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

School of Computer Science & Engineering
UESTC, China

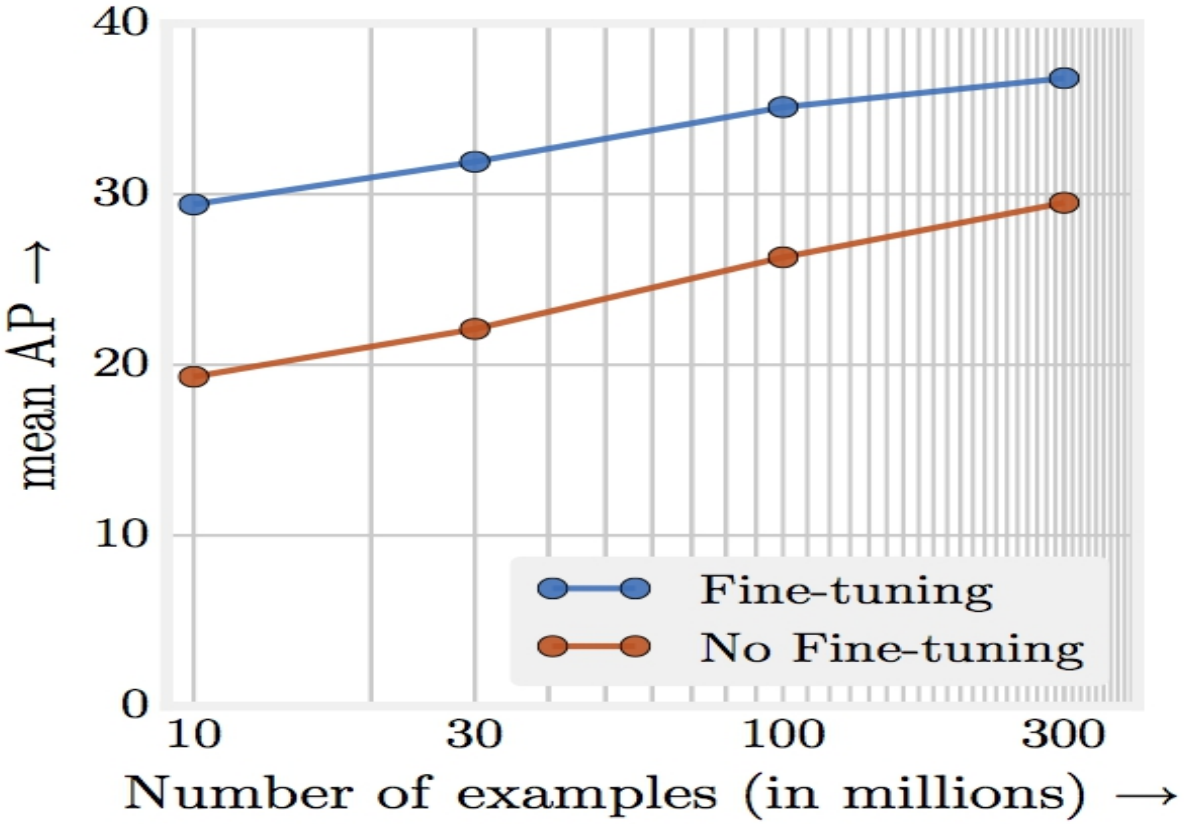
Guiyu Lai Mar 28, 2021

Contents

- **Introduction**
- **Approaches and limitations**
- ***Learning Loss for Active Learning***
- **Conclusion**
- **Inspiration**

Introduction

The performance of deep neural networks improves with more annotated data.



In the practical application scenarios, the problem is that the budget for annotation is limited...

Introduction

Howerver, sometimes labeled instances are very **difficult**, **time-consuming**, or **expensive** to obtain, e.g.:

- Annotating biomedical images

	X-Ray	CT
Time cost	4-5 min/image	15-20 min/image
Money cost	20-30 RMB/image	50-70 RMB/person

If we need to annotate 1K images, it will cost us about 3~4 days and 20,000~30,000 RMB for X-Ray, as well as 10~20 days and 50,000~70,000 RMB for CT images.

➔ One possible solution to this dilemma is ***active learning***.

Settles B. *Active learning*[J]. Synthesis lectures on artificial intelligence and machine learning, 2012, 6(1): 1-114.

Introduction

What is Active Learning?

- The key hypothesis:

  If the learning algorithm is allowed to choose the data from which it learns—to be "curious," if you will—it will perform better **with less training**.

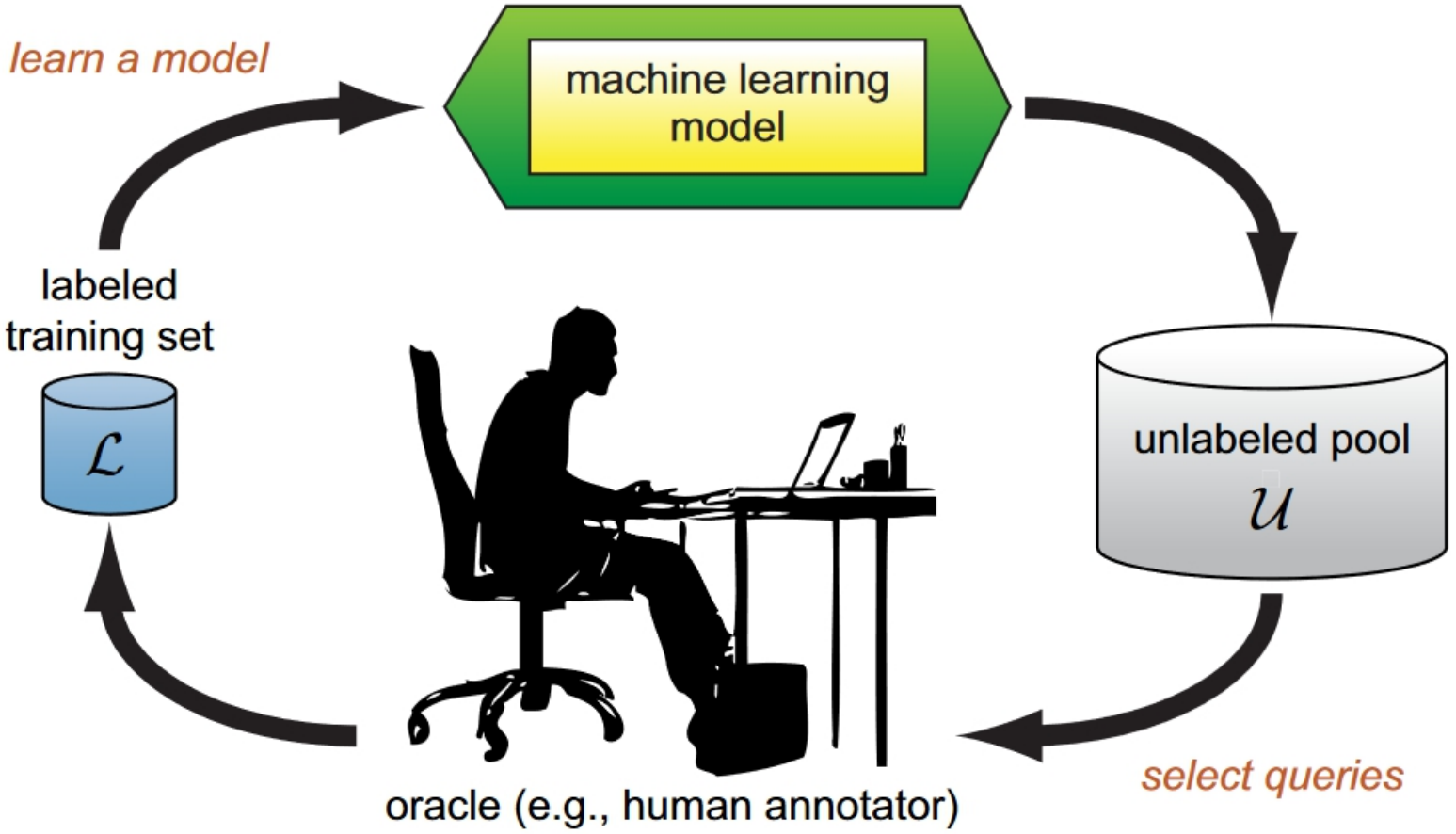
- The key idea:

  Active learning is that a machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which it learns.

Settles B. *Active learning literature survey*[J]. 2009.

Introduction

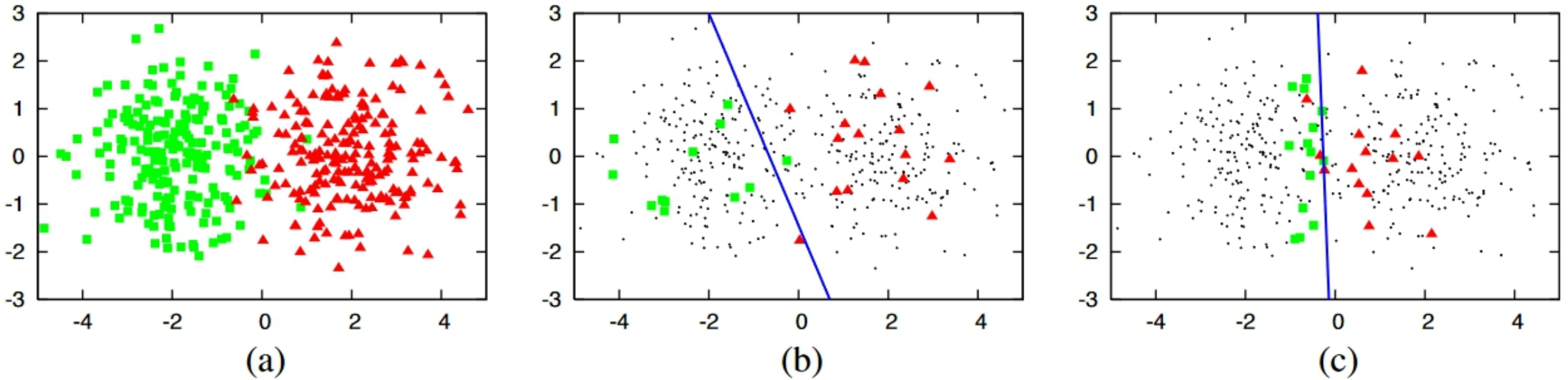
Taking the pool-based active learning cycle for example:



Settles B. *Active learning literature survey*[J]. 2009.

Introduction

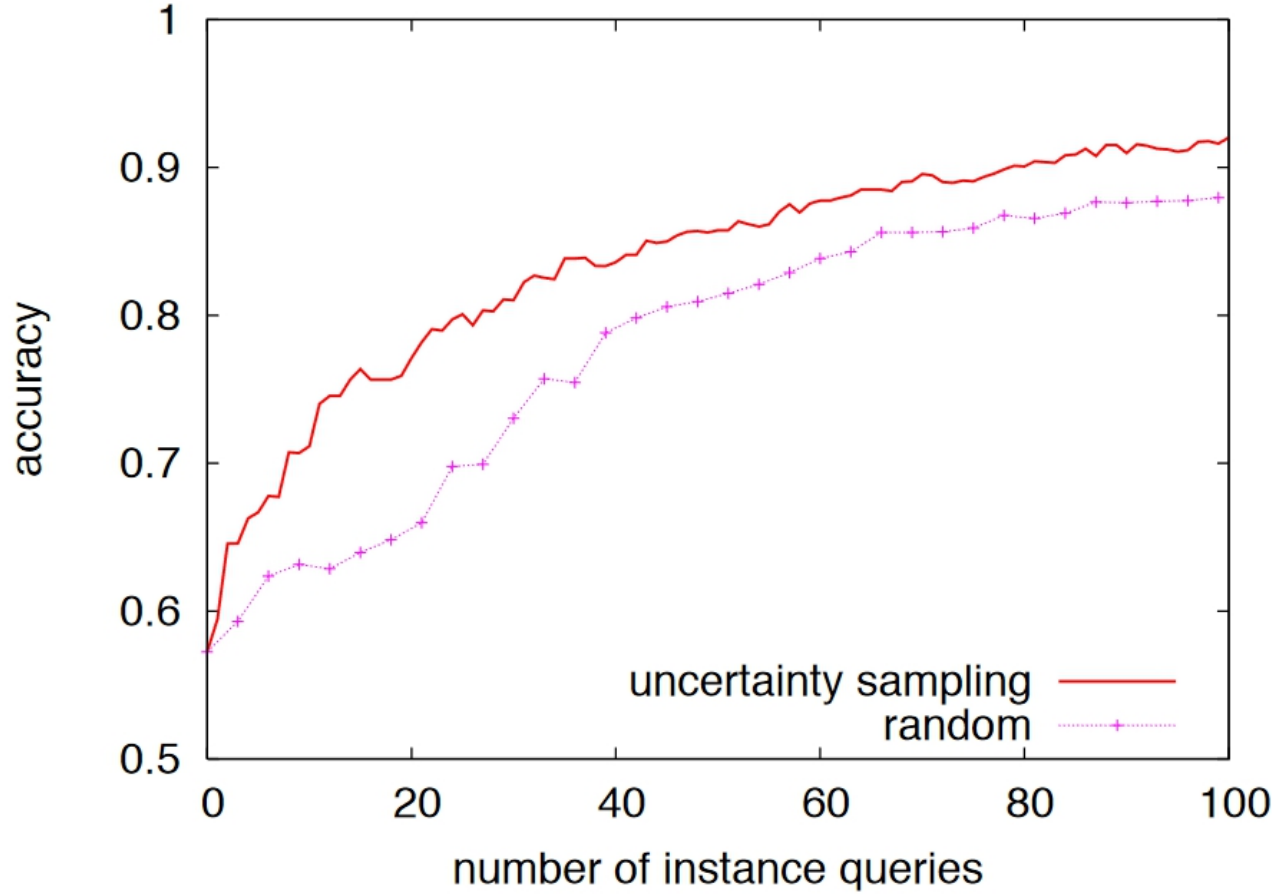
An illustrative example of pool-based *active learning*:



Settles B. *Active learning*[J]. Synthesis lectures on artificial intelligence and machine learning, 2012, 6(1): 1-114.

Introduction

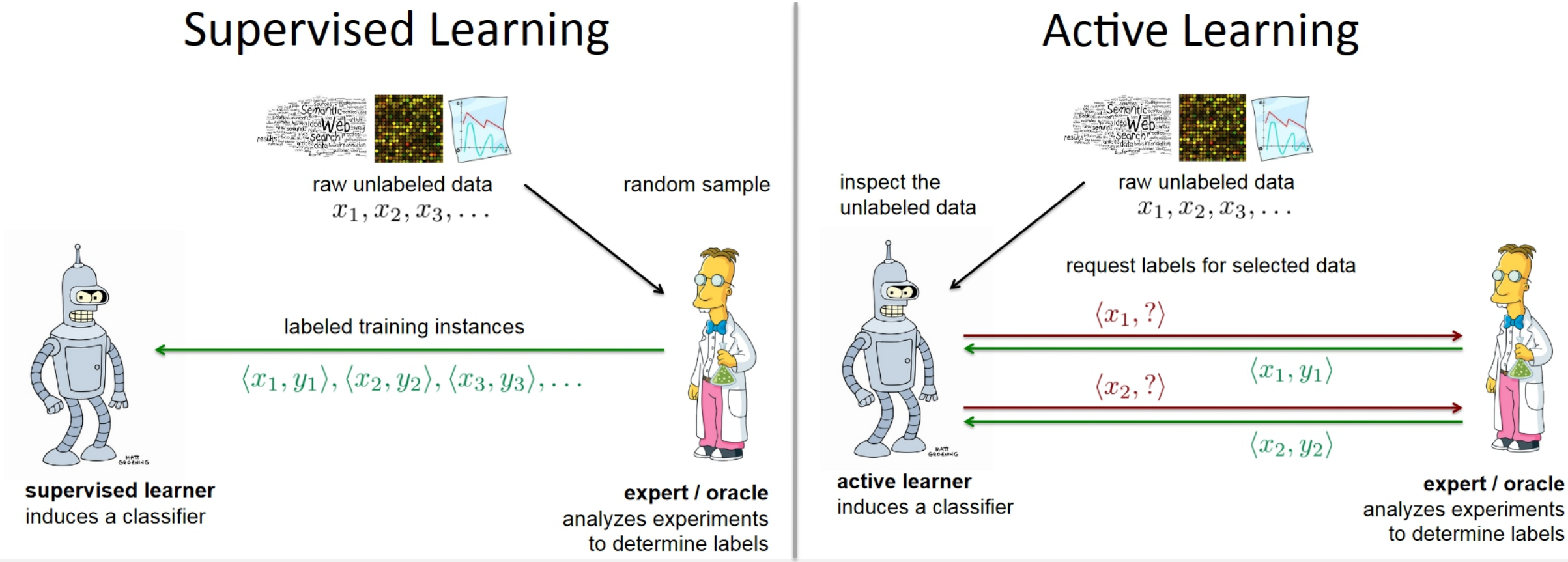
Active learning works:



Settles B. *Active learning*[J]. Synthesis lectures on artificial intelligence and machine learning, 2012, 6(1): 1-114.

Introduction

Active learning vs *Supervised learning*:



Settles B. *Active learning*[J]. Synthesis lectures on artificial intelligence and machine learning, 2012, 6(1): 1-114.

Introduction

Active learning vs *Semi-supervised learning*:

- In common:

  Both of them are aiming at improving the accuracy of classifier and reduce the workload of domain experts via selecting samples with **high value** from unlabeled samples and adding them to labeled samples.

- Differences:

  The selected unlabeled samples are **labeled artificially** in *active learning* while **no manual participation** is required in *semi-supervised learning*.

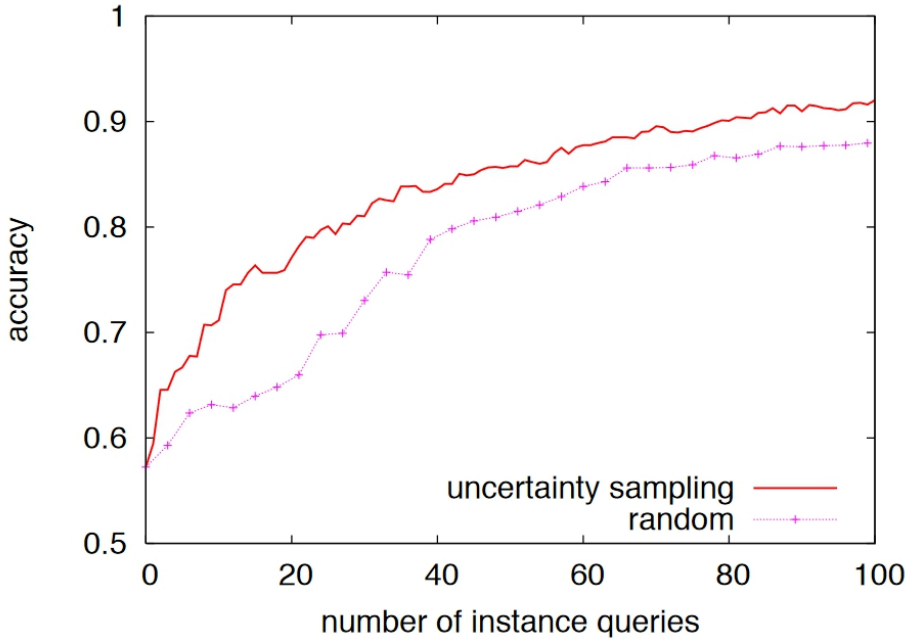
  However, the **accuracy** of the automatic annotation in *semi-supervised learning* not as good as that in *active learning*.

Settles B. *Active learning*[J]. Synthesis lectures on artificial intelligence and machine learning, 2012, 6(1): 1-114.

Introduction

However, it has been proved that given a fixed amount of data, the performance of the semi-supervised or unsupervised learning is still bound to that of fully-supervised learning.

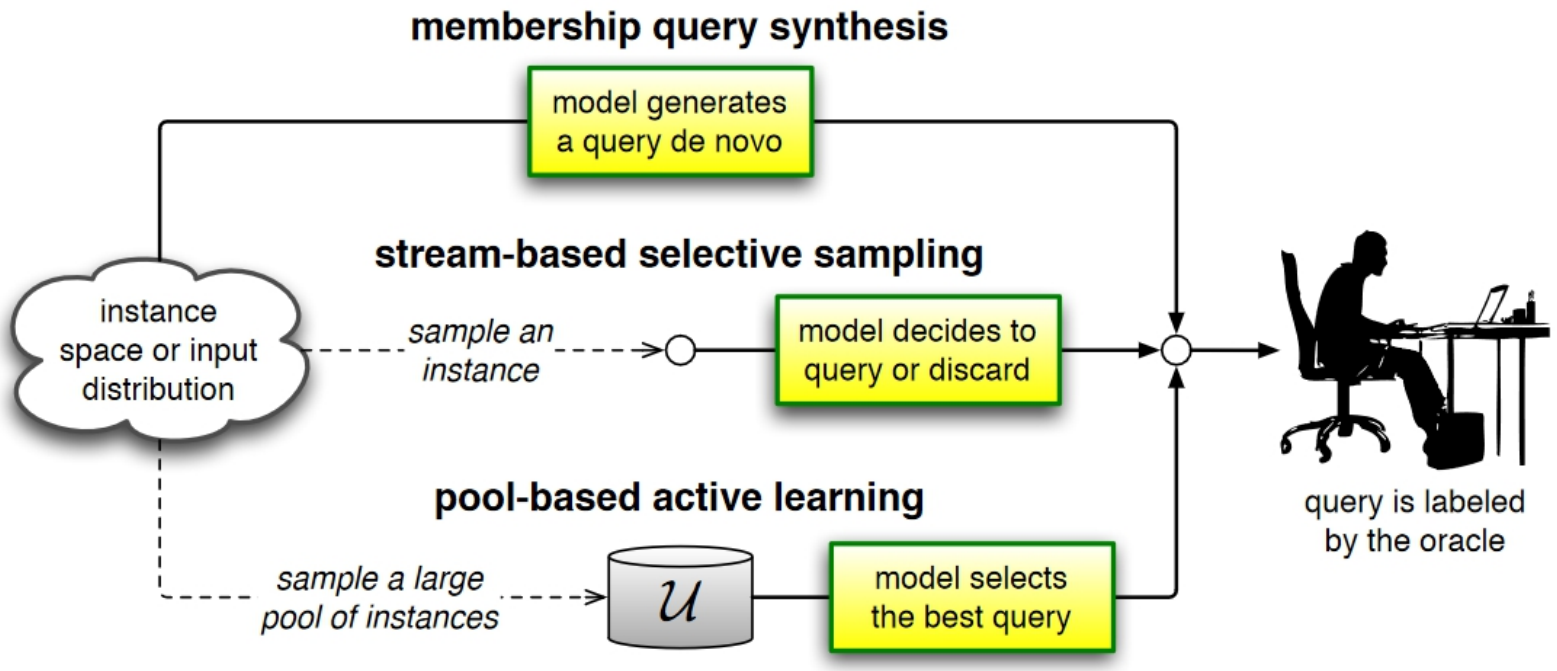
So, the budget for annotation is limited. What then is the most efficient use of the budget?



Settles B. *Active learning*[J]. Synthesis lectures on artificial intelligence and machine learning, 2012, 6(1): 1-114.

Approaches and limitations

Three main active learning scenarios:

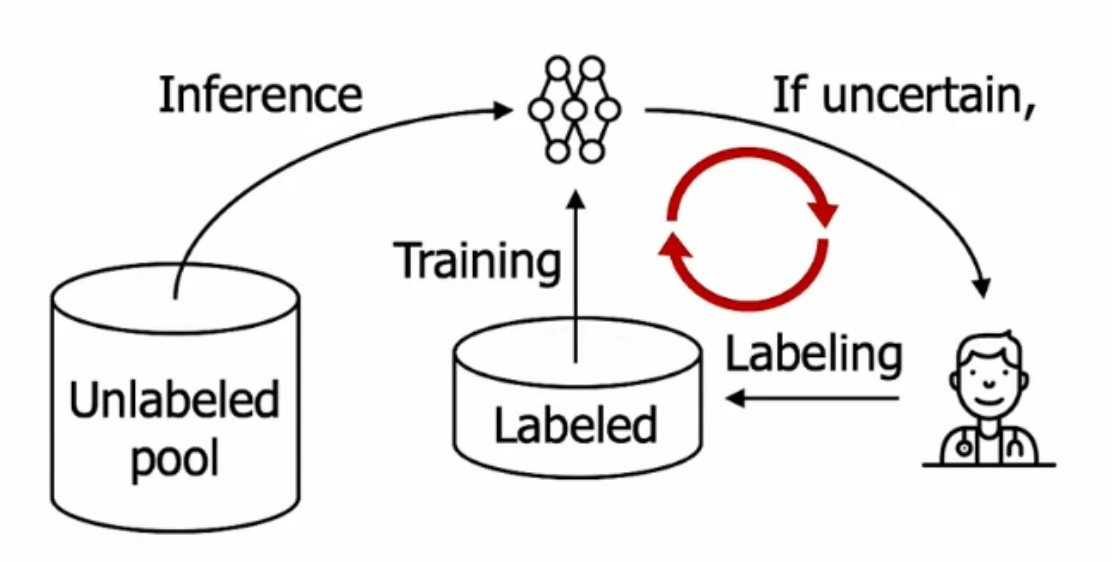


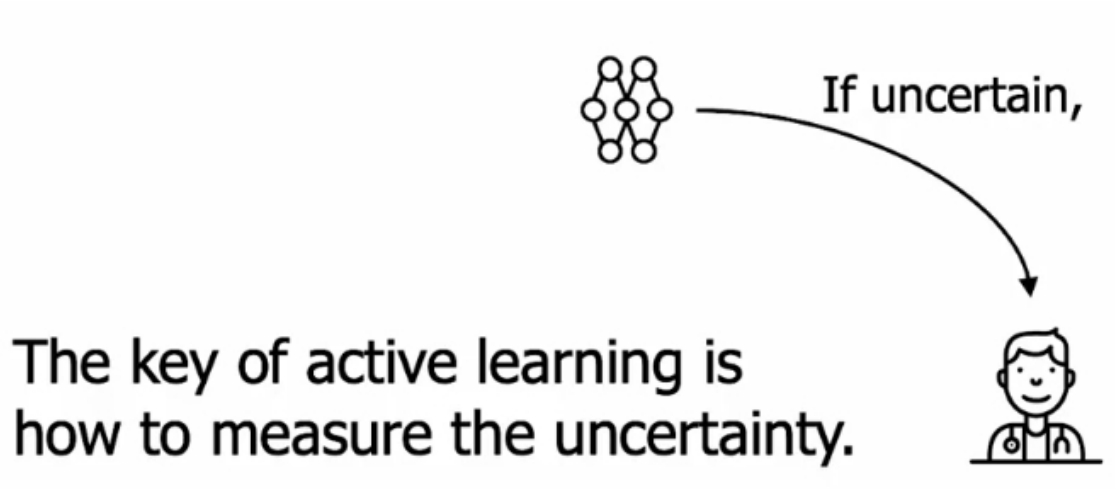
Settles B. *Active learning literature survey*[J]. 2009.

Approaches and limitations

Given a pool of unlabeled data, there have been three major approaches:

- Uncertainty-based approach
- Diversity approach
- Expected model change





Settles B. *Active learning literature survey*[J]. 2009.

Approaches and limitations

Limitations of previous approaches:

(-) task-specific

(-) computationally ineffient for large networks

The proposed novel active learning method:

(+) simple

(+) task-agnostic

(+) not heuristic, learning-based

(+) scalable to state-of-the-art networks and large data

Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Core idea

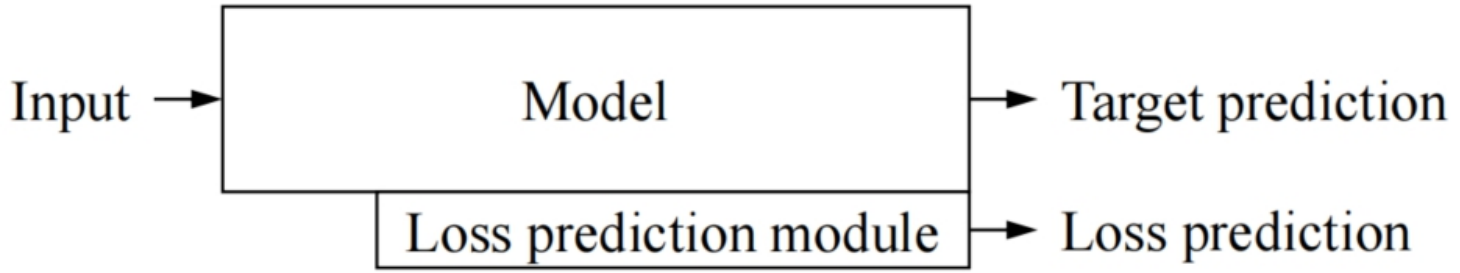
- A deep network is learned by minimizing a single loss.
- The most informative data point would be more beneficial to model improvement than a randomly chosen data point.
- The data points that have hign losses would be more informative to the current model.

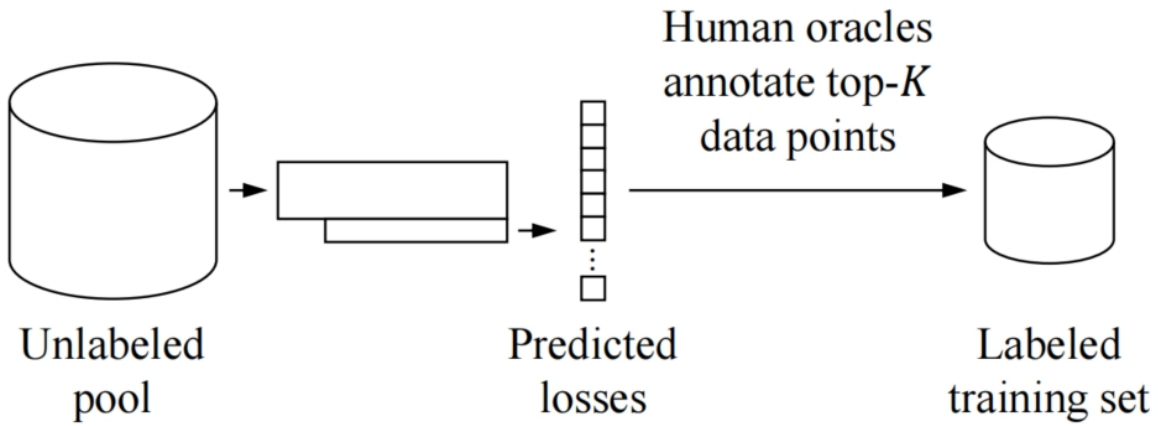
Predict the loss of an unlabeled data point
(regardless of what a task is, how many tasks there are, and how complex an architecture is)

Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Overview

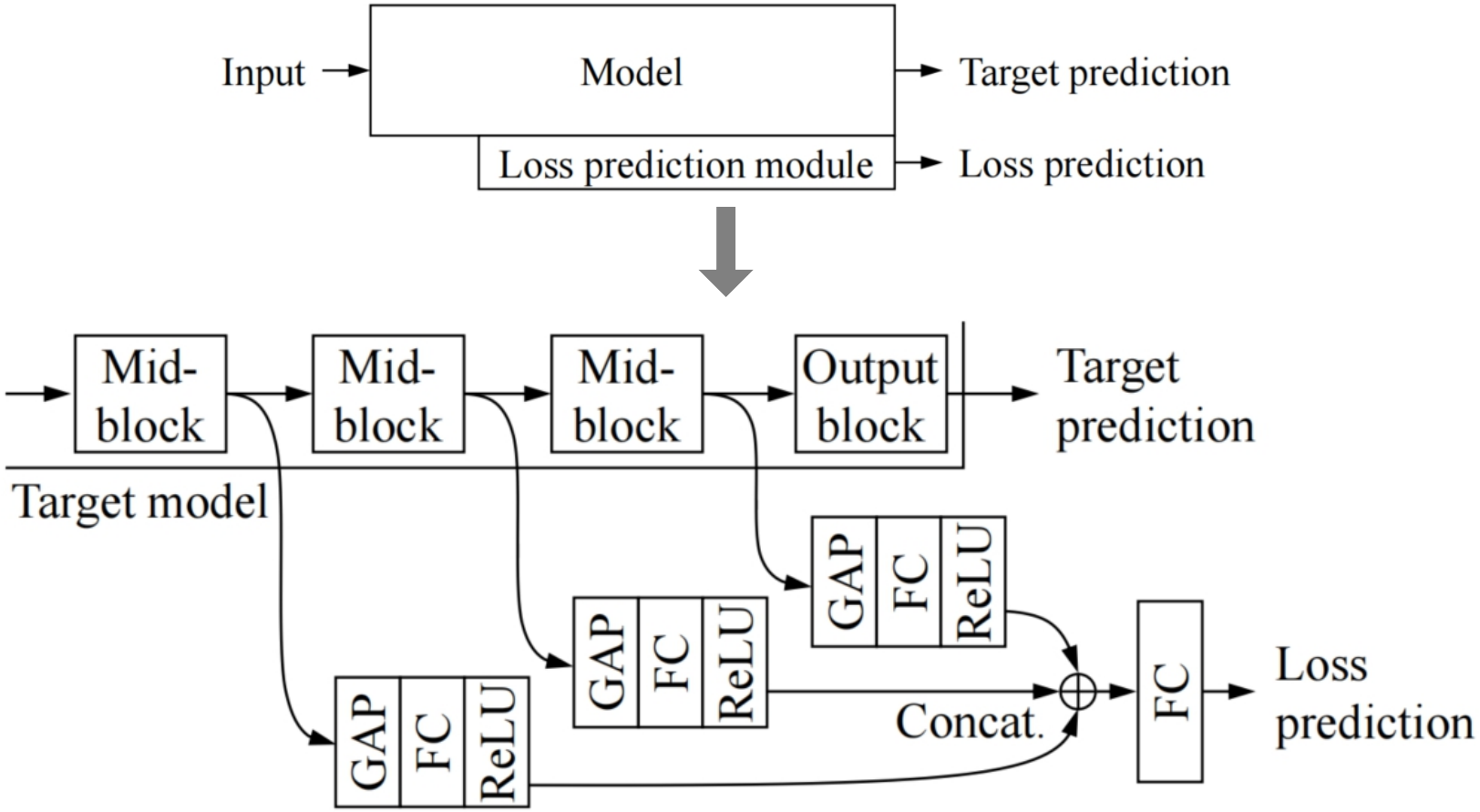




Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

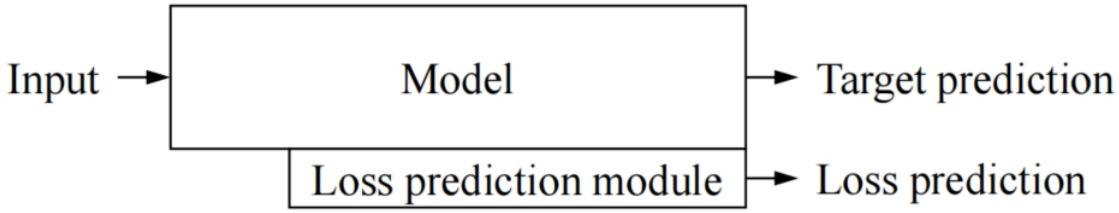
Loss Prediction Module



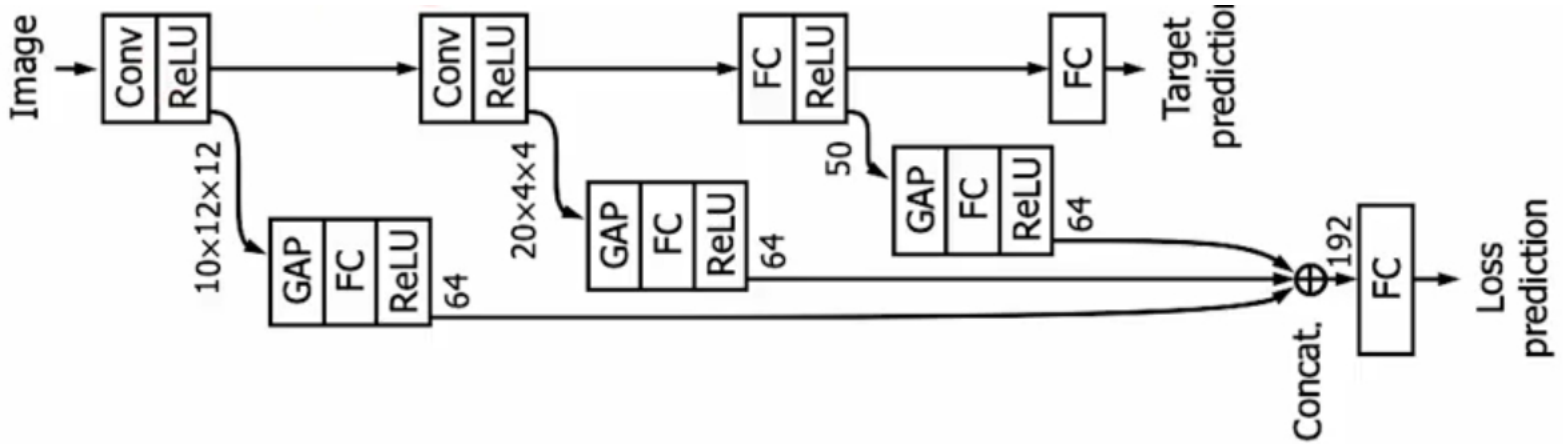
Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Loss Prediction Module



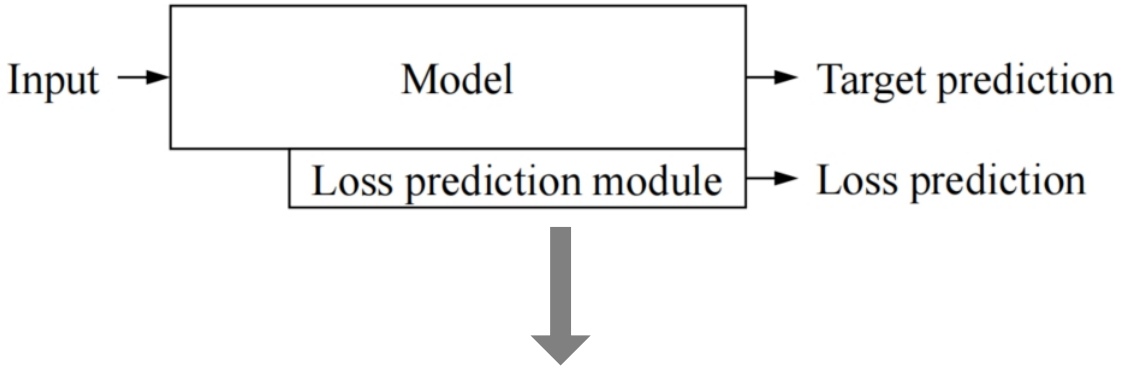
e.g. 1



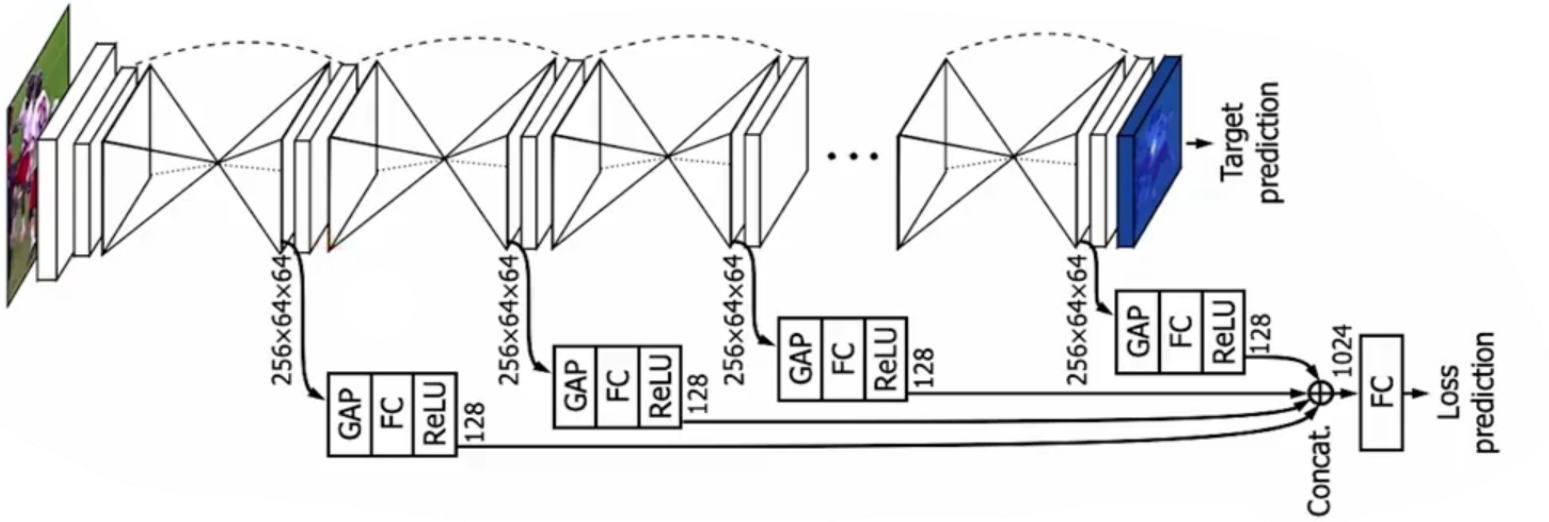
Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Loss Prediction Module



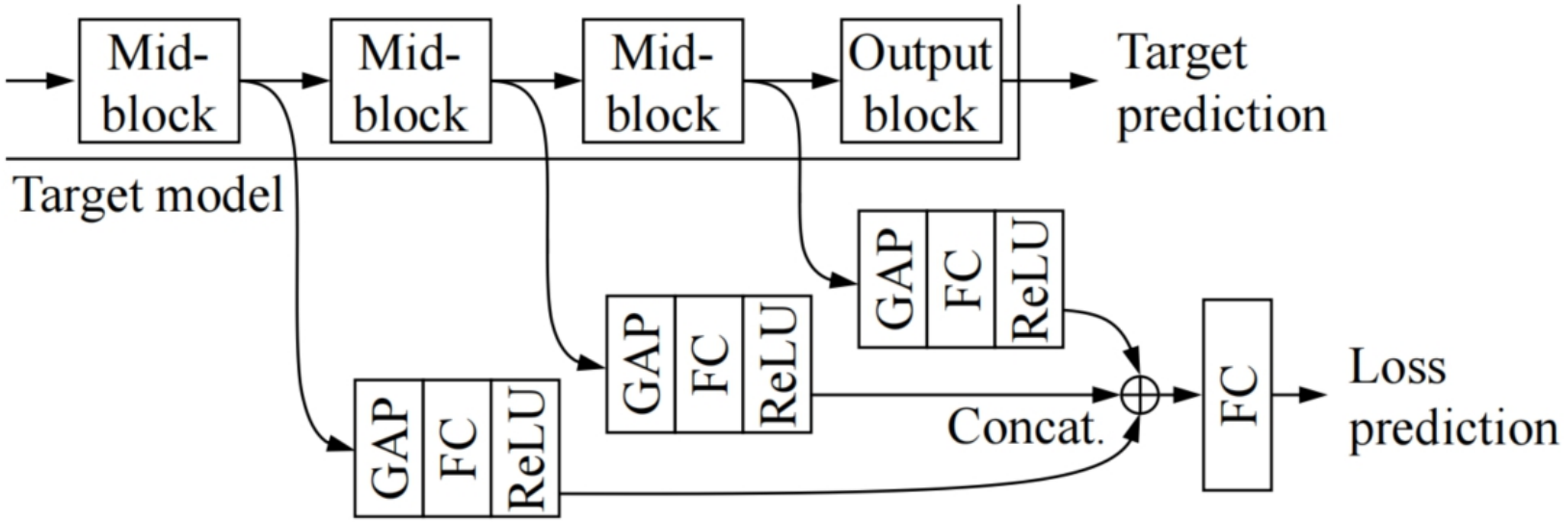
e.g. 2



Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Loss Prediction Module



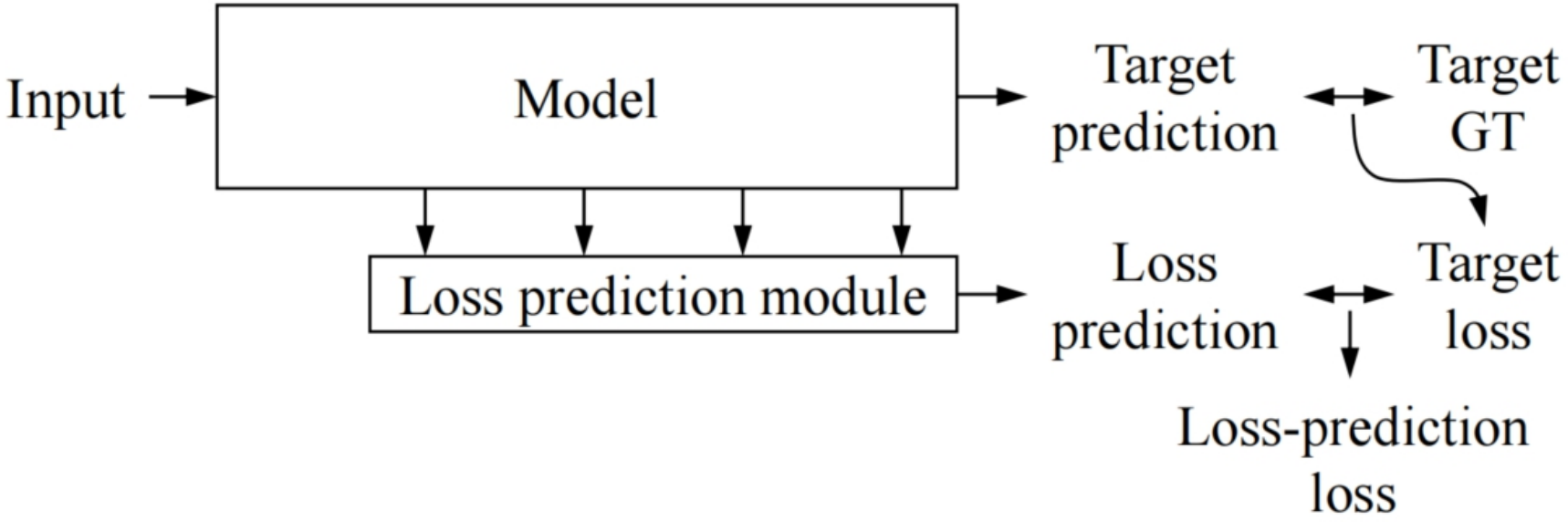
(+) Much smaller than the target model

(+) Jointly learned with the target model

Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

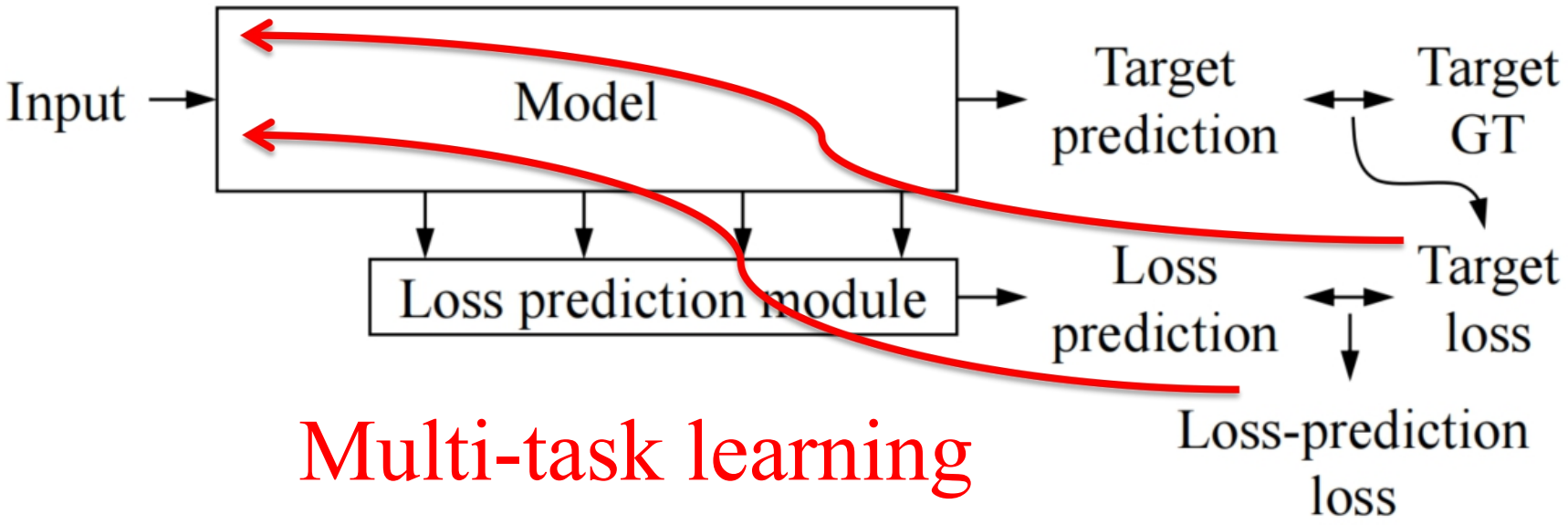
Learning Loss



Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

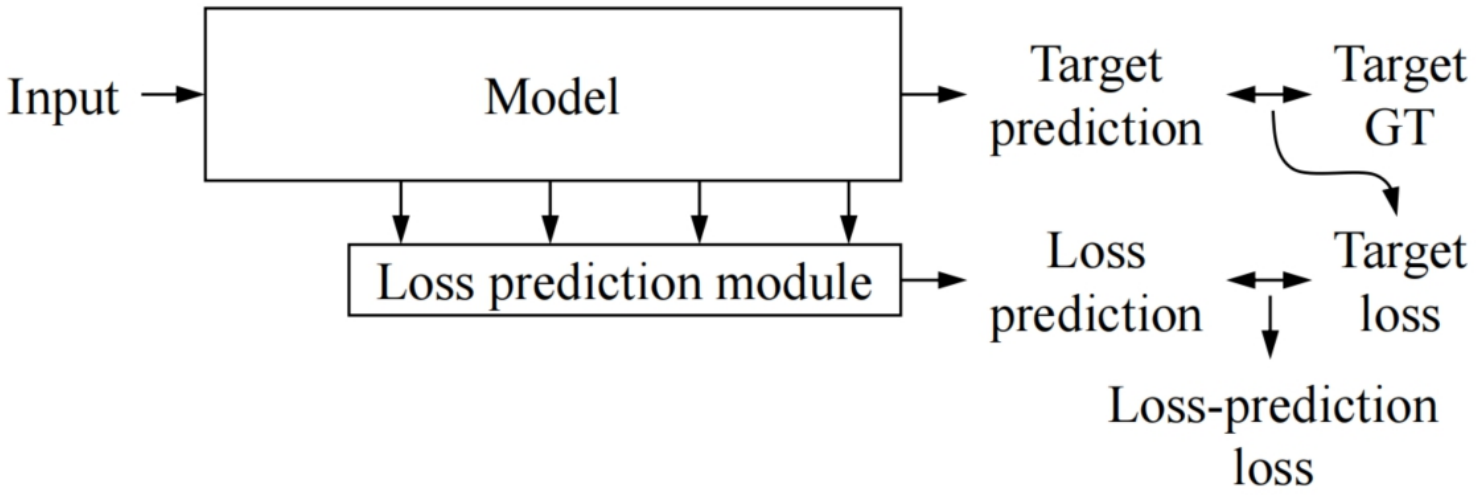
Learning Loss for Active Learning

Learning Loss



Learning Loss for Active Learning

Learning Loss



(+) Applicable to

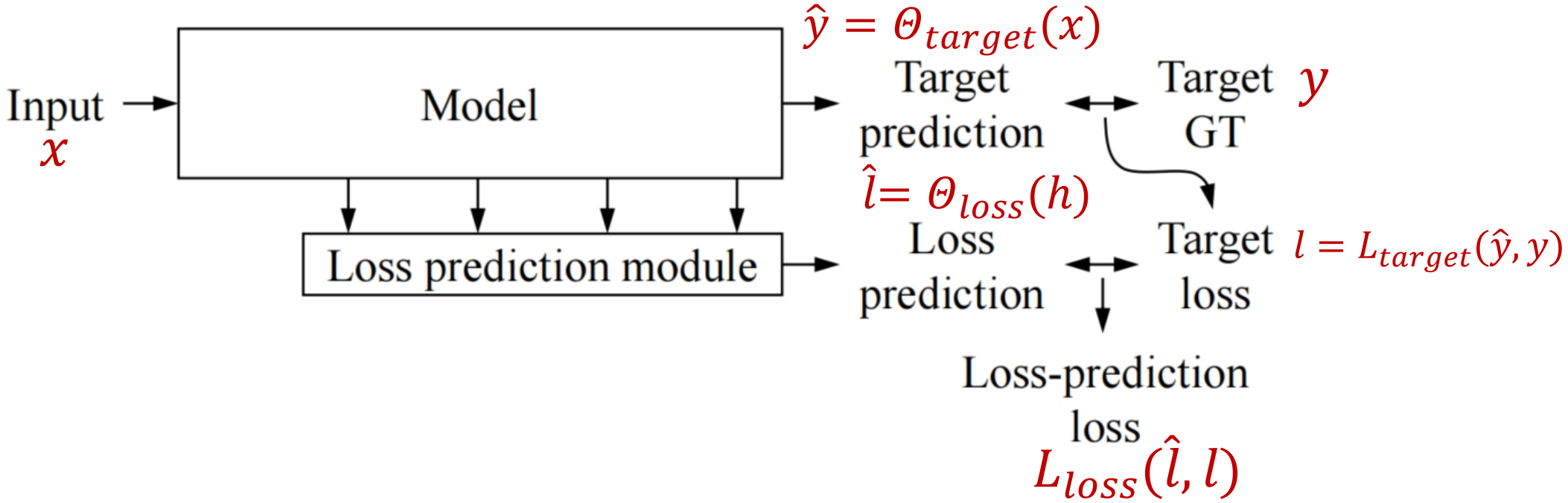
- any network and data
- any tasks

(+) Nearly zero cost

Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

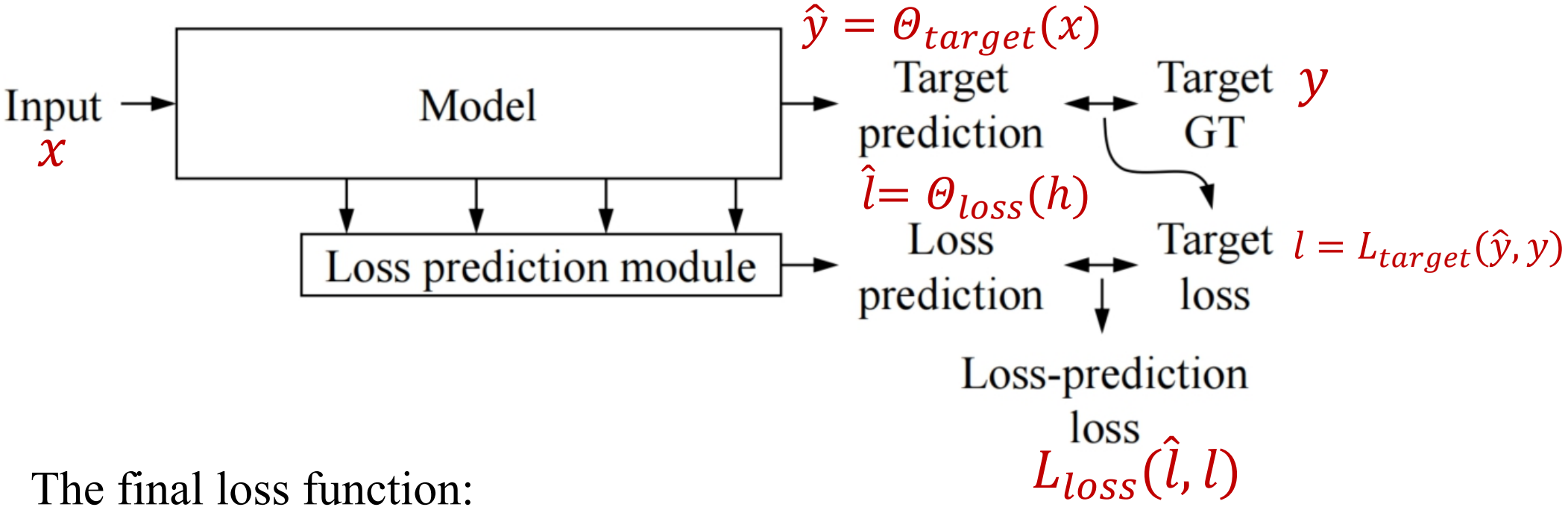
Learning Loss



Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Learning Loss



The final loss function:

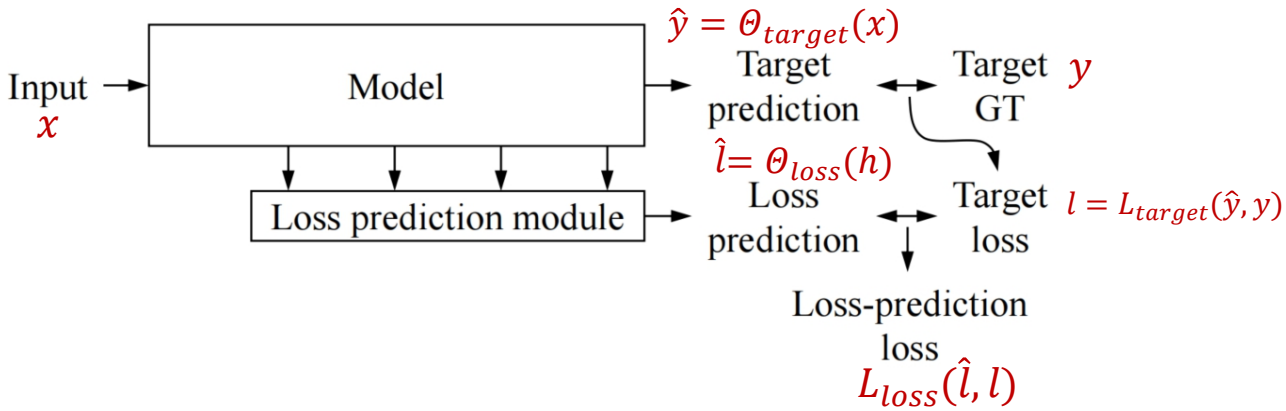
$$L_{\text{target}}(\hat{y}, y) + \lambda \cdot L_{\text{loss}}(\hat{l}, l)$$

Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Learning Loss

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



How to define the loss-prediction loss function?

- By the mean square error (MSE)?

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

(-) The scale of the real loss l changes as the learning of the target model progresses

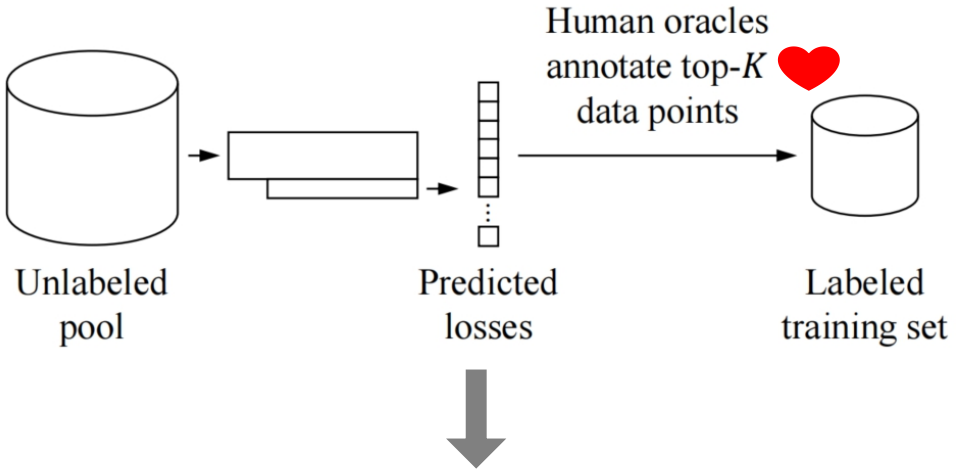
⬇

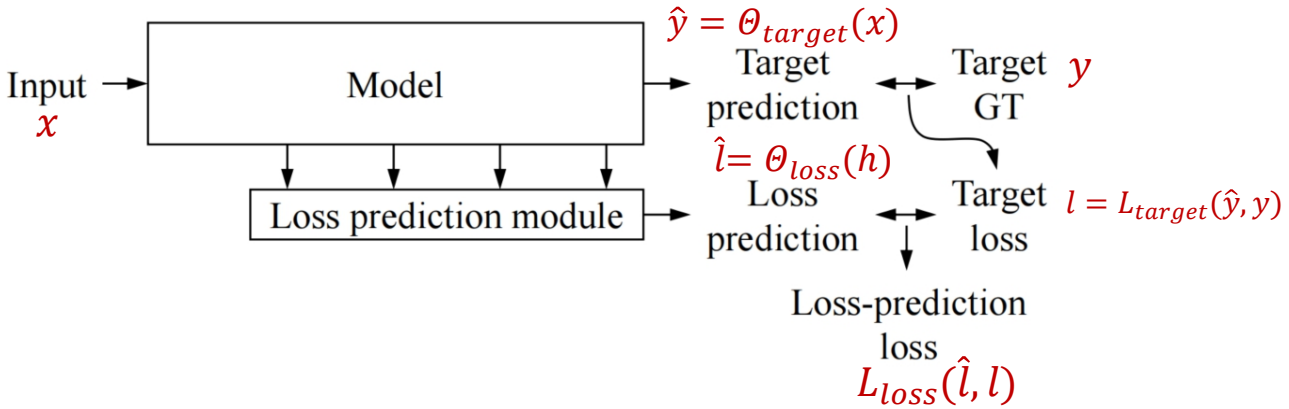
TODO: The loss-prediction loss function has to discard the overall scale of l.

Learning Loss for Active Learning

Learning Loss

TODO: The loss-prediction loss function has to discard the overall scale of l.





✖: exact value of the loss $\hat{l}$

✓: ranking of $\hat{l}$

Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Learning Loss for Active Learning

Learning Loss

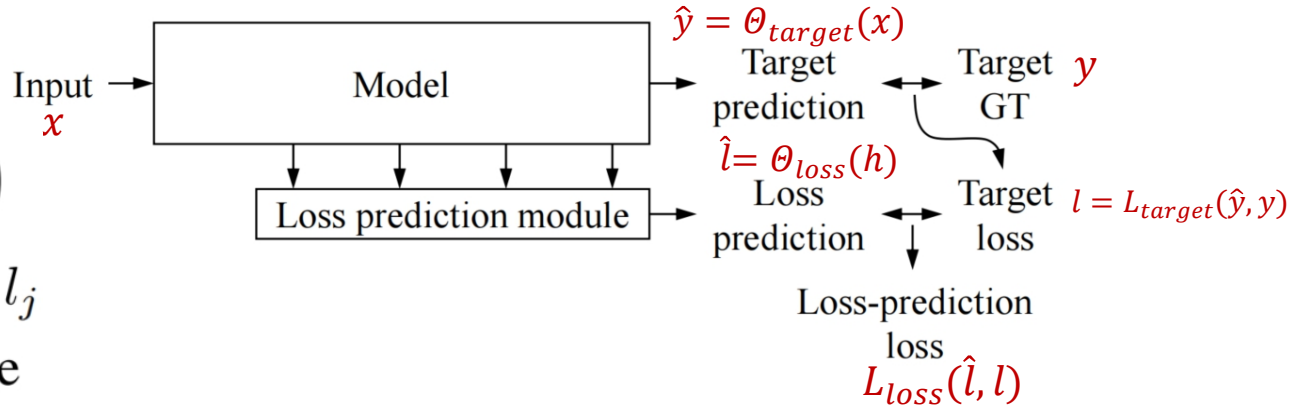
solution: pairwise

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

$\hat{l}^p = (\hat{l}_i, \hat{l}_j)$

$l^p = (l_i, l_j)$

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



Given a mini-batch $\mathcal{B}^s$ in the active learning stage s, the final loss function:

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

$$\hat{y} = \Theta_{\text{target}}(x)$$

$$\text{s.t.} \quad \hat{l}^p = \Theta_{\text{loss}}(h^p)$$

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

Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

Conclusion

- Introduced a novel active learning method that is
  - Simple
  - Task-agnostic
  - Works well with current deep networks
  - Computationally efficient with large networks
- Verified with
  - Three major visual recognition tasks
  - Three popular network architectures
- Limitations:
  - The diversity or density of data was not considered
  - The loss prediction accuracy was relatively low in complex tasks

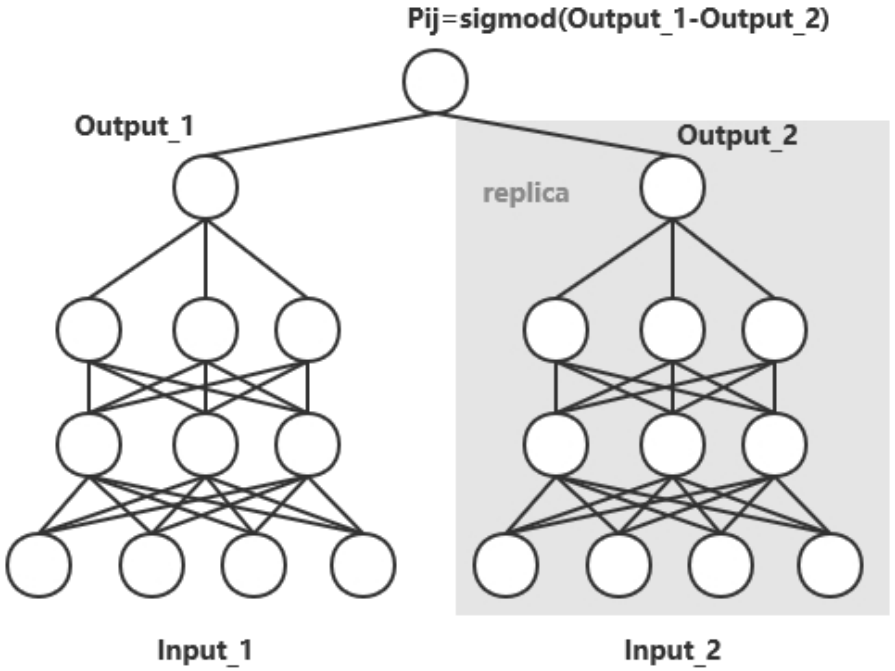
Yoo D, Kweon I S. *Learning loss for active learning*[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 93-102.

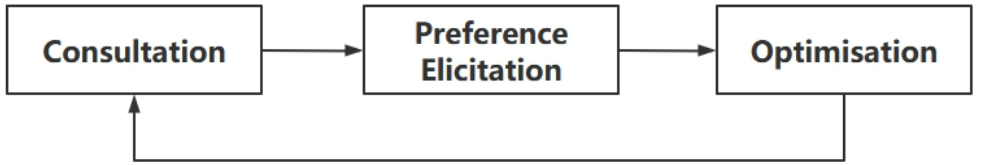
Inspiration

Algorithm 1: Interactive EMO via LTR

Input: maximum number of generations N, the number of generations between two consecutive consultation sessions τ, the number of pairwise comparisons in one consultation session μ;
Output: population of preferred solutions P;

```
initialize the population P;
t ← 0;
iterate the population with the plain EMO for τ generations;
while not converged ∧ t ⩽ N do
    if the DM needs to interact then
        present μ pairs of current individuals to the DM;
        obtain the the DM's preferred solution in each pair;
        pairwise preference learning
    end
    iterate the population for τ generations leaded by the preference model towards ROI;
    t ← t + 1;
end
return population P;
```





Workflow of IEMO/PL

The End

Thanks for Watching!