

Learning Loss for Active Learning

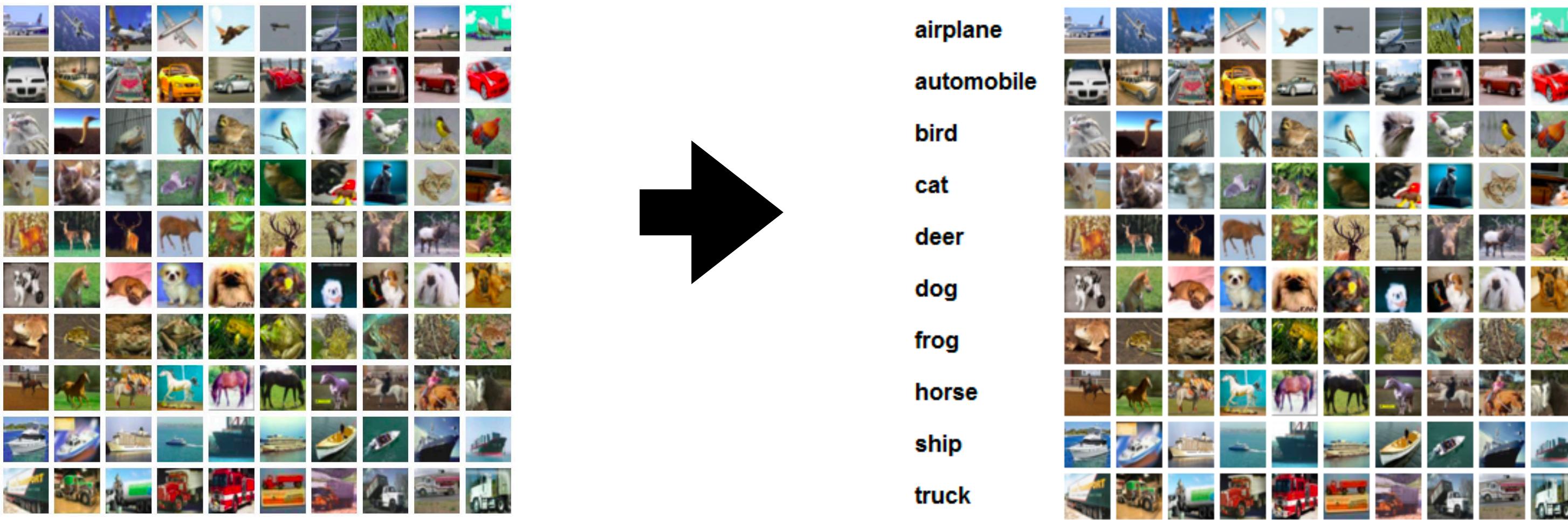
CVPR 2019

Minseon Kim

Preliminary knowledge

Exploit unlabeled Data

- Unlabeled Dataset을 수집하는 것은 상대적으로 적은 비용(cost)를 필요로 함
- 실제 data stream에서는 Annotation 과정에서 가장 많은 비용이 소요됨
-> Semi-supervised learning, Self-supervised learning 과 같은 많은 전략들이 제시되었다.

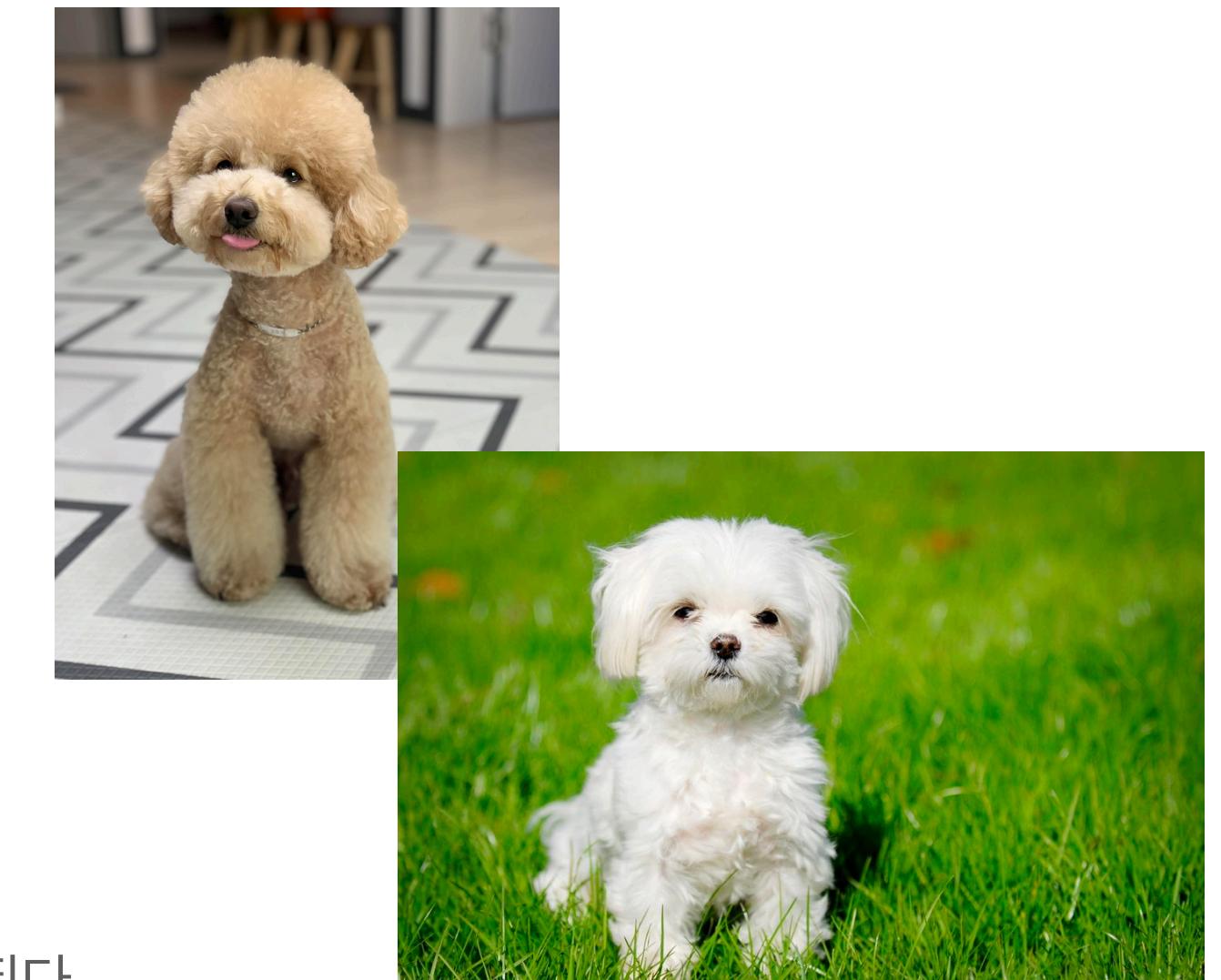
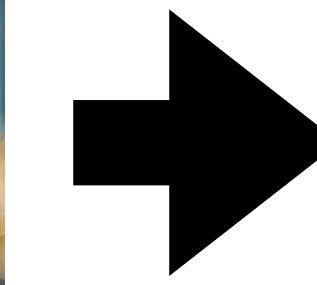
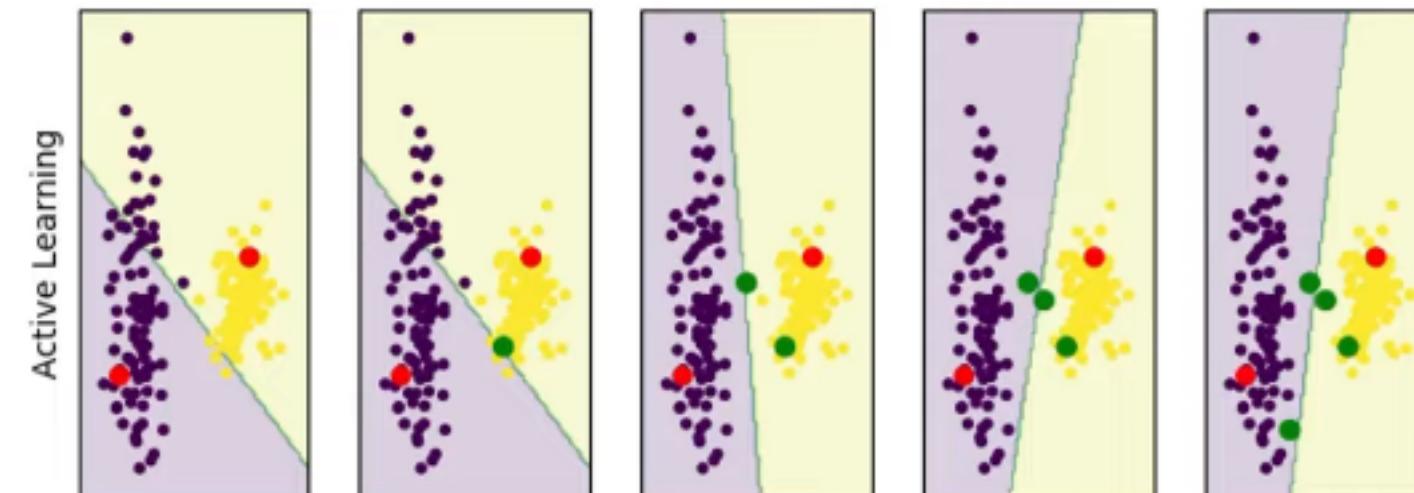


unlabeled data라고 하더라도 딥러닝 모델에서 활용가능한 다양한 feature 존재

Preliminary knowledge

Active Learning

- 기존의 이미지와 다른 이미지에 labeling 하는게 모델의 performance와 제한된 budget내의 비용 측면에서 효율적이다.
- 즉, active learning이란 기계학습 모델이 능동적으로 가장 유익한(informative) 데이터를 직접 선택하여 "우선적으로" labeling을 진행

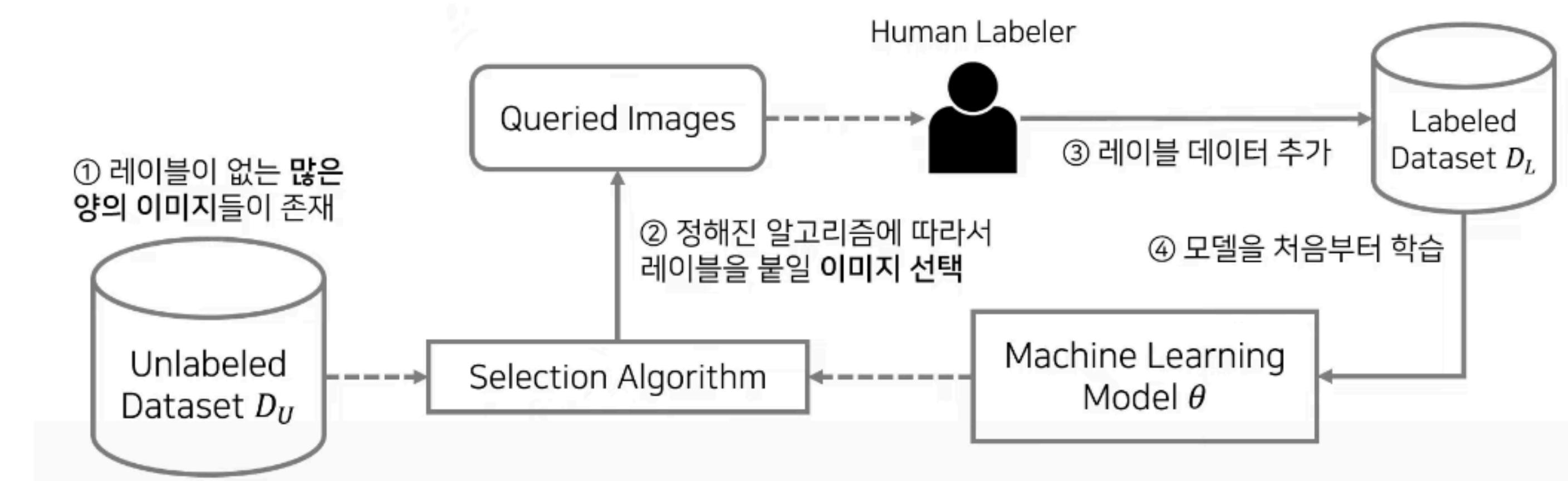


active learning을 이용할 때, decision boundary 근처의 불확실한 데이터를 선택하기 때문에, 보다 빠르게 올바른 학습이 수행된다.

Preliminary knowledge

Active Learning

- 다음과 같은 (2) ~ (4)의 과정을 반복하여 레이블링 비용을 줄인다.
 - 일반적인 active learning에서는 (4) 번의 과정에서는 변경된 labeled dataset을 이용하여 모델 가중치를 처음부터 다시 학습(entire model update)한다.



Related research efforts

Uncertainty Sampling Strategies

	Class 1	Class 2	Class 3
x_1	0.20	0.60	0.20
x_2	0.90	0.05	0.05
x_3	0.05	0.30	0.65

$1 - P_\theta(\hat{y} x)$
0.40
0.10
0.35

(1) Least Confident Sampling

$P_\theta(\hat{y}_1 x) - P_\theta(\hat{y}_2 x)$
0.40
0.85
0.35

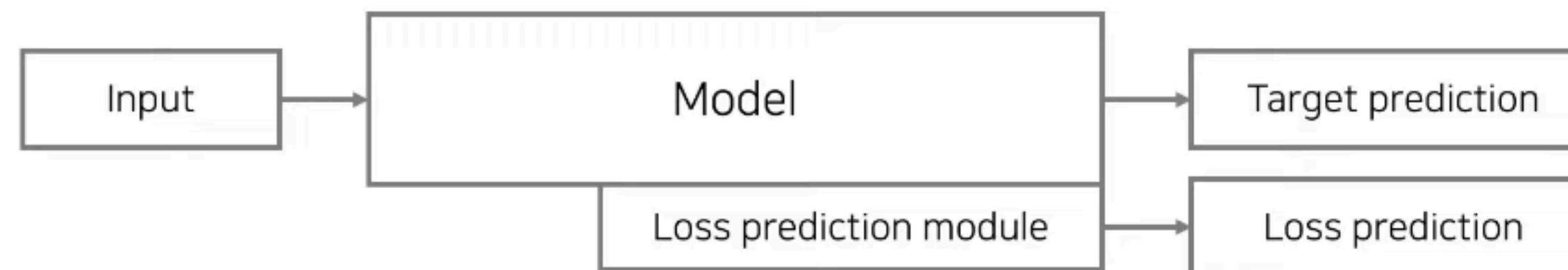
(2) Margin Sampling

$-\sum_i P_\theta(y_i x) \log P_\theta(y_i x)$
1.371
0.569
1.141

(3) Entropy Sampling
모든 클래스에 대한 확률값 고려
-> uniform distribution에 가까울 수록 높은 값

Contribution

- 기존 네트워크를 그대로 유지한 채(task와 관련없이)로 추가 가능한 Loss Prediction Module을 제안
(손실값이 높다 -> 불확실한 데이터)
- Unlabeled Data가 주어졌을때 손실 값을 예측하는 모듈



[Figure] A model with a loss prediction module.

Architecture

- Unlabeled Data에 대한 중요도(즉, 불확실성)평가 -> Loss prediction module에 의해 가장 불확실성이 높을것으로 예측된 top- K 개의 데이터를 label -> Labeled Data set에 추가
- 기존모델의 multi-level knowledge를 활용

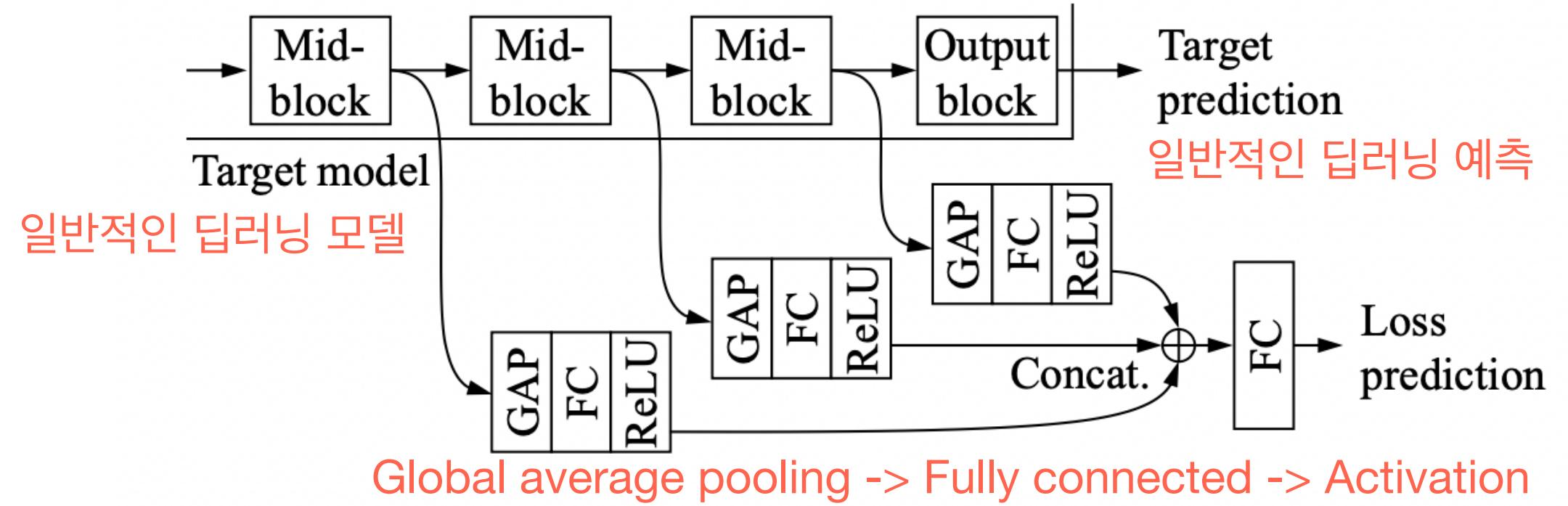
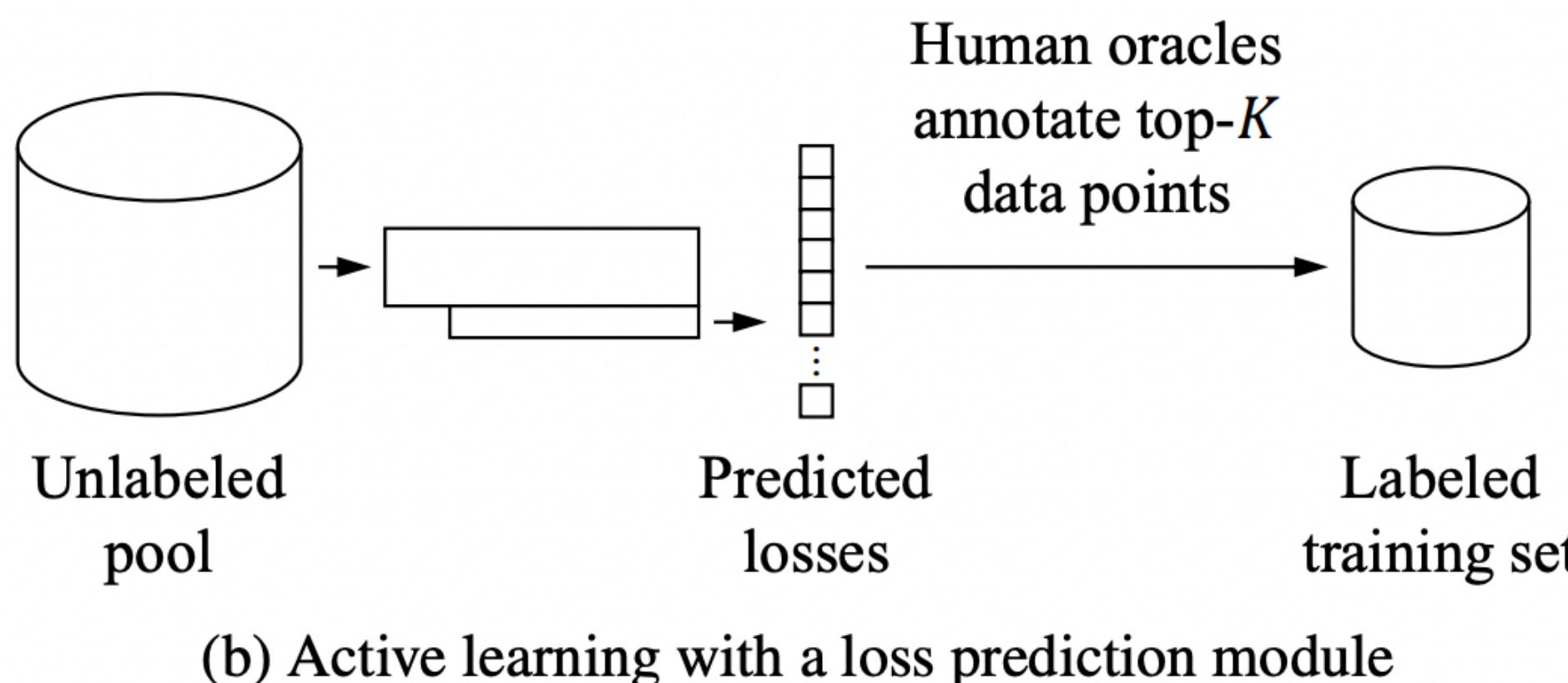


Figure 2. The architecture of the loss prediction module. This module is connected to several layers of the target model to take multi-level knowledge into consideration for loss prediction. The multi-level features are fused and map to a scalar value as the loss prediction.

Final Loss Function

$$L_{\text{target}}(\hat{y}, y) + \lambda \cdot L_{\text{loss}}(\hat{l}, l)$$

일반적인 딥러닝 모델학습시
사용되는 loss

image에 대해 loss값을 예측하고
그것이 실제 loss와
유사해질수있도록 loss를 구성해서
그것으로 학습을 진행

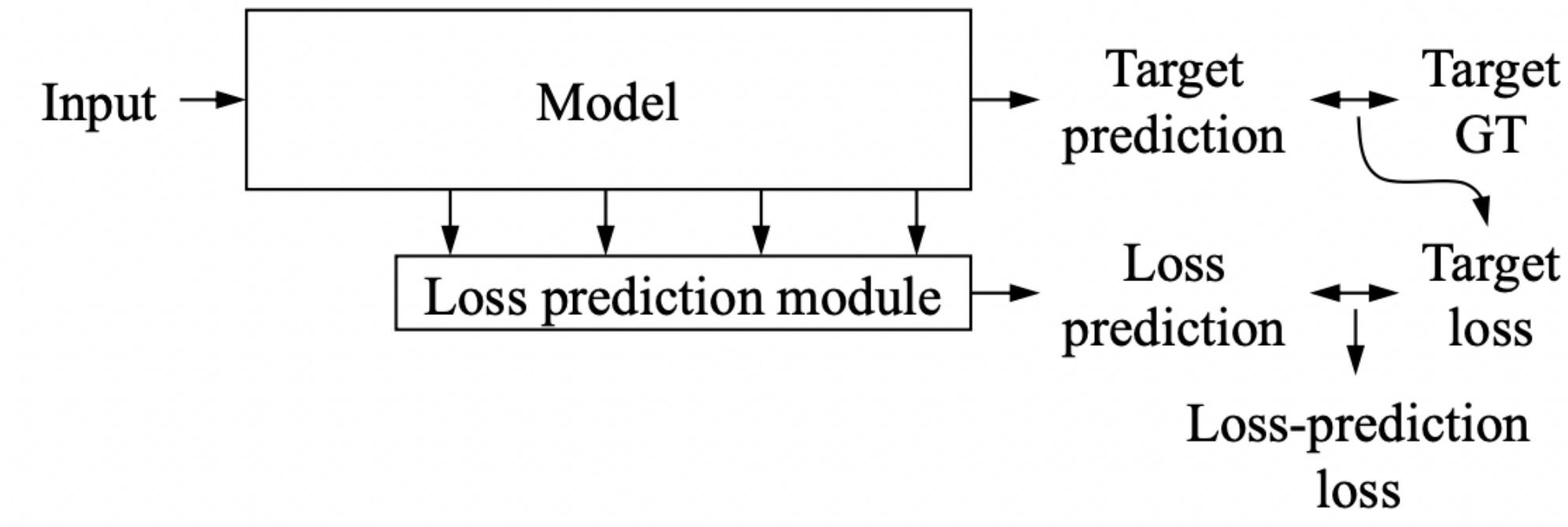


Figure 3. Method to learn the loss. Given an input, the target model outputs a target prediction, and the loss prediction module outputs a predicted loss. The target prediction and the target annotation are used to compute a target loss to learn the target model. Then, the target loss is regarded as a ground-truth loss for the loss prediction module, and used to compute the loss-prediction loss.

Loss Function for Loss Prediction Module

i 이미지에 대한 loss 값이 j 이미지에 대한 loss 값 보다 크다면 i에 대한 predicted loss를 j에 대한 predicted loss보다 충분히 크게(margin 값 이상으로) 만들어준다.

$$L_{\text{loss}}(\hat{l}^p, l^p) = \max \left(0, -\mathbb{1}(l_i, l_j) \cdot (\hat{l}_i - \hat{l}_j) + \xi \right)$$

$$\text{s.t. } \mathbb{1}(l_i, l_j) = \begin{cases} +1, & \text{if } l_i > l_j \\ -1, & \text{otherwise} \end{cases}$$

mini batch size B일때 총 B/2개의 데이터 pair를 생성할 수 있음
-> 모든 pair에 대해 margin loss 계산하여 줄이는 방식으로 학습진행

$$\frac{1}{B} \sum_{(x,y) \in \mathcal{B}^s} L_{\text{target}}(\hat{y}, y) + \lambda \frac{2}{B} \cdot \sum_{(x^p, y^p) \in \mathcal{B}^s} L_{\text{loss}}(\hat{l}^p, l^p)$$
$$\hat{y} = \Theta_{\text{target}}(x)$$

(xi, xj)에 대한 중간 layer로부터 추출되는 representation인 h_p 에 대한 predicted loss s.t. $\hat{l}^p = \Theta_{\text{loss}}(h^p)$ (3)

$$l^p = L_{\text{target}}(\hat{y}^p, y^p).$$

- MSE를 사용하지 않는 이유는 실제 loss / 값의 크기는 학습과정에 걸쳐 감소함
-> 정확한 loss를 예측하기 보다 크기변화에 적응하게 되는 한계 존재
- 이미지를 pairwise방식으로 예측된 두개의 손실값을 비교하는 **margin loss**를 활용

두개의 loss값의 차이를 벌리는 방식

Evaluation of 3 tasks

- initial K=1000 개의 random sampling한 데이터 세트에서 시작하여 각 round마다 1000개의 새로운 레이블을 얻어 모델을 학습하는 방식으로 모델을 평가

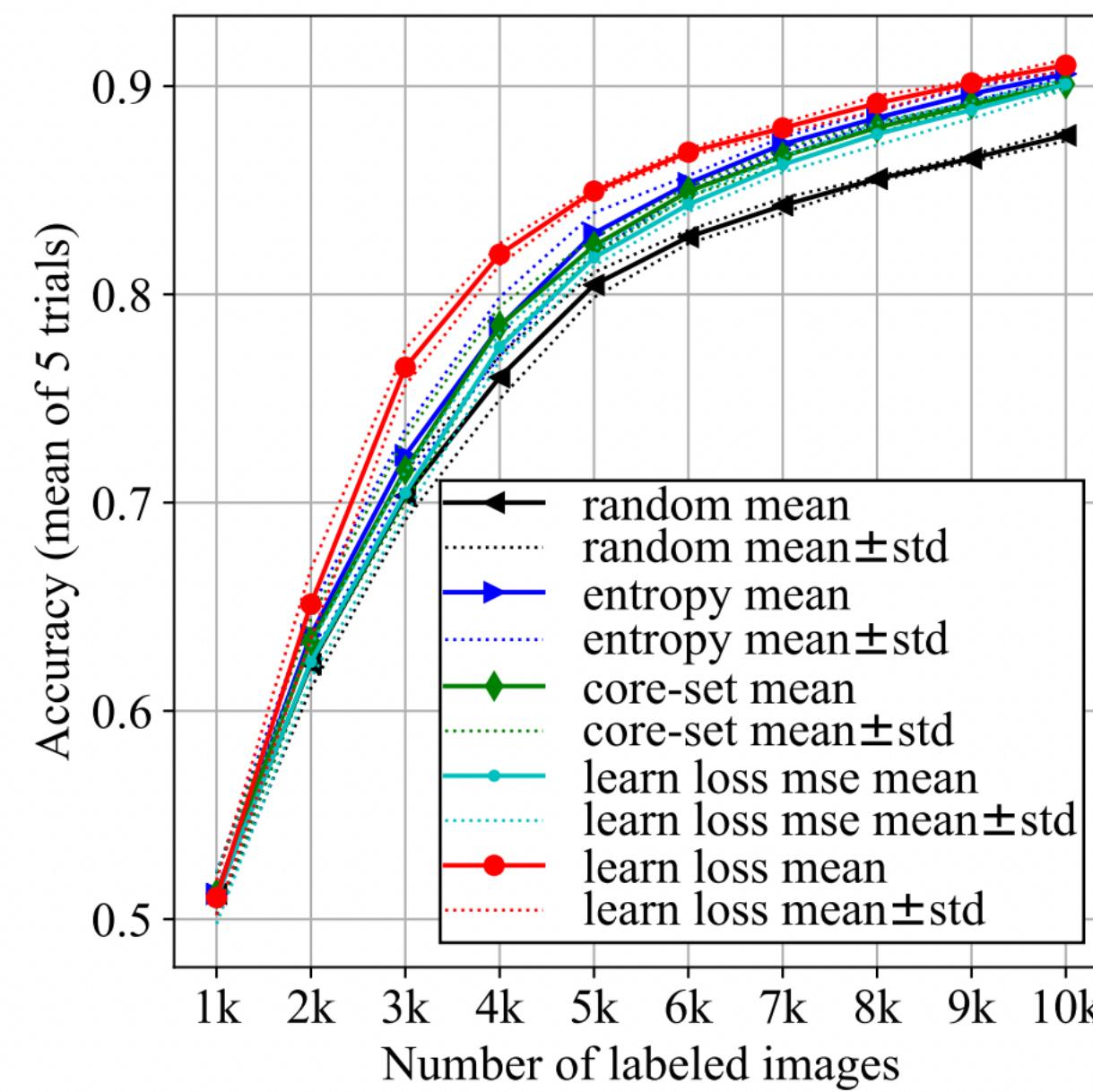


Figure 4. Active learning results of image classification over CIFAR-10.

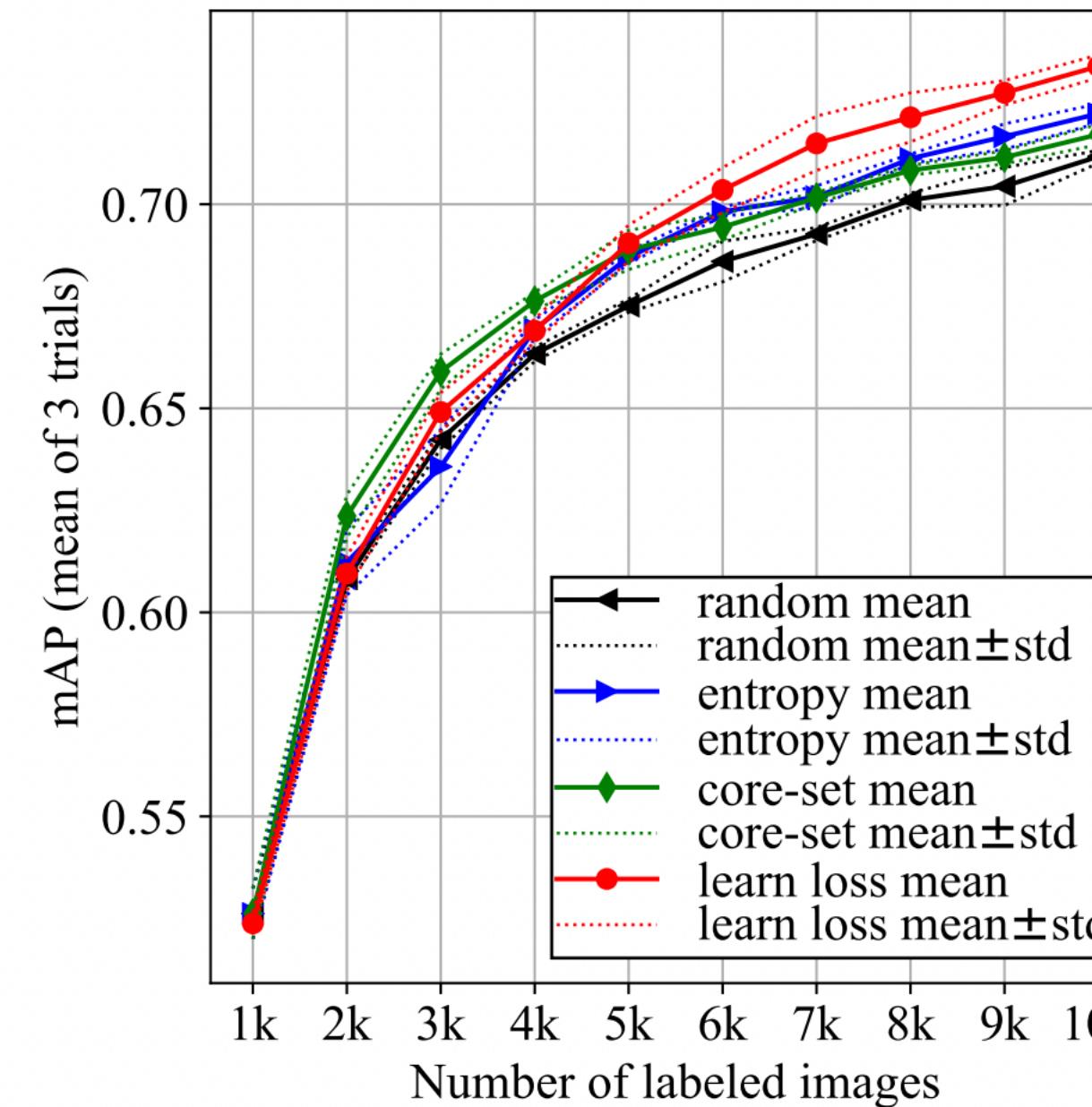


Figure 6. Active learning results of object detection over PASCAL VOC 2007+2012.

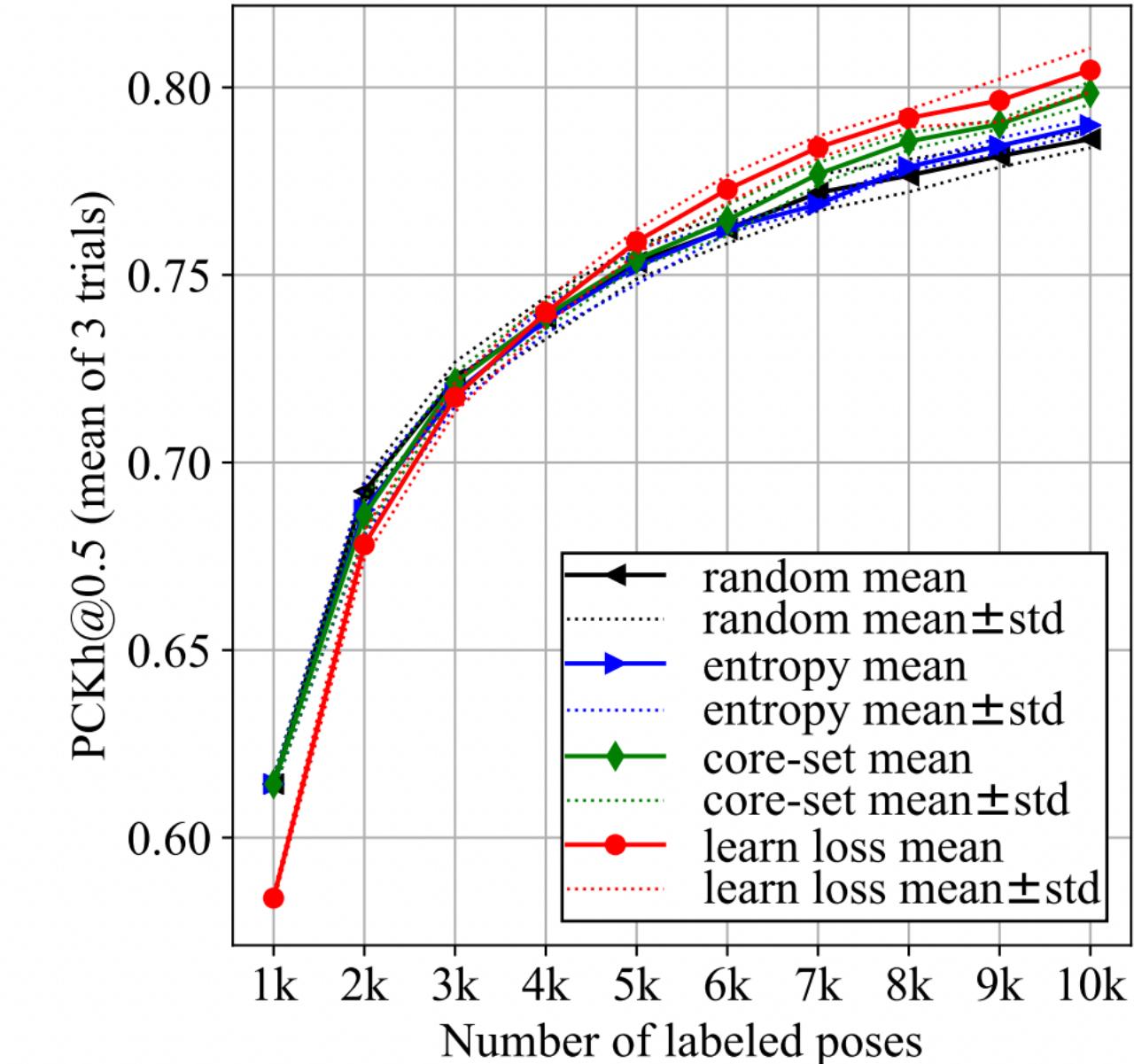


Figure 7. Active learning results of human pose estimation over MPII.

Evaluation of Loss Prediction Module

- 성공적인 loss를 예측하고 있는지 확인하기 위해 ranking accuracy를 이용함 -> MSE보다 높은 성능
- classification task에 비해 pose estimation, objection detection에서는 못미치는 예측 성능

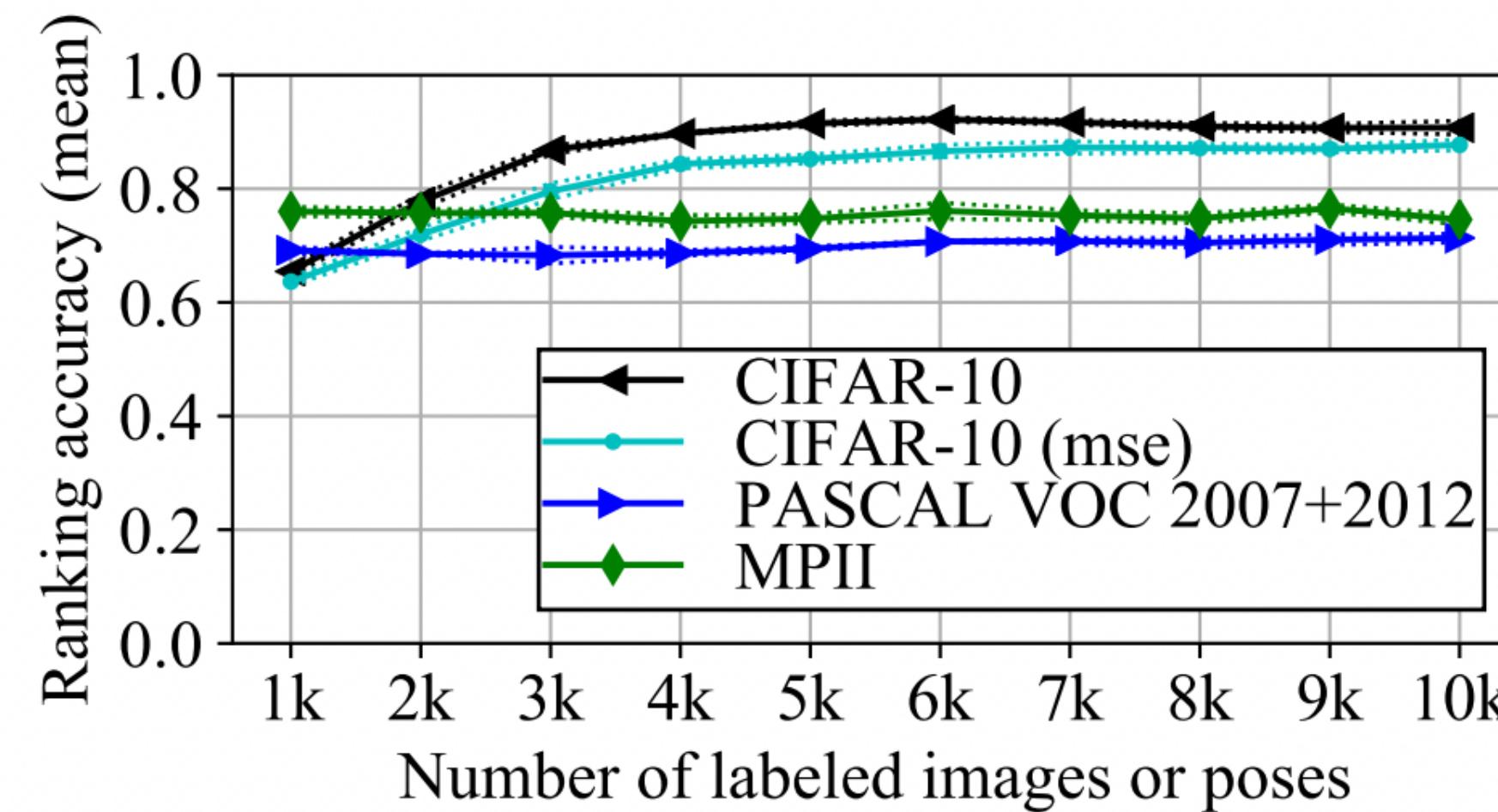
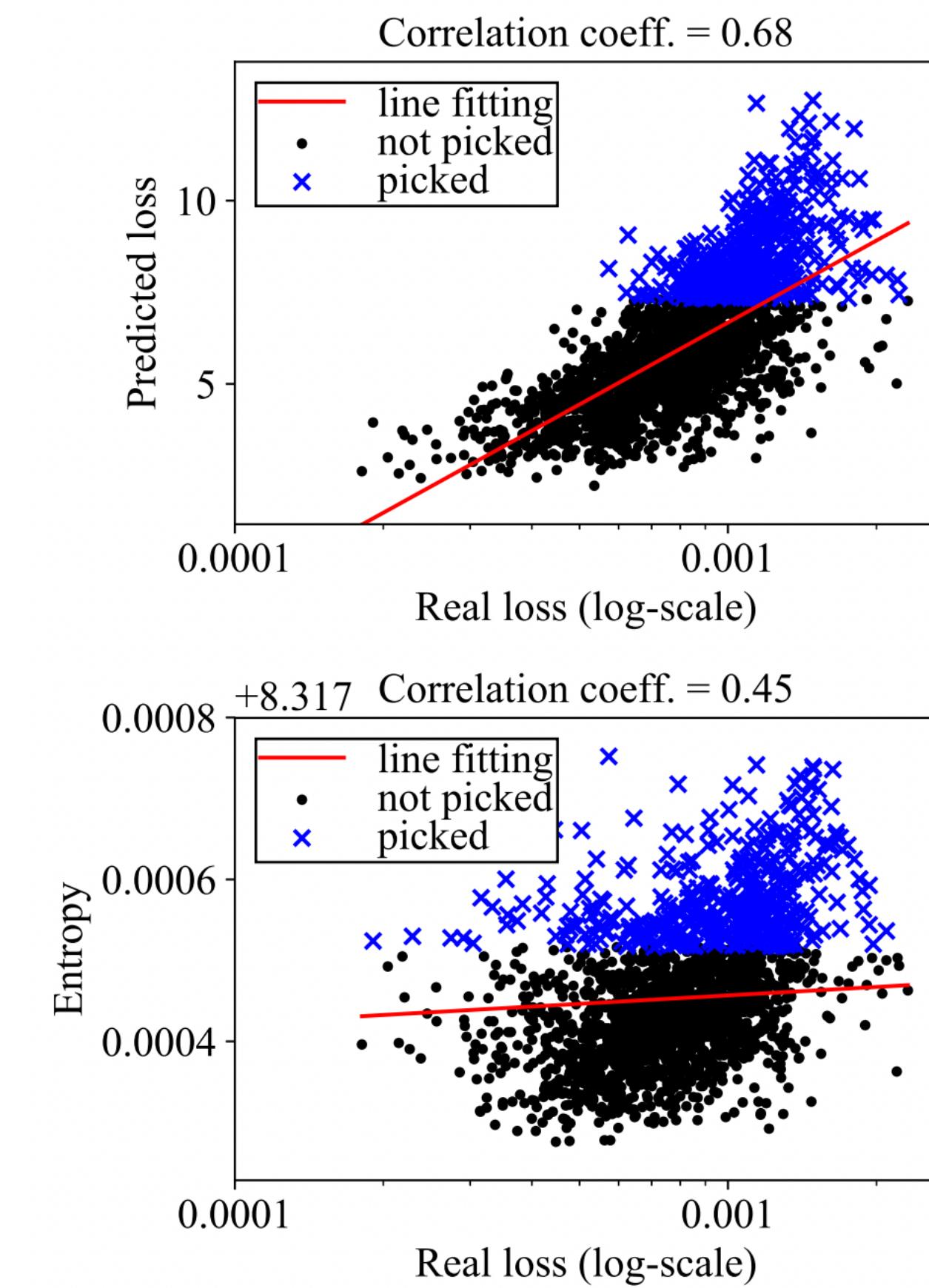


Figure 5. Loss-prediction accuracy of the loss prediction module.



제안하는 방법이 entropy방법보다 실제 정답 loss와 상관관계가 높다

Idea Development

- 본 논문의 한계점
 - 1) 하나의 쿼리 내의 이미지의 다양성은 고려하지 않음
 - 2) 복잡한 task에 대해서 loss prediction accuracy가 낮게 형성되었음. 따라서 데이터에 따라, 혹은 task에 따라 좀 더 효율적인 object function의 구성과 아키텍처를 고려해볼 수 있음.
 - > 즉, 다양한 동일한 continuous training이라는 목적을 가지고 있더라도(movie review / cyber attack related-tweet data와 같이 각 task (sentiment analysis, event classification)나 데이터의 특징에 따른 효율적인 설계에 초점을 향후 맞추어야할 것으로 보인다,

For me to consider ...

- Uncertain(informative)한 데이터에 대한 exemplar set을 유지하는것이 의미가 존재할까?
- CL에서 loss function을 정의하는 과정에서 기존의 데이터와 distinguished feature에 집중한 loss function을 정의할 수 있을까 (DNN이 아닌 CL의 concept을 attention mechanism에 적용?)
- Batch optimization을 predict하는 모듈(loss prediction module)과 Stable SGD의 전략관점에서 최적화 dynamic online learning objective function의 제안