EMNLP-IJCNLP 2019

Patient Knowledge Distillation for BERT Model Compression

Siqi Sun, Yu Cheng, Zhe Gan and Jingjing Liu
Microsoft Dynamics 365 AI Research

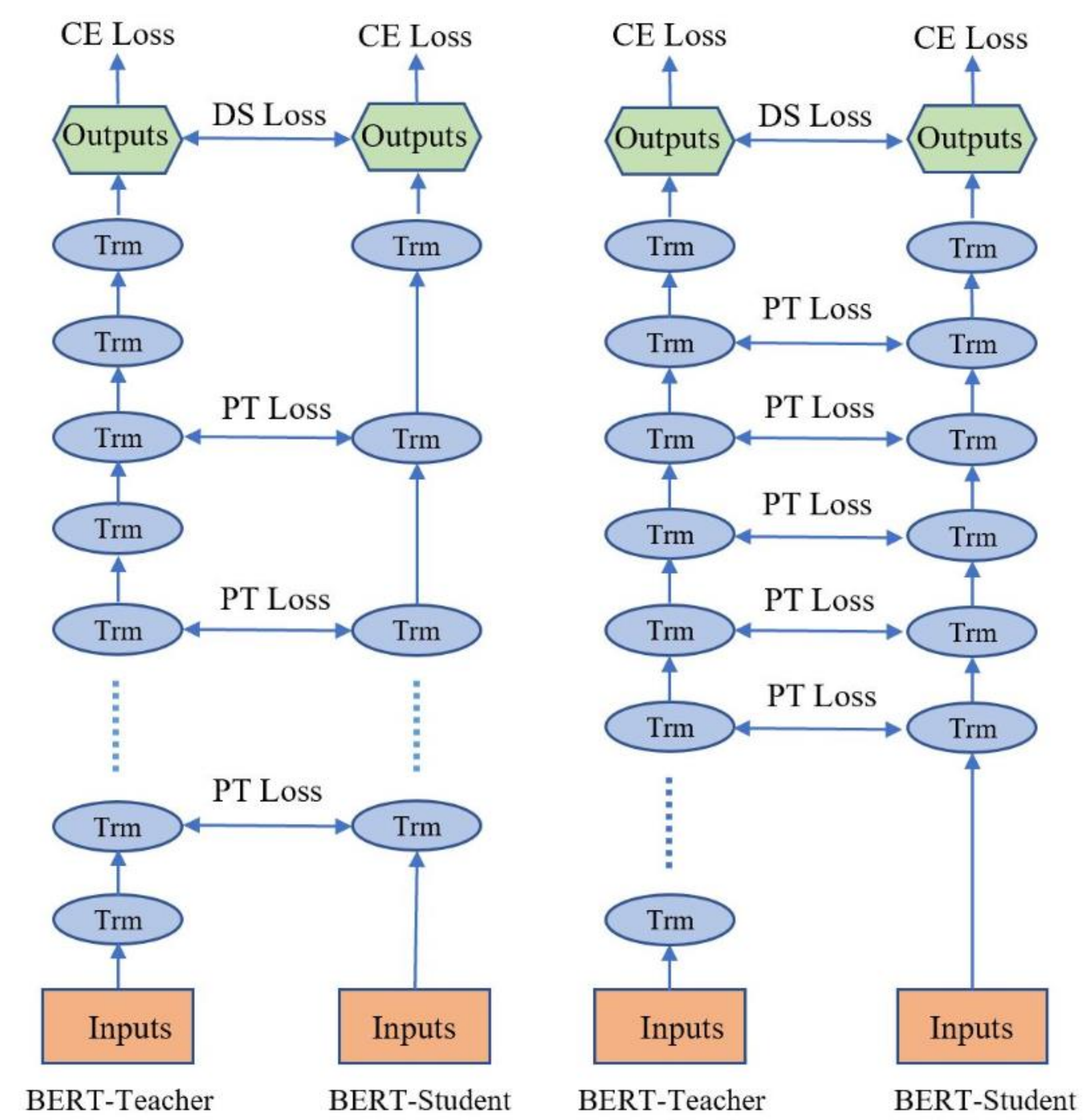
Motivation

- Pre-trained language model, such as BERT, has proven to be highly effective for downstream NLP tasks
- However, the high demand for computing resources during model training hinders their application in practice
- Knowledge Distillation (KD) is proven to be useful for model compression in previous work
- We propose Patient Knowledge Distillation, which learns knowledge from previous layers of the teacher network, and is more generalizable and effective than vanilla KD

Notations

- BERT-Teacher: BERT with 12 or 24 layers fine-tuned on downstream tasks
- BERT-Student: Transformer with 3 or 6 layers to be learned from the Teacher and downstream tasks
- CE-Loss: Cross-entropy loss
- DS-Loss: Distillation loss between teacher's and student's soft labels
- Embedding of [CLS]: $\mathbf{h}_i = [\mathbf{h}_{i,1}, \mathbf{h}_{i,2}, \ldots, \mathbf{h}_{i,k}] = \text{BERT}_k(\mathbf{x}_i) \in \mathbb{R}^{k \times d}$
- PT Loss on [CLS]: $L_{PT} = \sum_{i=1}^{N} \sum_{j=1}^{M} \left|\left| \frac{\mathbf{h}^s_{i,j}}{||\mathbf{h}^s_{i,j}||_2} - \frac{\mathbf{h}^t_{i,I_{pt}(j)}}{||\mathbf{h}^t_{i,I_{pt}(j)}||_2} \right|\right|_2^2$

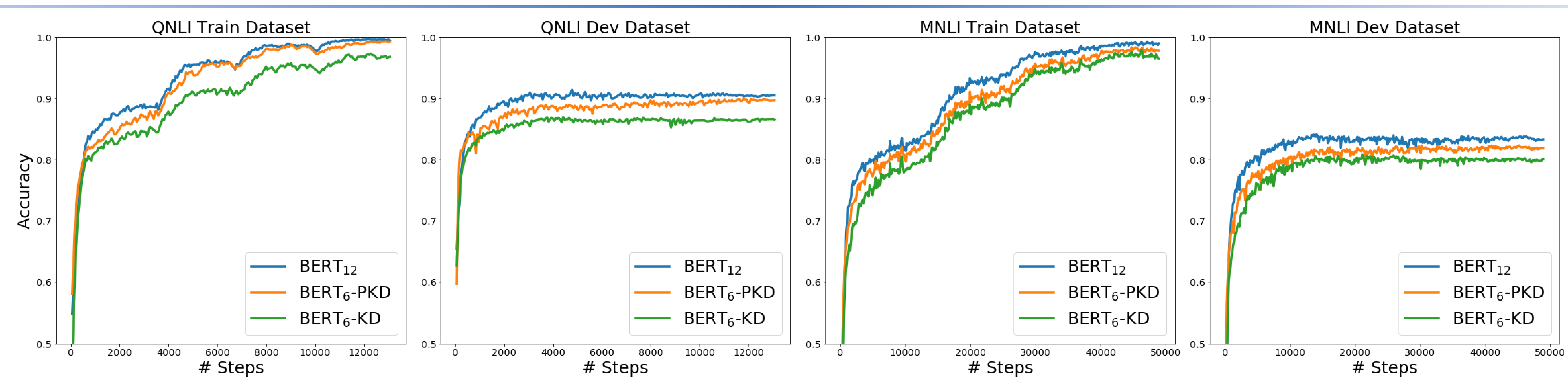
Patient Knowledge Distillation



(a) PKD-*Skip* (b) PKD-*Last*

- PKD-Skip: the Student learns the Teacher's outputs in ***every*** *T* layers
- PKD-Last: the Student learns the Teacher's outputs from the ***last*** *T* layers
- Final Loss: linear combination of task-specific CE loss, normal DS loss and proposed PT loss

Learning Curves on the Training and Dev sets of QNLI and MNLI



- Learning curves on QNLI and MNLI, two large-scale NLI datasets, where the Student network learned with vanilla KD ***quickly saturates*** on the dev set, while the proposed Patient-KD starts to plateau ***only in a later stage***

Experimental Results

Model	SST-2	MRPC (3.7k)	QQP (364k)	MNLI-m (393k)	MNLI-mm (393k)	QNLI (105k)	RTE (2.5k)
BERT_{12} (Google)	93.5	88.9/84.8	71.2/89.2	84.6	83.4	90.5	66.4
BERT_{12} (teacher)	94.3	89.2/85.2	70.9/89.0	83.7	82.8	90.4	69.1
BERT_6–FT	90.7	85.9/80.2	69.2/88.2	80.4	79.7	86.7	63.6
BERT_6–KD	91.5	**86.2/80.6**	70.1/88.8	80.2	79.8	88.3	64.7
BERT_6–PKD	**92.0**	85.0/79.9	**70.7/88.9**	**81.5**	**81.0**	**89.0**	**65.5**
BERT_3–FT	86.4	80.5/**72.6**	65.8/86.9	74.8	74.3	84.3	55.2
BERT_3-KD	86.9	79.5/71.1	67.3/87.6	75.4	74.8	84.0	56.2
BERT_3-PKD	**87.5**	**80.7**/72.5	**68.1/87.8**	**76.7**	**76.3**	**84.7**	**58.2**

- KD ***improves*** direct fine-tuning (FT)
- PKD-Skip almost always ***outperforms*** vanilla KD
- 6-layer Student trained via PKD performs ***comparable*** to Teacher on larger datasets
  - SST-2 (-2.3%), QQP (-0.1%), MNLI-m (-2.2%), MNLI-mm (-1.8%), and QNLI(-1.4%))

Model	SST-2	MRPC	QQP	MNLI-m	MNLI-mm	QNLI	RTE
BERT_6-PKD-Last	91.9	85.1/79.5	70.5/88.9	80.9	**81.0**	88.2	65.0
BERT_6-PKD-Skip	**92.0**	**85.0/79.9**	**70.7/88.9**	**81.5**	**81.0**	**89.0**	**65.5**

- PKD-Skip ***performs better*** than PKD-Last

Setting	Teacher	Student	SST-2	MRPC	QQP	MNLI-m	MNLI-mm	QNLI	RTE
N/A	N/A	BERT_{12}	94.3	89.2/85.2	70.9/89.0	83.7	82.8	90.4	69.1
N/A	N/A	BERT_{24}	94.3	88.2/84.3	71.9/89.4	85.7	84.8	92.2	72.8
#1	BERT_{12}	BERT_6[Base]-KD	91.5	86.2/80.6	70.1/88.8	79.7	79.1	88.3	64.7
#2	BERT_{24}	BERT_6[Base]-KD	91.2	86.1/80.7	69.4/88.6	80.2	79.7	87.5	65.7
#3	BERT_{24}	BERT_6[Large]-KD	89.6	79.0/70.0	65.0/86.7	75.3	74.6	83.4	53.7
#4	BERT_{24}	BERT_6[Large]-PKD	89.8	77.8/68.3	67.1/87.9	77.2	76.7	83.8	53.2

#1 vs. #2: there is ***not much difference*** between the Student's performance when changing teacher from BERT-Large to BERT-Base

#2 vs. #3: BERT_6 [Large] Student has 1.6 times more parameters than BERT_6[Base], but it ***performs much worse***

#3 vs. #4: PKD-Skip ***outperforms*** KD, which indicates PKD is a generic approach independent of the selection of the Teacher model

Initialization Mismatch

- Ideally, we should use pre-trained 6-layer BERT as initialization
- We are using ***first 6 layers*** of BERT-Base and BERT-Large because of computation limitation
- The first six layers of BERT-Large may not be able to capture high-level features, leading to worse KD performance