Knowledge Distillation in Time-series (1)

October 8, 2020

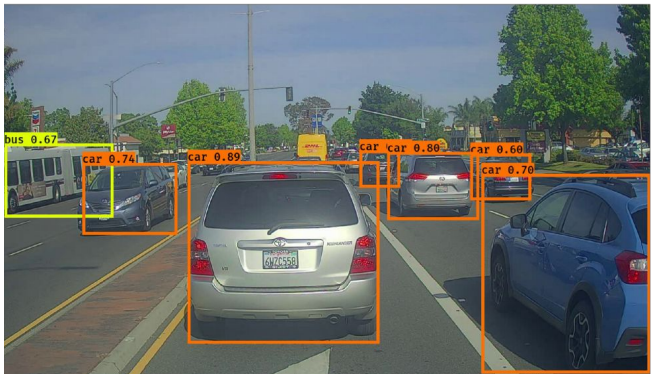
Seonyoung Kim

Model compression

Cloud computing

In the past, Cloud servers were used to run AI.

However, speed, privacy, and cost problems occurred, which is very important to some Applications.



Self driving cars

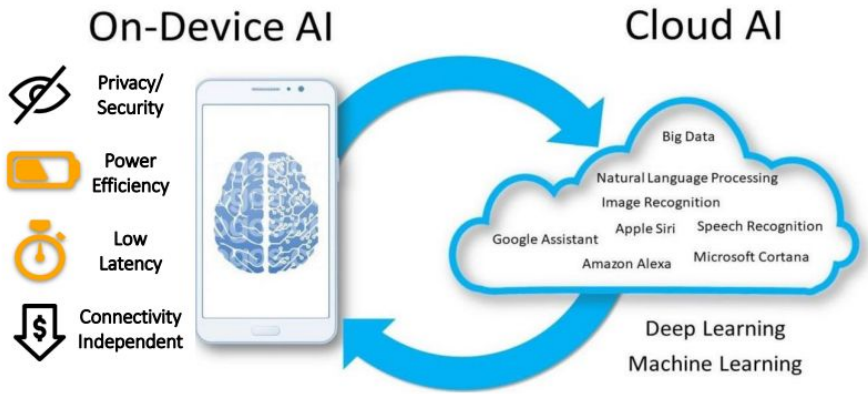
Security robots

On-device AI

AI without cloud server.

However, it is impossible to use only on-device AI.

→ Model training is executed in cloud server &
Pre-trained model is inference to device

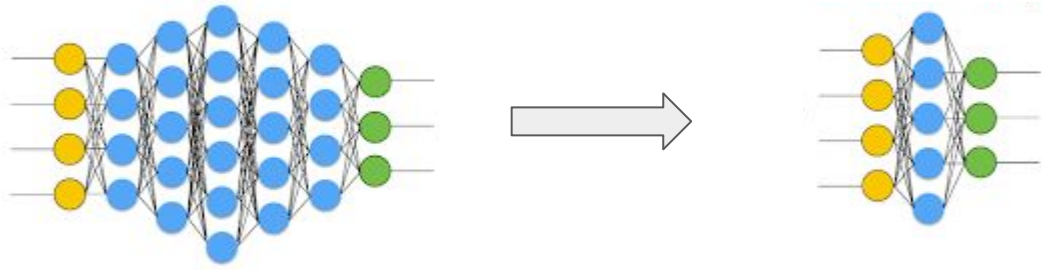


Model compression

Need to reduce the size of the model for fast speed, small storage space, and low battery consumption.

→ Model compression technique

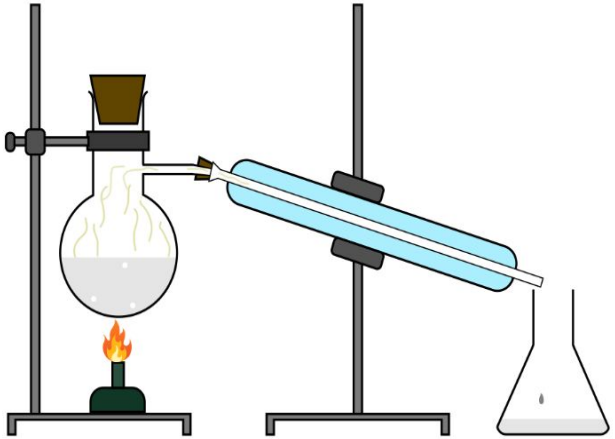
- Pruning, Quantization, Knowledge distillation, Low-rank approximation, Compact networks design



Original model on Cloud server

Compressed model on mobile device

Knowledge distillation(KD)



distillation 미국식 [dìstəléiʃən] 다른 뜻(1건)

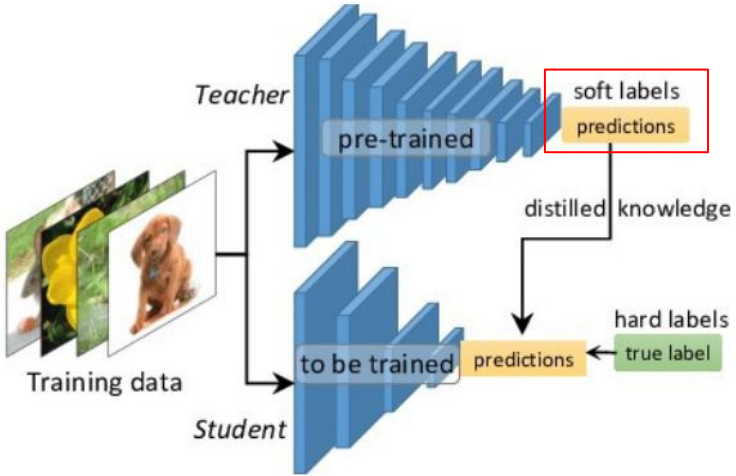
[U] 증류(법), [UC] 증류물, 정수

One of the techniques of model compression

A method of distilling important knowledge of a **large neural network** and delivering it to a **small neural network**

Knowledge distillation (KD)

Distilling the Knowledge in a Neural Network(Hinton Geoffrey, NIPS 2014 Workshop)



Large Network → Teacher Network

Small Network → Student Network

Teacher network's output(i.e., soft label) = Knowledge

cow	dog	cat	car
0	1	0	0

Hard label

cow	dog	cat	car
10^{-6}	.9	.1	10^{-9}

Soft label

By passing the output of the teacher network to the student network, the student network is trained on it.

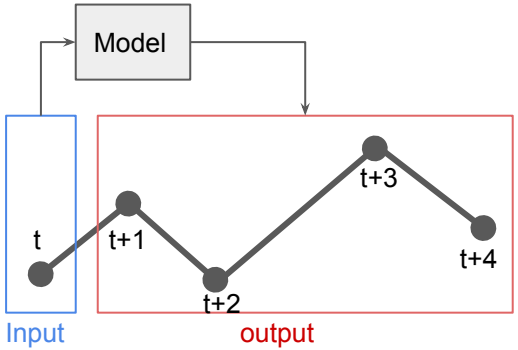
Knowledge distillation in Time-series data

Knowledge distillation in Time-series data

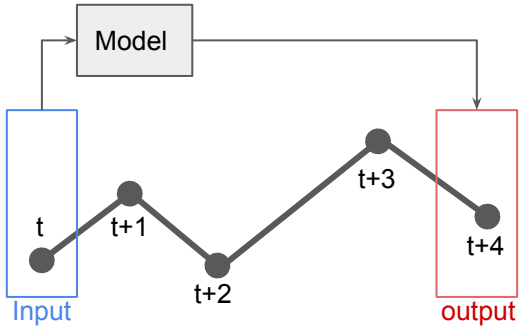
Long-Term Prediction of Small Time-Series Data Using Generalized Distillation(Hayashi et al., IJCNN 2019)

Time-series prediction

1. Multi-step prediction



2. Long-term prediction
   - Task to predict for the distant future
   - e.g., whether the stock price rise after 3 months?



Knowledge distillation in Time-series

Long-Term Prediction of Small Time-Series Data Using Generalized Distillation(Hayashi et al., IJCNN 2019)

Motivation

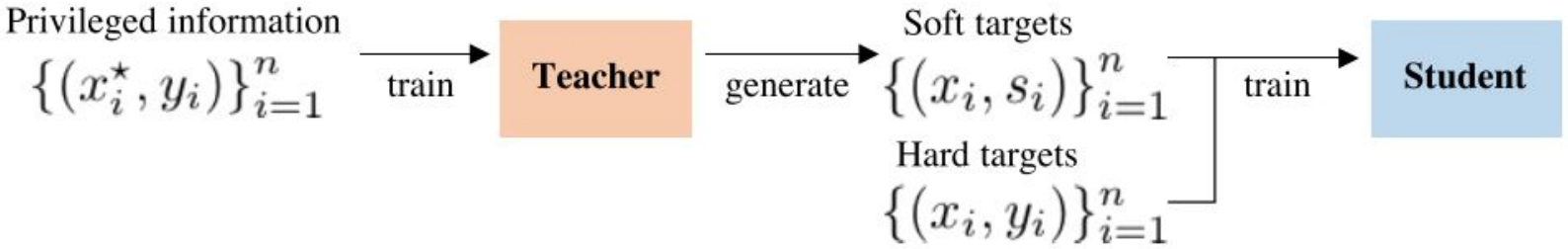
- There exist area where only limited amount of data are available(e.g., medical experiments, experiment with a small budget and etc.) → Cold-start problem
- Need to train model from small data

⇒ It use Knowledge distillation for **efficient learning, not model compression**

⇒ No difference in the size of Teacher network and Student network

Knowledge distillation in Time-series

Long-Term Prediction of Small Time-Series Data Using Generalized Distillation(Hayashi et al., IJCNN 2019)

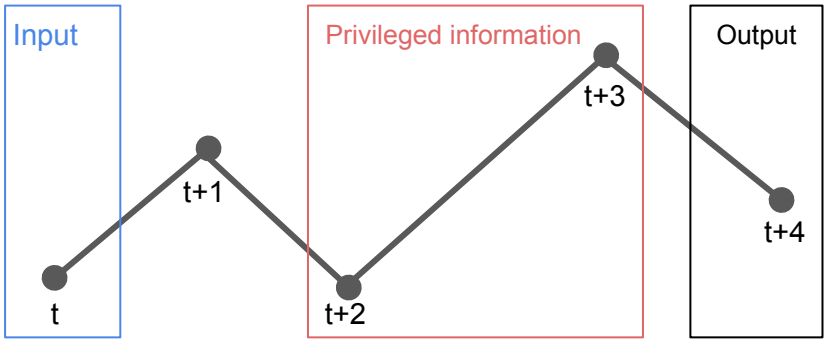


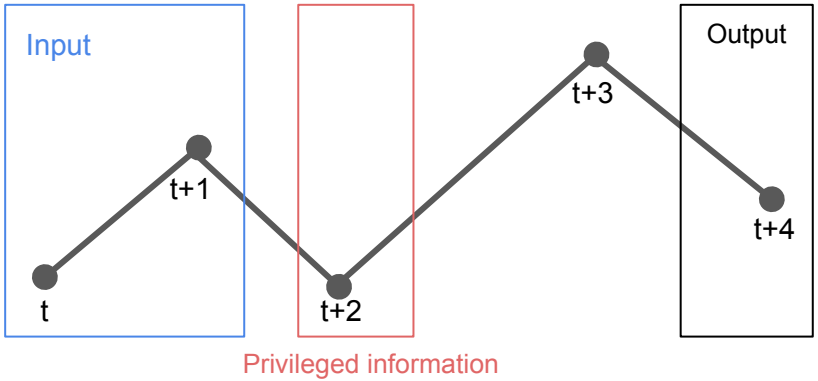
1. Train a teacher model using privileged information
2. Generate soft targets using the teacher model
3. Train a student model with a set of hard targets and a set of soft targets

Knowledge distillation in Time-series

Privileged information

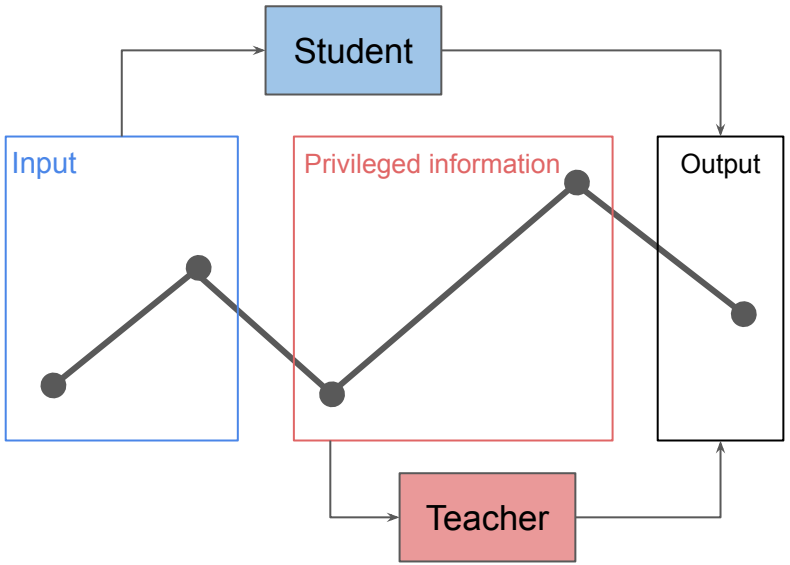
- *A new learning paradigm: Learning using privileged information(V. Vapnik et al., 2009)*
- Information is available at training time
- The selection of input & privileged information is arbitrary
- Example





Knowledge distillation in Time-series

Method



1. Train Teacher network using Privileged information
2. Train a 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*

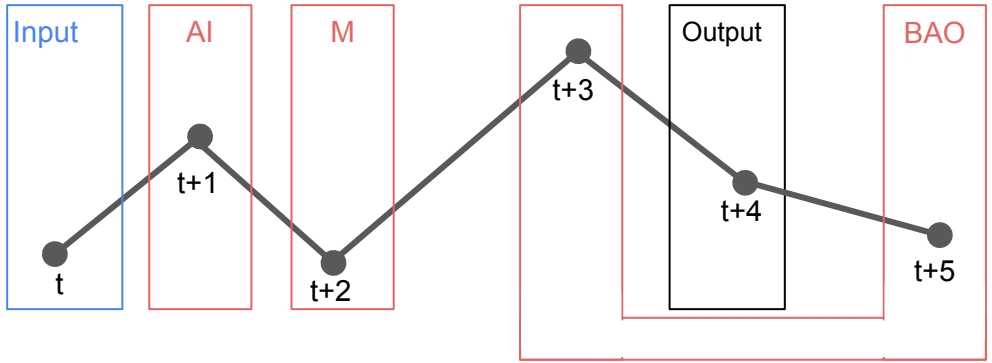
s_i : *soft target of Teacher model*

Knowledge distillation in Time-series

Experiment

- Dataset : Mackey-Glass data, Beijing PM2.5 Data
- Logistic regression model
- Binary classification

Privileged Information



1. AI(data after the input)
2. M(data at mid-time between the input and the output)
3. BAO(data before and after the output)

Knowledge distillation in Time-series

Experiment (1) - Mackey-Glass Data

- synthetic data : $$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -ax(t) + \frac{bx(t-t_0)}{1+x(t-t_0)^c}$$
$$a = 0.0, b = 0.18, c = 10, t_0 = 16, x(t_0) = 0.9$$
- task : whether the k-step-ahead value(x_{t+k}) larger or smaller than the current one(x_t)?
- y value : binary-label based on k-step-ahead value

$$y_t = I(x_t \leq x_{t+k})$$
$$I : the\ indicator\ function$$
$$k = 8$$

- Example $$x_0 = 0\ and\ x_8 = 10$$
$$then\ y_0 = 1$$

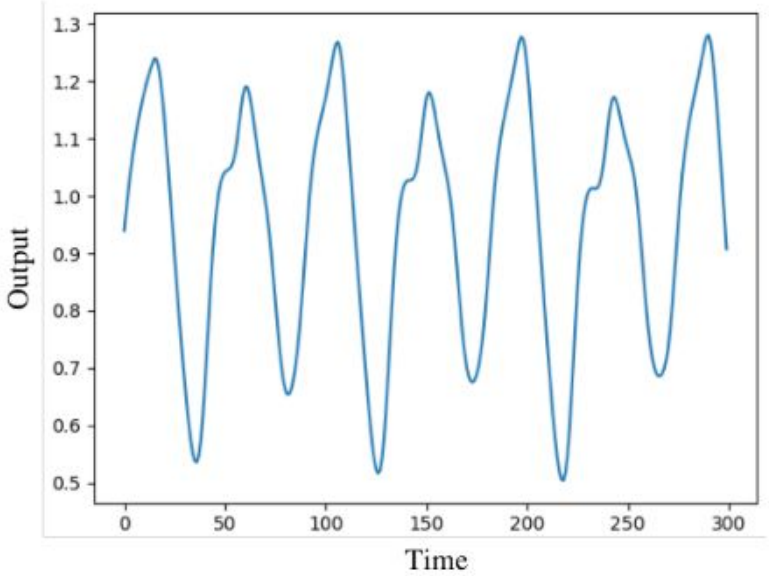
Knowledge distillation in Time-series

Experiment (2) - Beijing PM2.5 Data

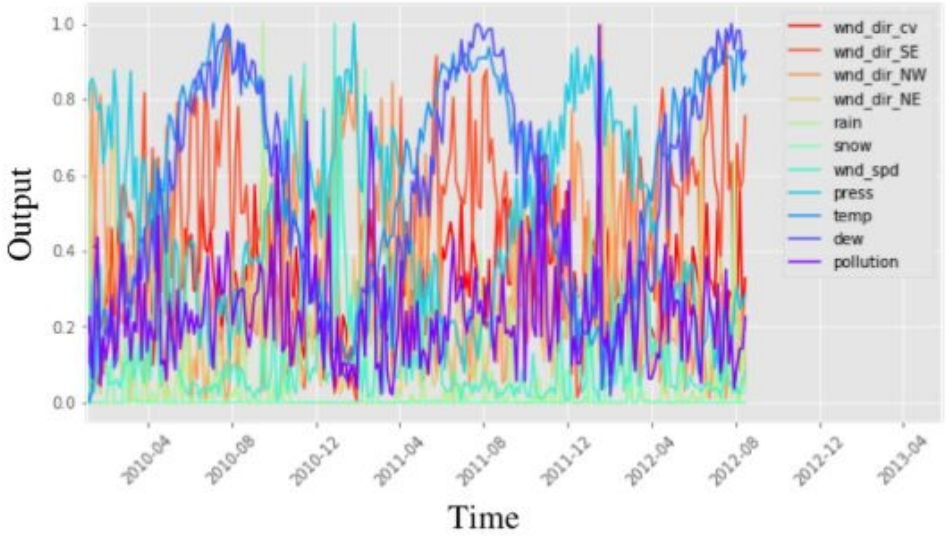
- real data & multivariate time-series data
- Eight-dimensional data containing the concentration of dust in air
- Use only one feature
- task : whether the k-step-ahead PM2.5 value larger or smaller than the current one?
- y value : binary-label based on k-step-ahead value

$$y_t = I(x_t \leq x_{t+k})$$
$$I : the\ indicator\ function$$
$$k = 15$$

Knowledge distillation in Time-series



Mackey-Glass Data



Beijing PM2.5 Data

Knowledge distillation in Time-series

Experiment Result

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

Table 1. Experimental setup

Method	Baseline (Logistic regression)	Proposed method
Mackey-Glass Data	0.971 (w/o regularizer)	0.932 (λ: 0.25, PI: BAO, T: 1.0, w/0 regularizer)
PM2.5-a Data	0.679 (L1: 1.0)	0.654 (λ: 0.5, PI: AI, T: 10.0, $L1$: 1.0)
PM2.5-b Data	0.690 (L1: 1.0)	0.716 (λ: 0.25, PI: BAO, T: 10.0, $L1$: 1.0)

Table 2. Experimental result

1. Mackey-Glass data
   - the problem might be too simple
2. PM2.5-a Data
   - the student input is too small

Knowledge distillation in Time-series

Conclusion

1. This paper proposed method about long-term prediction using privileged information

2. It showed that we should carefully select privileged information

3. It works well when data are multi-dimensional and hypothesis space is large

4. It need further experiment of more complex and difficult cases with higher-dimensional time-series data