Learning Loss for Active Learning

Donggeun Yoo[1,2] and In So Kweon[2]

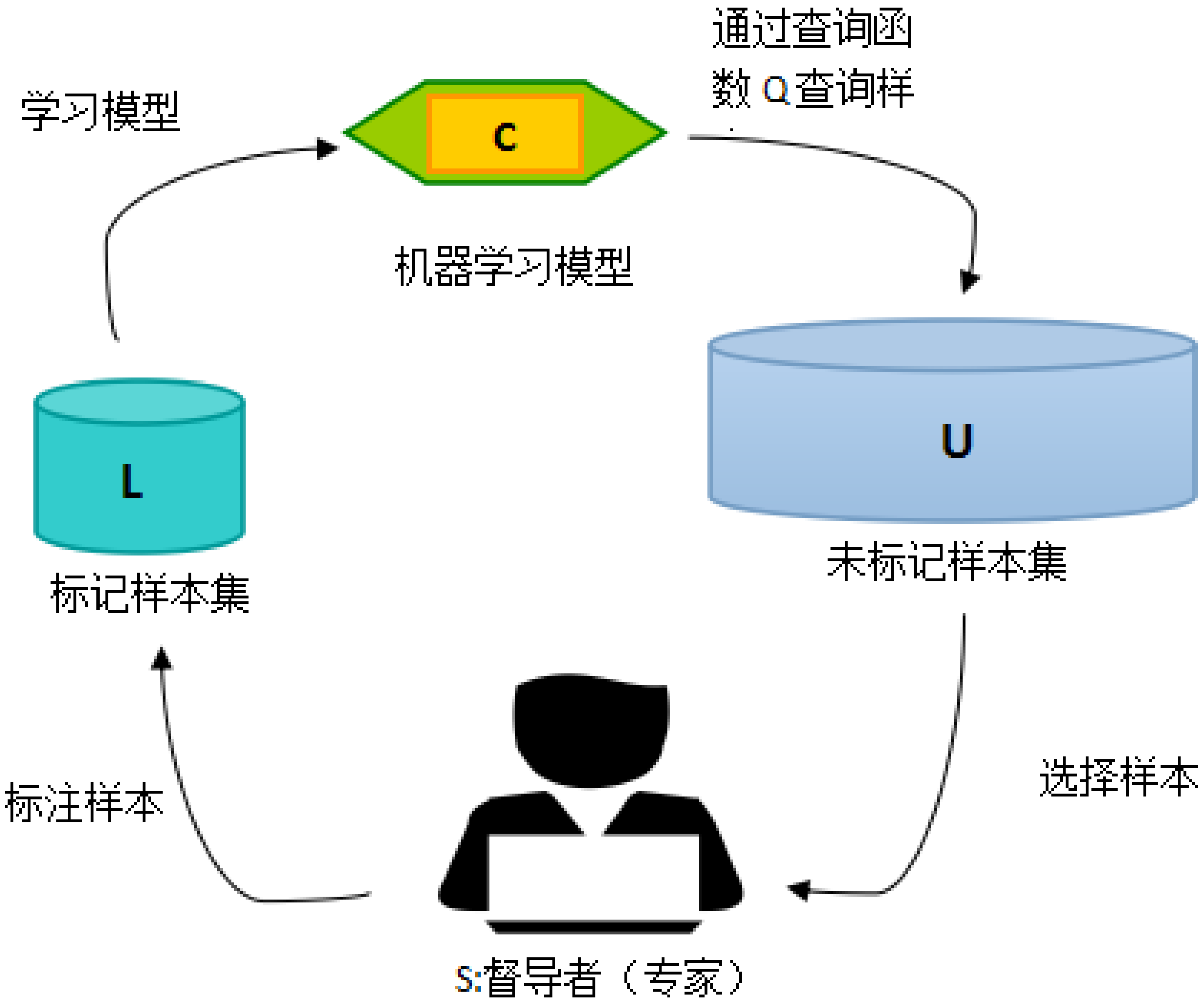
[1]Lunit Inc., Seoul, South Korea.

[2] KAIST, Daejeon, South Korea.

CVPR 2019

Content：

Introduction:



Approach：

uncertainty-based
diversity-based
expected model change
....

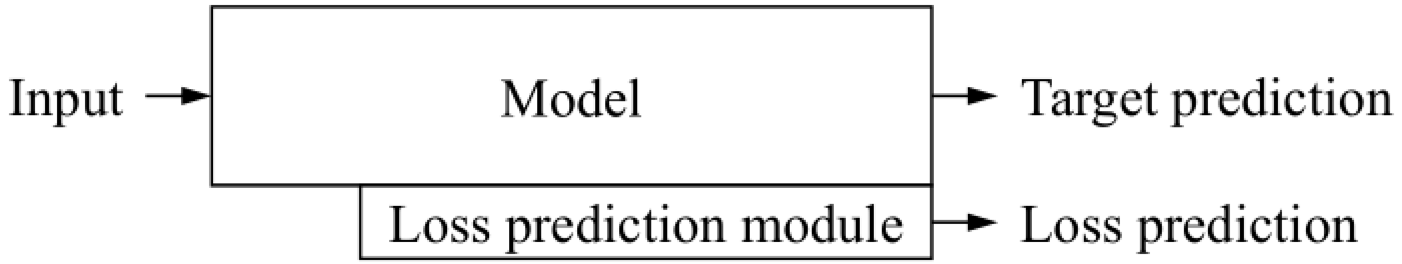
Problem:

Despite its simplicity, this approach has performed remarkably well in various scenarios. For more complex recognition tasks, it is required to redefine task specific uncertainty.

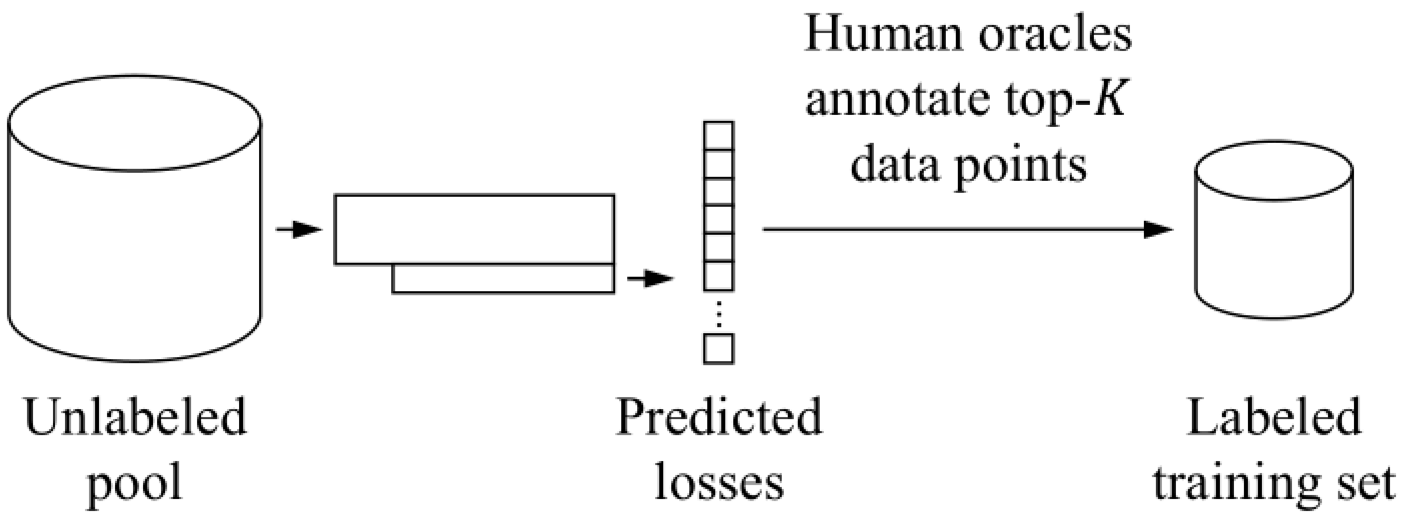
task-agnostic

If we can predict the loss of a data point, it becomes possible to select data points that are expected to have high losses.
The selected data points would be more informative to the current model.

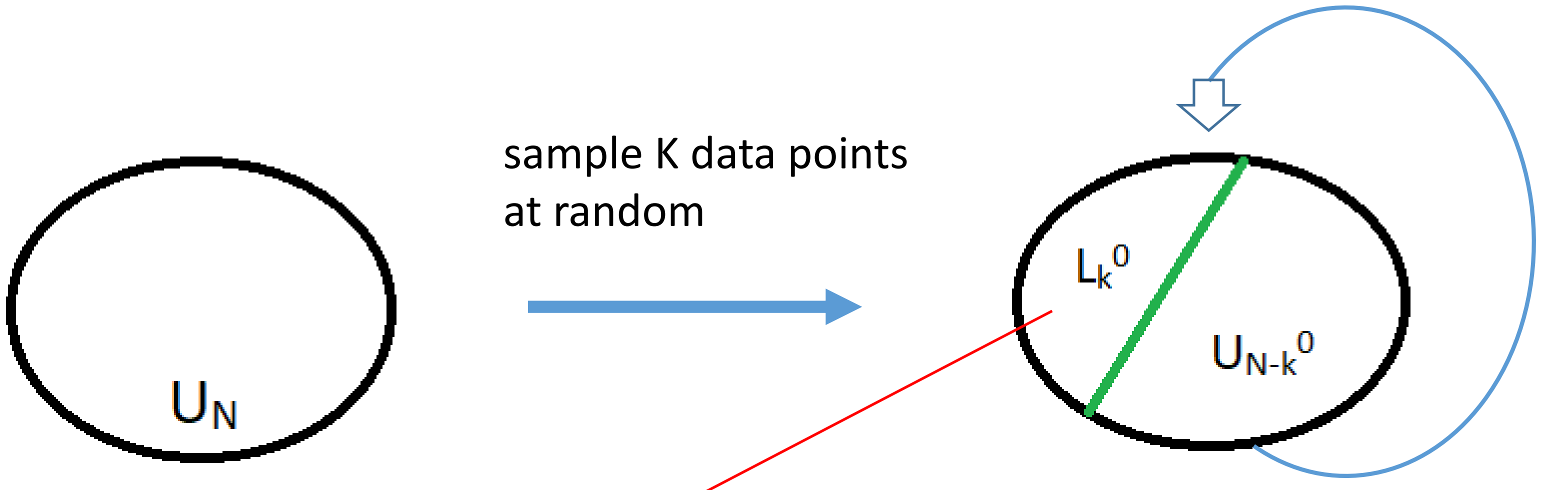
Approach



(a) A model with a loss prediction module



(b) Active learning with a loss prediction module



Notation	Explanation
Θ_{target}	target model 目标模型
Θ_{loss}	loss prediction module 损失预测模块
$\hat{y}=\Theta_{t\arg et}(x)$	目标模型的预测值
$l=L_{target}(\hat{y},y)$	目标模型预测的损失
h	x的特征集合，是Θ_{target}隐藏层的中间特征
$\hat{l}=\Theta_{loss}(h)$	损失预测模块对损失的预测值

Loss Prediction Module

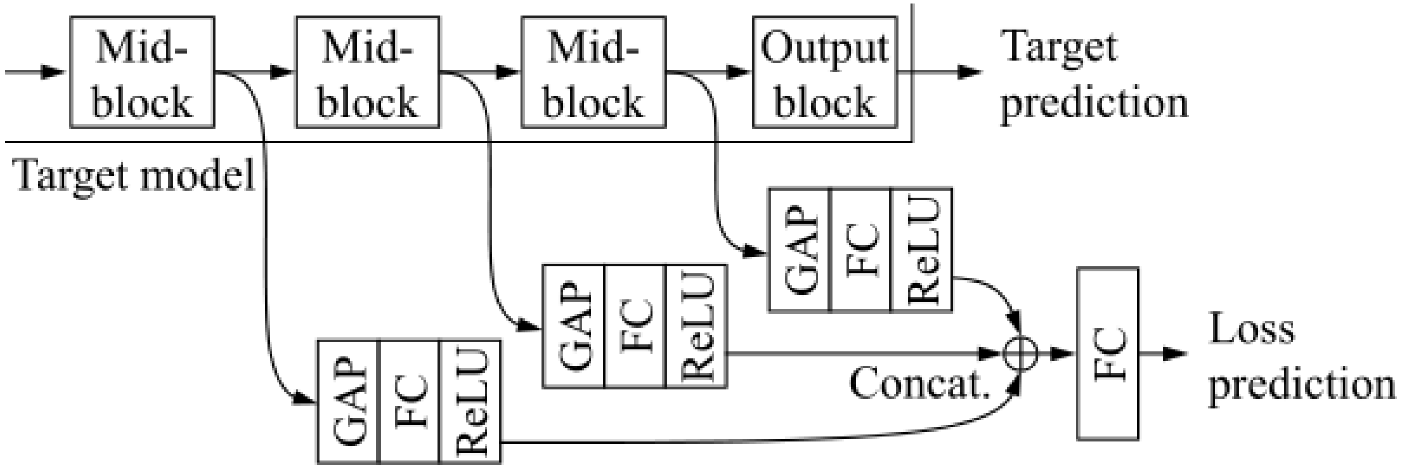


Figure 2. The architecture of the loss prediction module. This module is connected to several layers of the target model to take multi-level knowledge into consideration for loss prediction. The multi-level features are fused and map to a scalar value as the loss prediction.

(1) much smaller than the target model,
(2) jointly learned with the target model.

Learning Loss

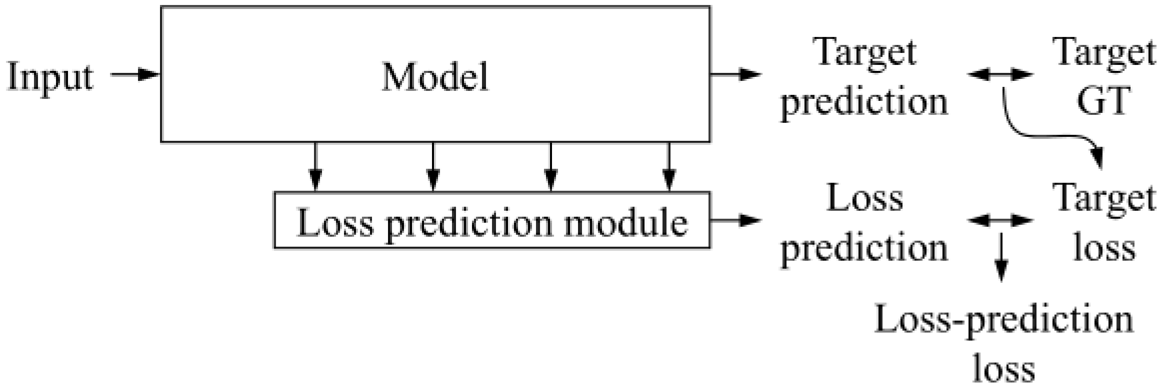


Figure 3. Method to learn the loss. Given an input, the target model outputs a target prediction, and the loss prediction module outputs a predicted loss. The target prediction and the target annotation are used to compute a target loss to learn the target model. Then, the target loss is regarded as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

The simplest way:

mean square error (MSE)?

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

L	具体任务采用的损失函数，比如说MSE $L_{loss}(\hat{l}, l) = (\hat{l} - l)^2$
λ	缩放常数，调整前后项大小关系

$$L_{loss}(\hat{l}, l) = (\hat{l} - l)^2$$

MSE is not a suitable choice for this problem since the scale of the real loss l changes (decreases in overall) as learning of the target model progresses. Minimizing MSE would let the loss prediction module adapt roughly to the scale changes of the loss l, rather than fitting to the exact value.

l	$\hat{l}$	MSE	误差比例
100	101	1	1%
50	50.5	0.25	1%
1	0.8	0.04	20%
0.5	0.6	0.01	20%

Solution

In the mini-batch whose size is B, we can make B/2 data pairs such as $\{x^p = (x_i, x_j)\}$.

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

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

case	L_{loss}	$= 0$
$l_i > l_j$	$max(0, \hat{l}_j - \hat{l}_i + \xi)$	$\hat{l}_i > \hat{l}_j + \xi$
$l_j > l_i$	$max(0, \hat{l}_i - \hat{l}_j + \xi)$	$\hat{l}_j > \hat{l}_i + \xi$

The final loss function:

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

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

$$\begin{aligned} & \hat{y} = \Theta_{\text{target}}(x) \\ \text{s.t.} \quad & \hat{l^p} = \Theta_{\text{loss}}(h^p) \\ & l^p = L_{\text{target}}(\hat{y^p}, y^p). \end{aligned}$$

Experiments and Results:

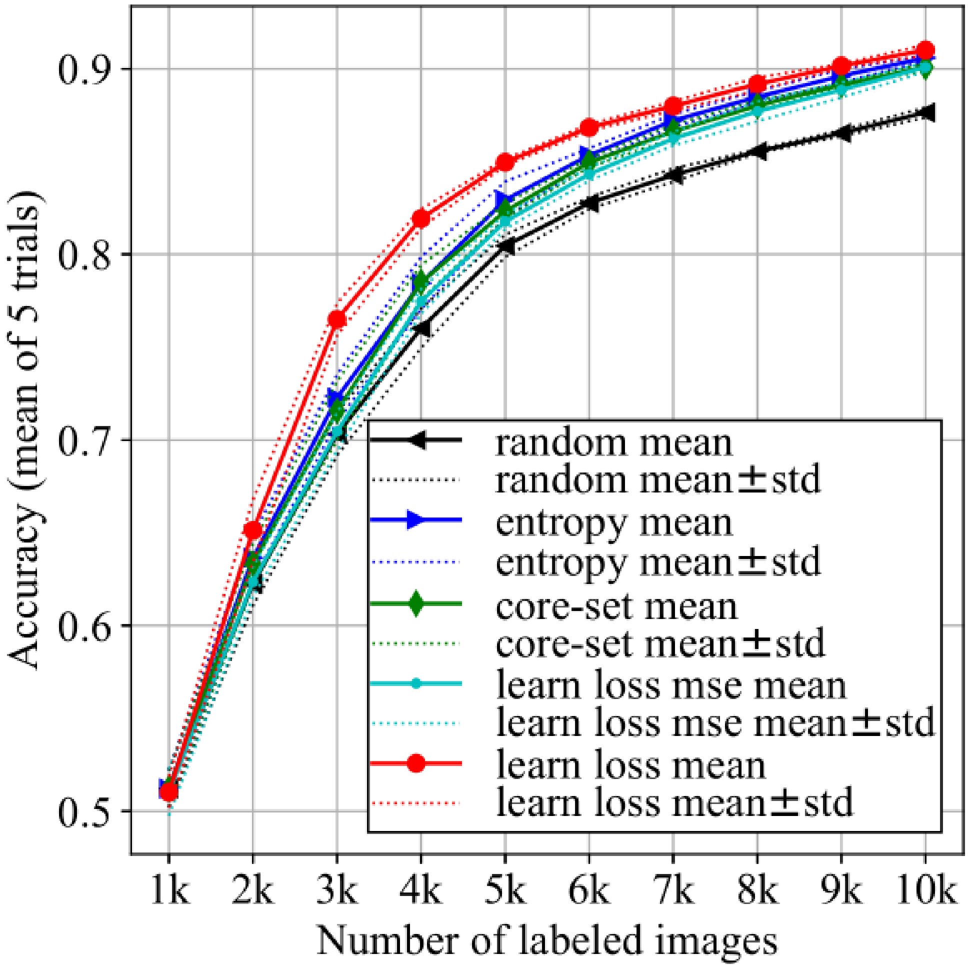


Figure 4. Active learning results of image classification over CIFAR-10.

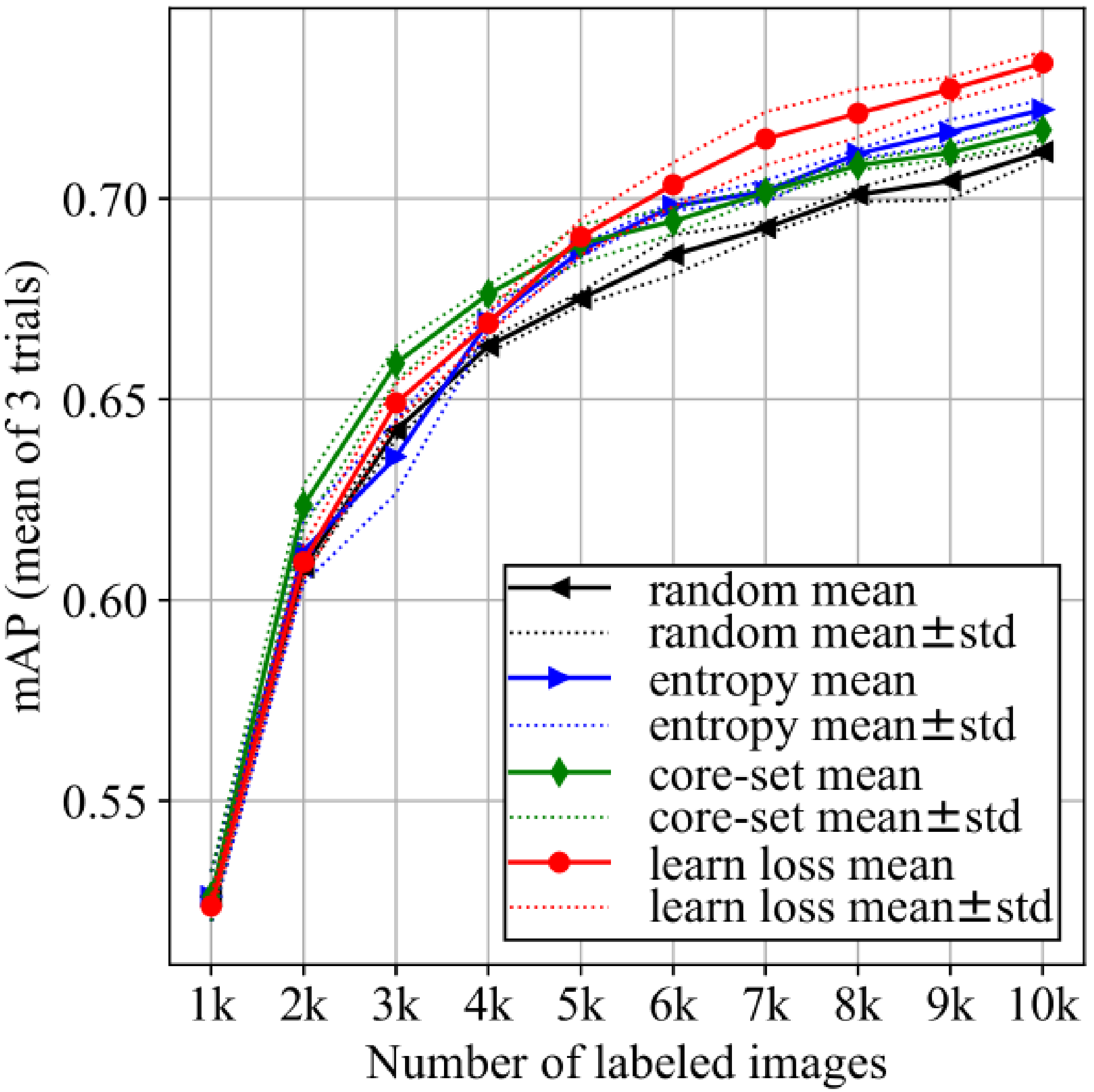


Figure 6. Active learning results of object detection over PASCAL VOC 2007+2012.

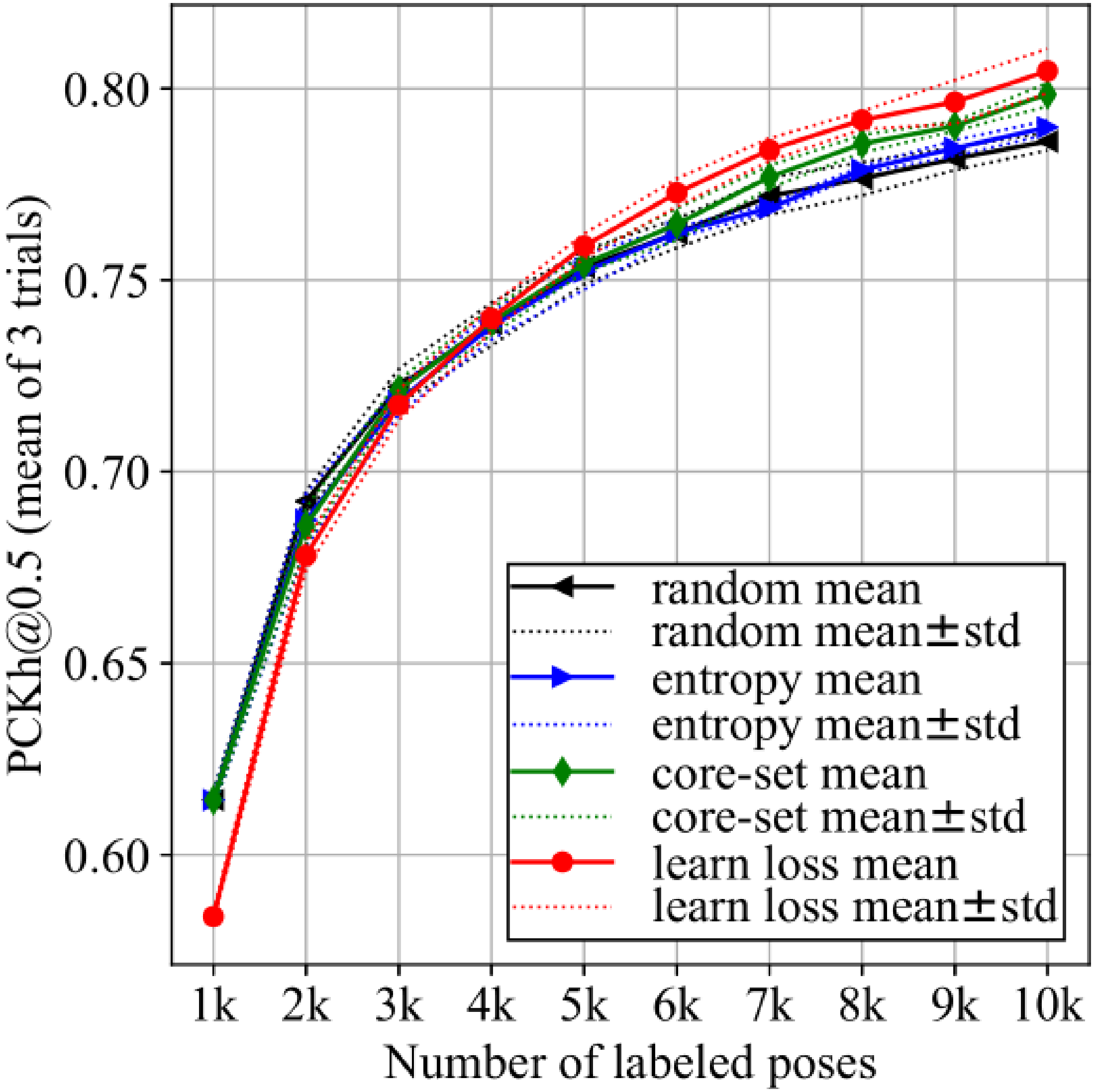


Figure 7. Active learning results of human pose estimation over MPII.

Conclusion:

Major contributions are

1. Proposing a simple but efficient active learning method with the loss prediction module, which is directly applicable to any tasks with recent deep networks.

2. Evaluating the proposed method with three learning tasks including classification, regression, and a hybrid of them, by using current network architectures.

Limitation:

1.the diversity or density of data was not considered.

2.the loss prediction accuracy was relatively low in complex tasks such as object detection and human pose estimation.