Knowledge Distillation

November 12, 2020

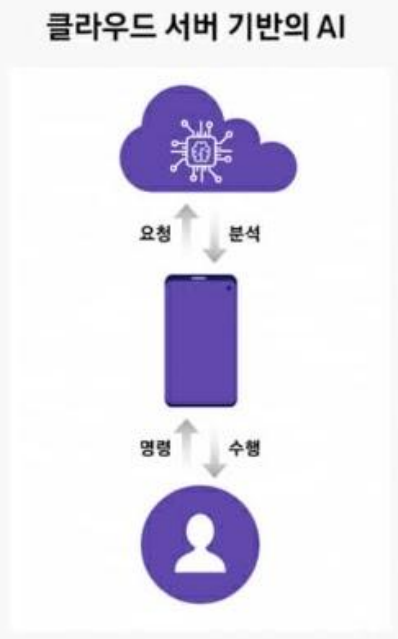
Seonyoung Kim

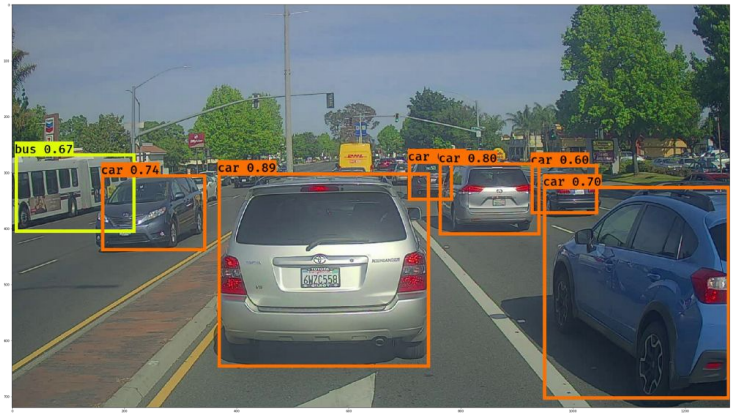
Contents

- Cloud computing
- On-device AI
- Model compression
- Knowledge distillation
- Related work
- Future work

Cloud computing

- In the past, cloud servers were used to train and run AI models
- Recently, as AI models started to be widely used, the long inference time becomes an issue for running the models
- In case of some applications like self driving cars and security robots, the long inference time is a main problem to use the trained AI models



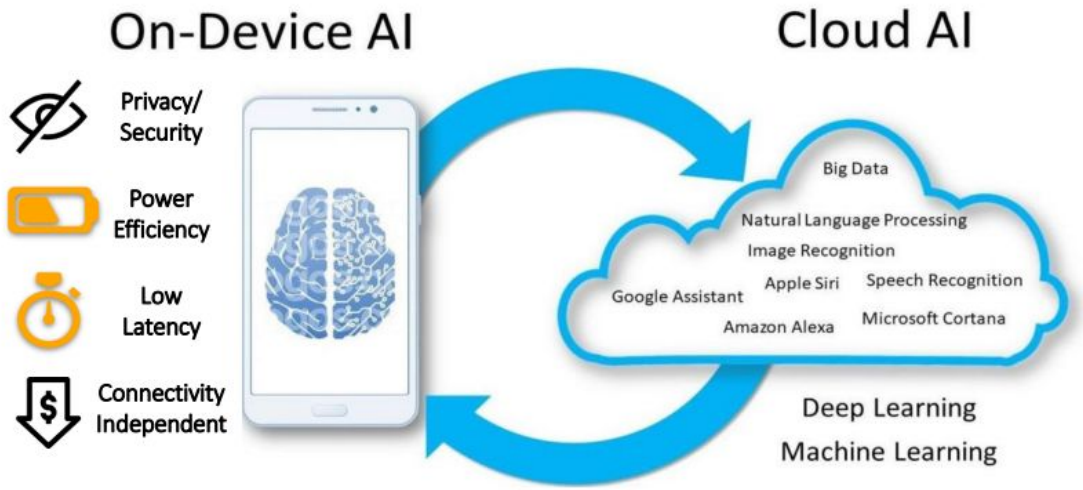


Self driving cars

Security robots

On-device AI

- On-device AI
  - The method to use AI models without cloud servers
  - However, it is inefficient to train and use the AI models on devices
- Popular solutions
  - Train the model in cloud servers
  - Inference with trained models on devices

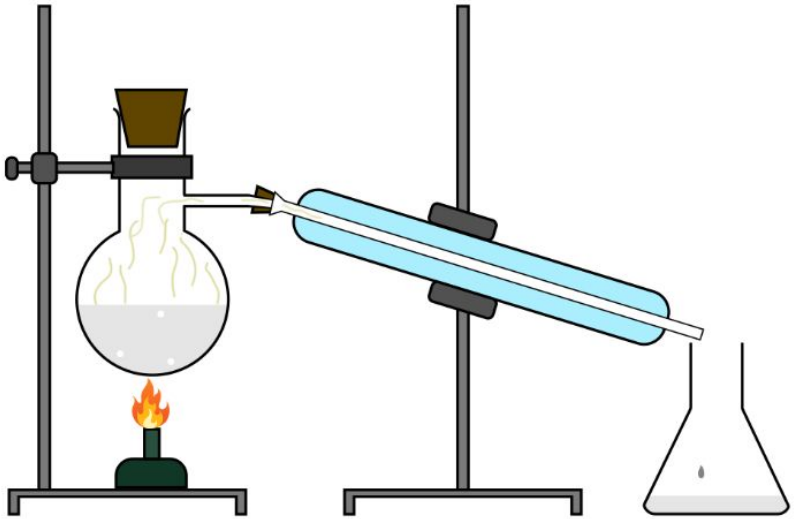


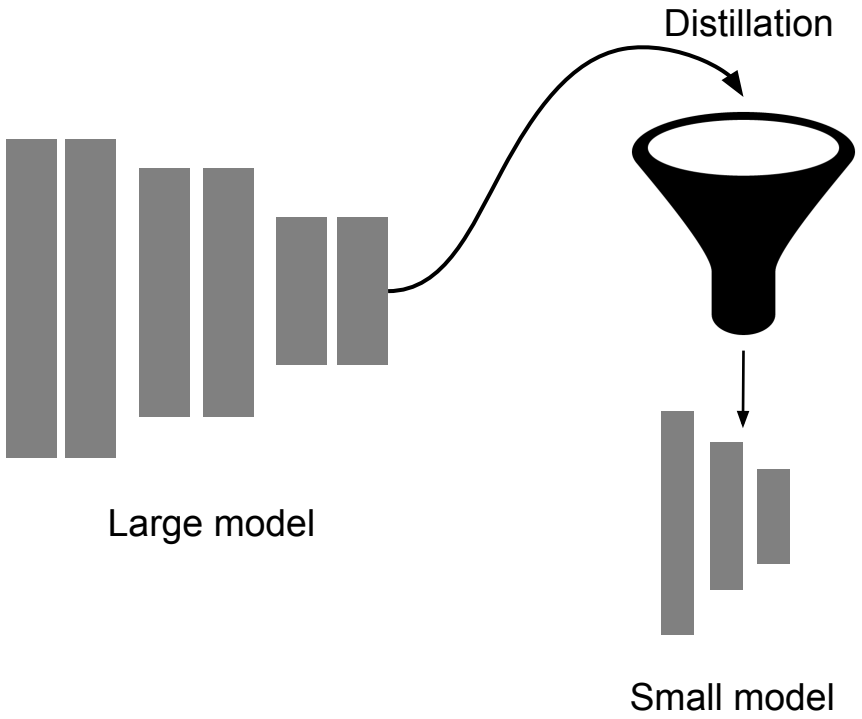
Model compression

- Need to **reduce the size of the model** for fast inferences, small storage spaces, and low battery consumption
- It is a challenging task to retain **the same accuracy after compressing the model**
- To address this challenge, some techniques have been proposed
- Model compression techniques
  - Pruning
  - Quantization
  - Low-rank approximation
  - Compact network design
  - Knowledge distillation

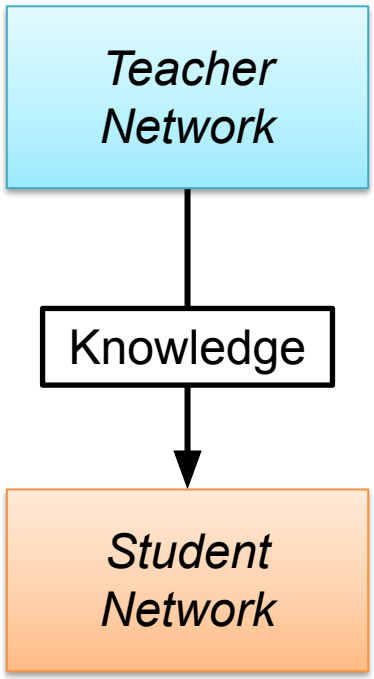
Knowledge distillation

- One of the techniques for model compression
- A method of distilling **important knowledge of a large neural network** and delivering it to a small neural network
- It retain the same or similar performance after compression





Knowledge distillation



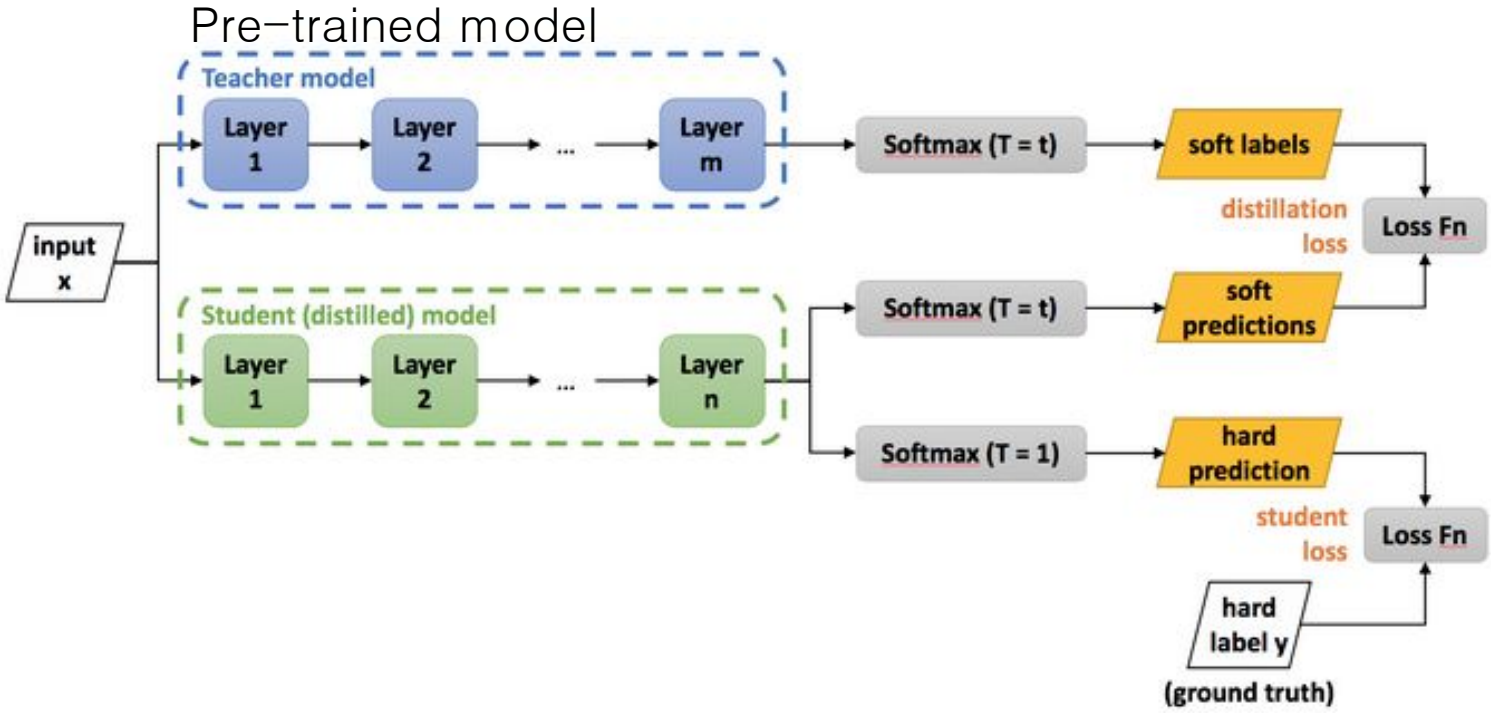
1. Teacher network
   - Cumbersome model

   e.g. ensemble / a large generalized model
   - (pros) excellent performance
   - (cons) computationally expensive
   - can not be deployed on devices with limited resources

2. Student network
   - Smaller model
   - (pros) fast inference
   - (cons) lower performance than the teacher network
   - suitable for deployment on devices

Related work (1)

- Distilling the knowledge in a neural network(Hinton Geoffrey et al., NIPS 2014)
  - The baseline knowledge distillation method
  - Knowledge is softmax outputs of the teacher networks



Related work (1)

- Distilling the knowledge in a neural network(Hinton Geoffrey et al., NIPS 2014)
  - Softmax output
    - Probability distribution as the output
    - It highlights the larger value, but loses **the relativeness with other value**
  - Soft label

$$p_i = \frac{exp(\frac{z_j}{T})}{\sum_j exp(\frac{z_j}{T})}$$

p_i : *Probability of class*
z_j : *Logits of class*
T : *Temperature hyperparameter*

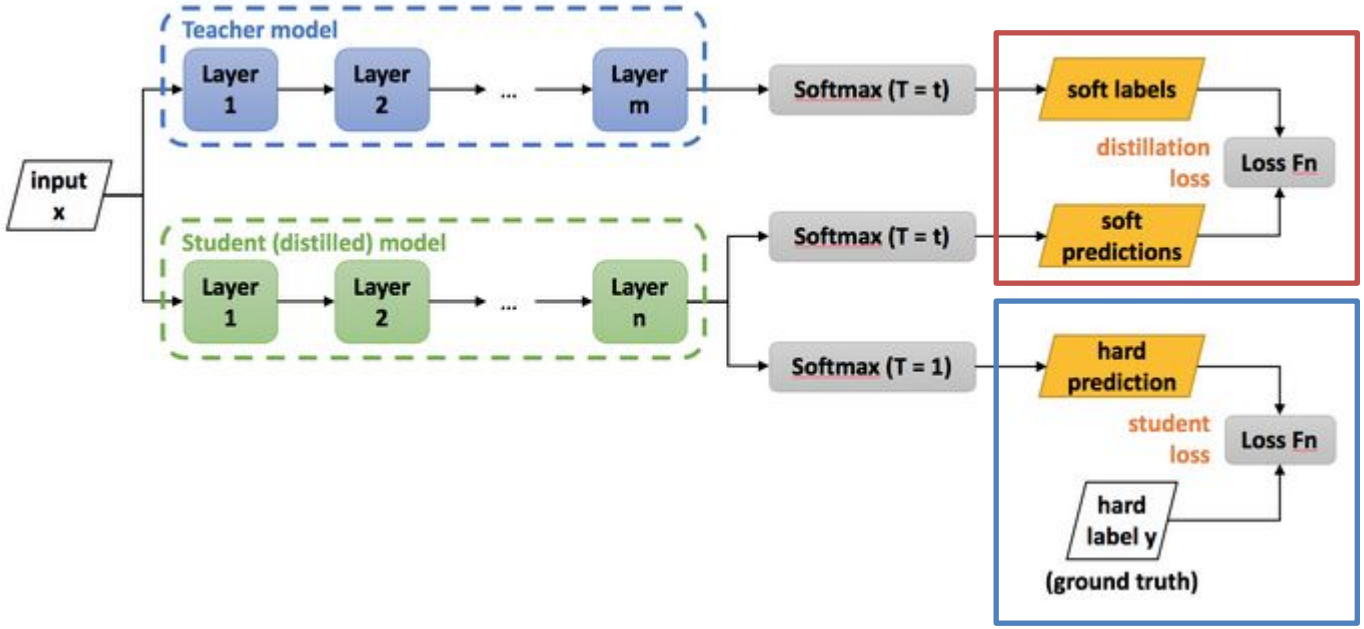
- Make the value of logits smaller before passing them to softmax
- It may be **a smoother probability distribution**
- Can get the relativeness with other value

dog

Cow	Dog	Cat	Car	
0	1	0	0	Ground Truth(i.e., hard label)
0.005	0.9	0.084	0.001	Softmax output
0.096	0.61	0.20	0.083	Soft label

Related work (1)

- Distilling the knowledge in a neural network(Hinton Geoffrey et al., NIPS 2014)



$$Total\ Loss = (1-\alpha)L_{CE}(\sigma(Z_s),\ \hat{y}) + 2\alpha T^2 L_{CE}(\sigma(\frac{Z_s}{T}),\sigma(\frac{Z_s}{T}))$$

Z_s: *Output logits of Student network*
Z_t: *Output logits of Teacher network*
$\hat{y}$: *Ground truth(one-hot)*
α: *Balancing parameter*
T: *Temperature hyperparameter*

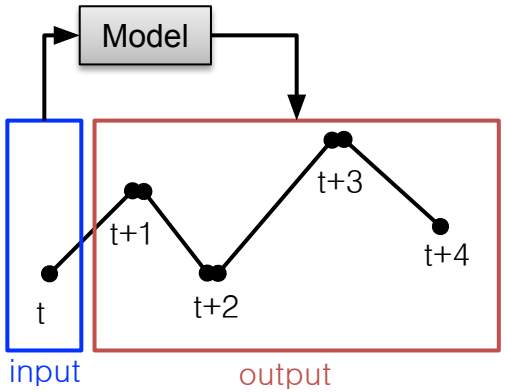
Related work (2)

- Long-term prediction of small time-series data using generalized distillation(Hayashi et al., IJCNN 2019)
  - Motivation
    - There exists area where **only limited amount of data** are available such like medical experiments, experiments with a small budget
    - Need to train the model with small data

  - They use knowledge distillation **for efficient learning, not model compression**
  - No difference in the size of the teacher network and the student network

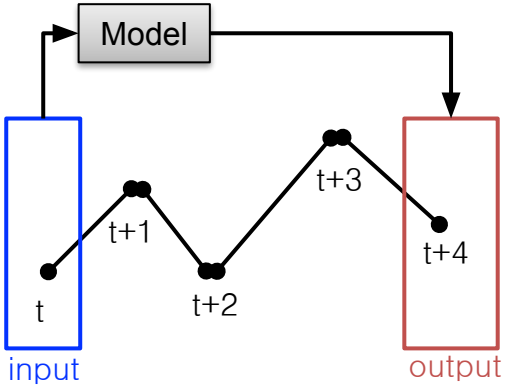
Related work (2)

- Long-term prediction of small time-series data using generalized distillation(Hayashi et al., IJCNN 2019)
  - Time-series prediction
    1. Multi-step prediction



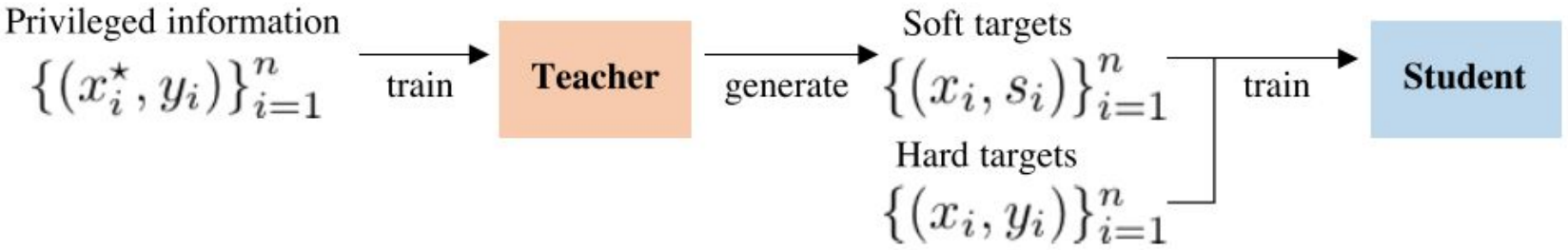
2. Long-term prediction

- Task to predict for the distant future
- Example) whether the stock price rise after 3 months?



Related work (2)

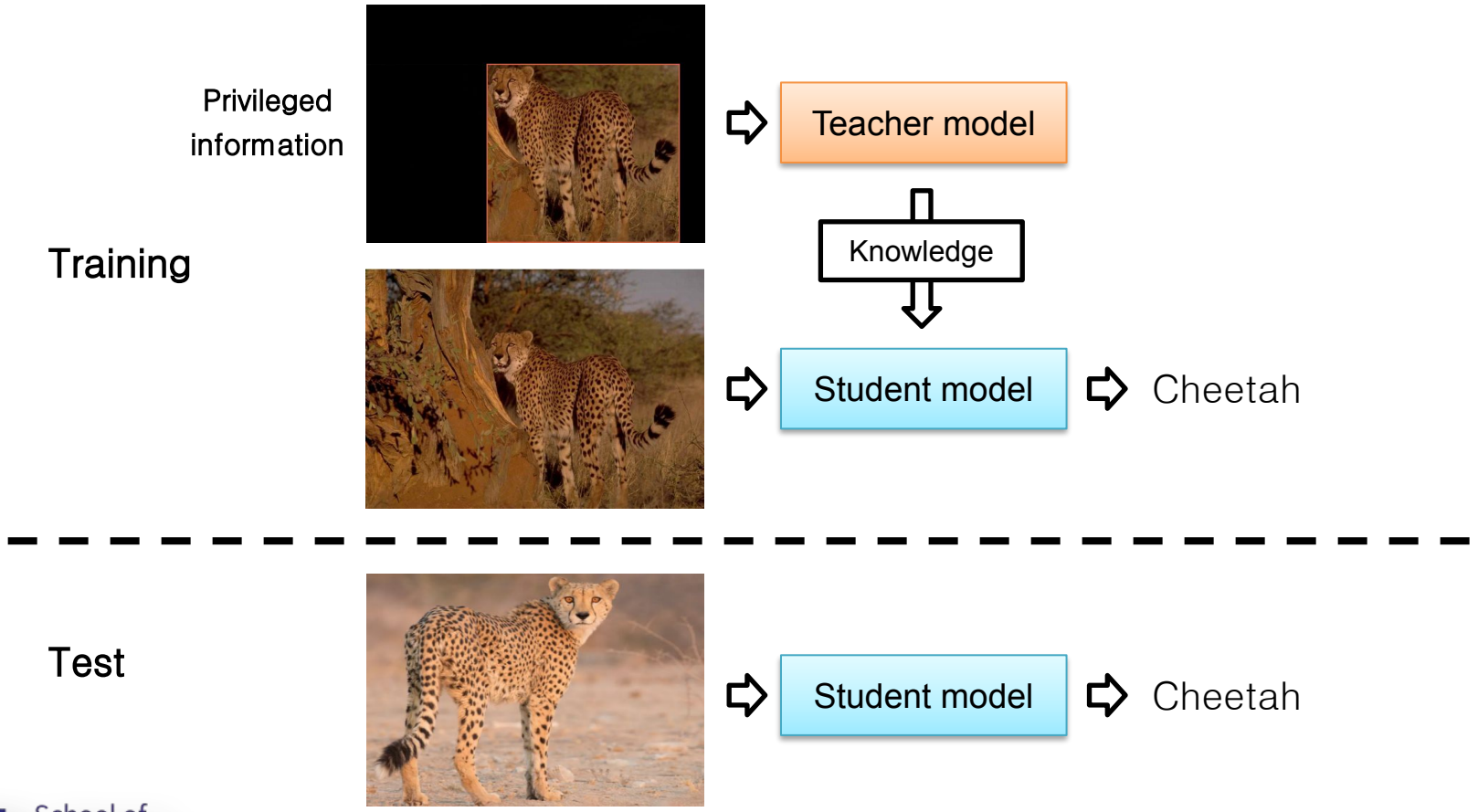
- Long-term prediction of small time-series data using generalized distillation(Hayashi et al., IJCNN 2019)
  - Privileged information
    - *Unifying distillation and privileged information(V. Vapnik et al., 2016)*
    - Information is available at training time, but not available at testing time



1. Train the teacher model using **privileged information**
2. Generate soft targets using the teacher model
3. Train the student model with a set of hard targets and a set of soft targets

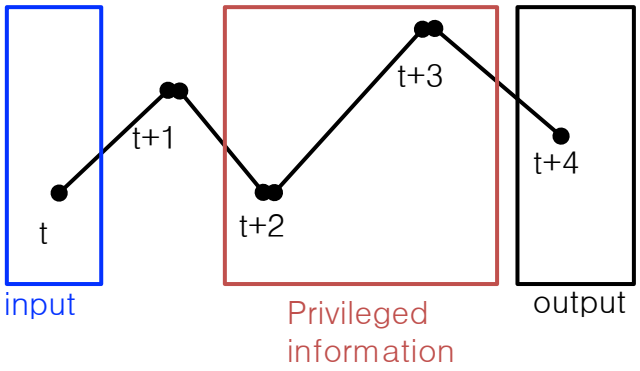
Related work (2)

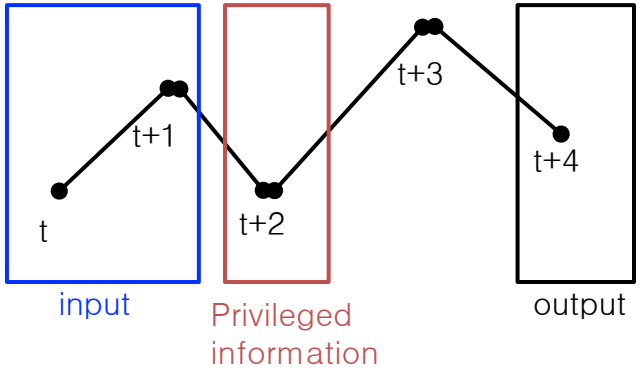
- Long-term prediction of small time-series data using generalized distillation(Hayashi et al., IJCNN 2019)
  - Example
    - Deep learning under privileged information using heteroscedastic dropout(J Lambert et at., CVPR 2018)



Related work (2)

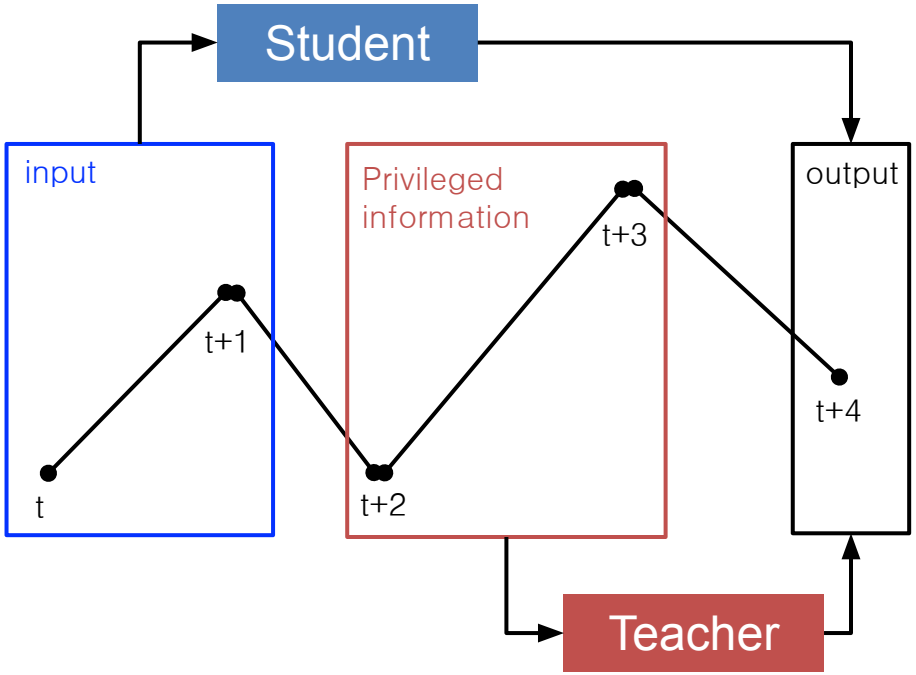
- Long-term prediction of small time-series data using generalized distillation(Hayashi et al., IJCNN 2019
  - Privileged information
    - The selection of input & privileged information is arbitrary
    - Example





Related work (2)

- Long-term prediction of small time-series data using generalized distillation(Hayashi et al., IJCNN 2019)
  - Method



1. Train the teacher network using privileged information
2. Train the student model with a set of hard targets and a set of soft targets

$$f_s = \arg\min_{f \in F_s} \frac{1}{n} \sum_{i=1}^{n} [(1-\lambda) l(y_i, \sigma(f(x_i))) + \lambda l(s_i, \sigma(f(x_i)))]$$

f_s : *student model*

λ : *imitation parameter* $\in [0, 1]$

l : *cross entropy loss function*

s_i : *soft target of Teacher model*

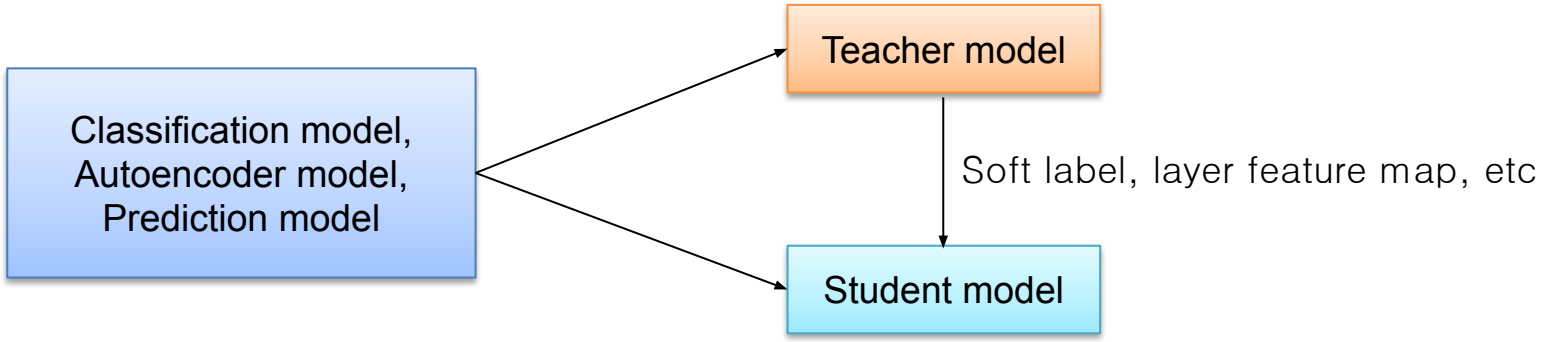
Future work

- To use knowledge distillation and privileged information for anomaly detection in time series data

- For future research, there are two directions
  - Network Compression
  - Predictive models in limited resource environment

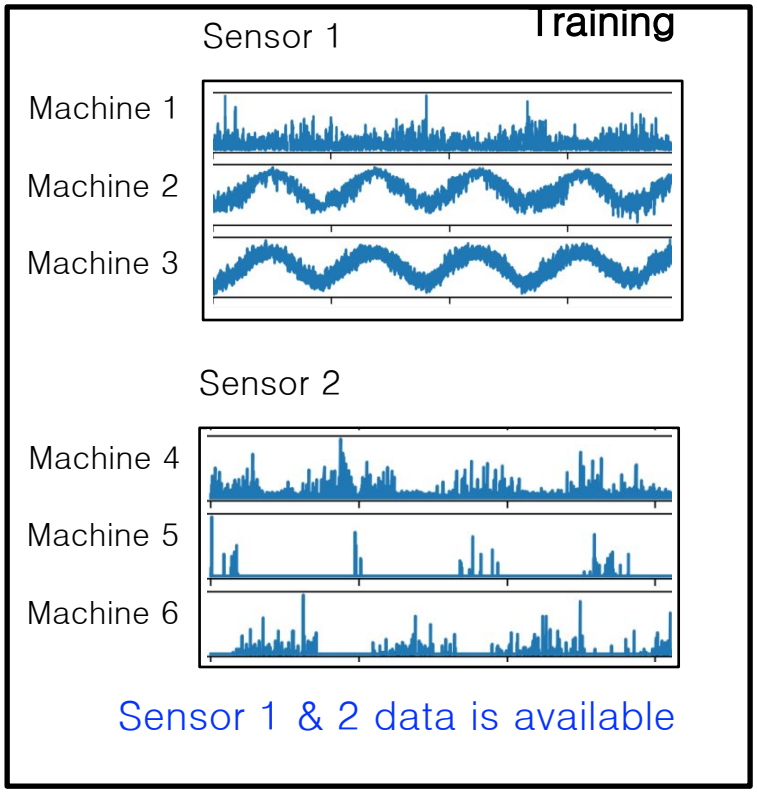
Future work (1)

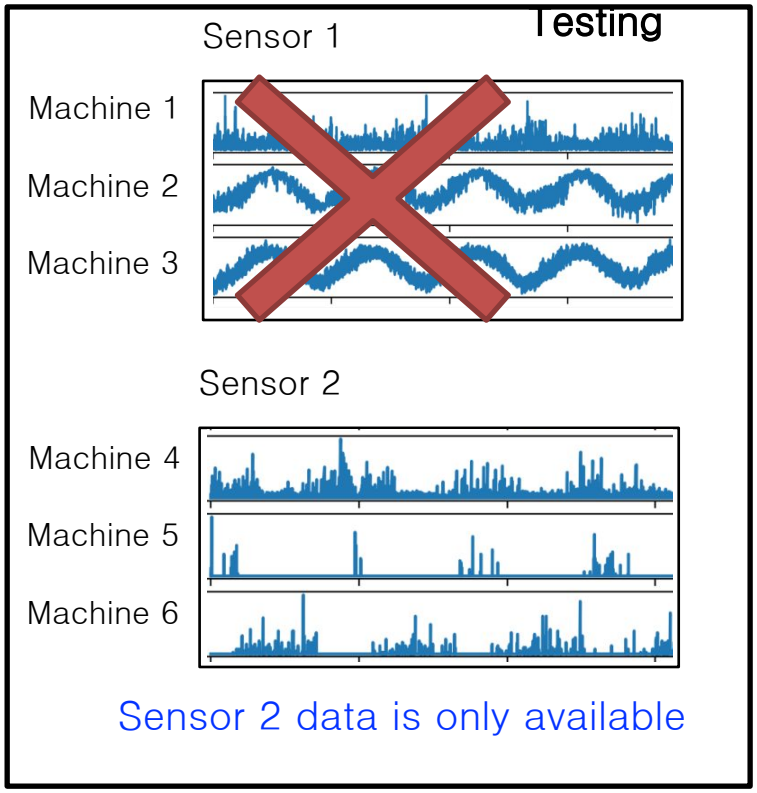
- Network compression
  - There are situations in which **the network needs to be compressed** for high speed and fast inference in time-series domain
  - Experiment with classification model, autoencoder model, and prediction model to determine which model performs well
  - Consider various knowledge such as layer feature map as well as soft label



Future work (2)

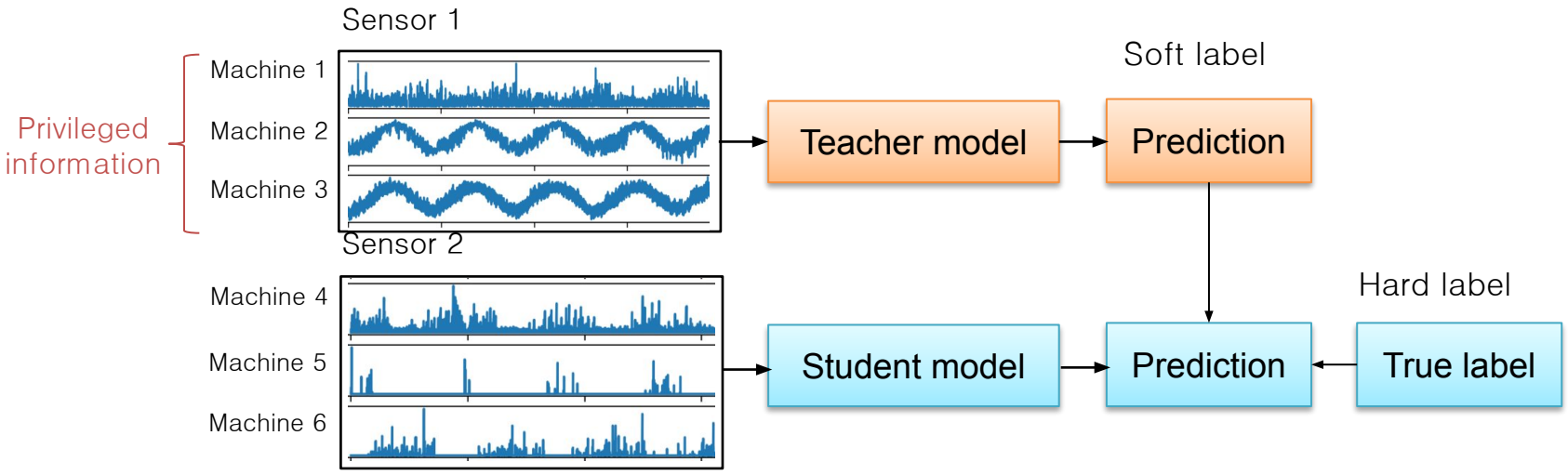
- Predictive models in limited resource environment
  - Depending on the situation, **resources may not be available**
  - Problems can arise when prediction results with only limited resources
  - Example



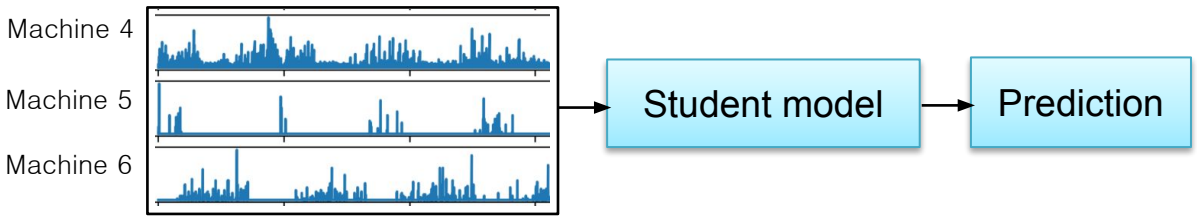


Future work (2)

- Predictive models in limited resource environment
  - Training
    - The teacher model is trained with data that can only be used in the training, and the student model is trained with data that can be used in both the training and testing



  - Testing
    - The student model detects anomaly using only data available for both train and testing



Reference

[1] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).

[2] Kim, Jangho, SeongUk Park, and Nojun Kwak. "Paraphrasing complex network: Network compression via factor transfer." *Advances in neural information processing systems*. 2018.

[3] Vapnik, Vladimir, and Akshay Vashist. "A new learning paradigm: Learning using privileged information." *Neural networks* 22.5-6 (2009): 544-557.

[4] Lee, Seung Hyun, Dae Ha Kim, and Byung Cheol Song. "Self-supervised knowledge distillation using singular value decomposition." *European Conference on Computer Vision*. Springer, Cham, 2018.

[5] Lopes, Raphael Gontijo, Stefano Fenu, and Thad Starner. "Data-free knowledge distillation for deep neural networks." *arXiv preprint arXiv:1710.07535* (2017).

[6] Hou, Yuenan, et al. "Learning lightweight lane detection cnns by self attention distillation." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[7] Meng, Fanqing, et al. "B-mode ultrasound based diagnosis of liver cancer with CEUS images as privileged information." *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018.

[8] Lambert, John, Ozan Sener, and Silvio Savarese. "Deep learning under privileged information using heteroscedastic dropout." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Appendix

- Related work (1)
  - Experiment (1)
    - MNIST dataset
      - Training dataset : 60,000
      - Test dataset : 10,000
    - Neural networks
      - Fully connected layer
      - Teacher network : 784 □ 1200 □ 1200 □ 10
      - Student network : 784 □ 800 □ 800 □ 10

Appendix

- Related work (1)
  - Experiment result

Model	# Parameters	Compression ratio	Test error
Teacher model	2,395,210	1.0	67
Student model without Distillation	1,276,810	0.533	146
Student model with Distillation	1,276,810	0.533	74

Appendix

- Related work (1)
  - Experiment (2)
    - When training student network, remove all images representing 3
    - Student network only receive information of 3 through teacher network's soft label
  - Result
    - Student network only makes 109 errors of which 14 are on 3s

Appendix

- Related work (2)
  - Experiment
    - Dataset
      - Mackey-Glass Data
      - PM2.5a-Data
      - PM2.5b-Data

	Input length(student)	Input dim(student)	Input length(teacher)	Input dim(teacher)	Prediction time k	Training length	Validation length	Test length
Mackey-Glass	4	4	2	2	8	100	100	10000
PM2.5-a	3	33	2	22	15	110	100	184
PM2.5-b	7	77	2	22	15	110	100	184

Appendix

- Related work (2)
  - Experiment
    - Model
      - Logistic regression model
  - Result
    - Accuracy on the test data

Method	Baseline (Logistic regression)	Proposed method
Mackey-Glass	0.971	0.932
PM2.5-a	0.679	0.654
PM2.5-b	0.690	0.716