Knowledge Distillation

Xiachong Feng

Pic https://data-soup.gitlab.io/blog/knowledge-distillation/

Outline

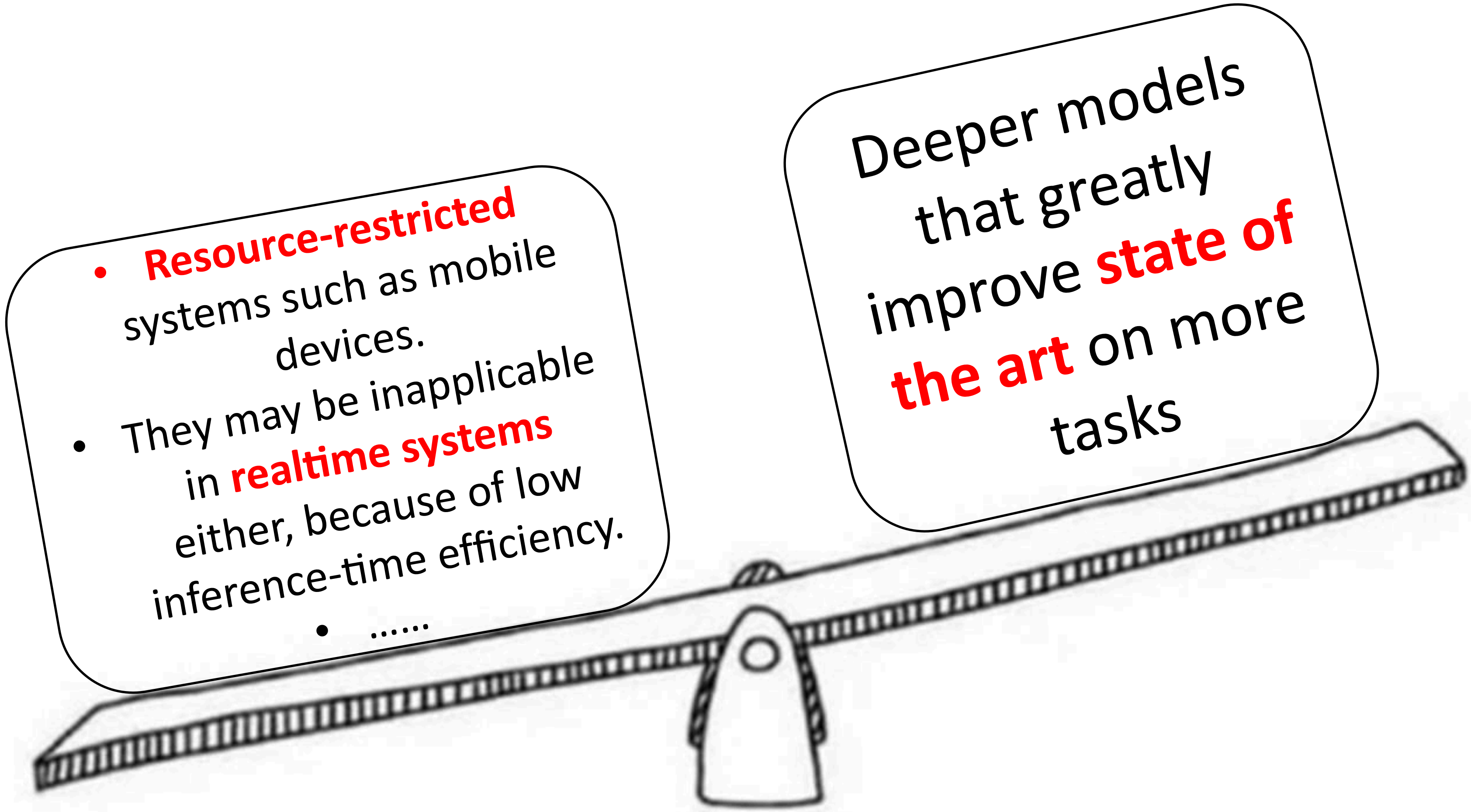
- **Why Knowledge Distillation?**
- **Distilling the knowledge in a neural network** *NIPS2014*
- **Model Compression**
  - Distilling Task-Specific Knowledge from BERT into Simple Neural Networks *arxiv 2018*
- **Multi-Task Setting**
  - Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding *arxiv*
  - BAM! Born-Again Multi-Task Networks for Natural Language Understanding
- **Seq2Seq NMT**
  - Sequence level knowledge distillation *EMNLP16*
- **Cross Lingual NLP**
  - Cross-lingual Distillation for Text Classification *ACL17*
  - Zero-Shot Cross-Lingual Neural Headline Generation *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018*
- **Variant**
  - Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection *AAAI19*
- **Paper List**
- **Reference**
- **Conclusion**

Cost

- **BERT$_{large}$**
  - 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

XLNet训练成本6万美元，顶5个BERT，大模型「身价」惊人
https://zhuanlan.zhihu.com/p/71609636?utm_source=wechat_session&utm_medium=social&utm_oi=71065644564480&from=timeline&isappinstalled=0&s_r=0

Trade-Off



Distilling Task-Specific Knowledge from BERT into Simple Neural Networks

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.

Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding

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/

Hot Topic

6、**提供一个轻量级的 BERT 替代方案 BERB**：Bidirectional Encoder Representation from BiRNN。大家都惊叹于 BERT 所需的巨大的计算资源。但实际上，假如采用一个真双向 RNN（就是高层可以同时看到底层正反向的信息的那种），堆个 4 层或者 6 层（而不需要像本文一样弄 24 层），然后同样使用 MLM 和 NSP 两个目标来训练，需要的计算资源应该会少很多，并且完全用到了 BERT 模型核心的改进点。至于效果的话我预期会比 BERT 差一些，但是应该会比现有的其他方法好。RNN 堆的层数深了以后可能会难以训练，所以可能需要加 residual connection 或者 layer normalization。这里也带出了 Transformer 的另一个优越之处，那就是自带各种 normalization，堆很多层照样能稳定训练。当然，把预训练好的 BERT 蒸馏成 BERB 也是可以的。（更新：现在已经有人这么做了。）

Towser 如何评价BERT模型

https://www.zhihu.com/question/298203515/answer/509923837

3. 更快的BERT

BERT另外一大挑战是如此大，如此重的模型，如何上线?

用小模型，如三层transformer就获得十二层transformer的效果？这是模型蒸馏的角度。

用更少的参数？剪去多余无效的参数？这是模型剪枝。

用更低的精度呢？INT8行不行？三值网络行不行？二值网络行不行?

精简transformer呢？研究更高效的transformer?

霍华德 BERT模型在NLP中目前取得如此好的效果，那下一步NLP该何去何从？

https://www.zhihu.com/question/320606353/answer/658786633

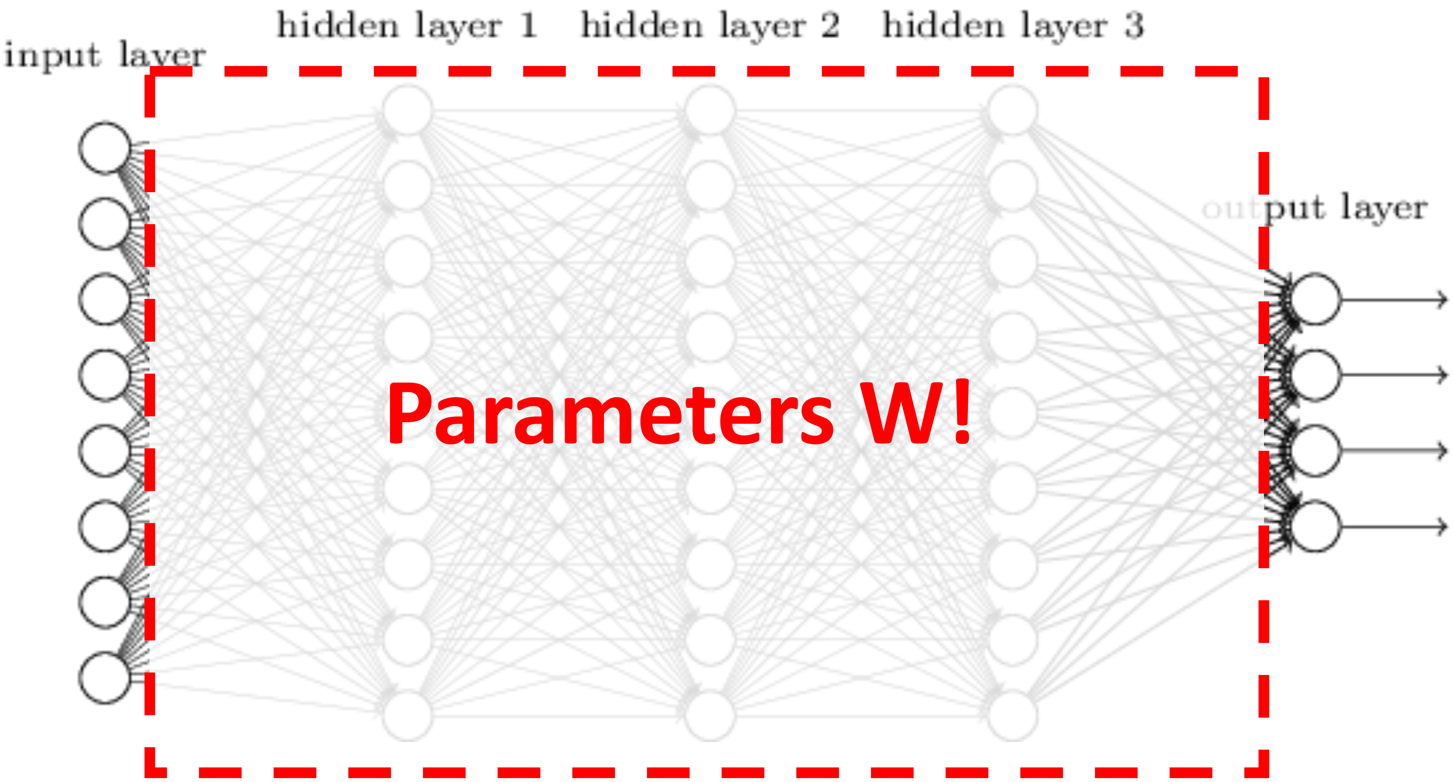
Distilling the Knowledge in a Neural Network

Hinton
NIPS 2014 Deep Learning Workshop

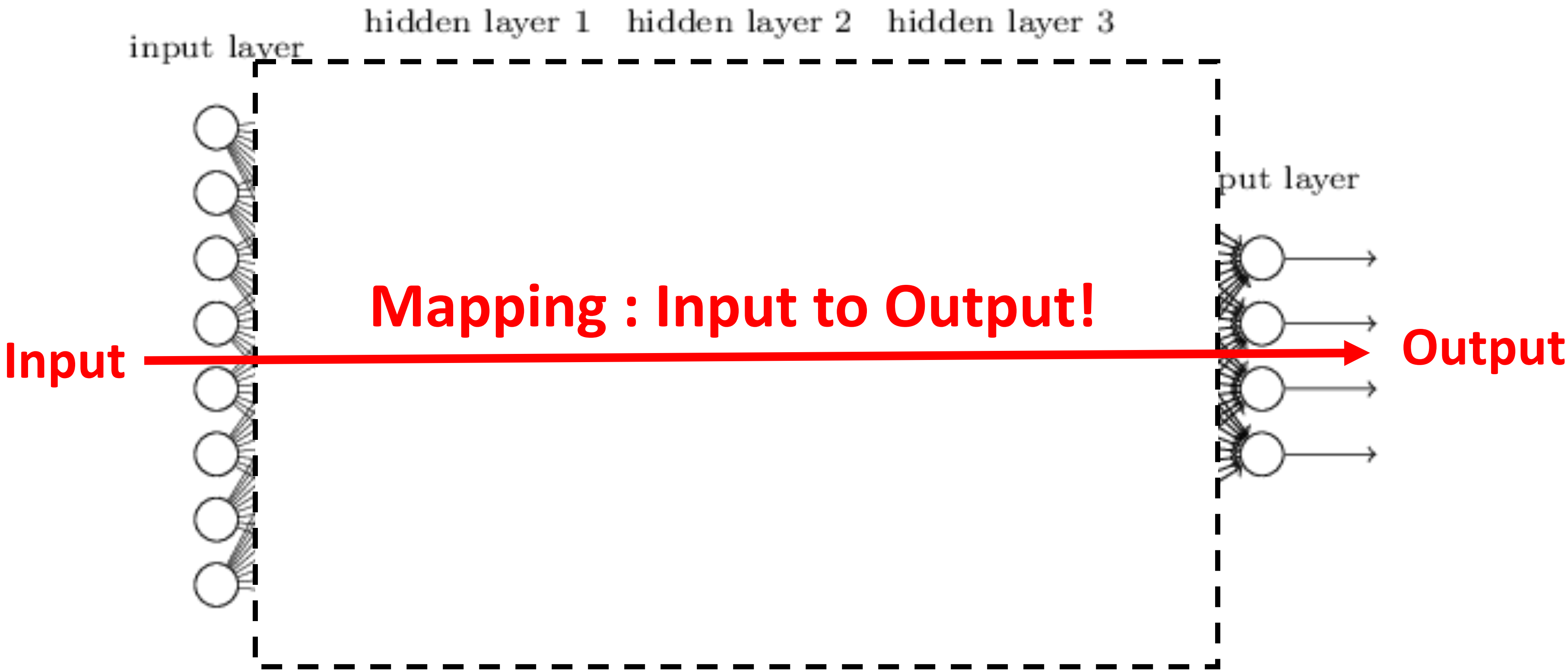
Model Compression

- **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

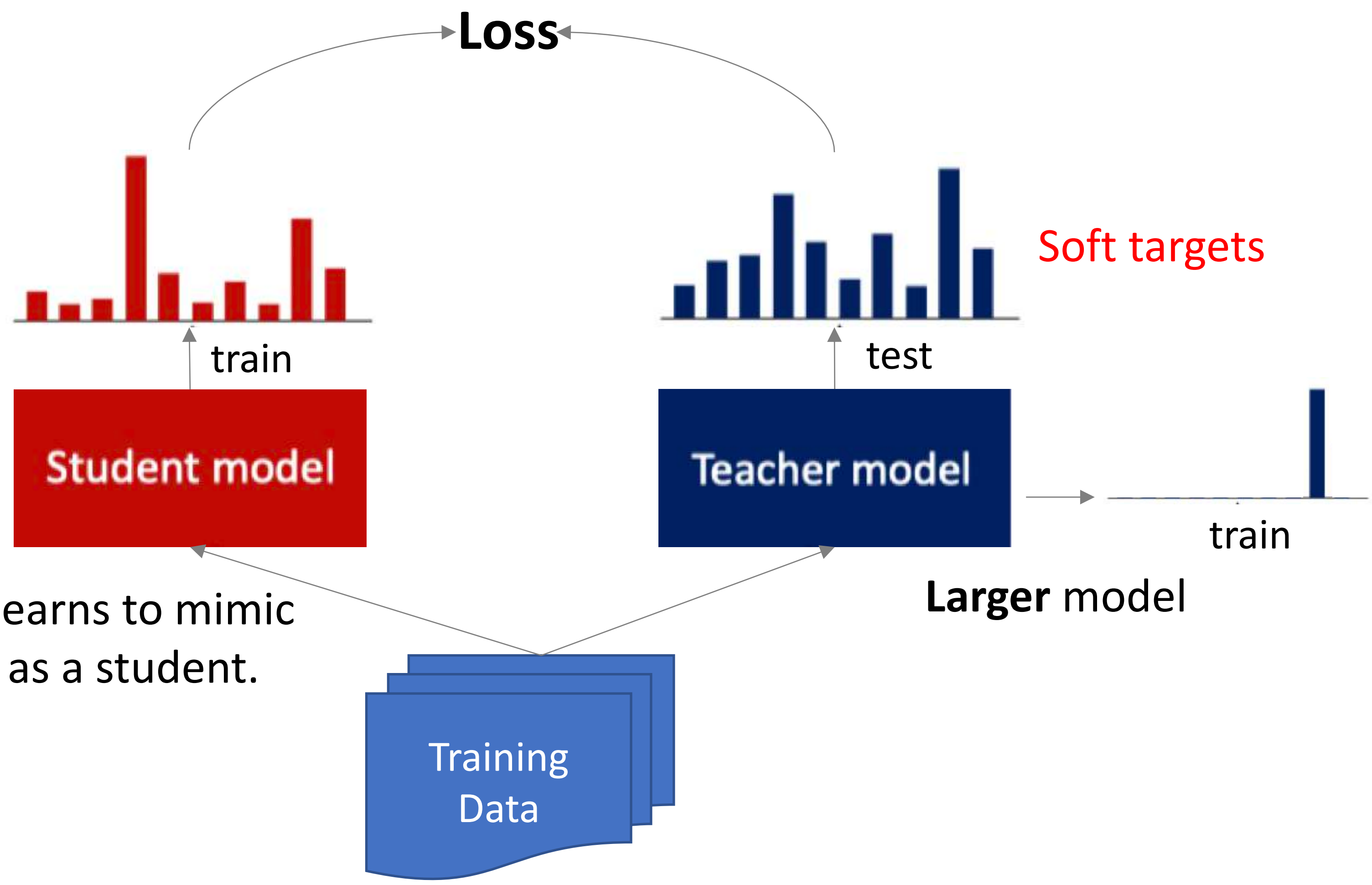


What is Knowledge? 2



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



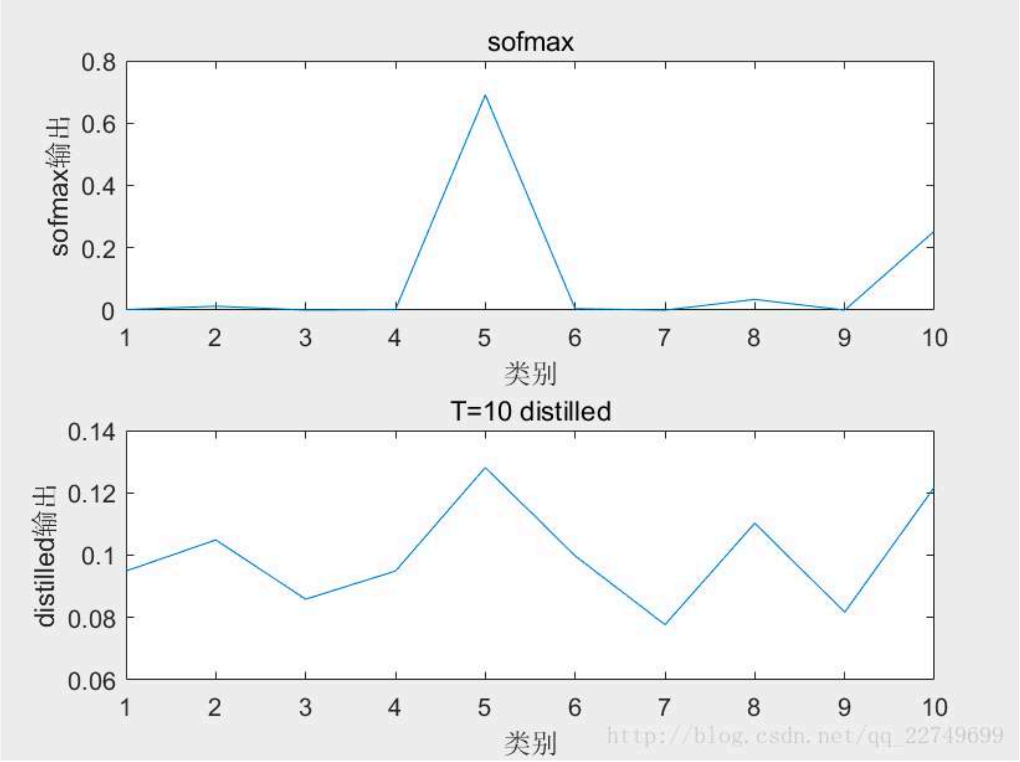
Small model learns to mimic the teacher as a student.

Softmax With Temperature

Logits

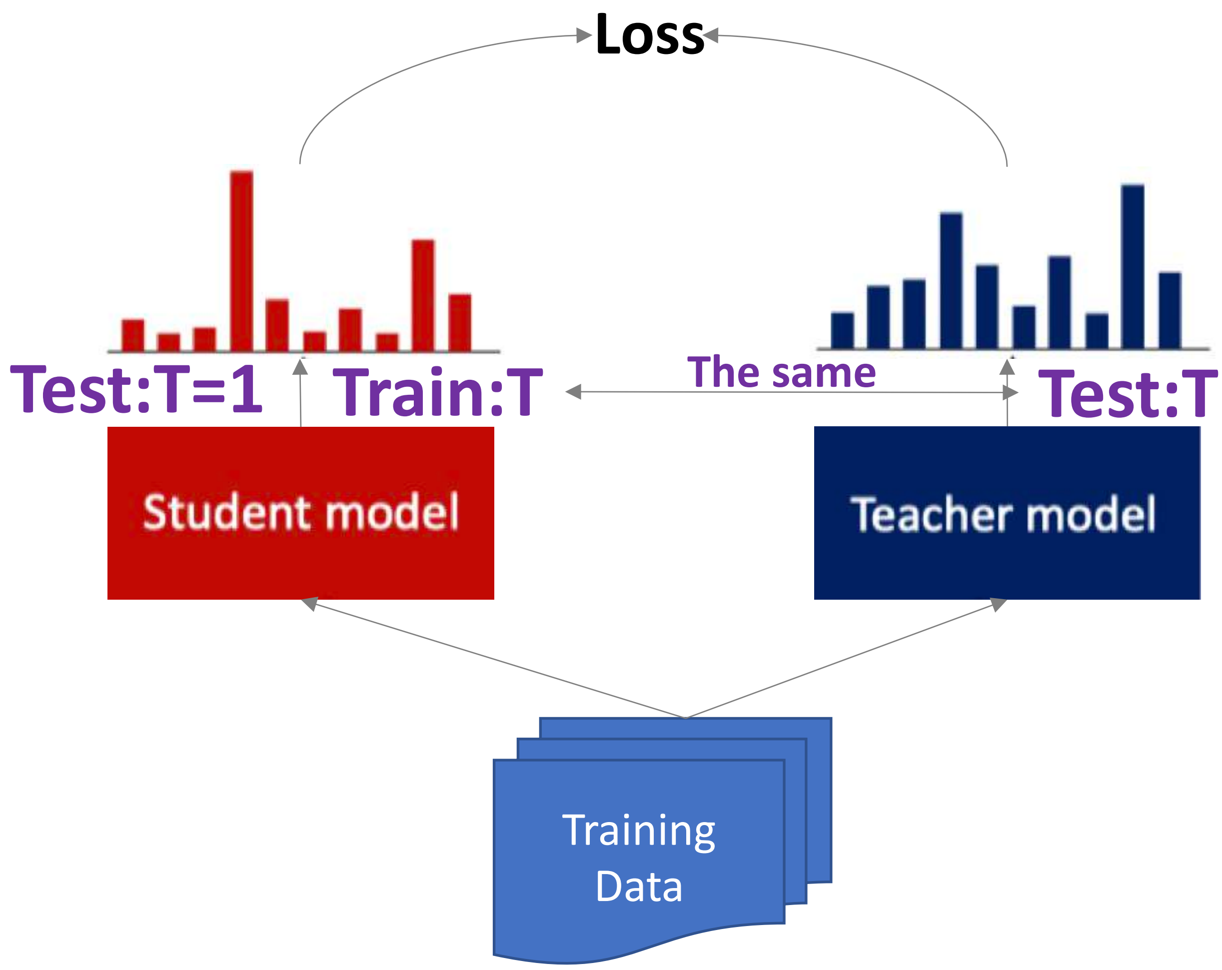
$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

Temperature

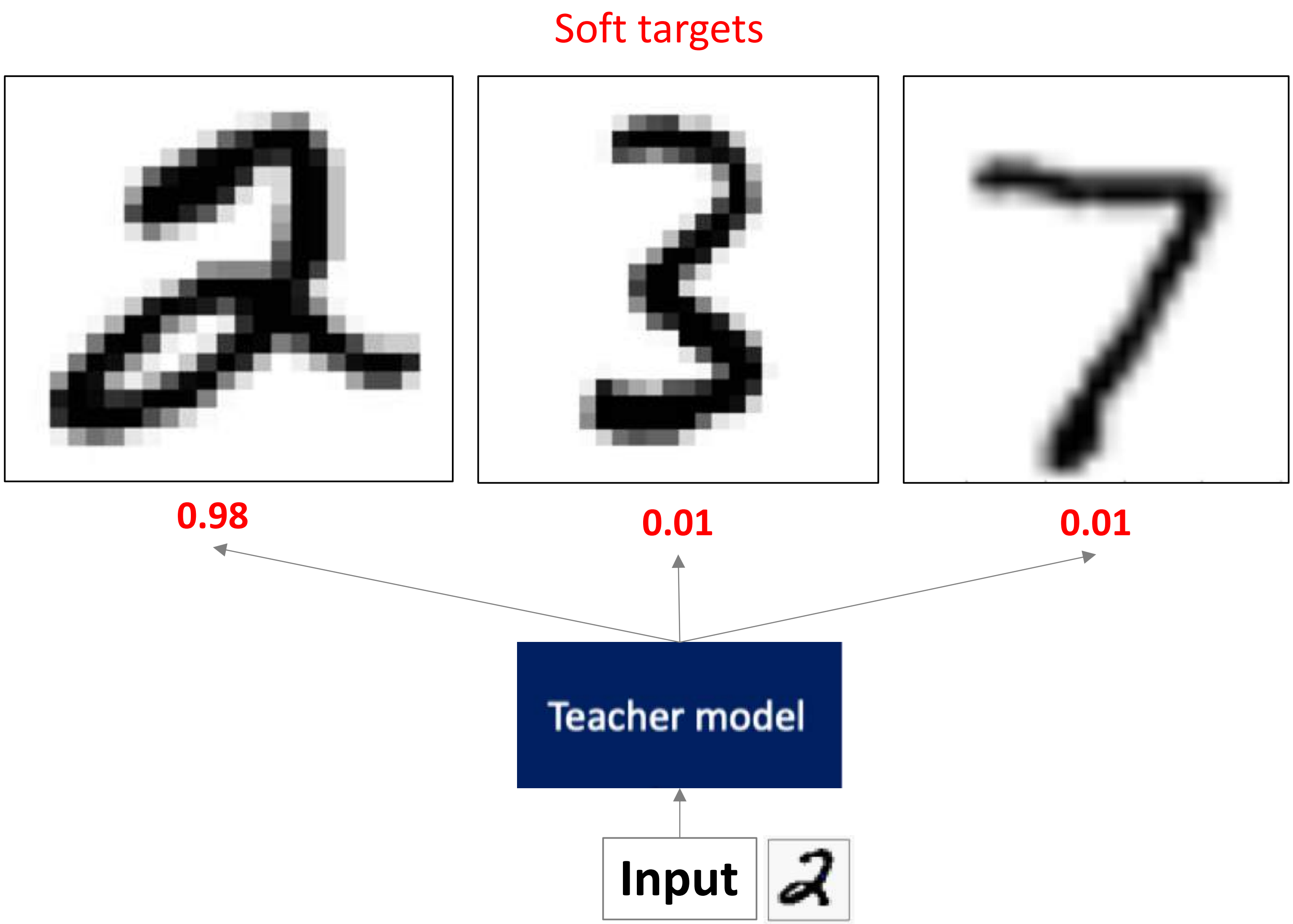


https://blog.csdn.net/qq_22749699/article/details/79460817

Note

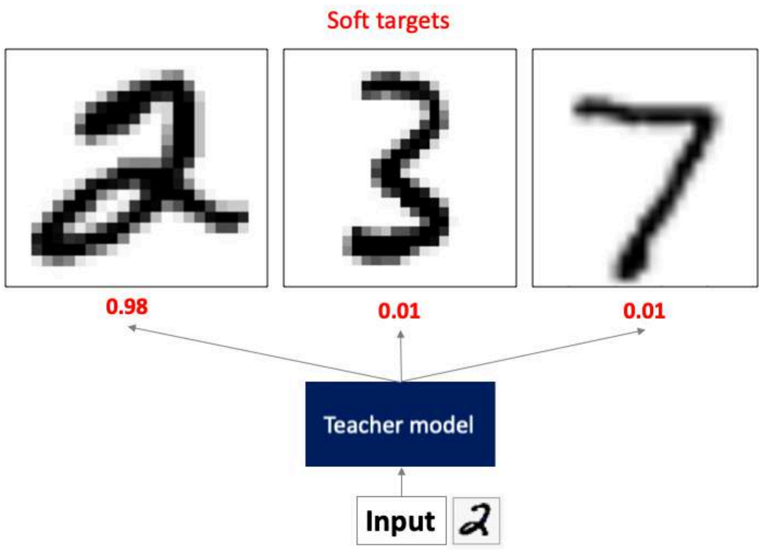

Loss
Test:T=1
Train:T
The same
Test:T
Student model
Teacher model
Training
Data

Soft Targets



Supervisory signals

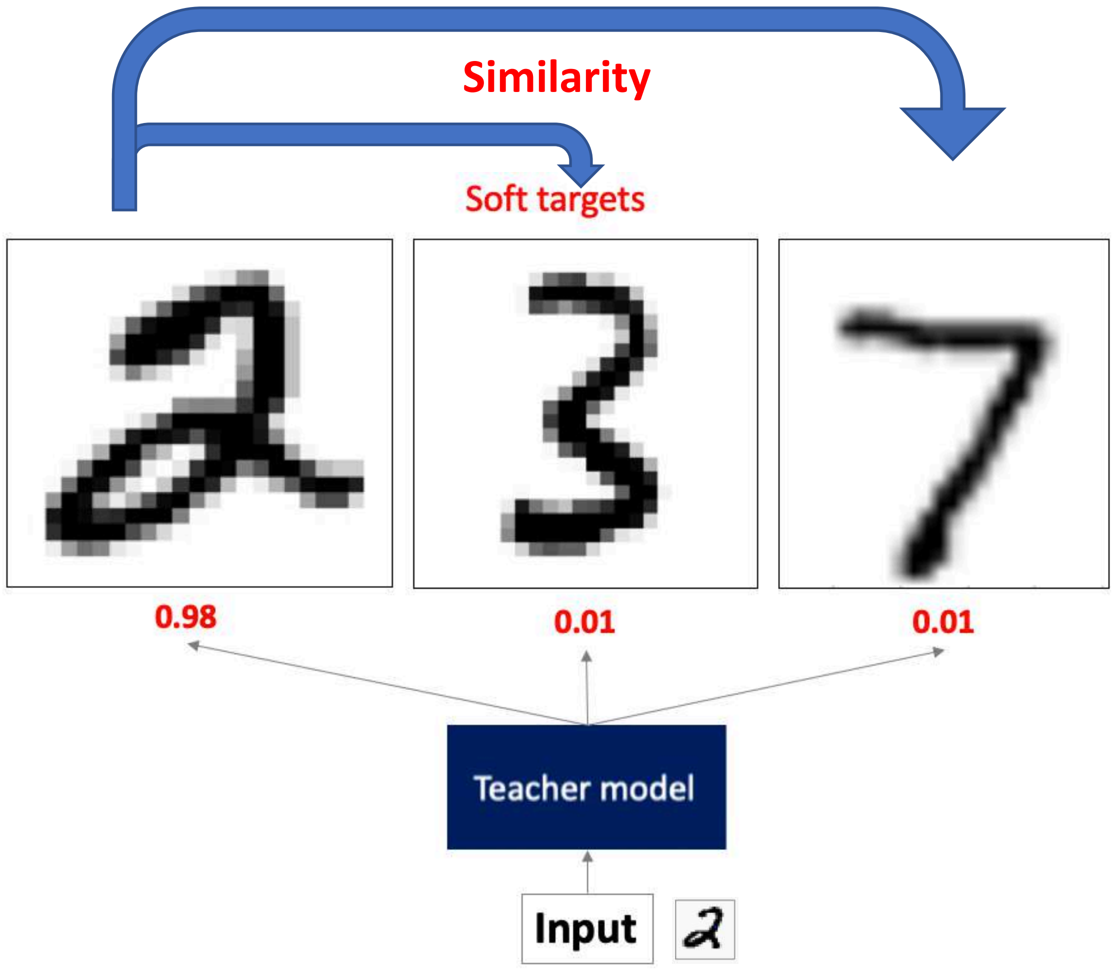
Soft target	One-hot
• 2 is similar to 3 and 7	• 2 independent of 3 and 7.
• Contiguous distribution	• Discrete distribution
• **Inter-Class variance** ✓	• **Inter-Class variance**
• **Between-Class distance** ✓	• **Between-Class distance**



Soft targets have high entropy！

Naiyan Wang https://www.zhihu.com/question/50519680/answer/136363665

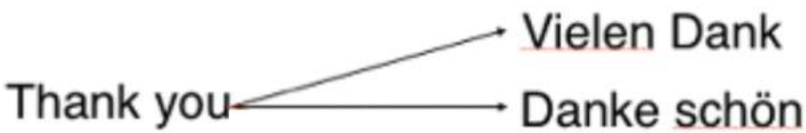
2 Data augmentation



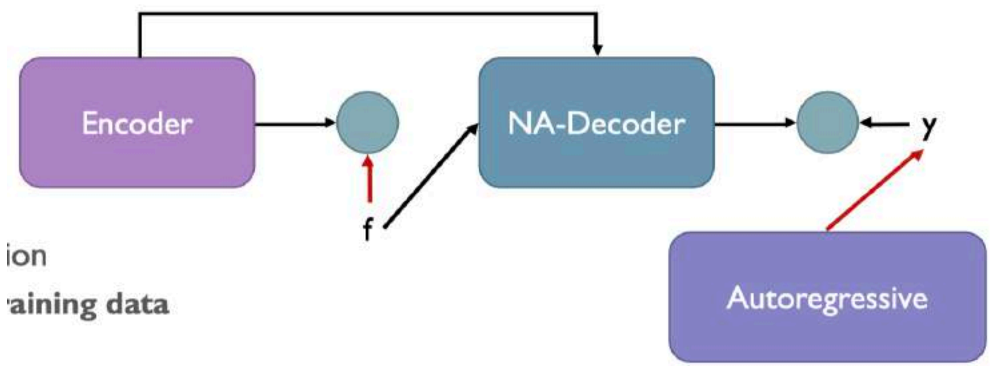
周博磊 https://www.zhihu.com/question/50519680/answer/136359743

3 Reduce Modes

- NMT : Real translation data has many modes.



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

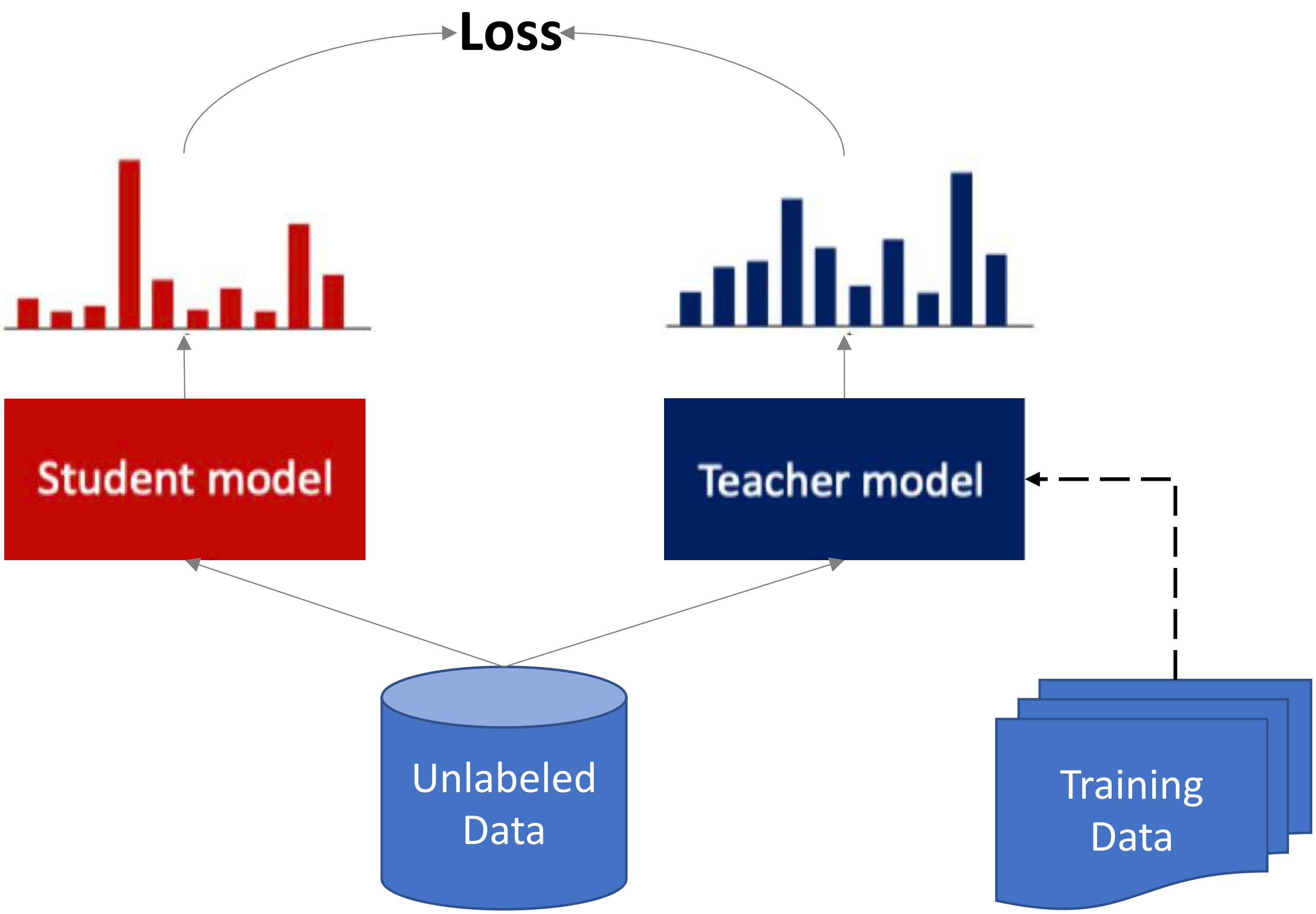


Jiatao Gu Non-Autoregressive Neural Machine Translation
https://zhuanlan.zhihu.com/p/34495294

Soft Targets

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

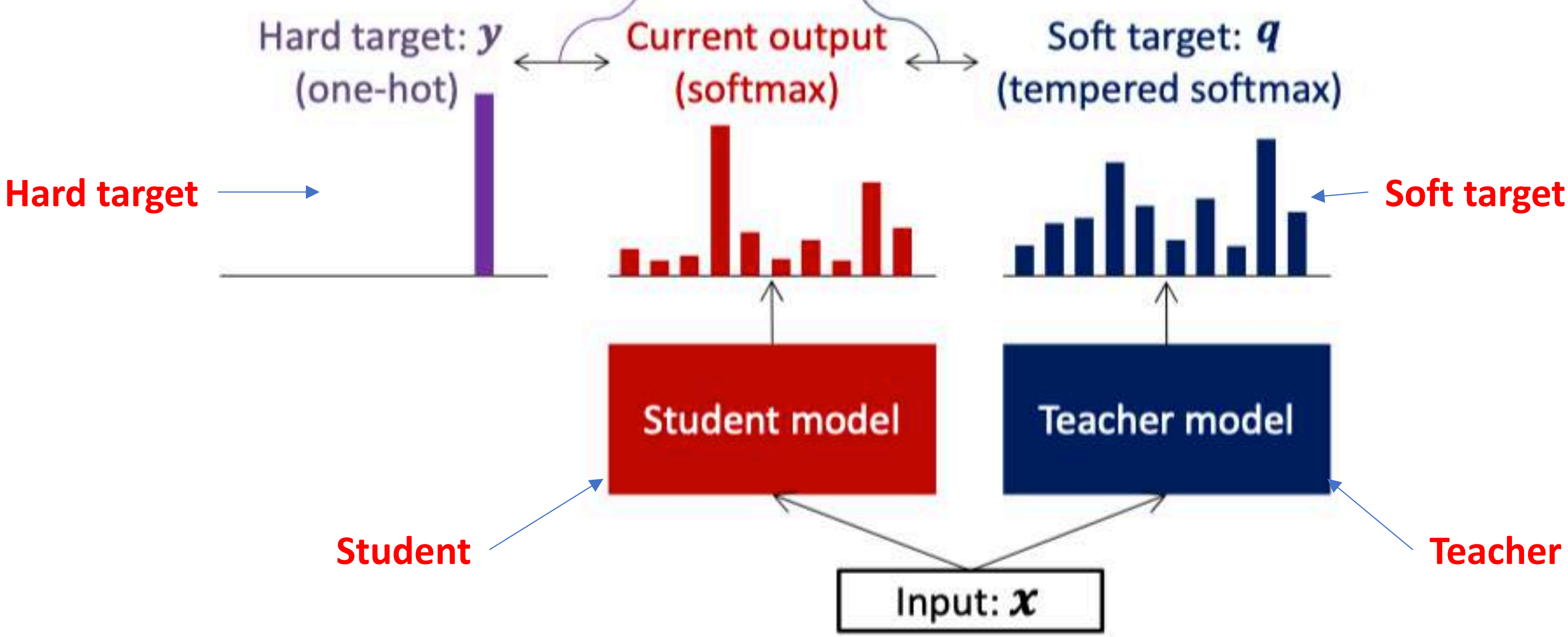
How to use unlabeled data?



Loss function

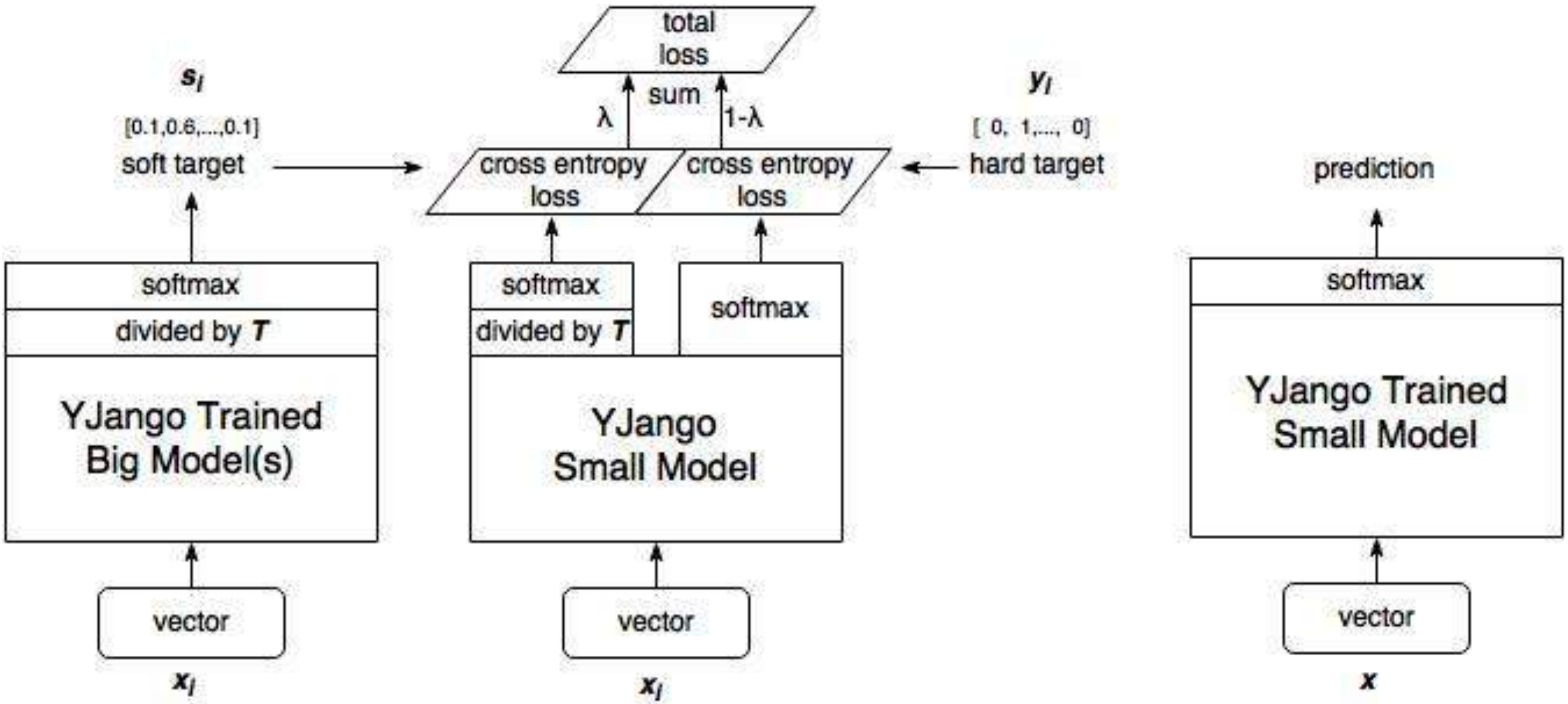
Transfer set = unlabeled data + original training set

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



DOMAIN ADAPTATION OF DNN ACOUSTIC MODELS USING KNOWLEDGE DISTILLATION 2017 ICASSP

Knowledge Distillation



如何理解soft target这一做法？Yjango https://www.zhihu.com/question/50519680?sort=created

Outline

- **Why Knowledge Distillation?**
- **Distilling the knowledge in a neural network *NIPS2014***
- **Model Compression**
  - Distilling Task-Specific Knowledge from BERT into Simple Neural Networks *arxiv 2018*
- **Multi-Task Setting**
  - Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding *arxiv*
  - BAM! Born-Again Multi-Task Networks for Natural Language Understanding
- **Seq2Seq NMT**
  - Sequence level knowledge distillation *EMNLP16*
- **Cross Lingual NLP**
  - Cross-lingual Distillation for Text Classification *ACL17*
  - Zero-Shot Cross-Lingual Neural Headline Generation *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018*
- **Variant**
  - Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection *AAAI19*
- **Paper List**
- **Reference**
- **Conclusion**

Distilling Task-Specific Knowledge from BERT into Simple Neural Networks

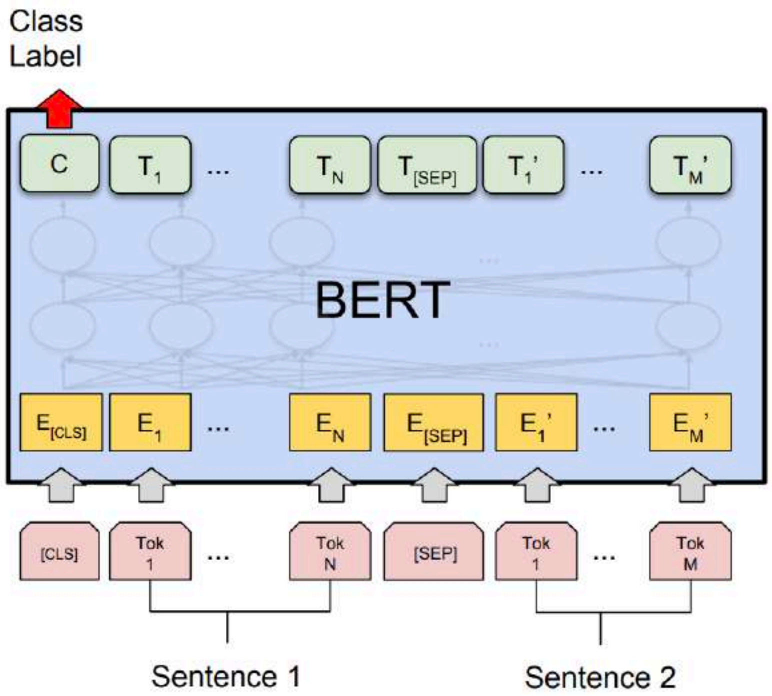
University of Waterloo
arxiv

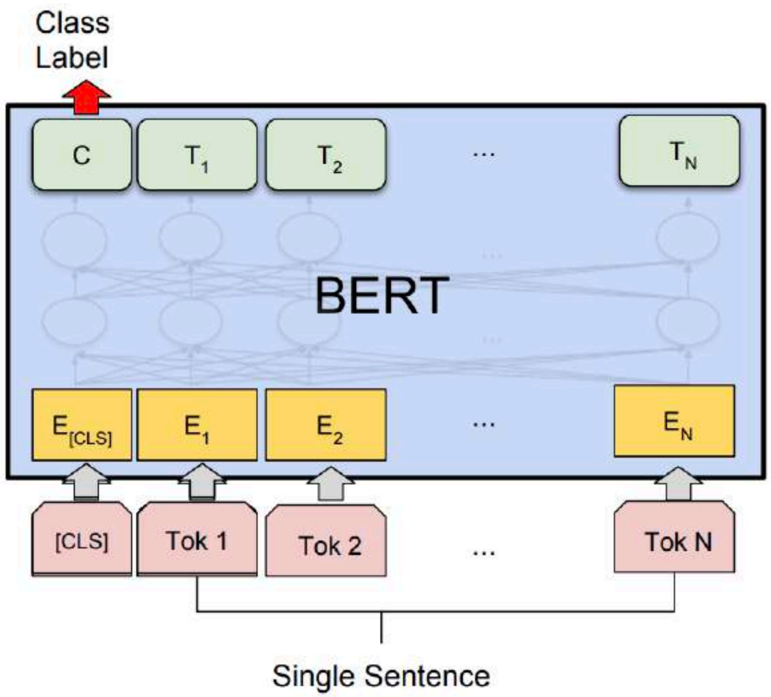
Overview

- 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

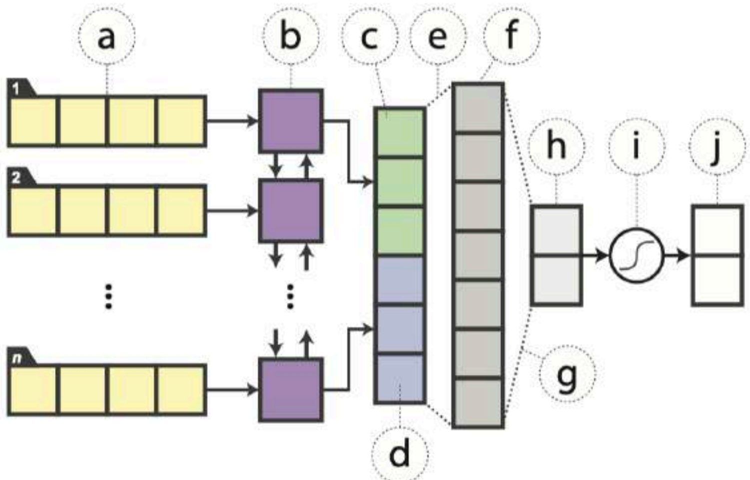
- **Teacher Model:** $BERT_{large}$

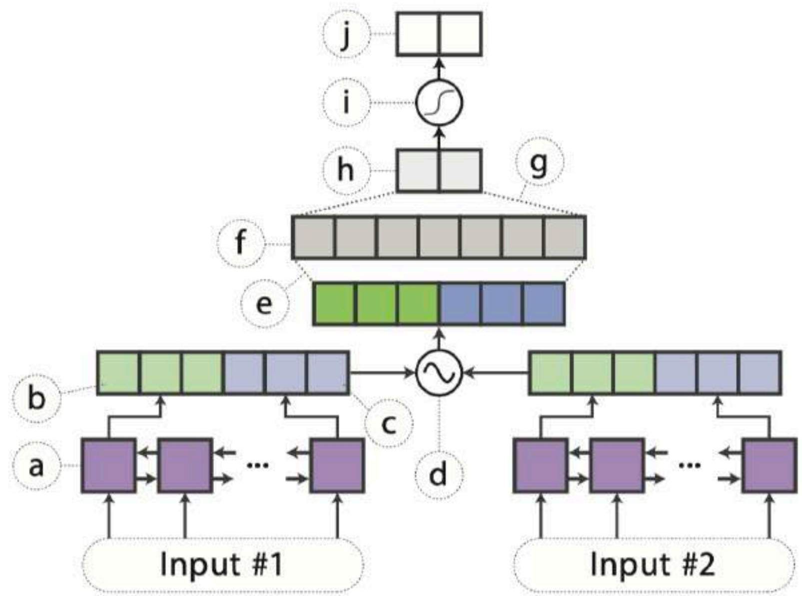




Student Model

- **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	$BERT_{LARGE}$ (Devlin et al., 2018)	94.9	72.1/89.3	86.7	85.9
2	$BERT_{BASE}$ (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 $BiLSTM_{SOFT}$	**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^{*}

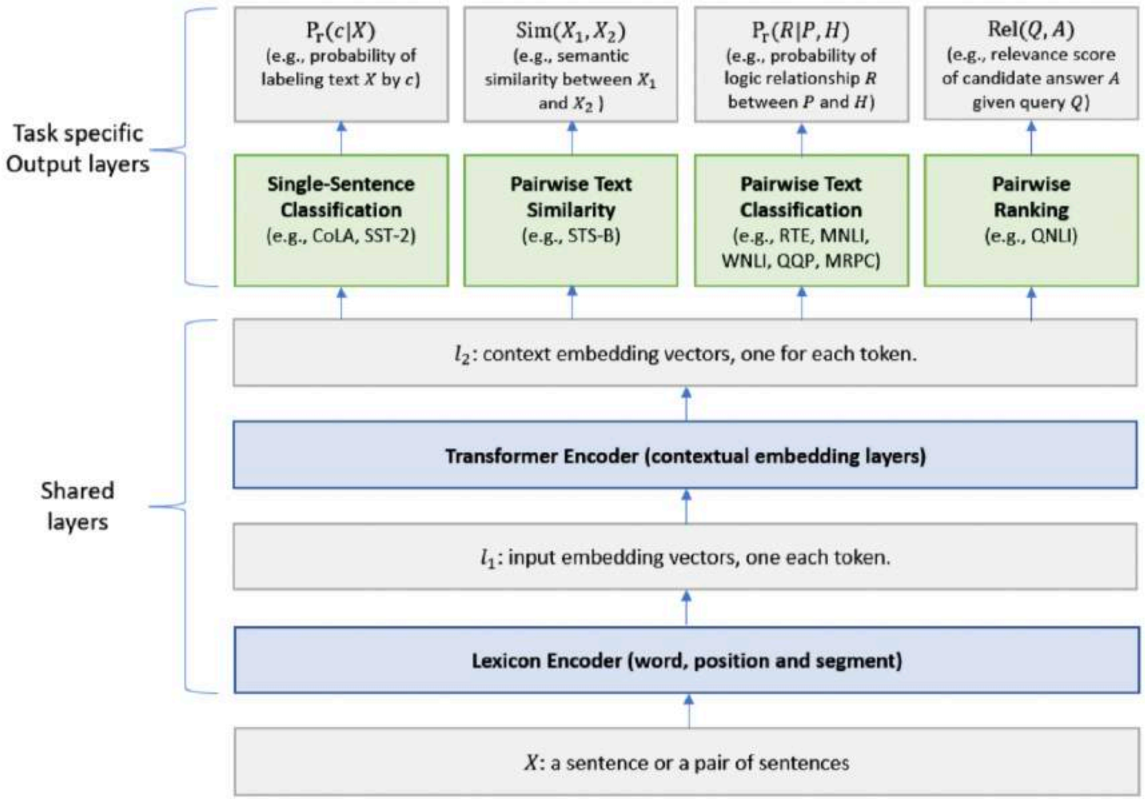
Outline

- **Why Knowledge Distillation?**
- **Distilling the knowledge in a neural network *NIPS2014***
- **Model Compression**
  - Distilling Task-Specific Knowledge from BERT into Simple Neural Networks *arxiv 2018*
- **Multi-Task Setting**
  - Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding *arxiv*
  - BAM! Born-Again Multi-Task Networks for Natural Language Understanding
- **Seq2Seq NMT**
  - Sequence level knowledge distillation *EMNLP16*
- **Cross Lingual NLP**
  - Cross-lingual Distillation for Text Classification *ACL17*
  - Zero-Shot Cross-Lingual Neural Headline Generation *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018*
- **Variant**
  - Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection *AAAI19*
- **Paper List**
- **Reference**
- **Conclusion**

Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding

Microsoft

MT-DNN



pre-training stage

Algorithm 1: Training a MT-DNN model.

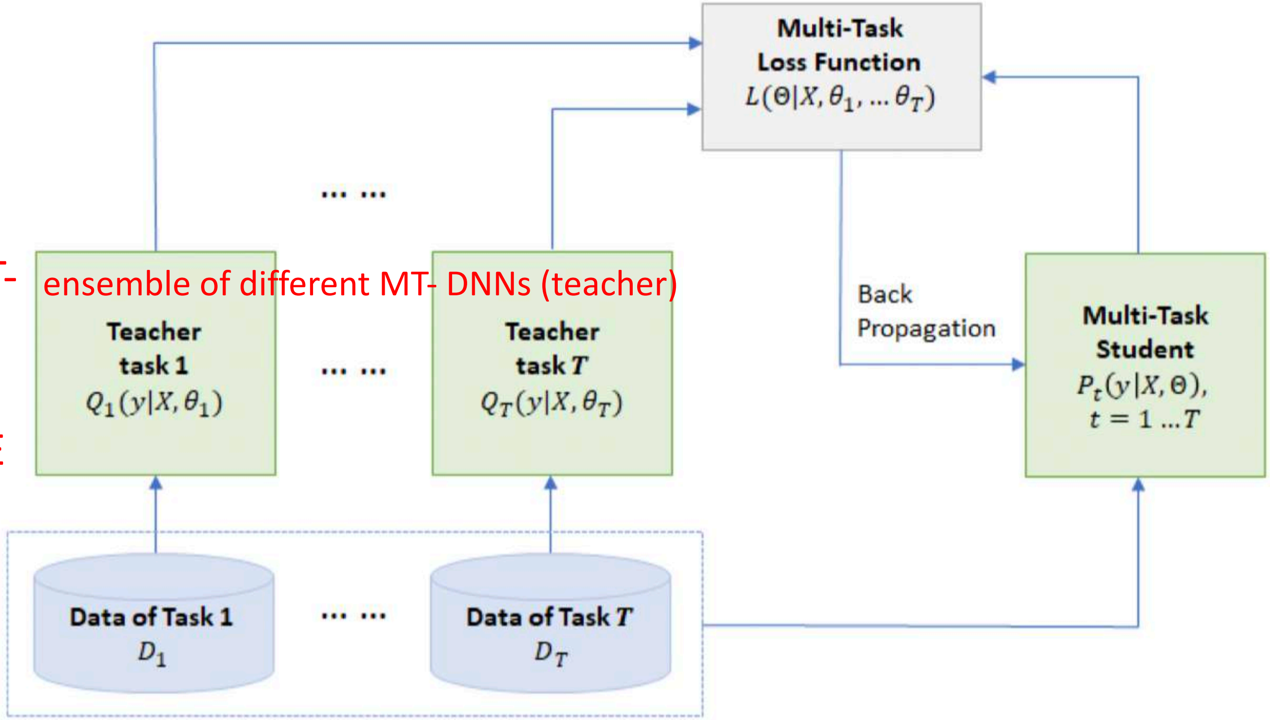
Initialize model parameters Θ randomly.
Initialize the shared layers (i.e., the lexicon encoder and the transformer encoder) using a pre-trained BERT model.
Set the max number of epoch: $epoch_{max}$.
//Prepare the data for T tasks.
for t *in* $1, 2, ..., T$ **do**
    Pack the dataset t into mini-batch: D_t.
end
for *epoch in* $1, 2, ..., epoch_{max}$ **do**
    1. Merge all the datasets:
    $D = D_1 \cup D_2 ... \cup D_T$
    2. Shuffle D
    for b_t *in* D **do**
        //b_t is a mini-batch of task t.
        3. Compute task-specific loss : $L_t(\Theta)$
        4. Compute gradient: $\nabla(\Theta)$
        5. Update model: $\Theta = \Theta - \epsilon\nabla(\Theta)$
    end
end

MTL stage

Multi-task deep neural networks for natural language understanding

Distillation

correct targets + soft targets



Initialized using the MT-DNN model pre-trained on the GLUE dataset

ensemble of different MT- DNNs (teacher)

Initialized using the pre-trained BERT

- The parameters of its shared layers are initialized using the MT-DNN model pre-trained on the GLUE dataset via MTL, as in Algorithm 1, and the parameters of its task-specific output layers are randomly initialized.
- Disttilled MT-DNN significantly outperforms the original MT-DNN on **7 out of 9 GLUE tasks**(single model).

Teacher Annealing

- BAM! Born-Again Multi-Task Networks for Natural Language Understanding
- **Born Again** : the student has the same model architecture as the teacher.

$$CE(\lambda y_\tau^i + (1-\lambda) p_\tau(y|x_\tau^i, \theta_\tau), p_\tau(y|x_\tau^i, \theta))$$

λ is linearly increased from 0 to 1

Outline

- **Why Knowledge Distillation?**
- **Distilling the knowledge in a neural network *NIPS2014***
- **Model Compression**
  - Distilling Task-Specific Knowledge from BERT into Simple Neural Networks *arxiv 2018*
- **Multi-Task Setting**
  - Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding *arxiv*
  - BAM! Born-Again Multi-Task Networks for Natural Language Understanding
- **Seq2Seq NMT**
  - Sequence level knowledge distillation *EMNLP16*
- **Cross Lingual NLP**
  - Cross-lingual Distillation for Text Classification *ACL17*
  - Zero-Shot Cross-Lingual Neural Headline Generation *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018*
- **Variant**
  - Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection *AAAI19*
- **Paper List**
- **Reference**
- **Conclusion**

Sequence level knowledge distillation

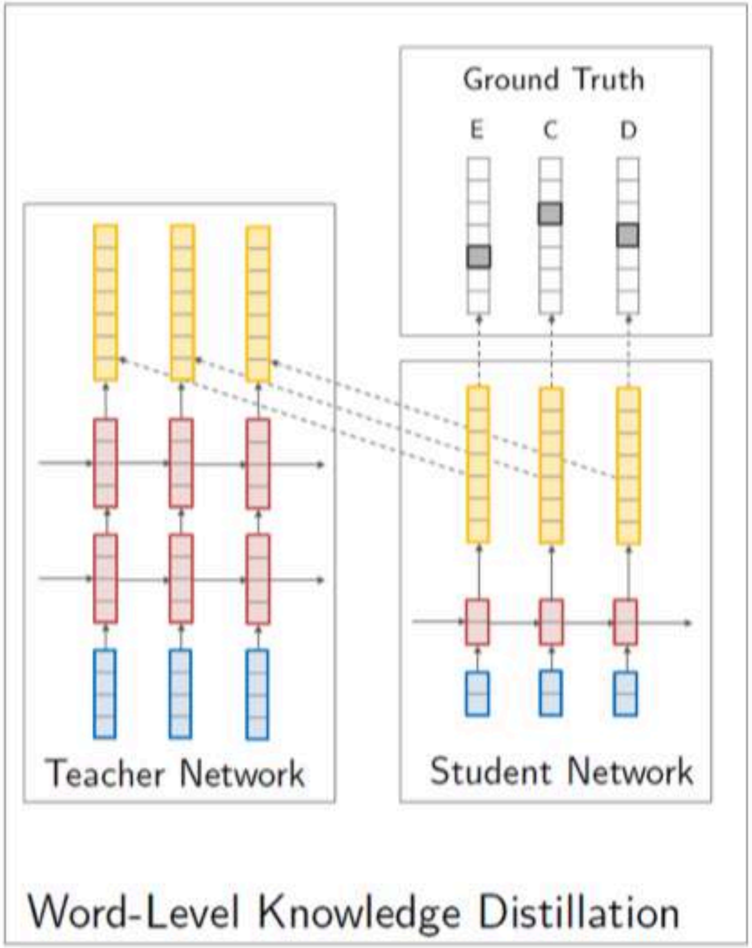
EMNLP16
Yoon Kim Harvard

Seq2Seq

- **Non-recurrent models** in the multiclass prediction setting

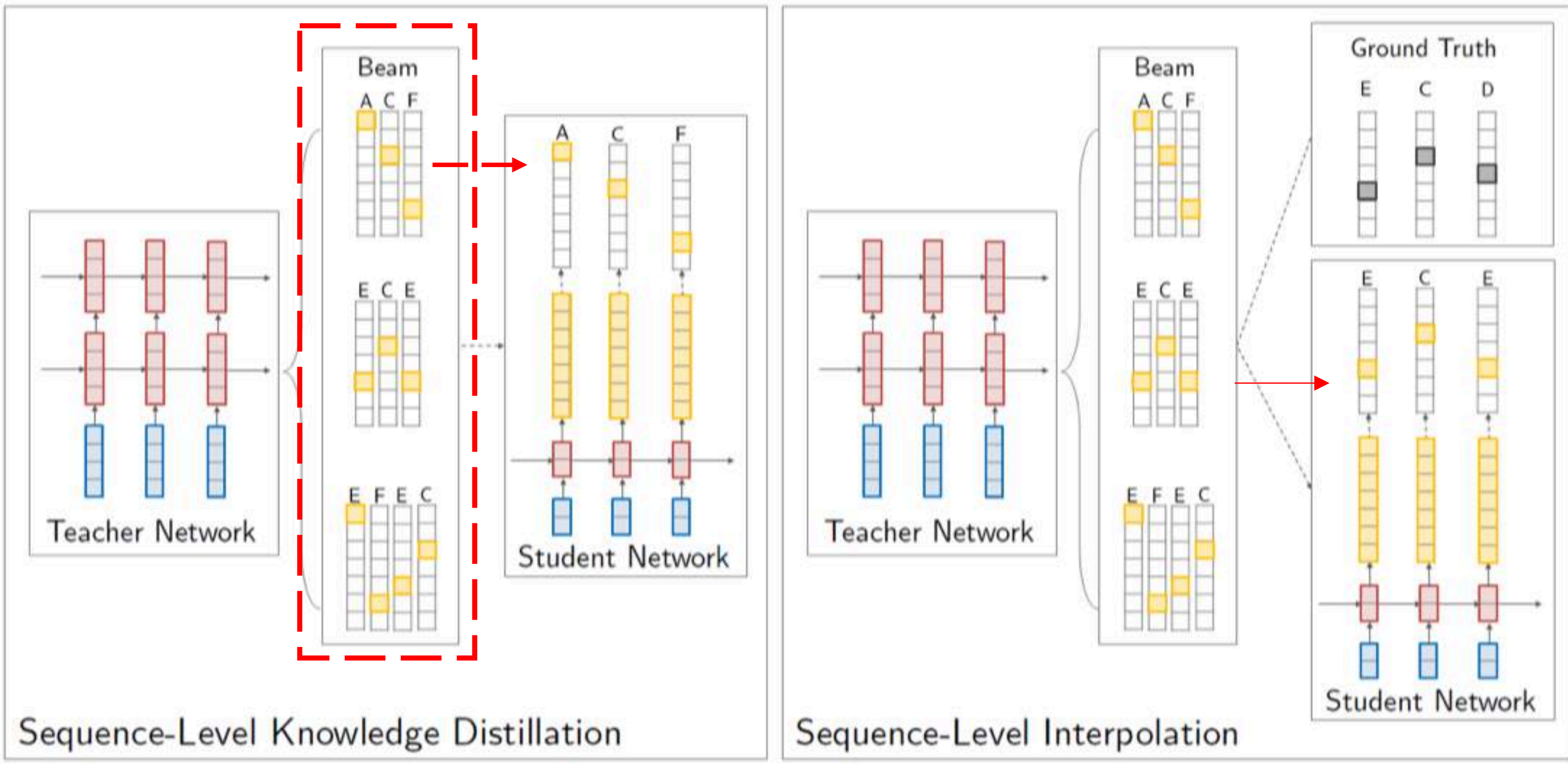
- **Method**
  - **Word-level** Distillation
  - Two novel **sequence-level** versions of knowledge distillation
    - Sequence-Level Knowledge Distillation
    - Sequence-Level Interpolation

Word-Level



Ground Truth
E
C
D
Teacher Network
Student Network
Word-Level Knowledge Distillation

Sentence Level



Result

- Large state-of-the-art 4×1000 LSTM
  - → 2×500 LSTM
- Not requiring any beam search at test-time. As a result we are able to perform greedy decoding on the 2×500 model 10 times faster than beam search on the 4×1000 model with comparable performance.

Outline

- **Why Knowledge Distillation?**
- **Distilling the knowledge in a neural network *NIPS2014***
- **Model Compression**
  - Distilling Task-Specific Knowledge from BERT into Simple Neural Networks *arxiv 2018*
- **Multi-Task Setting**
  - Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding *arxiv*
  - BAM! Born-Again Multi-Task Networks for Natural Language Understanding
- **Seq2Seq NMT**
  - Sequence level knowledge distillation *EMNLP16*
- **Cross Lingual NLP**
  - Cross-lingual Distillation for Text Classification *ACL17*
  - Zero-Shot Cross-Lingual Neural Headline Generation *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018*
- **Variant**
  - Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection *AAAI19*
- **Paper List**
- **Reference**
- **Conclusion**

Cross-lingual Distillation for Text Classification

ACL17

CMU

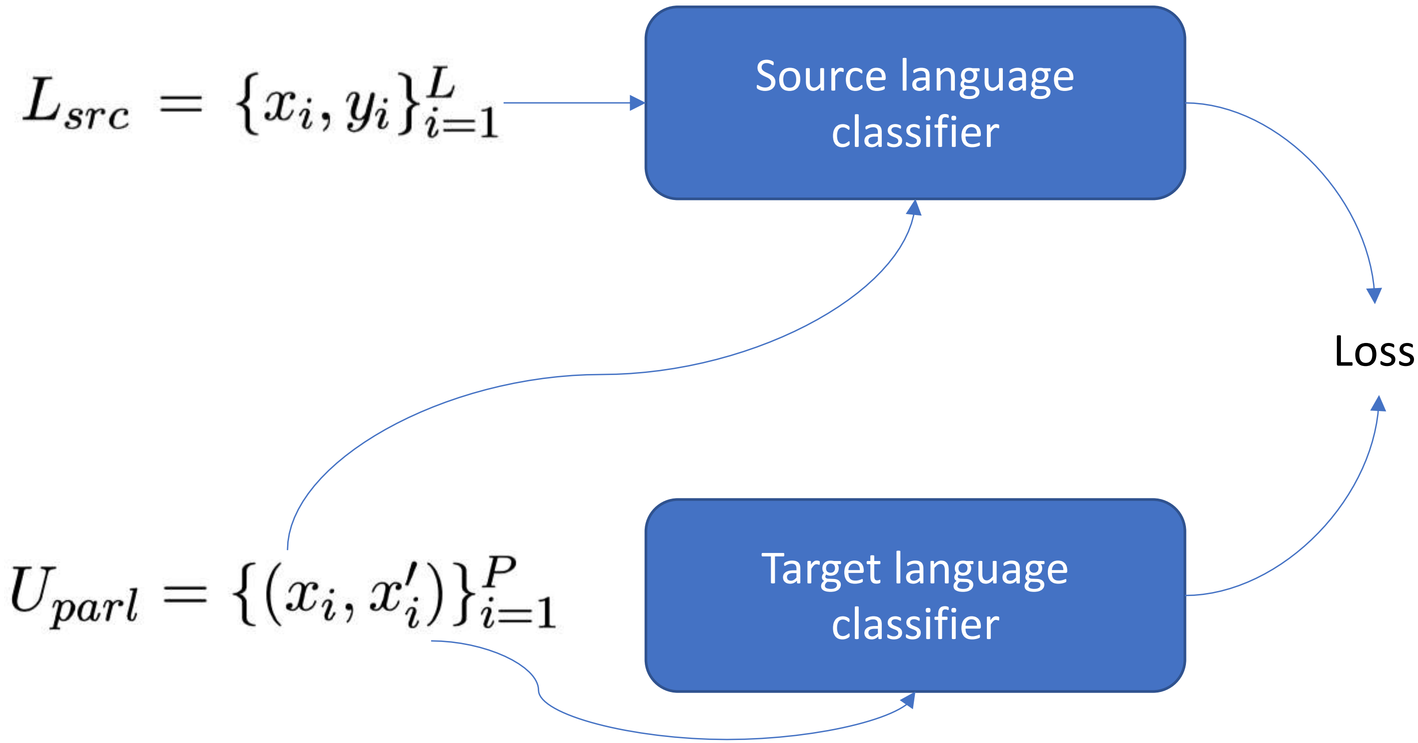
Overview

- **Task**
  - Cross-lingual text classification(CLTC) is the task of classifying documents written in different languages into the same taxonomy of categories.
- **Problem**
  - How can we effectively leverage the trained classifiers in a label-rich source language to help the classification of documents in other label-poor target languages?
- **Method**
  - Vanilla version
  - Distillation with Adversarial Feature Adaptation

Vanilla version

- The **first** step of our framework is to train the source-language classifier on labeled source documents $L_{src} = \{x_i, y_i\}_{i=1}^{L}$.
- In the **second** step, the knowledge captured in θ_{src} is transferred to the distilled model in the target language by training it on the **parallel corpus**.
$U_{parl} = \{(x_i, x_i')\}_{i=1}^{P}$

Vanilla version



- **Intution**
  - The intuition is that paired documents in parallel corpus should have the **same distribution of class** predicted by the source model and target model.

Problem

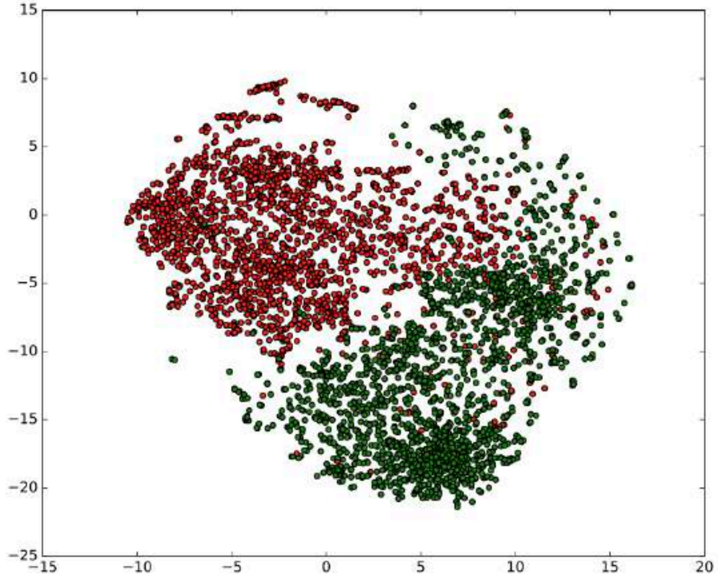


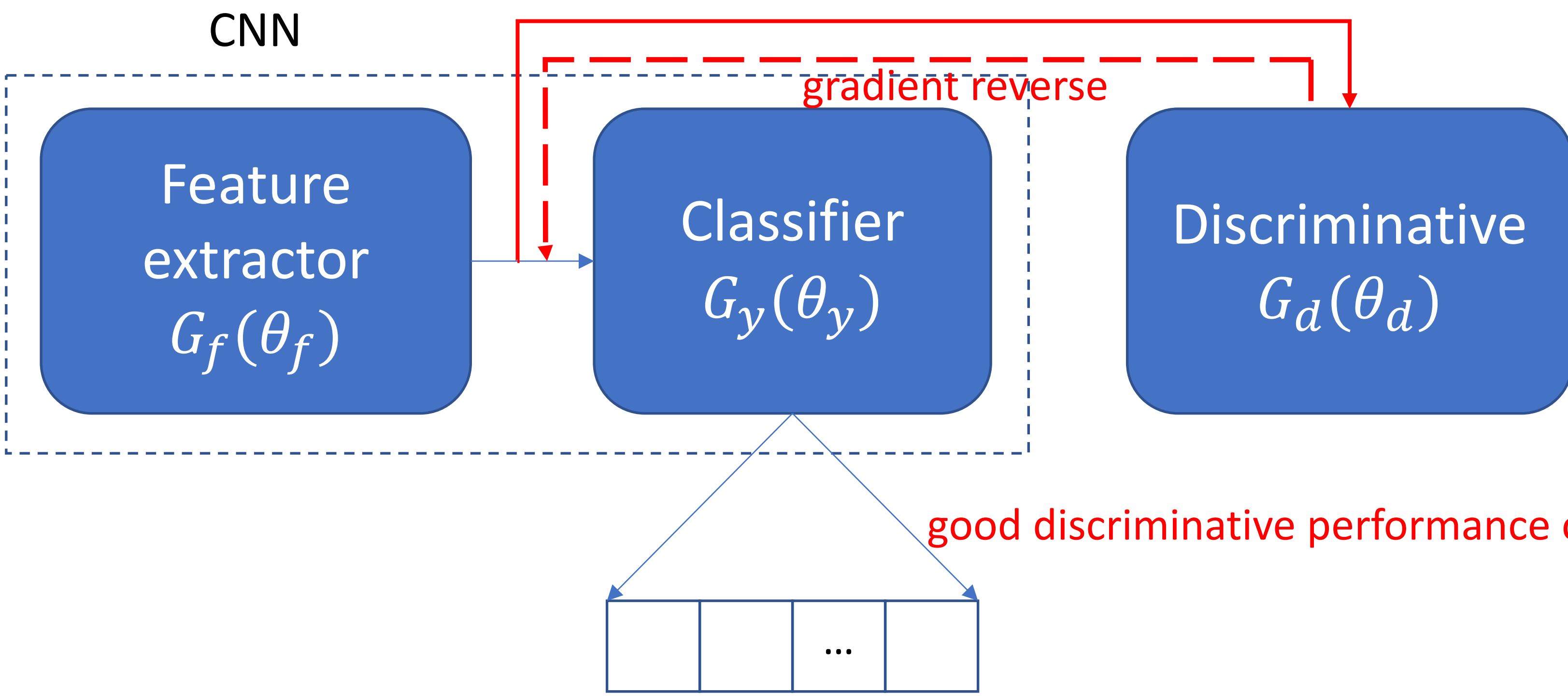
Figure 1: Extracted features for source-language documents in the English-Chinese Yelp Hotel Review dataset. Red dots represent features of the documents in L_{src} and green dots represent the features of documents in U_{parl}, which is a general-purpose parallel corpus.

Distillation with Adversarial Feature Adaptation

extracts features which have similar distributions on L and U

$$-\alpha \sum_{x_i \in L} L_d(0, G_d(G_f(x_i, \theta_f), \theta_d))$$

$$-\alpha \sum_{x_j \in U} L_d(1, G_d(G_f(x_j, \theta_f), \theta_d))$$



good discriminative performance on L

$$\sum_{x_i, y_i \in L} L_y(y_i, G_y(G_f(x_i, \theta_f), \theta_y))$$

Zero-Shot Cross-Lingual Neural Headline Generation

Ayana, Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Mao-song Sun

IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 26, NO. 12, DECEMBER 2018

Cross-lingual headline generation

- **Task**
  - Produce a headline in a target language (e.g., Chinese) given a document in a different source language (e.g., English).
- **Problem**
  - **Lack of those parallel corpora** of direct source language articles and target language headlines,
  - **Error propagation** in the translation and summarization phases.

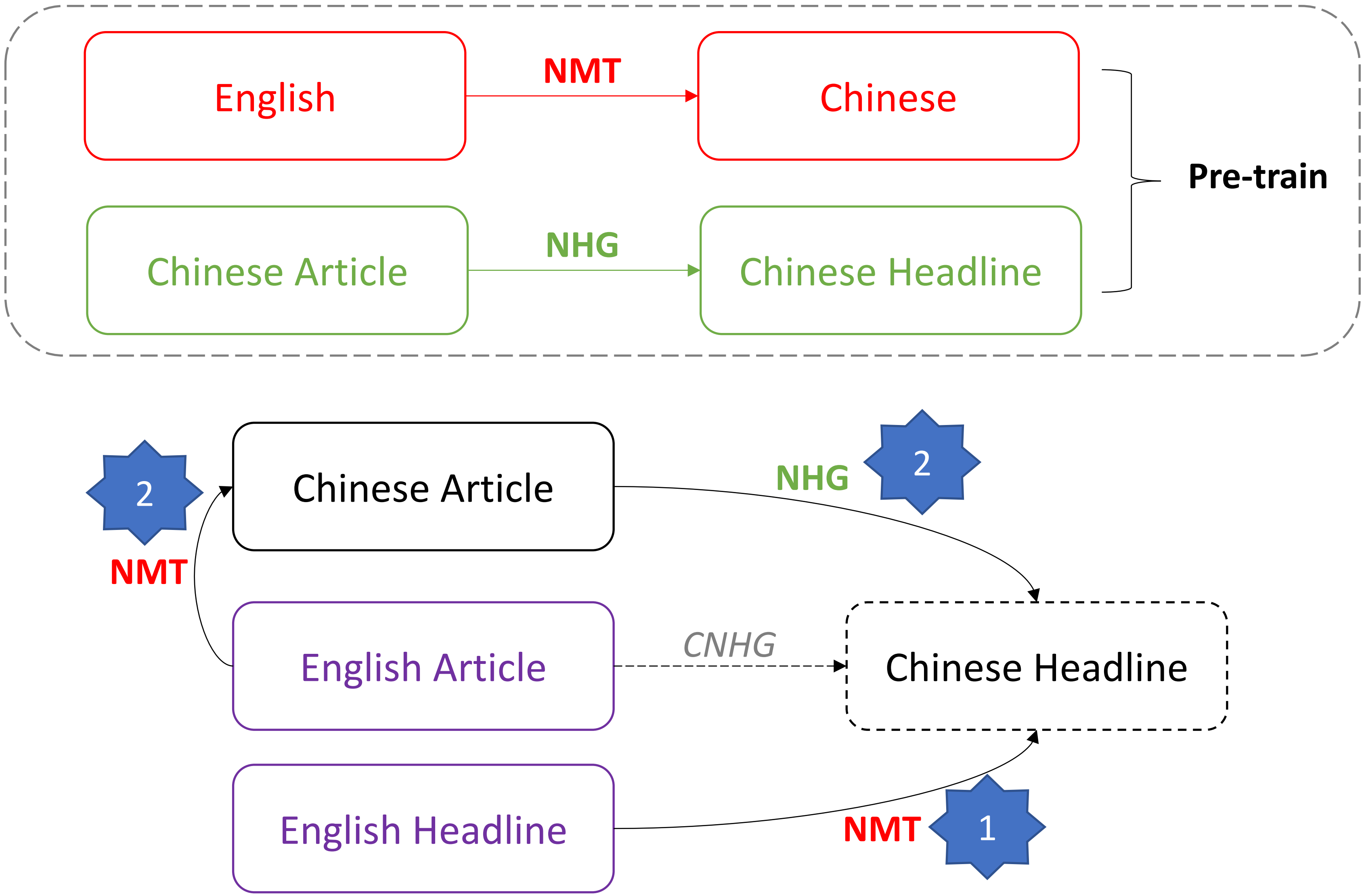
Asian-Pacific summit faces major economic and political challenges
亚 太 首脑 会议 面临 重大 经济 和 政治 挑战

The last time the Asia-Pacific region held its annual summit to promote free trade, Japan's prime minister assured everyone that his economy wouldn't be the next victim of Asia's financial crisis …

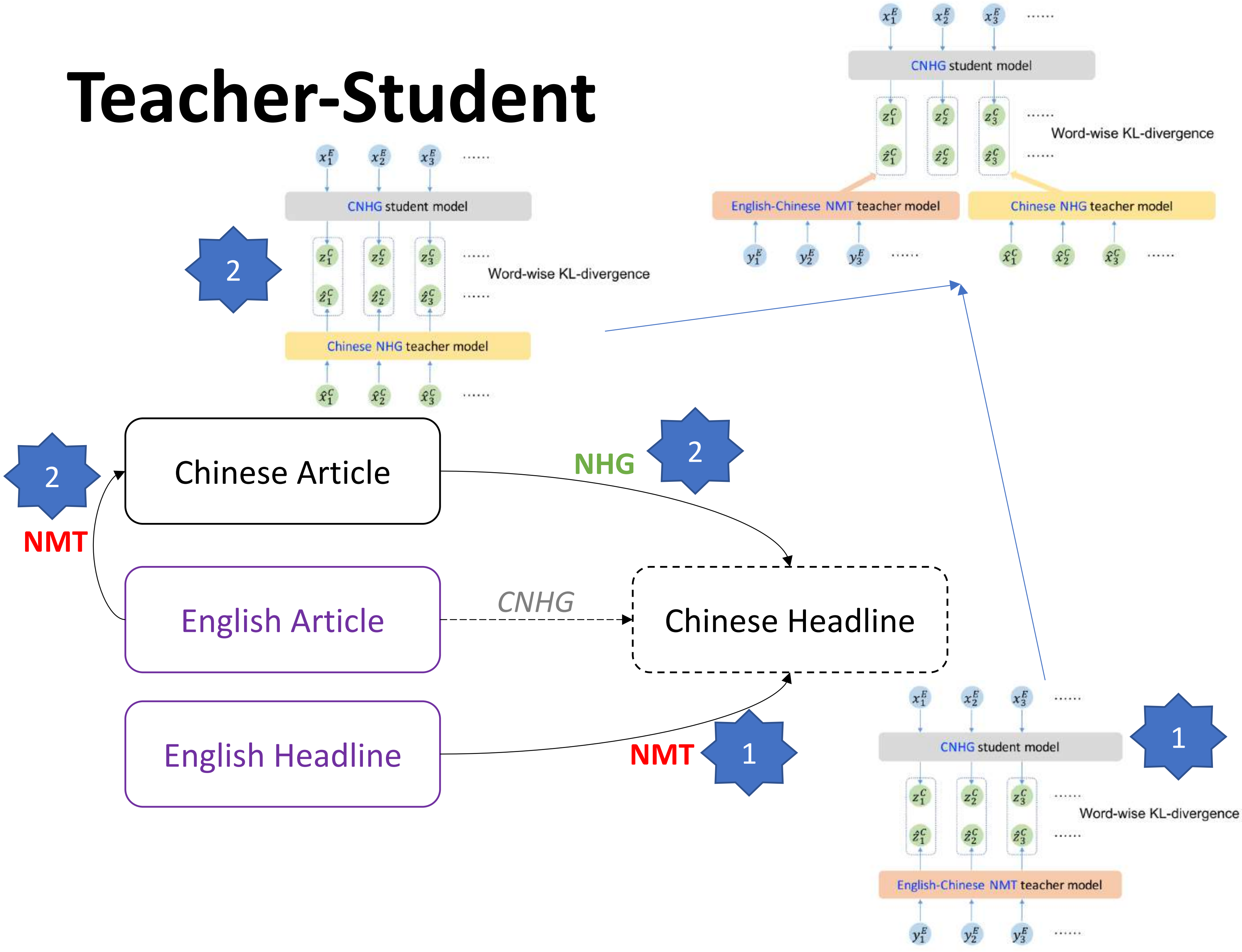
Corpus

- **English headline generation**
  - Gigaword
- **Chinese headline generation**
  - LCSTS
- **English-Chinese translation**
  - LDC2002E18, LDC2003E07, LDC2003E14, part of LDC2004T07, LDC2004T08 and LDC2005T06.

Model


English
NMT
Chinese
Chinese Article
NHG
Chinese Headline
Pre-train
Chinese Article
NHG
2
2
NMT
English Article
CNHG
Chinese Headline
English Headline
NMT
1

Teacher-Student



Outline

- **Why Knowledge Distillation?**
- **Distilling the knowledge in a neural network *NIPS2014***
- **Model Compression**
  - Distilling Task-Specific Knowledge from BERT into Simple Neural Networks *arxiv 2018*
- **Multi-Task Setting**
  - Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding *arxiv*
  - BAM! Born-Again Multi-Task Networks for Natural Language Understanding
- **Seq2Seq NMT**
  - Sequence level knowledge distillation *EMNLP16*
- **Cross Lingual NLP**
  - Cross-lingual Distillation for Text Classification *ACL17*
  - Zero-Shot Cross-Lingual Neural Headline Generation *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018*
- **Variant**
  - Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection *AAAI19*
- **Paper List**
- **Reference**
- **Conclusion**

Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection

AAAI19

Jian Liu , Yubo Chen , Kang Liu

National Laboratory of Pattern Recognition, Institute of Automation

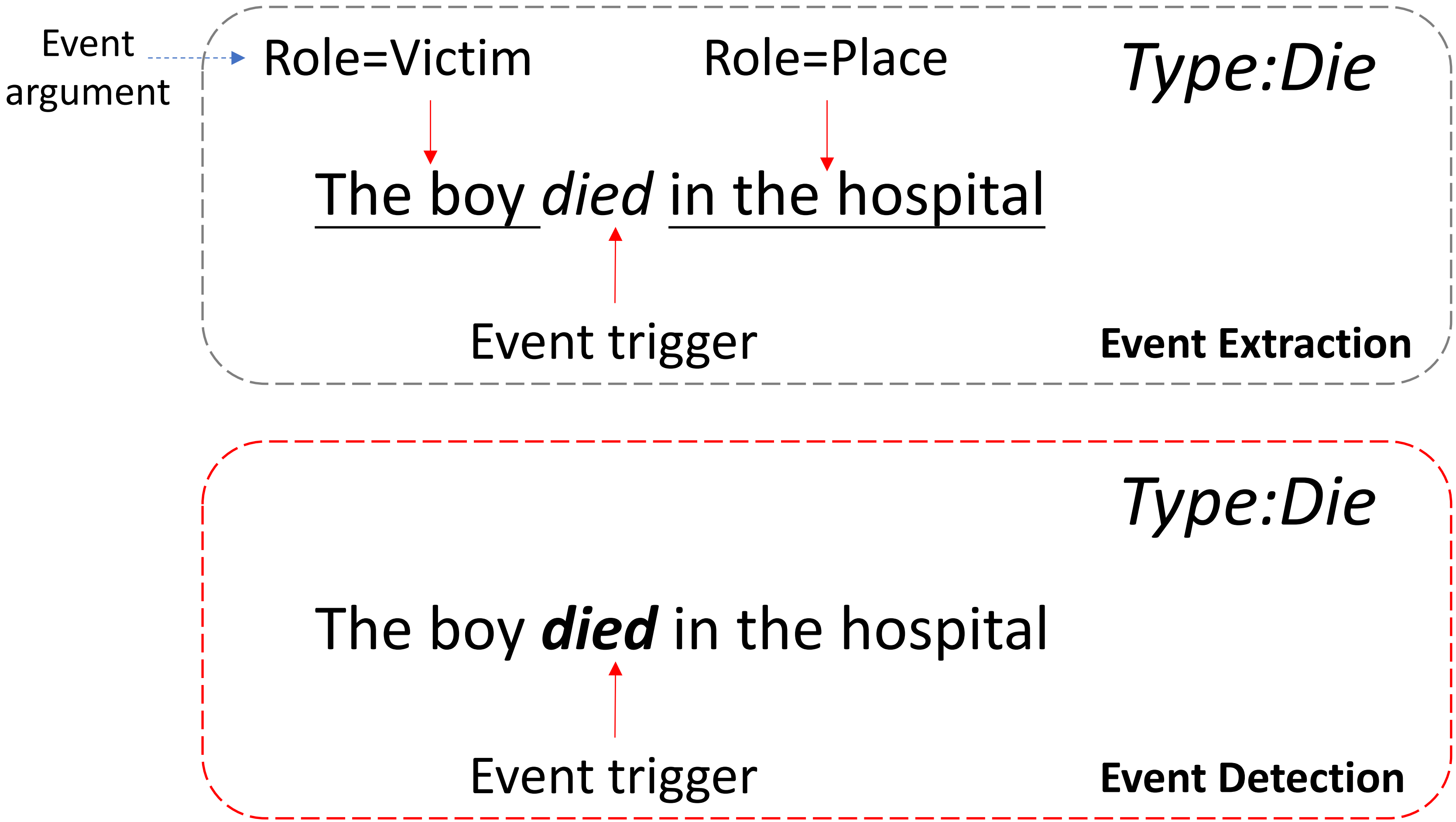
Author

陈玉博
Associate Professor
2017 赵军
Event Extraction , Relation Extraction and Knowledge Graph Construction .

刘康
Associate Professor
Sentiment Analysis, Information Extraction, Question Answering

Event Detection

- Event Detection ∈ Event Extraction



Problem

- **Ambiguity**
  - The same event can be expressed in a wide variation
  - Depending on the context, the same expression might refer to entirely different events.

Transfer-Money

*S*1: *The European Unit* is set to **release** *20 million* euros to Iraq.

*S*2: The government reports that *Anwar* 's earliest **release** date is *April 14*.

Release-Parole

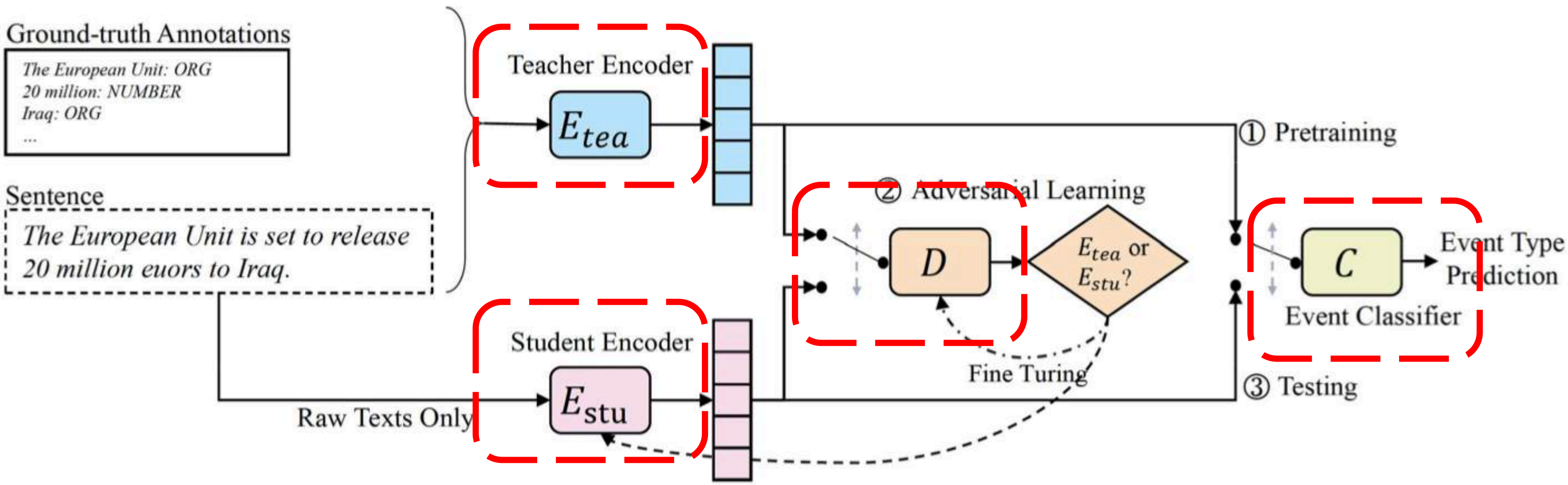
Previous

- Chunk knowledge corresponding to the sentences can provide evidence for event type disambiguation

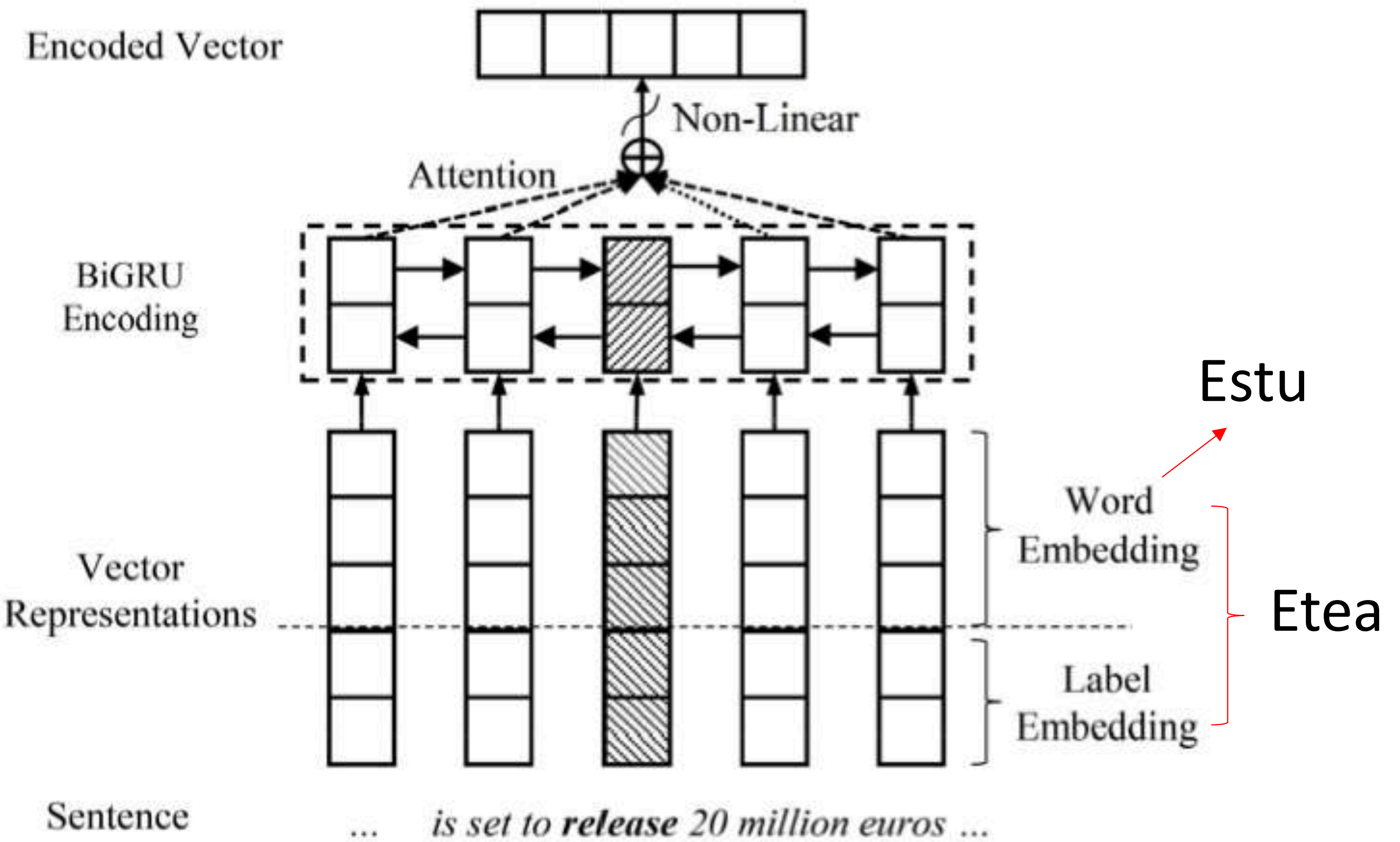
The European Unit: ORG
20 million: NUMBER
Iraq: ORG
...

- Problem
  - In the real test scenario where the ground-truth annotations are missing.
  - Pipline Error propagation

Model


Ground-truth Annotations
The European Unit: ORG
20 million: NUMBER
Iraq: ORG
...
Sentence
The European Unit is set to release 20 million euors to Iraq.
Raw Texts Only
Teacher Encoder
E_{tea}
Student Encoder
E_{stu}
① Pretraining
② Adversarial Learning
D
E_{tea} or E_{stu}?
Fine Turing
③ Testing
C
Event Classifier
Event Type Prediction

Attention Based Encoder



Binary Classification-Based Discriminator

- Input

$$f^{(w_t)} \text{ (either } f_{tea}^{(w_t)} \text{ or } f_{stu}^{(w_t)}\text{)}$$

- Output
  - A probability *p* that indicates the probability that D thinks f(wt) comes from Etea .

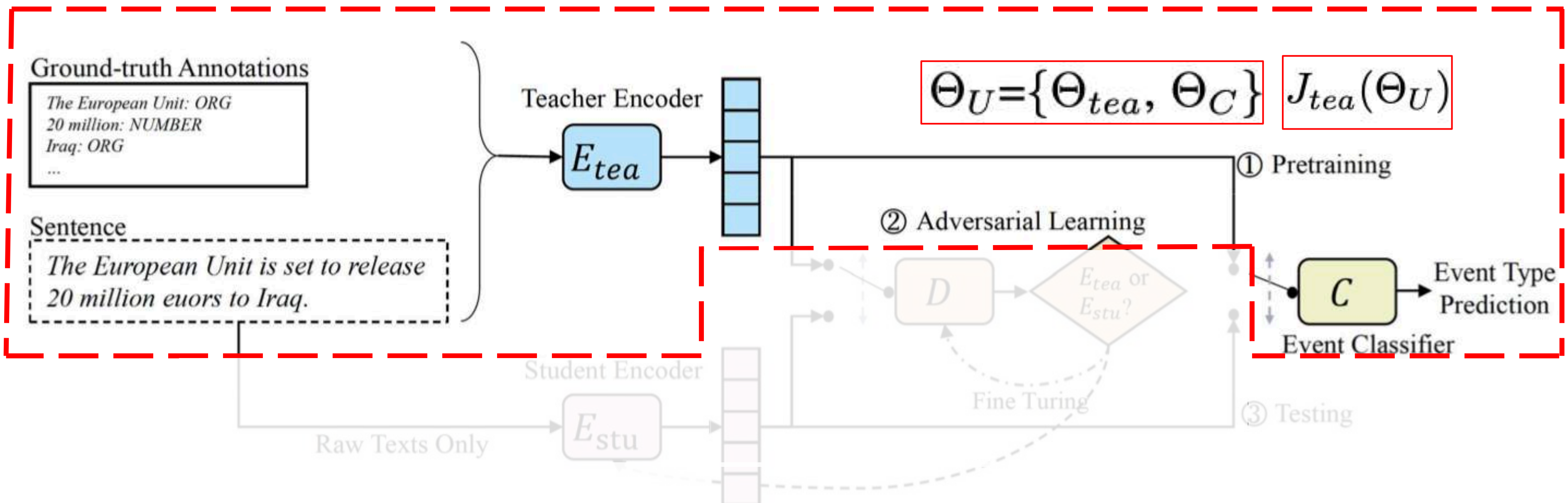
$$p = D(f^{(w_t)}) = \sigma(W_h(tanh(W_x f^{(w_t)} + b_x)) + b_h)$$

Multi-class Event Classifier

$$out = softmax(W_o \cdot f^{(w_t)} + b_o))$$
$$P(l|f, \Theta) = out_{(l)}$$

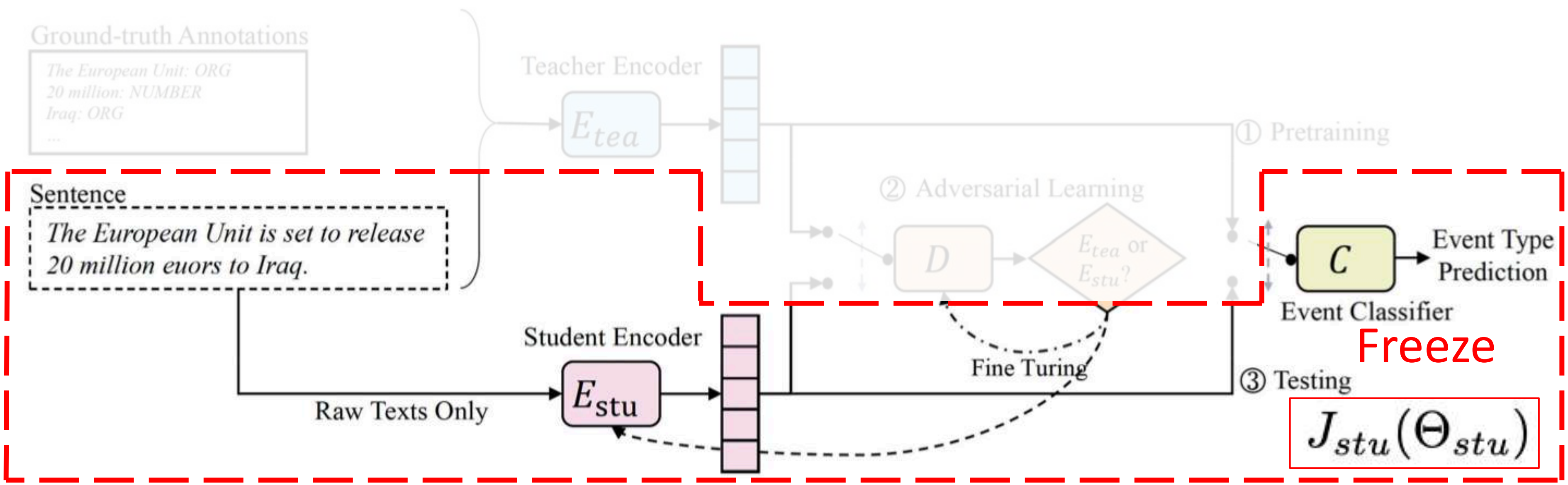
The Adversarial Imitation Strategy

- In the **Pretraining Stage:**
  - concatenate Etea and C to form an event detector



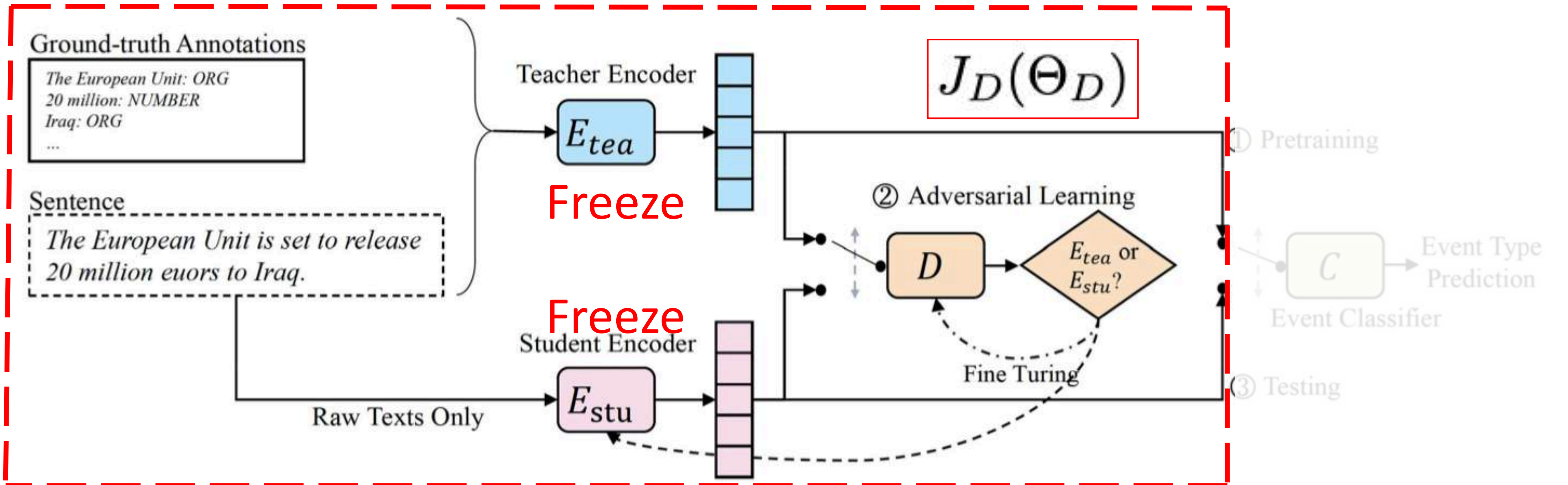
The Adversarial Imitation Strategy

- In the **Pretraining Stage:**
  - concatenate Etea and C to form an event detector
  - freeze the event classifier C, and we concatenate Estu and C to build a raw-sentences event detector.



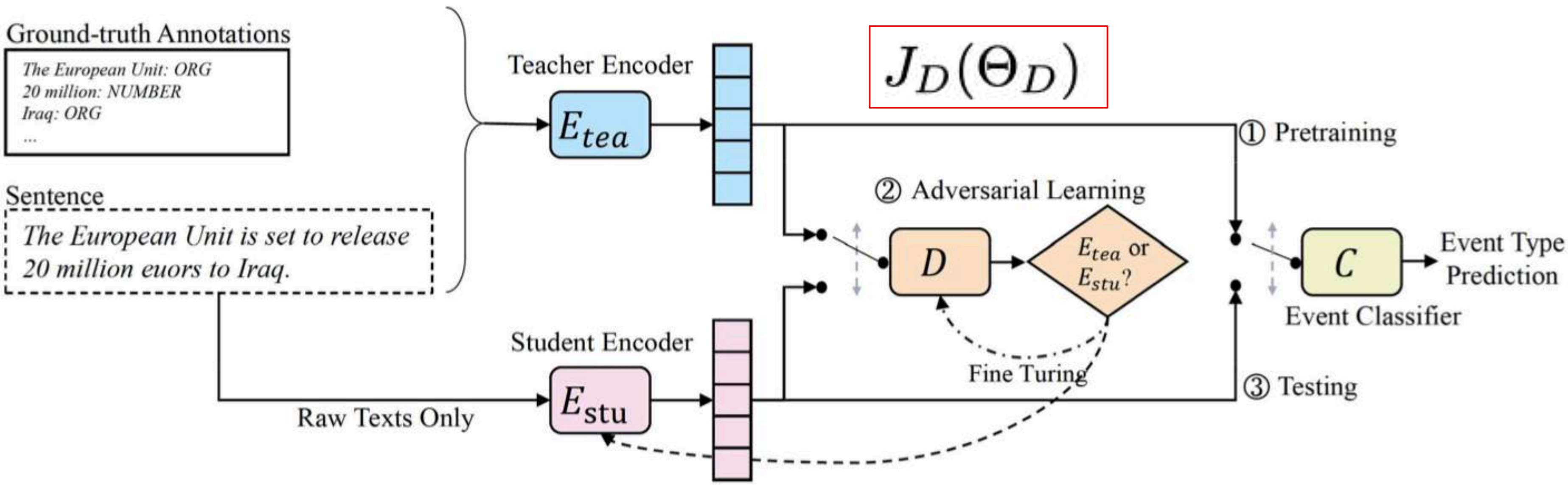
The Adversarial Imitation Strategy

- In the **Pretraining Stage:**
  - concatenate Etea and C to form an event detector
  - freeze the event classifier C, and we concatenate Estu and C to build a raw-sentences event detector.
  - freeze both Etea and Estu , outputs of Etea as positive examples (labeled as 1s) and the outputs of Estu as negative examples (labeled as 0s) to pretrain D.



The Adversarial Imitation Strategy

- In the **Adversarial learning** Stage



$$J_{stu\_adv}(\Theta_{stu}) = J_{stu}(\Theta_{stu}) + \lambda * J_{adv}(\Theta_{stu})$$

whether Estu has successfully fooled D

final classification error

$$J_{adv}(\Theta_{stu}) = \sum_{k=1}^{K} log(1 - D(f_{stu}^{(w_k)})) = -\sum_{k=1}^{K} log(D(f_{stu}^{(w_k)}))$$

Experiments

- ACE 2005 corpus
- 34-class classification problem （33+None）

Performance on Gold-truth Annotations

Model	**P**	**R**	**F_1**
CrossEntity (Hong et al.)	72.9	64.3	68.3
CNNED (Nguyen and Grishman)	71.8	66.4	69.0
DLRNN (Duan, He, and Zhao)	77.2	64.9	70.5
ArgATT (Liu et al.)	**78.0**	66.3	71.7
Teacher + emb word	71.9	66.0	68.8
Teacher + emb + ety entity	71.6	69.1	70.3
Teacher + emb + agt event-argument	76.3	72.4	74.2
Teacher + emb + ety + agt	76.8	**72.9**	**74.8**

Table 1: Experimental results on the ACE 2005 English set. Bold indicates the best performance with respective to each evaluation metric.

Performance in the Real Testing Scenario

Setting	Model	P	R	F_1
Golden	*CNNED*‡	71.8	66.4	69.0
	ArgATT‡	78.0	66.3	71.7
	Teacher	76.8	72.9	74.8
LSTM-CRF taggers Predicted	*CNNED*‡	71.9	63.8	67.6
	ArgATT	**76.1**	66.0	70.7
	Teacher	72.4	68.9	70.6
Adv	*Student-Final*	73.4	**69.1**	**71.2**

Table 2: Experimental results on ACE 2005 English corpus. *Golden/Predicted* means resorting to golden/predicted annotations. ‡ indicates taken from the original paper. Bold indicates the best performance.

Paper List

Knowledge Distillation

Paper	Conference
Distilling Task-Specific Knowledge from BERT into Simple Neural Networks	
BAM! Born-Again Multi-Task Networks for Natural Language Understanding	
Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding	
Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection	AAAI19
Distilling Knowledge for Search-based Structured Prediction	ACL18
On-Device Neural Language Model based Word Prediction	COLING18
Zero-Shot Cross-Lingual Neural Headline Generation	IEEE/ACM TRANSACTIONS 18
Cross-lingual Distillation for Text Classification	ACL17
DOMAIN ADAPTATION OF DNN ACOUSTIC MODELS USING KNOWLEDGE DISTILLATION	ICASSP17
Sequence-Level Knowledge Distillation	EMNLP16
Distilling Word Embeddings: An Encoding Approach	CIKM16
Distilling the Knowledge in a Neural Network	NIPS14 Deep Learning Workshop

Reference

1. WHAT IS KNOWLEDGE DISTILLATION? https://data-soup.gitlab.io/blog/knowledge-distillation/
2. 李如【DL】模型蒸馏 Distillation https://zhuanlan.zhihu.com/p/71986772
3. Towser 如何评价BERT模型https://www.zhihu.com/question/298203515/answer/509923837
4. 霍华德 BERT模型在NLP中目前取得如此好的效果，那下一步NLP该何去何从？ https://www.zhihu.com/question/320606353/answer/658786633
5. Andrej Karpathy A Recipe for Training Neural Networks http://karpathy.github.io/2019/04/25/recipe/
6. XLNet训练成本6万美元，顶5个BERT，大模型「身价」惊人 https://zhuanlan.zhihu.com/p/71609636?utm_source=wechat_session&utm_medium=social&utm_oi=71065644564480&from=timeline&isappinstalled=0&s_r=0
7. Naiyan Wang https://www.zhihu.com/question/50519680/answer/136363665
8. 周博磊 https://www.zhihu.com/question/50519680/answer/136359743
9. https://blog.csdn.net/qq_22749699/article/details/79460817
10. 如何理解soft target这一做法？Yjango https://www.zhihu.com/question/50519680?sort=created
11. Jiatao Gu Non-Autoregressive Neural Machine Translation https://zhuanlan.zhihu.com/p/34495294

Thanks!