Some slides referred from a lecture note of
CUHK (Bei Yu, CMSC 5743)

ML Coding Practice
Lecture 03-1
Knowledge Distillation

Prof. Jongwon Choi
Chung-Ang University
Fall 2022

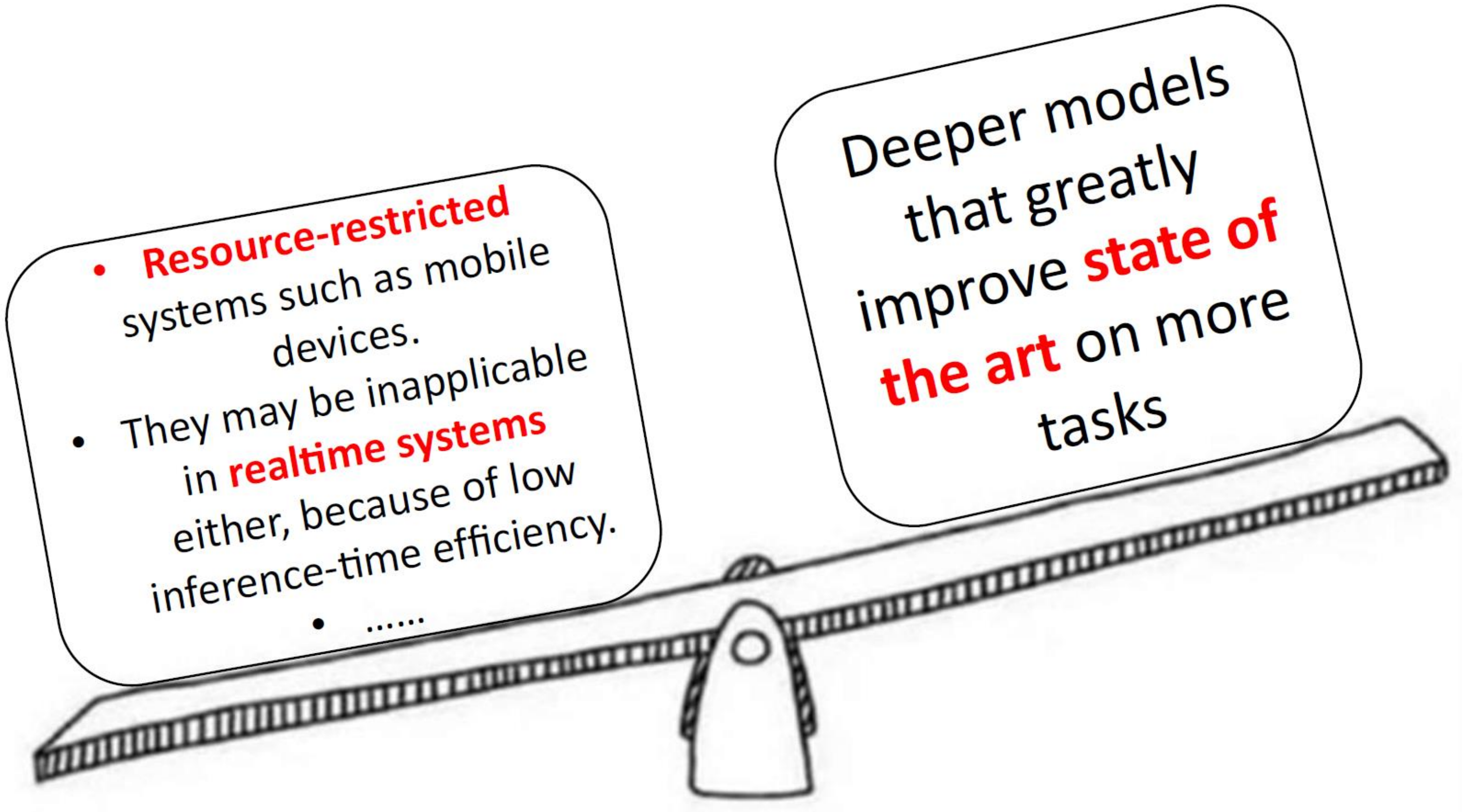
Today’s Lecture

- **What’s Knowledge Distillation?**
- **NIPSW – Knowledge Distillation**
- Knowledge Modeling
- Distillation Method
- Knowledge Distillation Scenarios

Cost

- **$\text{BERT}_{\text{large}}$**
  - Contains 24 transformer layers with 344 million parameters
  - 16 Cloud TPU | 4 days
  - 12000 dollars
- **GPT-2**
  - Contains 48 transformer layers with 1.5 billion parameters
  - 64 Cloud TPU v3 | one week
  - 43000 dollars
- **XLNet**
  - 128 Cloud TPU v3 | Two and a half days
  - 61000 dollars

Trade-off


• Resource-restricted systems such as mobile devices.
• They may be inapplicable in realtime systems either, because of low inference-time efficiency.
• ……
Deeper models that greatly improve state of the art on more tasks

Knowledge Distillation

Knowledge distillation is a process of distilling or transferring the knowledge from a (set of) large, cumbersome model(s) to a lighter, easier-to-deploy single model, without significant loss in performance.

Hot Topic

ensembles. Model ensembles are a pretty much guaranteed way to gain 2% of accuracy on anything. If you can't afford the computation at test time look into distilling your ensemble into a network using dark knowledge.

Andrej Karpathy

A Recipe for Training Neural Networks

http://karpathy.github.io/2019/04/25/recipe/

Distilling the Knowledge in a Neural Network

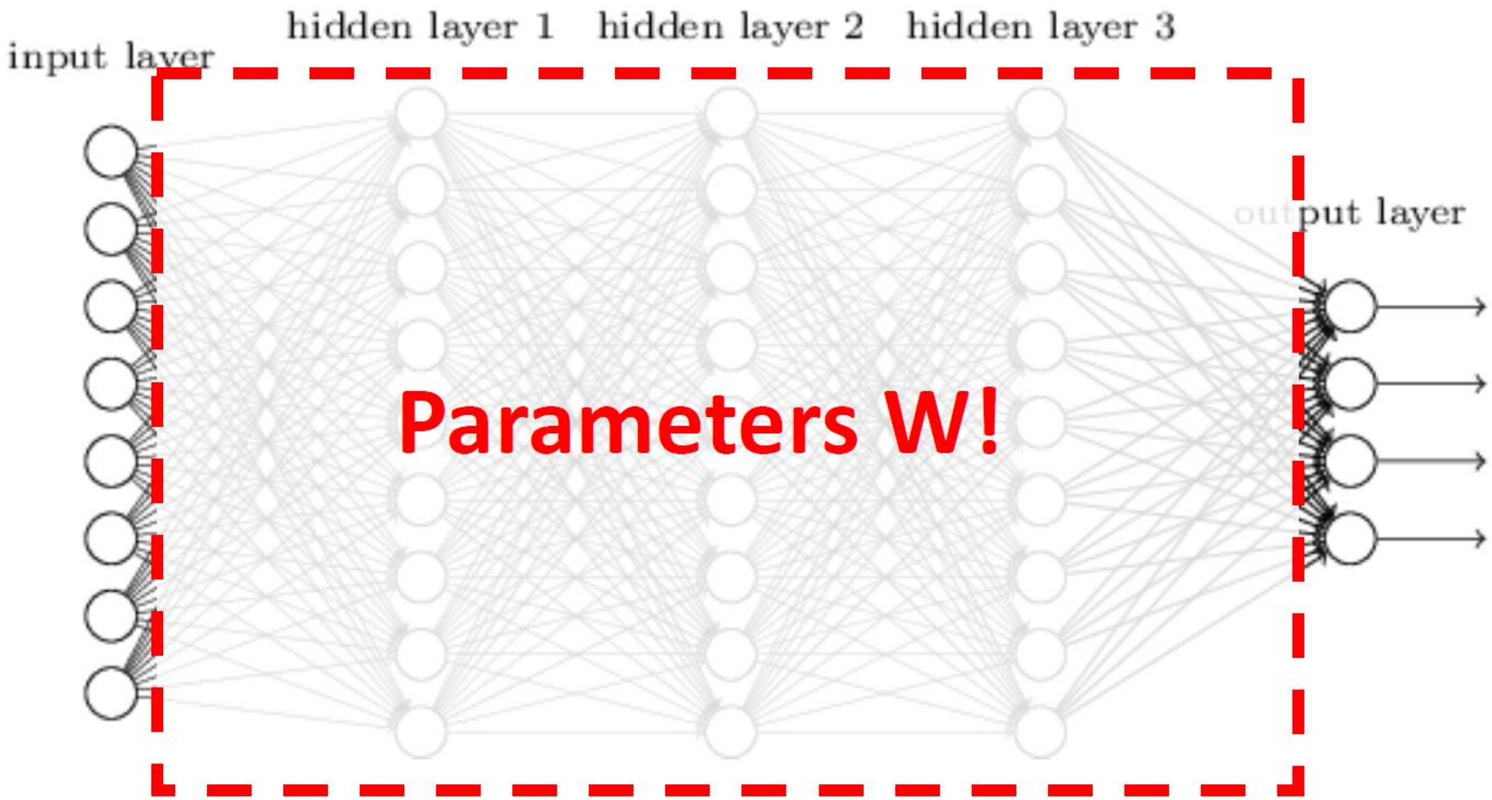
Hinton

NIPS 2014 Deep Learning Workshop

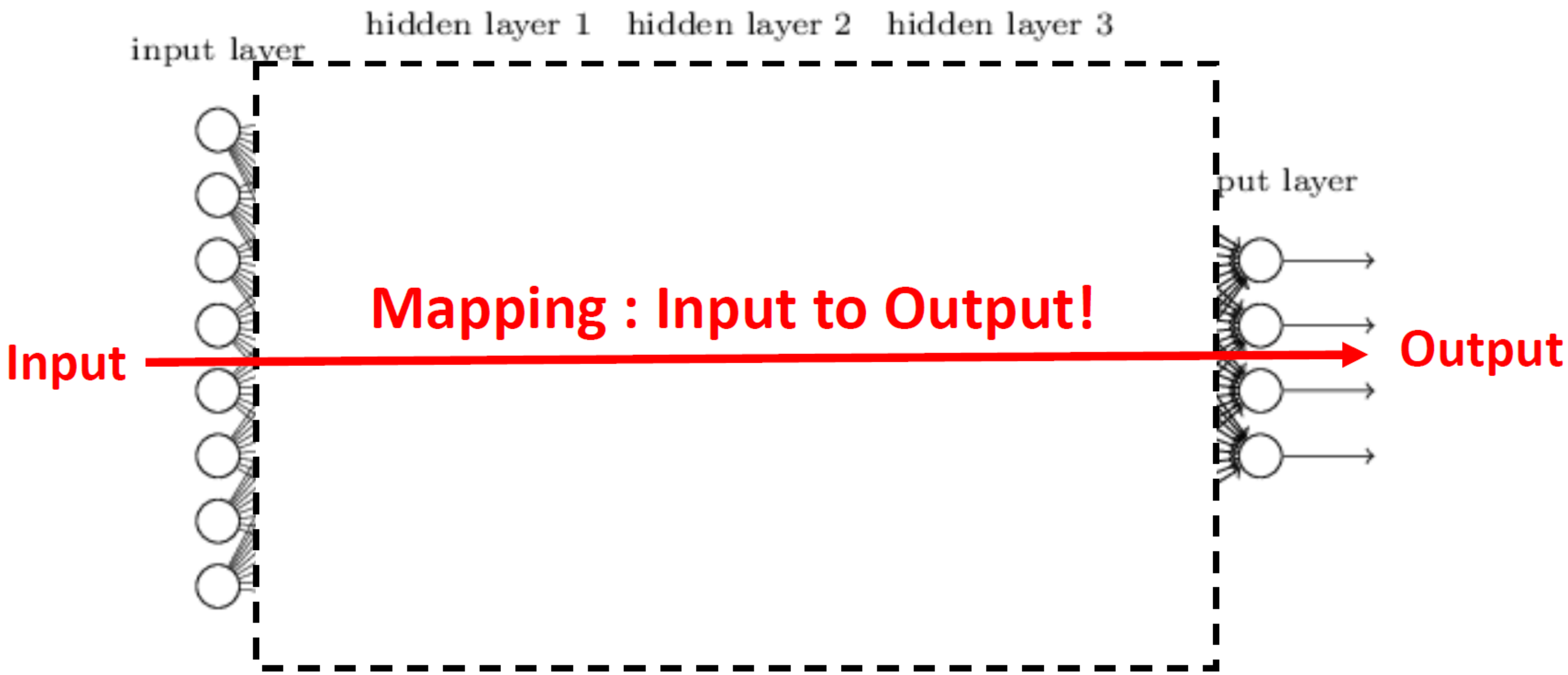
Model Compression

- **Ensemble model**
  - Cumbersome and may be too computationally expensive
- **Solution**
  - The knowledge acquired by a large ensemble of models can be transferred to a single small model.
  - We call “**distillation**” to **transfer** the knowledge from the cumbersome model to a small model that is more suitable for deployment.

What is Knowledge? - 1

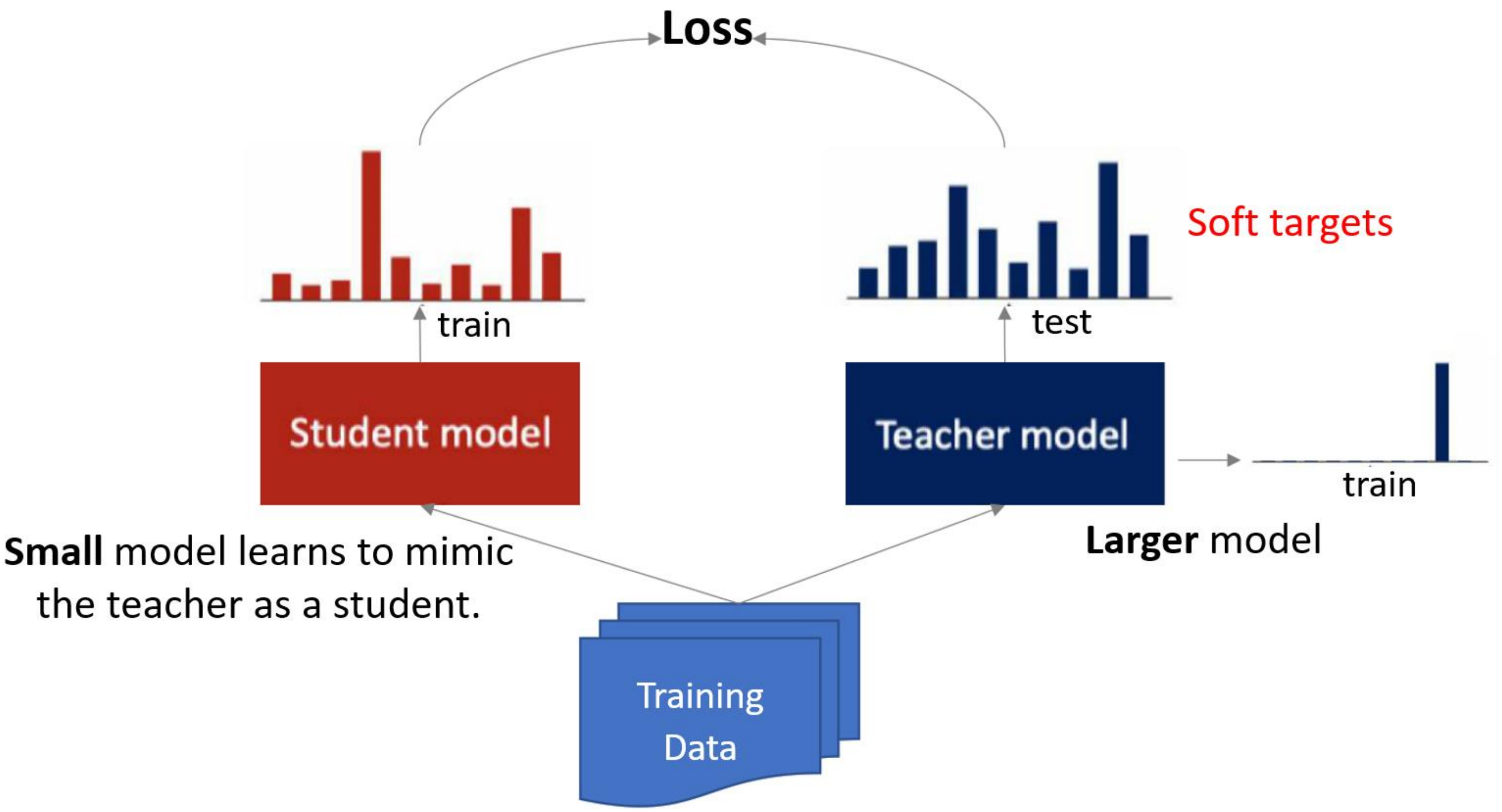


What is Knowledge? - 2



A more abstract view of the knowledge, that frees it from any **particular instantiation**, is that it is a learned mapping from input vectors to output vectors.

Knowledge Distillation

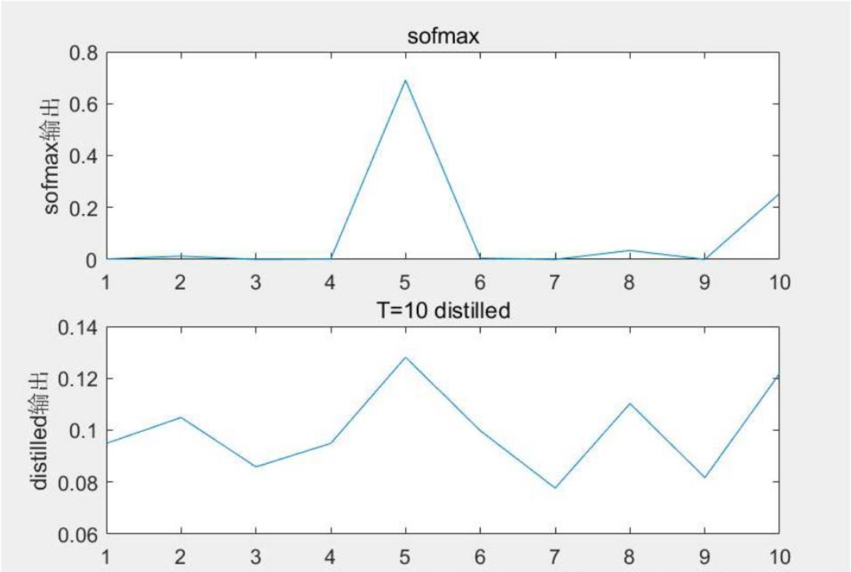


Softmax with Temperature

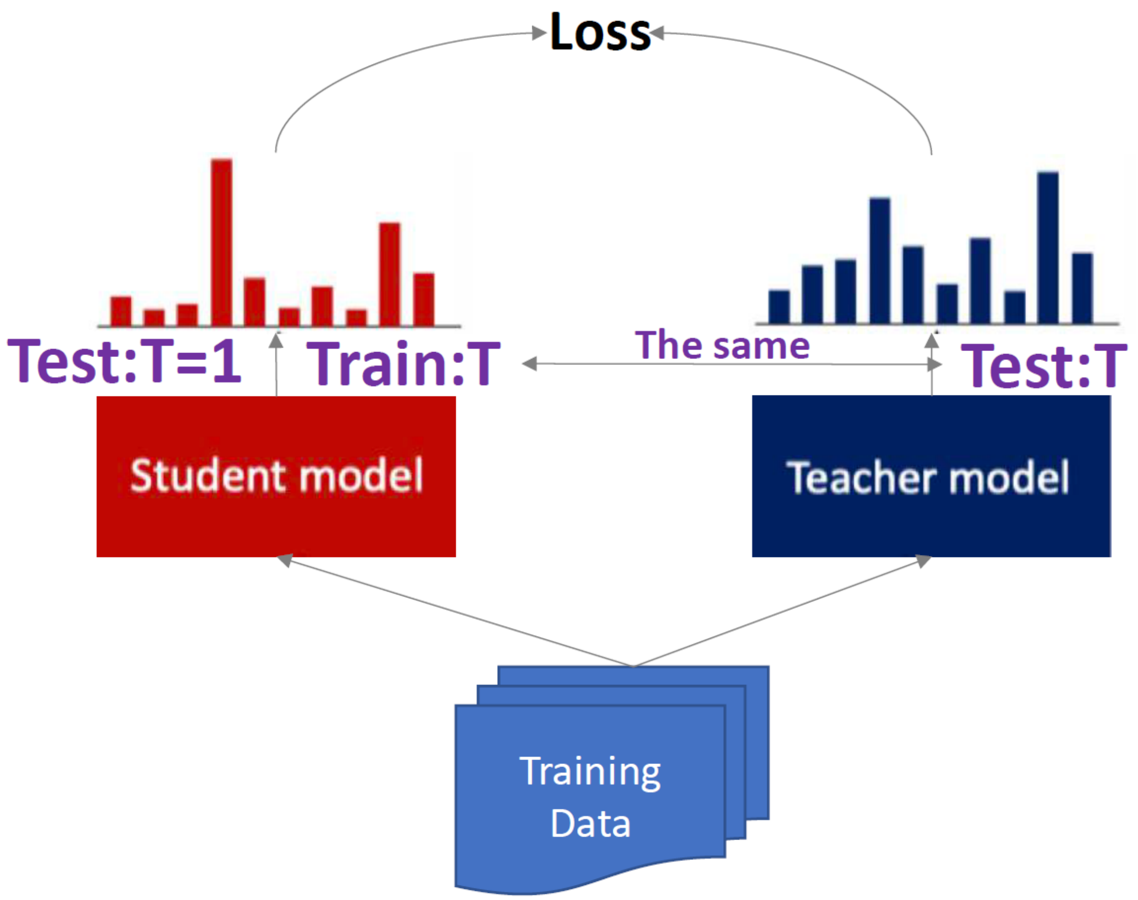
Logits

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

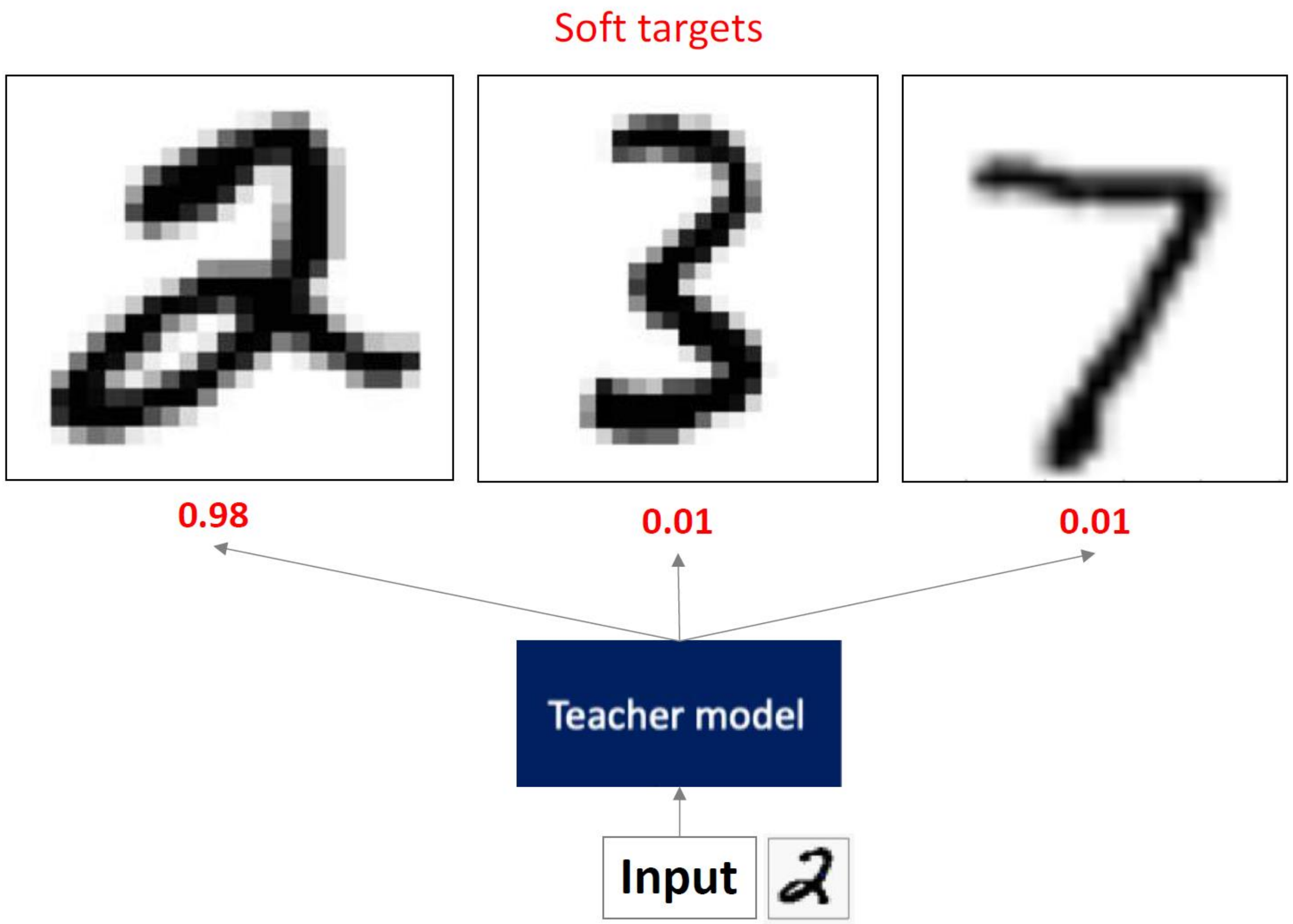
Temperature



Softmax with Temperature

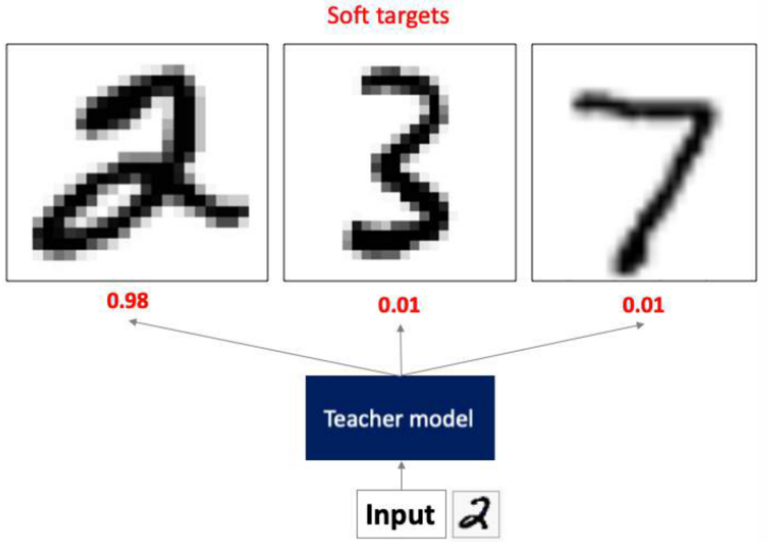


Softmax with Temperature



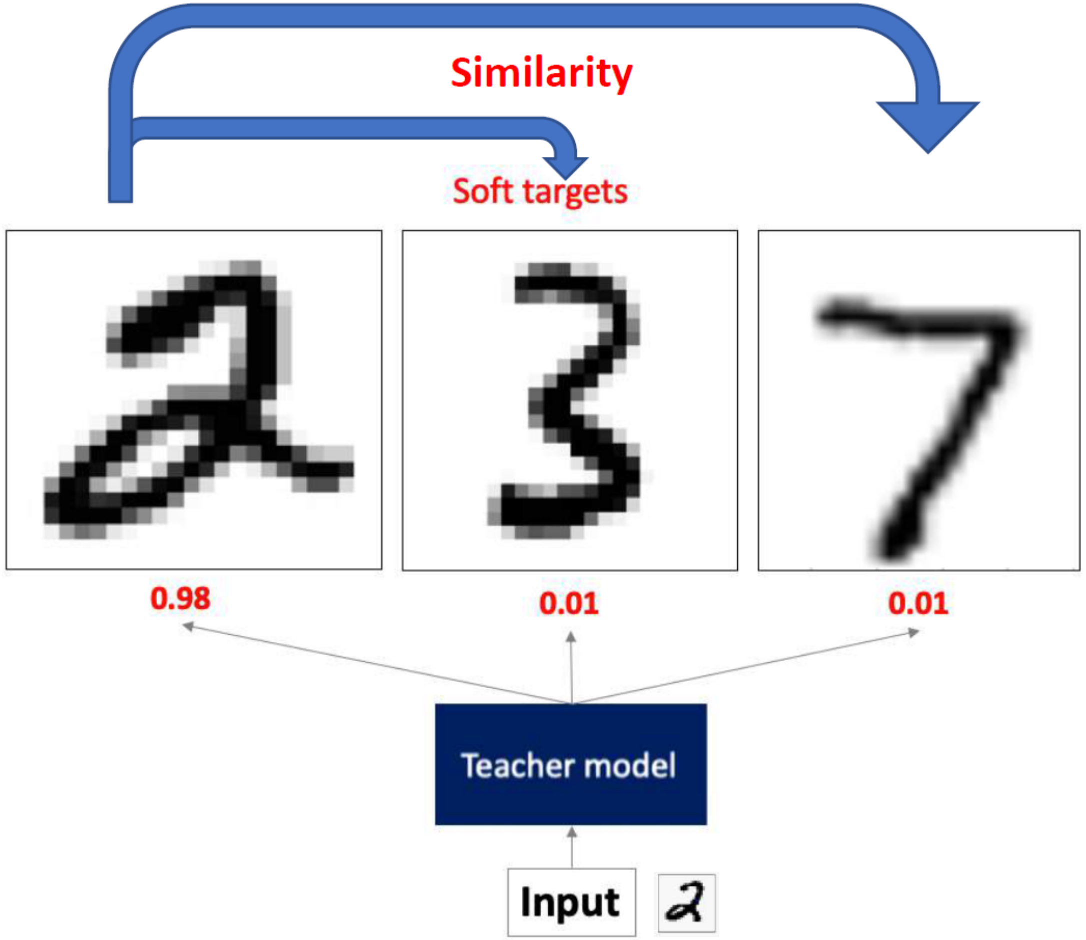
Supervisory Signals

Soft target		**One-hot**
• 2 is similar to 3 and 7	→	• 2 independent of 3 and 7.
• Contiguous distribution	→	• Discrete distribution
• **Inter-Class variance** ✓	→	• **Inter-Class variance**
• **Between-Class distance** ✓	→	• **Between-Class distance**



Soft targets have high entropy！

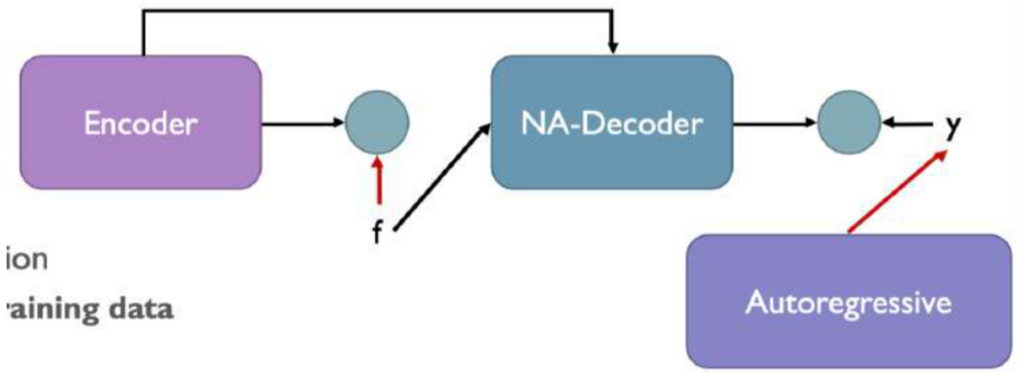
Data Augmentation



Reduce Modes

- NMT : Real translation data has many modes.

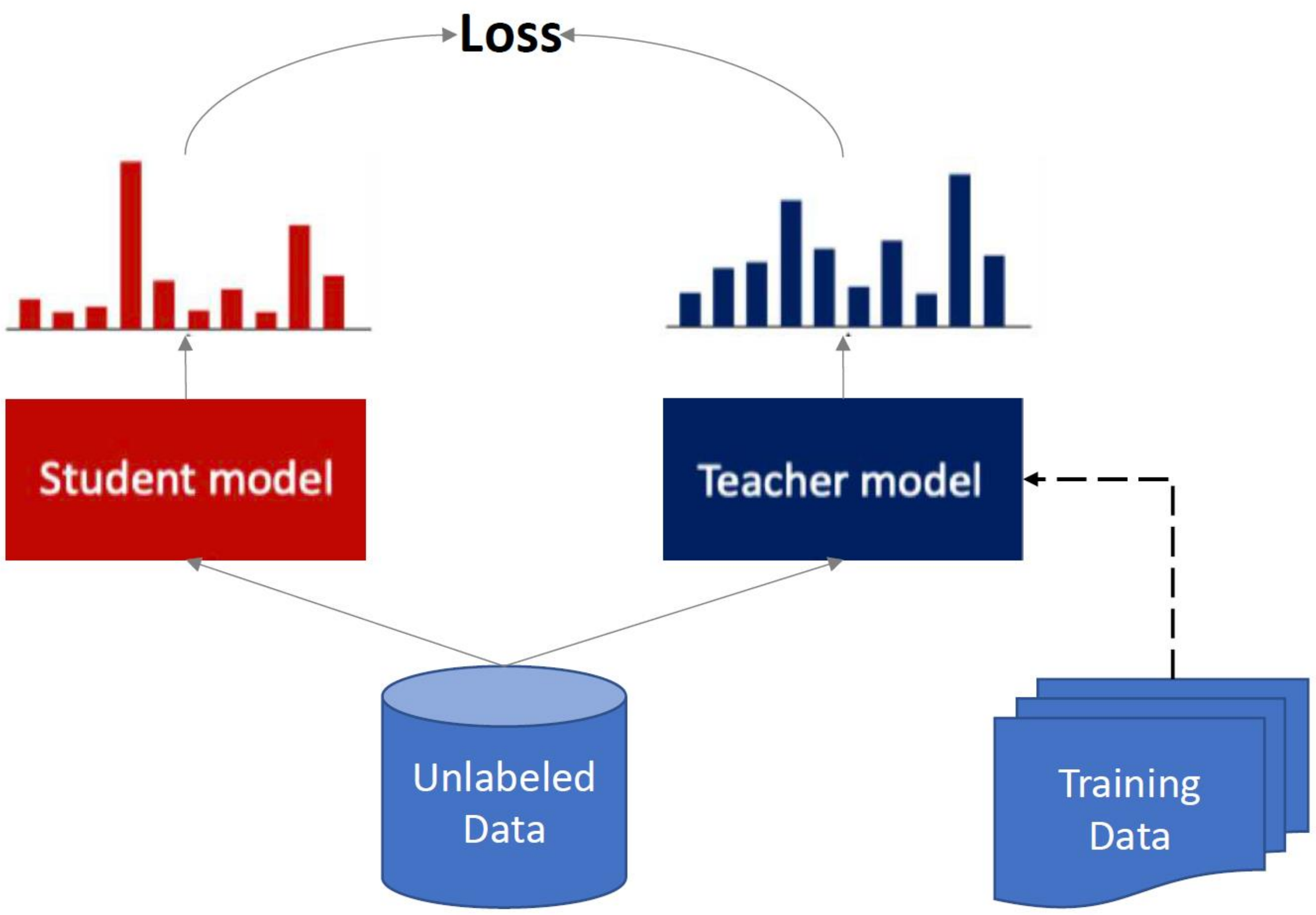
- MLE training tends to use a single-mode model to cover multiple modes.



Soft Targets

1. Supervisory signals
2. Data augmentation
3. Reduce Modes

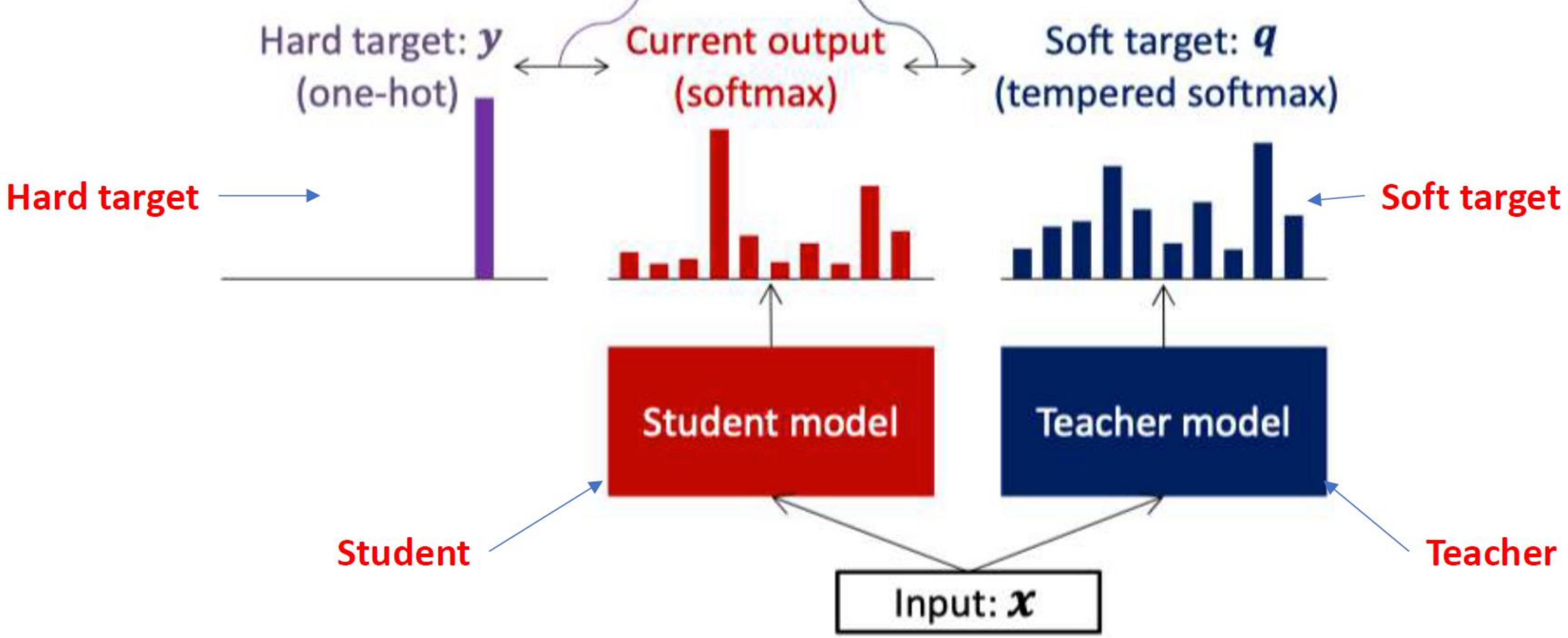
How to Use Unlabeled Data?



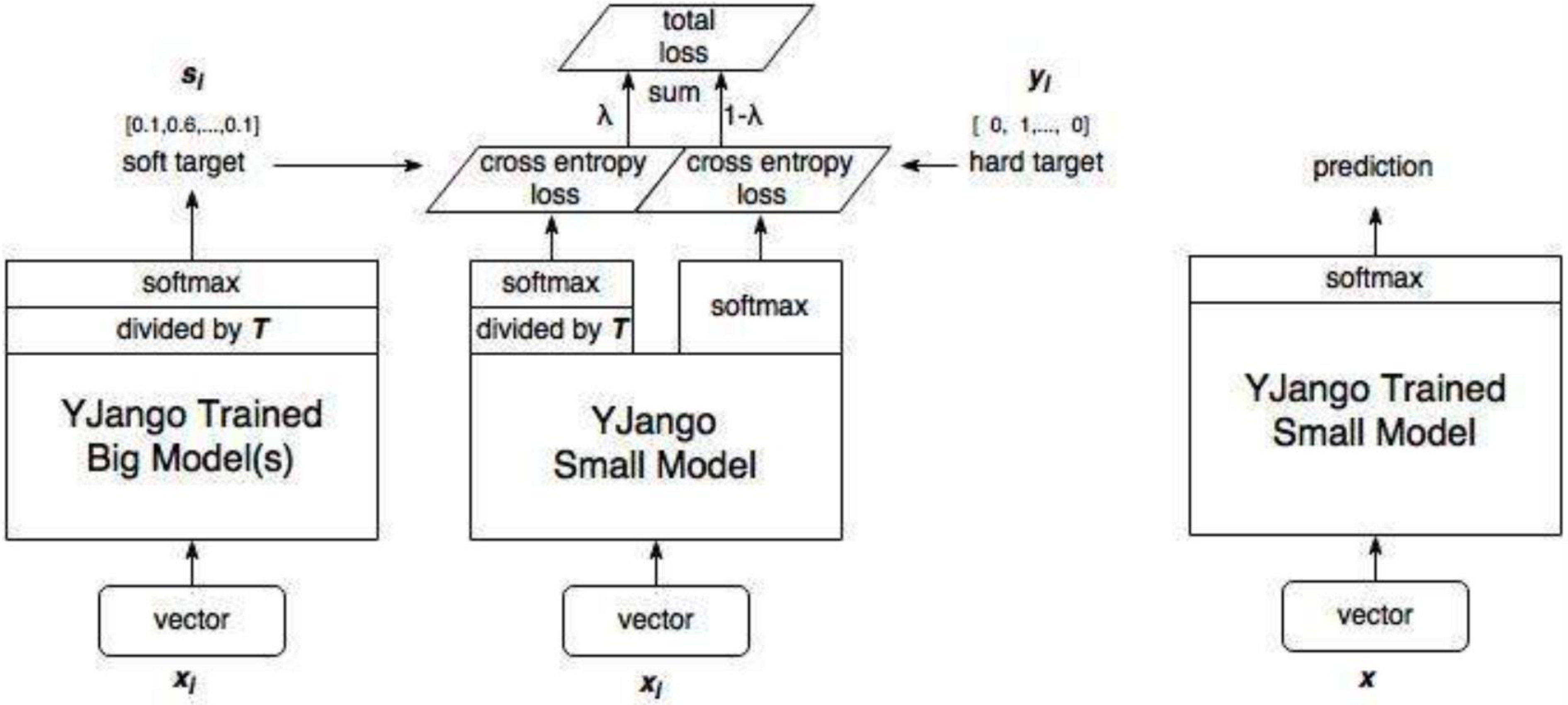
Loss Function

- Transfer set = Unlabeled data + original training set

$$L = (1-\rho)C_{hard} + \rho C_{soft}$$



Knowledge Distillation



Distilling Task-Specific Knolwedge from BERT into Simple Neural Networks

University of Waterloo

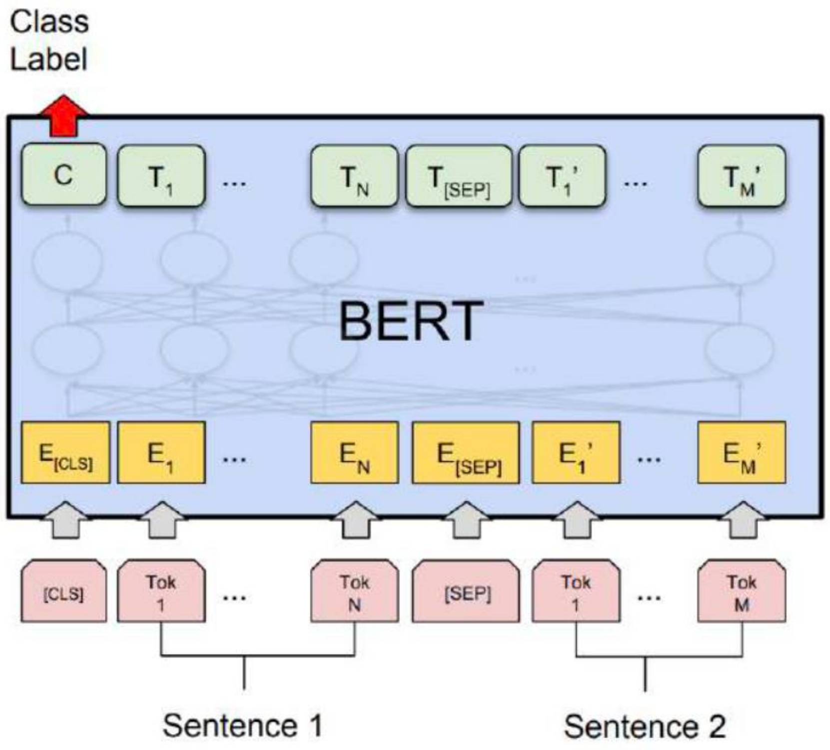
arxiv

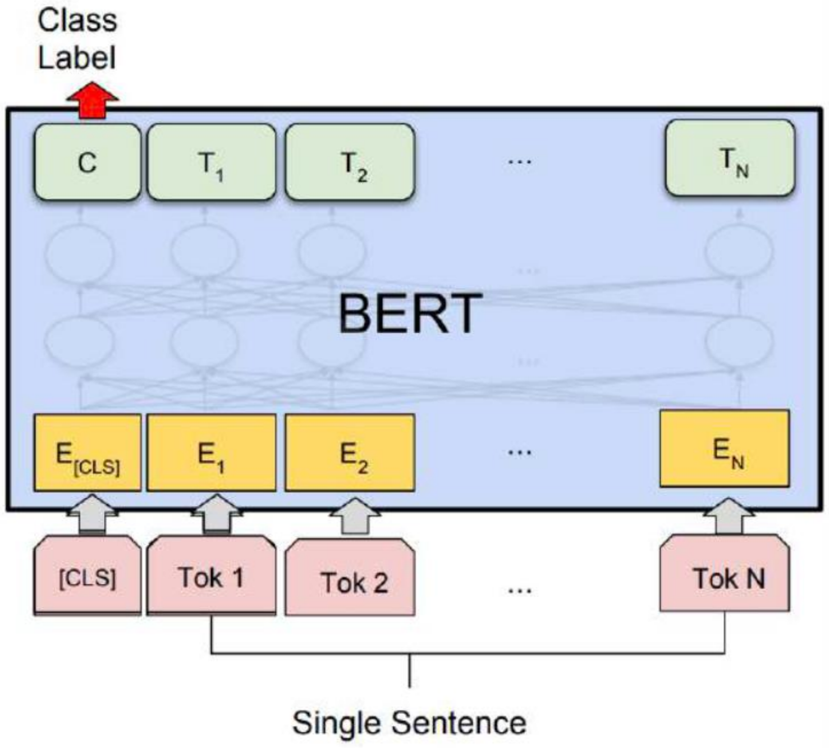
Overview

- Distill knowledge from BERT, a state-of-the-art language representation model, into a single-layer BiLSTM
- **Task**
  1. Binary sentiment classification
  2. Multi-genre Natural Language Inference
  3. Quora Question Pairs redundancy classification
- Achieve comparable results with ELMo, while using roughly 100 times fewer parameters and 15 times less inference time.

Teacher Model

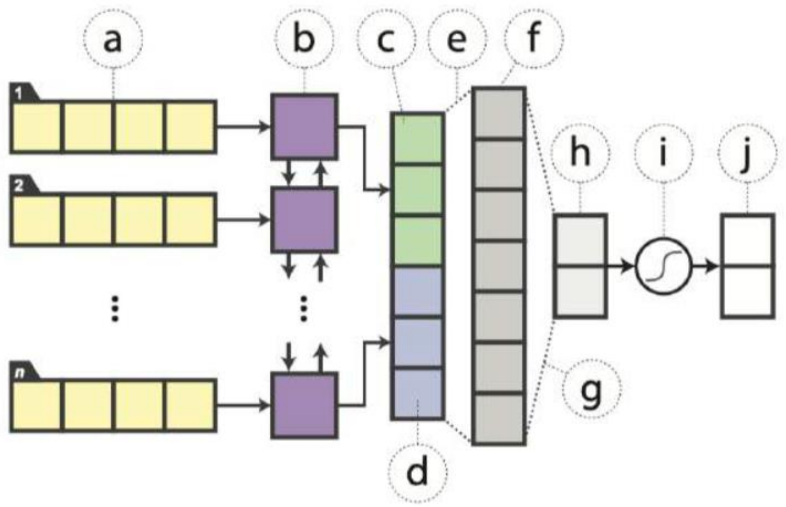
- **Teacher Model:** $BERT_{large}$

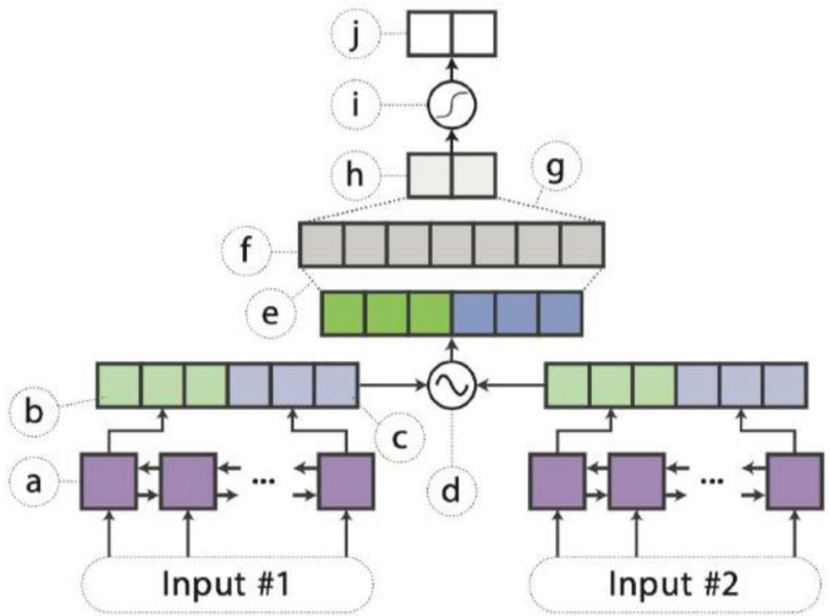




Student Model

- **Student Model :** Single-layer Bi-LSTM with a non-linear classifier





Data Augmentation for Distillation

- In the distillation approach, a small dataset may not suffice for the teacher model to fully express its knowledge. Augment the training set with a large, unlabeled dataset, with pseudo-labels provided by the teacher
- **Method**
  - **Masking**. With probability pmask , we randomly replace a word with [MASK],
  - **POS-guided word replacement**. With probability ppos , we replace a word with another of the same POS tag.
  - **n-gram sampling.** With probability png , we randomly sample an n-gram from the example, where n is randomly selected from {1, 2, . . . , 5}.

Distillation Objective

- **Mean-squared-error (MSE)** loss between the student network's logits against the teacher's logits.
- MSE to perform slightly better.

Teacher's logits Student's logits

$$\mathcal{L}_{\text{distill}} = ||\boldsymbol{z}^{(B)} - \boldsymbol{z}^{(S)}||_2^2$$

$$\begin{aligned}\mathcal{L} &= \alpha \cdot \mathcal{L}_{\text{CE}} + (1-\alpha) \cdot \mathcal{L}_{\text{distill}} \\ &= -\alpha \sum_i t_i \log y_i^{(S)} - (1-\alpha)||\boldsymbol{z}^{(B)} - \boldsymbol{z}^{(S)}||_2^2\end{aligned}$$

Result

#	Model	SST-2	QQP	MNLI-m	MNLI-mm
		Acc	F_1/Acc	Acc	Acc
1	$\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2018)	94.9	72.1/89.3	86.7	85.9
2	$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2018)	93.5	71.2/89.2	84.6	83.4
3	OpenAI GPT (Radford et al., 2018)	91.3	70.3/88.5	82.1	81.4
4	BERT ELMo baseline (Devlin et al., 2018)	90.4	64.8/84.7	76.4	76.1
5	GLUE ELMo baseline (Wang et al., 2018)	90.4	63.1/84.3	74.1	74.5
6	Distilled $\text{BiLSTM}_{\text{SOFT}}$	**90.7**	**68.2/88.1**	**73.0**	**72.6**
7	BiLSTM (our implementation)	86.7	63.7/86.2	68.7	68.3
8	BiLSTM (reported by GLUE)	85.9	61.4/81.7	70.3	70.8
9	BiLSTM (reported by other papers)	$87.6^{\dagger}$	– /$82.6^{\ddagger}$	66.9^{*}	66.9^{*}

Today’s Lecture

- **What’s Knowledge Distillation?**
- **NIPSW – Knowledge Distillation**
- Knowledge Modeling
- Distillation Method
- Knowledge Distillation Scenarios