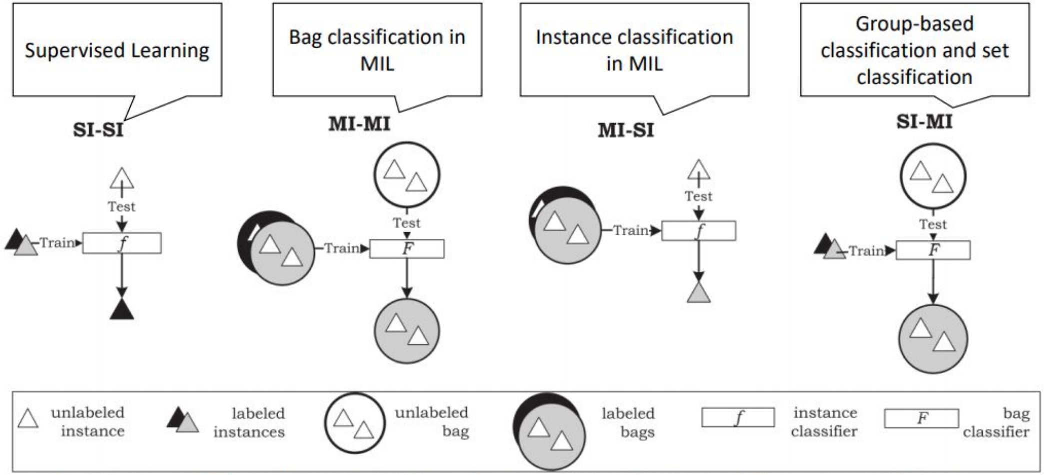


Image from: V. Cheplygina, D. M. J. Tax, and M. Loog, "On classification with bags, groups and sets," *Pattern Recognition Letters*, vol. 59, pp. 11–17, Jul. 2015.

Multiple Instance Learning (MIL)

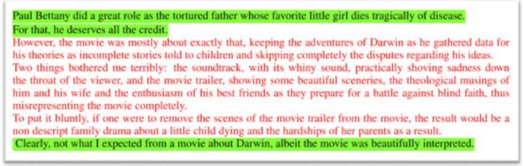
Sentiment Analysis in Text

Objective: Predict if a text/sentence expresses positive or negative sentiment.

Bags: Texts/paragraphs.

Instances: Sentences.

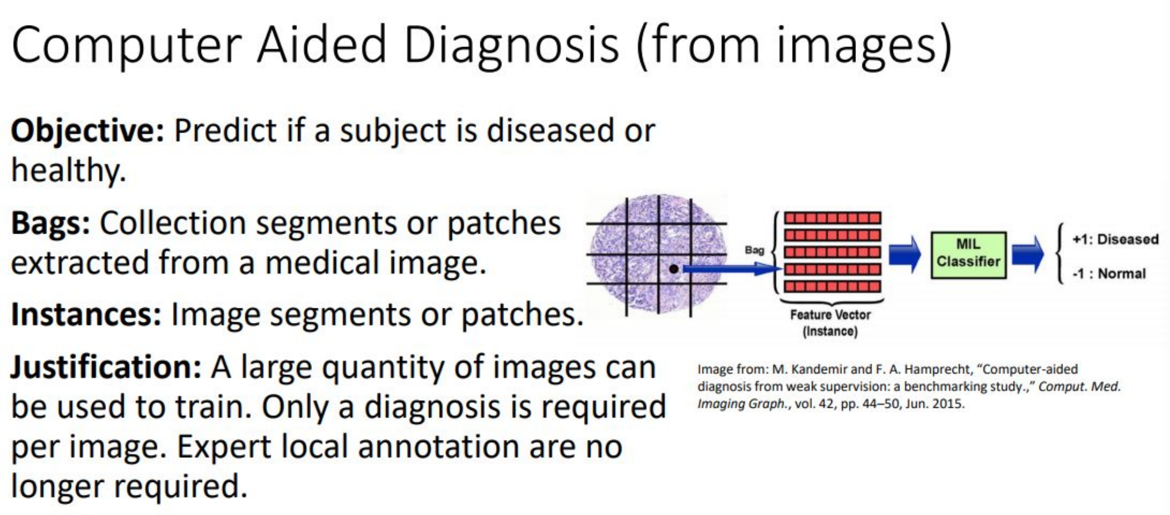
Justification: Large quantity of text can be harvested from the web. A sentiment is usually given to a complete text while it may contain positive and negative sentences.



Paul Bettany did a great role as the tortured father whose favorite little girl dies tragically of disease.
For that, he deserves all the credit.
However, the movie was mostly about exactly that, keeping the adventures of Darwin as he gathered data for his theories as incomplete stories told to children and skipping completely the disputes regarding his ideas.
Two things bothered me terribly: the soundtrack, with its whiny sound, practically shoving sadness down the throat of the viewer, and the movie trailer, showing some beautiful sceneries, the theological musings of him and his wife and the enthusiasm of his best friends as they prepare for a battle against blind faith, thus misrepresenting the movie completely.
To put it bluntly, if one were to remove the scenes of the movie trailer from the movie, the result would be a non descript family drama about a little child dying and the hardships of her parents as a result.
Clearly, not what I expected from a movie about Darwin, albeit the movie was beautifully interpreted.

Image from: D. Kotzias, M. Denil, P. Blunsom, and N. de Freitas, “Deep Multi-Instance Transfer Learning,” *CoRR*, vol. abs/1411.3, 2014.

Multiple Instance Learning (MIL)



Computer Aided Diagnosis (from images)

Objective: Predict if a subject is diseased or healthy.

Bags: Collection segments or patches extracted from a medical image.

Instances: Image segments or patches.

Justification: A large quantity of images can be used to train. Only a diagnosis is required per image. Expert local annotation are no longer required.

Image from: M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: a benchmarking study.," *Comput. Med. Imaging Graph.*, vol. 42, pp. 44–50, Jun. 2015.

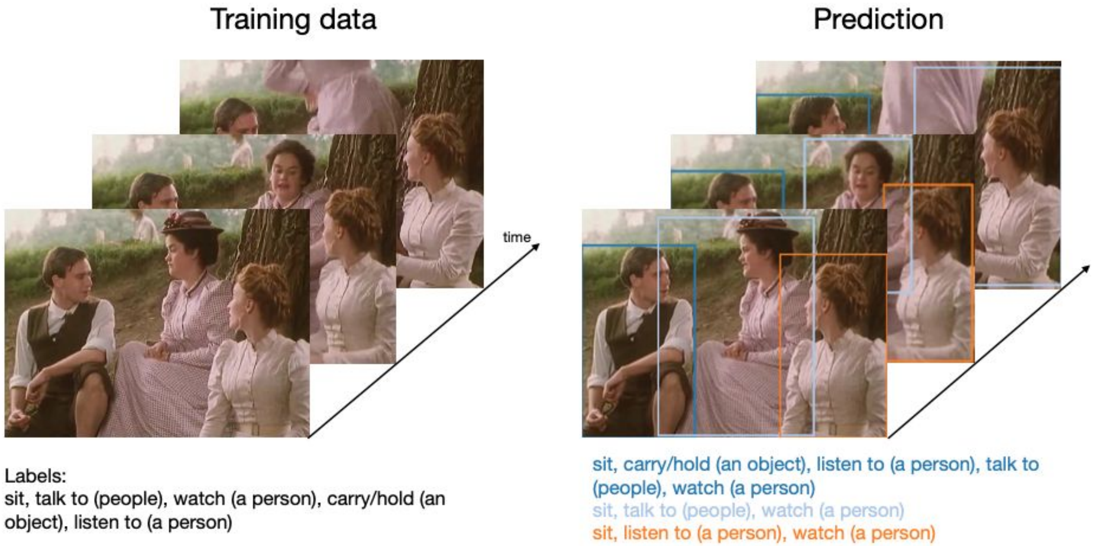

Training data
Prediction
time
Labels:
sit, talk to (people), watch (a person), carry/hold (an object), listen to (a person)
sit, carry/hold (an object), listen to (a person), talk to (people), watch (a person)
sit, talk to (people), watch (a person)
sit, listen to (a person), watch (a person)

Fig. 1. We propose a method to train a spatio-temporal action detector using only weak, video-level labels on challenging, real-world datasets. Note that the video-level labels that we have may apply to multiple people in the video, and that these labels may only be active for an unannotated time interval of the input clip.

Multiple Instance Learning (MIL)

Goal: Instance-level classifier

x: 한 개의 bag에 들어있는 instance $x = \{x_1, x_2, \ldots, x_N\}$

y: label 벡터 $y \in \mathbb{R}^C$ $y_l = 1$

$y_l = 0$

what to predict:
Instance j가 label yl를 가질 확률 $p(y_l = 1|x_j),$

Labels:
sit, talk to (people), watch (a person), carry/hold (an object), listen to (a person)

Prediction

sit, carry/hold (an object), listen to (a person), talk to (people), watch (a person)
sit, talk to (people), watch (a person)
sit, listen to (a person), watch (a person)

Multiple Instance Learning (MIL)

Goal: Instance-level classifier

하지만 우리는 Bag-level label만 가지고 있음.

-> Instance level 확률 집합 {Pij}를 bag level 확률 Pi로 aggregation 하고 싶다. (function g 이용)
-> 여기서 function g는 sigmoid/softmax와 같은 activation function

$$p(y_l = 1|x_j),$$

$$p(y_{il} = 1|x_1, x_2, \ldots, x_N) = g(p_{i1}, p_{i2}, \ldots, p_{iN}). \tag{1}$$

Multiple Instance Learning (MIL)

Goal: Instance-level classifier

Bag-level prediction을 얻었으면, standard classification loss 사용이 가능하다.

$$\mathcal{L}_{ce}(x, y) = -\sum_{i}^{N_b}\sum_{l}^{C} y_{il}\log p_{il} + (1 - y_{il})\log(1 - p_{il}) \qquad (2)$$

MIL - Aggregation

NN의 Global Pooling 으로 구현이 가능하다.

Permutation-invariant pooling functions

- Max-pooling
- Generalized mean-pooling
- Log-Sum-Exponential pooling (LSE)

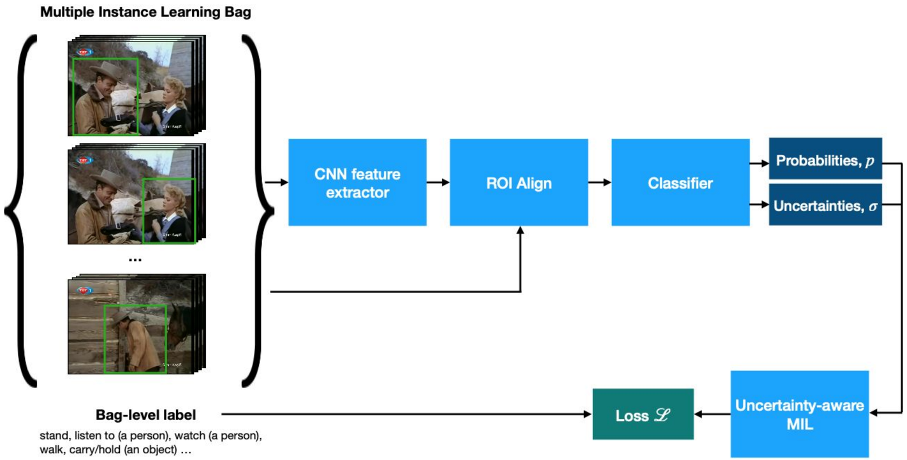
$$g(\{p_j\}) = \max_j p_j \tag{3}$$

$$g(\{p_j\}) = \left(\frac{1}{|j|} \sum_j p_j^r \right)^{\frac{1}{r}} \tag{4}$$

$$g(\{p_j\}) = \frac{1}{r} \log \left(\frac{1}{|j|} \sum_j e^{r \cdot p_j} \right) \tag{5}$$

Spatio-temporal action recognition as MIL

1. Person detector를 이용해서 k개의 연속 프레임에 대한 person tublet을 생성한다.
2. K개의 프레임과 person tublet을 사용하여, 각 detection의 중앙에서 발생하는 동작을 주변 K-1프레임의 temporal context를 고려하여 분류한다.
3. Bag size: 1. video의 길이 2. 감지된 사람의 수에 따라 달라진다.



Spatio-temporal action recognition as MIL

Video 전체가 T frame으로 되어있을 때, T - K + 1개의 Video clip이 생기게 되고,

프레임 당 한 사람만 있을 경우, T-K+1개의 person tublet이 생긴다.

GPU 메모리의 한계 때문에, 학습시 bag-level label를 유지하되, instance를 샘플링 하게 된다.

이 과정에서 noise가 발생하게 된다.

Label noise / violation of standard MIL

1. Instance sampling
   <bag-level label에 아무것도 해당하지 않은 instance를 샘플링할 경우>
   :전체 비디오에서 해당 액션의 길이에 반비례한다.

2. Missing detection
   <bag-level label에 해당하는 instance가 없을 경우>

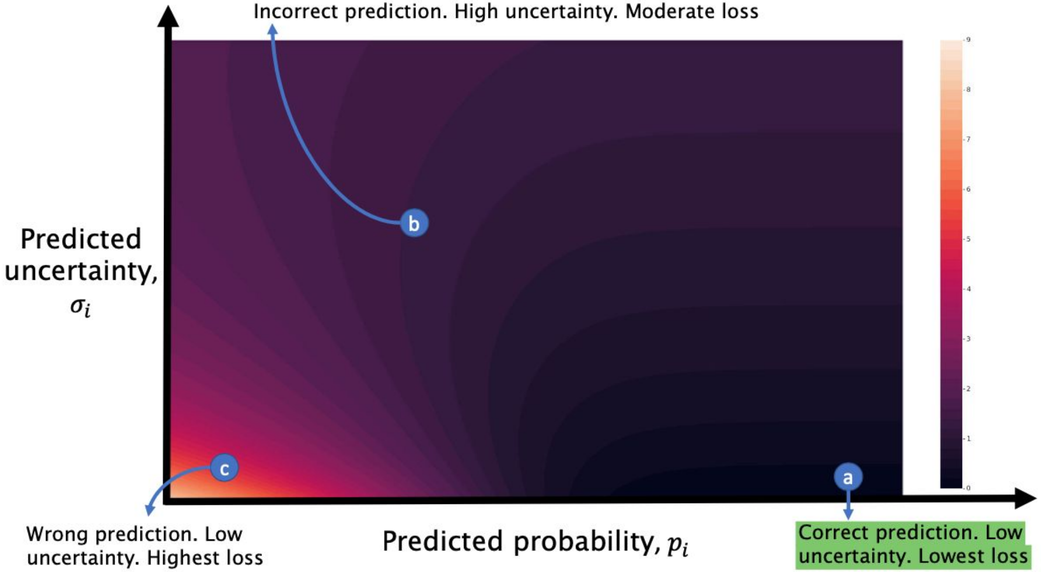
Label noise / violation of standard MIL

네트워크는

1. Uncertainty가 낮은 방향으로 학습을 하거나
2. Uncertainty가 아예 큰 방향으로 학습할 수 있다.
   (label에 해당하는 tublet이 없는 noisy한 bag일 경우 uncertainty가 클 것이다)

$$\mathcal{L}(x, y, \sigma) = \frac{1}{\sigma^2}\mathcal{L}_{ce}(x, y) + \log \sigma^2 \tag{6}$$

Uncertainty-base loss의 loss surface

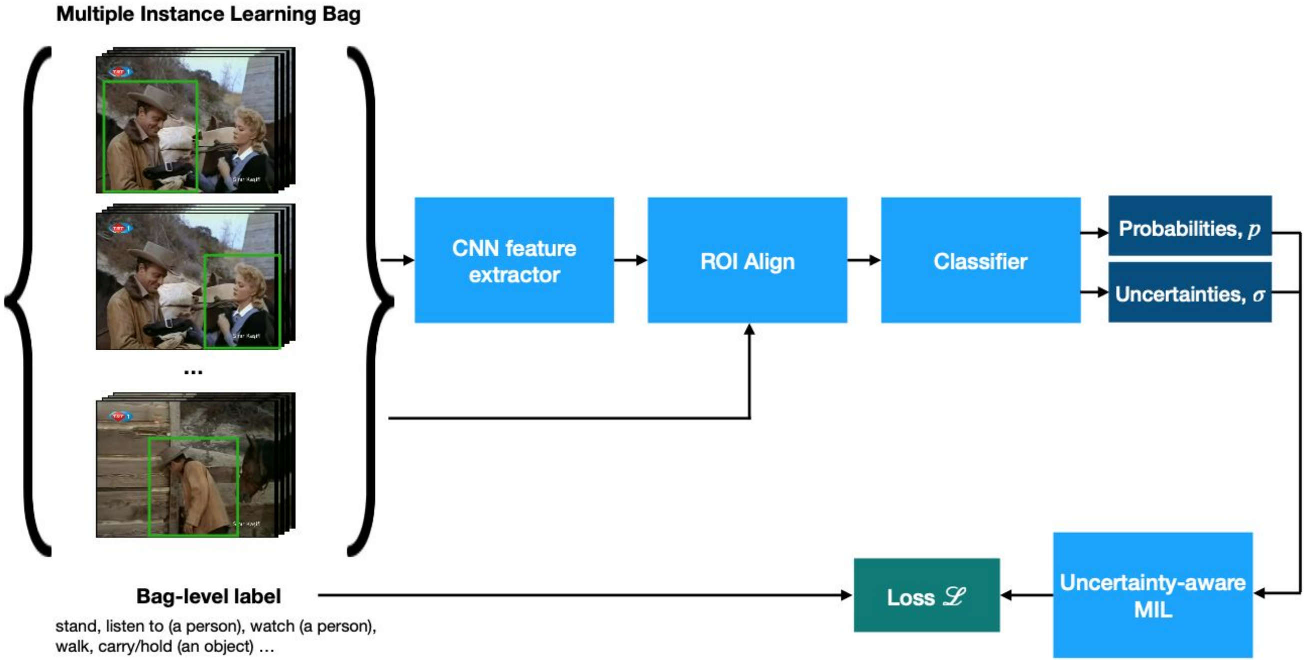


Incorrect prediction. High uncertainty. Moderate loss

Wrong prediction. Low uncertainty. Highest loss

Correct prediction. Low uncertainty. Lowest loss

Implementation details



Implementation details

각 tublet에 대해 C binary label을 가지는지에 대한 uncertainty σ ∈ RC 를 예측

$$\mathcal{L}(x, y, \sigma) = \frac{1}{\sigma^2}\mathcal{L}_{ce}(x, y) + \log \sigma^2 \tag{6}$$

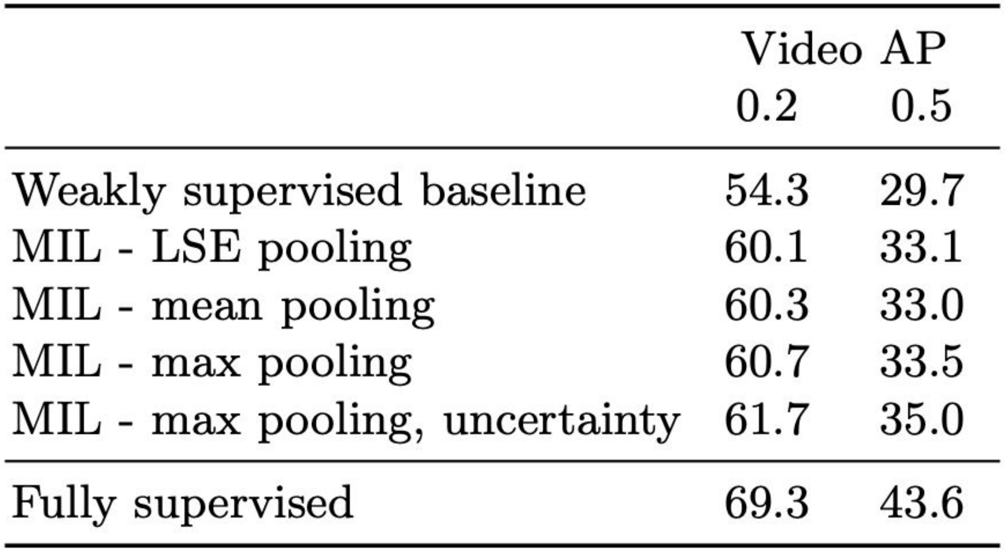
항상 양의 값을 갖게 하기 위해 Soft plus 사용

$$v := \log \sigma^2 \qquad f(x) = \log(1 + \exp(-x))$$

0으로 나눠지는 경우가 없도록

$$\frac{1}{\sigma^2} = \exp(-v)$$

Ablation study on UCF101-24



	Video AP	
	0.2	0.5
Weakly supervised baseline	54.3	29.7
MIL - LSE pooling	60.1	33.1
MIL - mean pooling	60.3	33.0
MIL - max pooling	60.7	33.5
MIL - max pooling, uncertainty	61.7	35.0
Fully supervised	69.3	43.6

The effect of tublet sampling

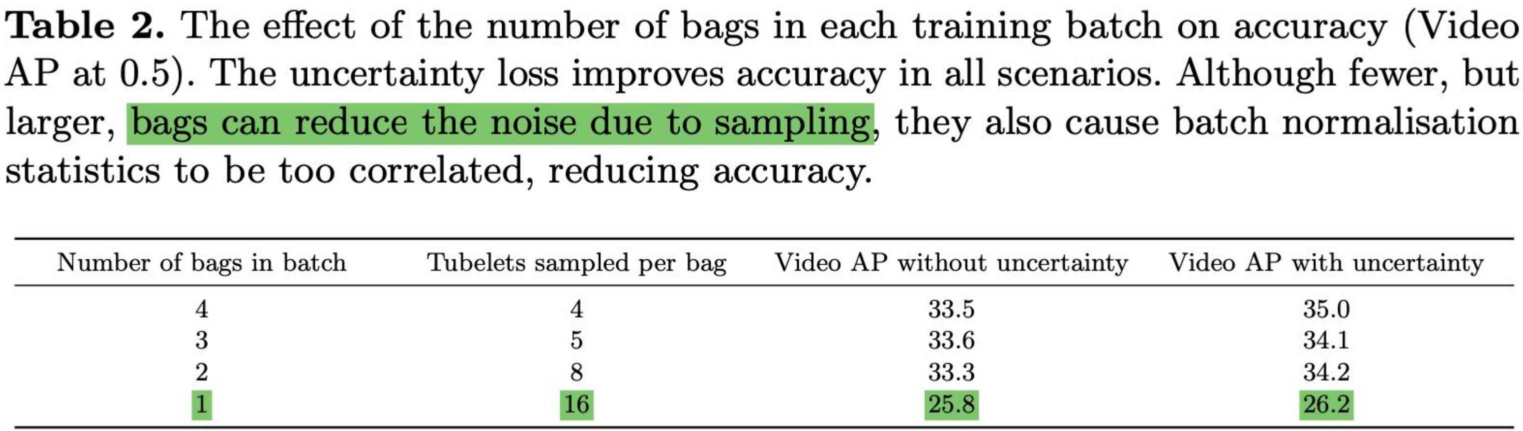


Table 2. The effect of the number of bags in each training batch on accuracy (Video AP at 0.5). The uncertainty loss improves accuracy in all scenarios. Although fewer, but larger, bags can reduce the noise due to sampling, they also cause batch normalisation statistics to be too correlated, reducing accuracy.

Number of bags in batch	Tubelets sampled per bag	Video AP without uncertainty	Video AP with uncertainty
4	4	33.5	35.0
3	5	33.6	34.1
2	8	33.3	34.2
1	16	25.8	26.2

:Tublet 수가 많으면 sampling으로 인한 noise가 줄어들 수 있으나,
Batch norm에 의한 correlation이 accuracy의 하락을 야기했다.

Comparison to state-of-the-art

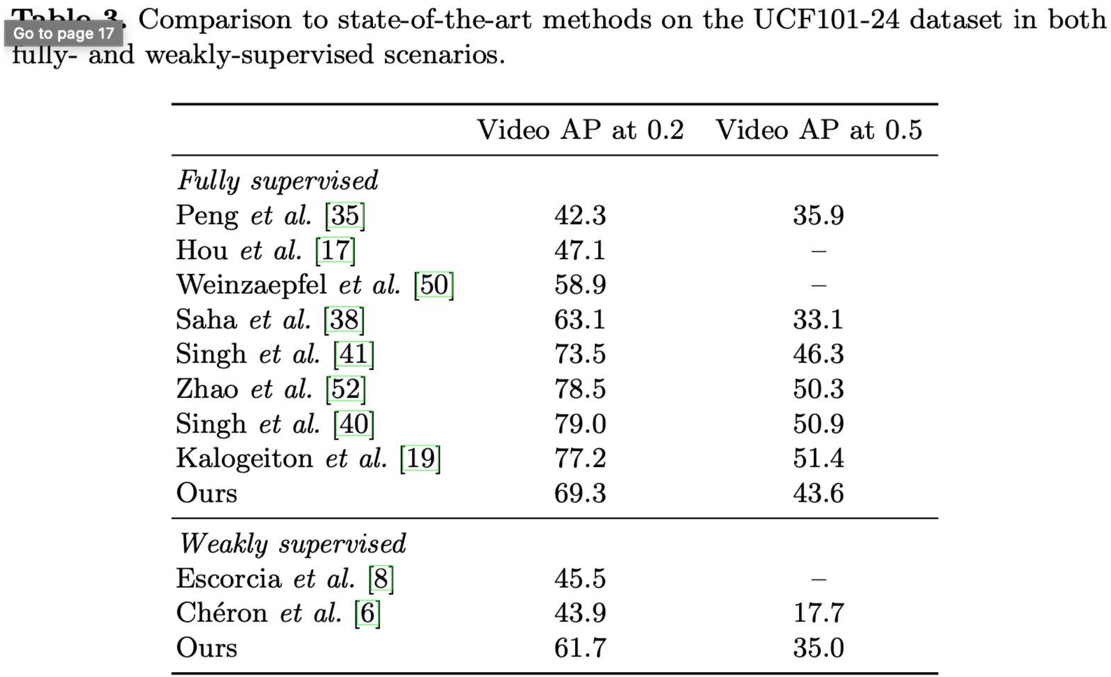

Go to page 17

Table 3. Comparison to state-of-the-art methods on the UCF101-24 dataset in both fully- and weakly-supervised scenarios.

	Video AP at 0.2	Video AP at 0.5
Fully supervised		
Peng *et al.* [35]	42.3	35.9
Hou *et al.* [17]	47.1	–
Weinzaepfel *et al.* [50]	58.9	–
Saha *et al.* [38]	63.1	33.1
Singh *et al.* [41]	73.5	46.3
Zhao *et al.* [52]	78.5	50.3
Singh *et al.* [40]	79.0	50.9
Kalogeiton *et al.* [19]	77.2	51.4
Ours	69.3	43.6
Weakly supervised		
Escorcia *et al.* [8]	45.5	–
Chéron *et al.* [6]	43.9	17.7
Ours	61.7	35.0

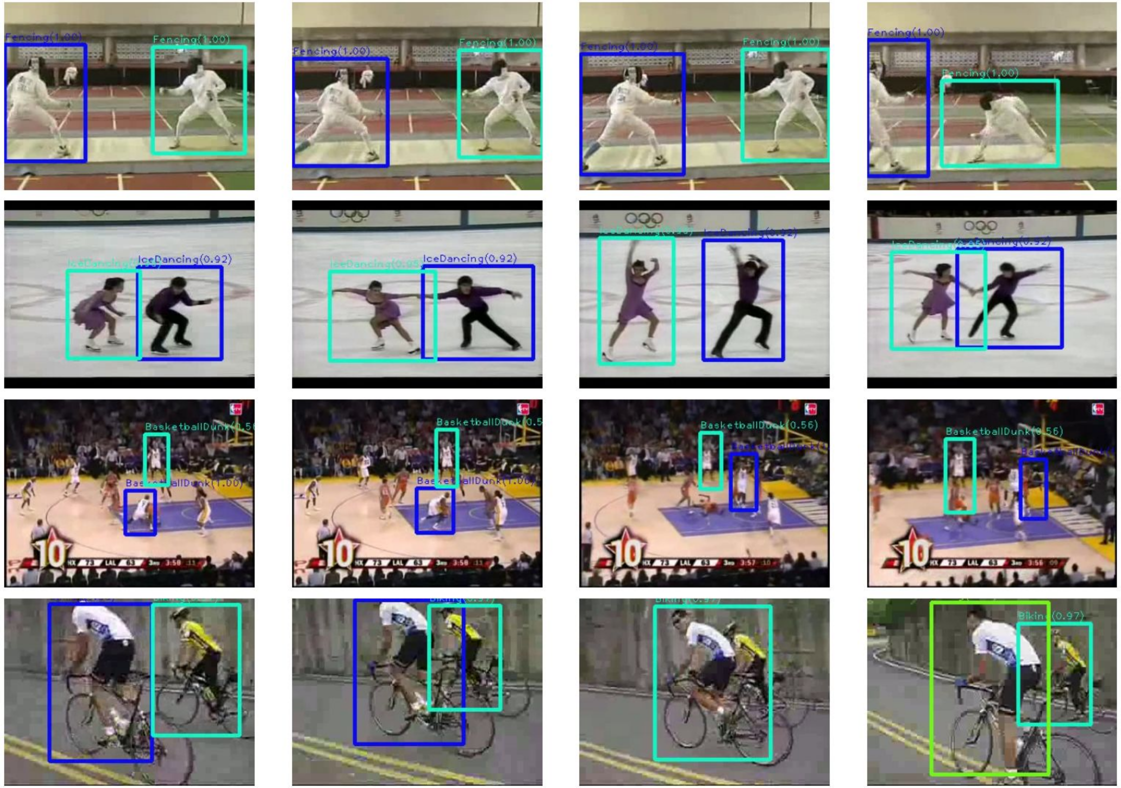


Fig. 4. Qualitative examples on UCF101-24. Note that the bounding boxes are coloured according to the identity of the track. The action label, and tube score are labelled from the top-left of the bounding box. Further discussion is included in the text.

Results on AVA dataset

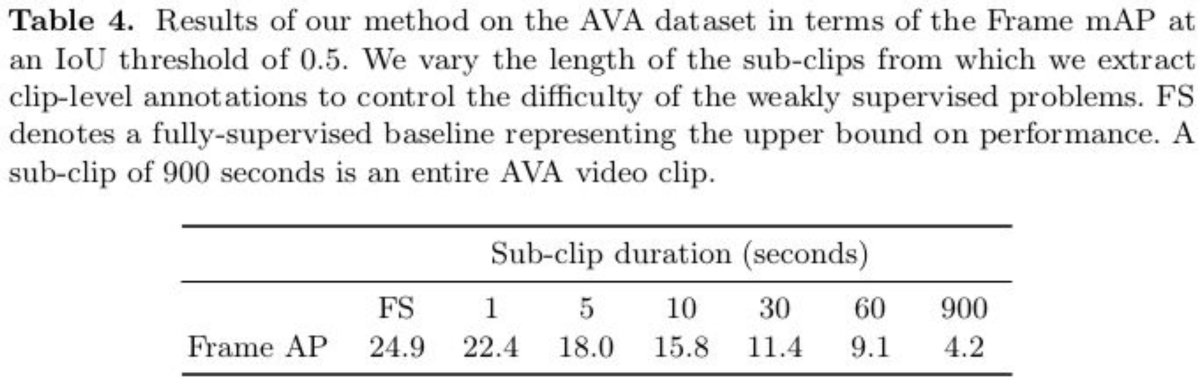


Table 4. Results of our method on the AVA dataset in terms of the Frame mAP at an IoU threshold of 0.5. We vary the length of the sub-clips from which we extract clip-level annotations to control the difficulty of the weakly supervised problems. FS denotes a fully-supervised baseline representing the upper bound on performance. A sub-clip of 900 seconds is an entire AVA video clip.

	Sub-clip duration (seconds)						
	FS	1	5	10	30	60	900
Frame AP	24.9	22.4	18.0	15.8	11.4	9.1	4.2

:Frame수가 많아질 수록 포함된 action label의 수도 많아지므로,
더 어려운 weakly-supervised 문제가 된다.

Results on AVA dataset

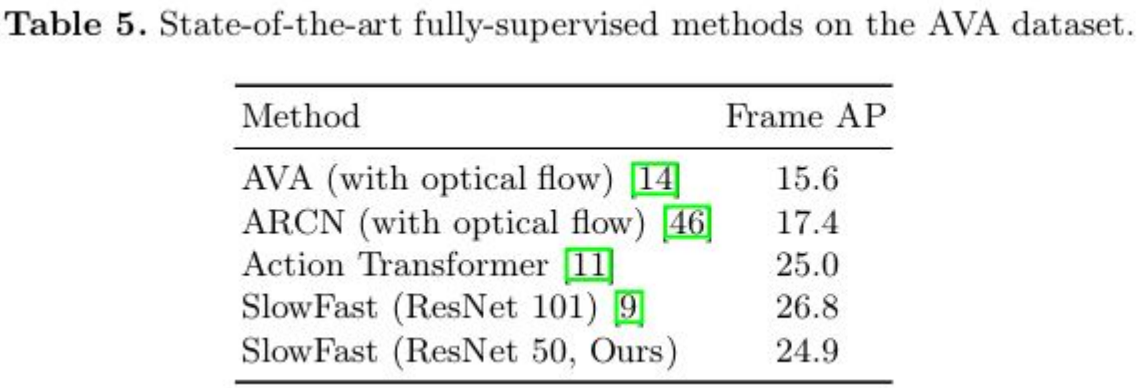


Table 5. State-of-the-art fully-supervised methods on the AVA dataset.

Method	Frame AP
AVA (with optical flow) [14]	15.6
ARCN (with optical flow) [46]	17.4
Action Transformer [11]	25.0
SlowFast (ResNet 101) [9]	26.8
SlowFast (ResNet 50, Ours)	24.9