Dynamic Knowledge Distillation for Pre-trained Language Models

2022 01 30

주세준

- Dynamic Teacher Adoption
- Dynamic Data Selection
- Dynamic Supervision Adjustment

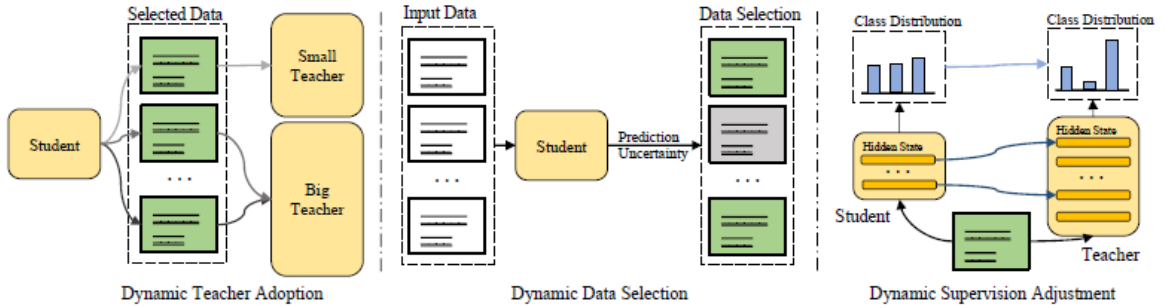


Figure 1: The three aspects of dynamic knowledge distillation explored in this paper. Best viewed in color.

Dynamic Teacher Adoption

- Larger teacher model is not necessarily better
  - The predicted logits of the teacher model become **less soft** as the teacher model becomes larger and more confident about its prediction
  - The **capacity gap** between the teacher and student model increases as the teacher becomes larger.

Method	RTE	IMDB	CoLA	Avg.
$\text{BERT}_{\text{BASE}}$	67.8	89.1	54.2	70.4
$\text{BERT}_{\text{LARGE}}$	72.6	90.4	60.1	74.4
No KD	63.7	86.3	39.0	63.0
KD w/ $\text{BERT}_{\text{BASE}}$	64.9	86.9	39.4	63.7
KD w/ $\text{BERT}_{\text{LARGE}}$	64.5	86.5	38.2	63.1
KD w/ Ensemble	64.9	86.7	39.9	63.8
Uncertainty-Hard	**66.9***	86.3	**42.7***	**65.3**
Uncertainty-Soft	66.4*	**87.1***	41.0	64.8

Table 1: We find that bigger teacher with better performance raises a worse student model. Results are average of 3 seeds on the validation set. * denotes statistically significant improvement over the best performing baseline with $p < 0.05$.

Dynamic Teacher Adoption

- Teacher selection
  - Hard selection
    - Instances in one batch are sorted according to uncertainty (entropy)
    - Evenly divided into two batches (confident, less confident)
    - Less confident queried by small teacher
    - Confident queried by large teacher
  - Soft selection

$$u_x = \text{Entropy}\left(\sigma\left(S\left(x\right)\right)\right)$$

$$w_1 = \frac{u_x}{U}, \quad w_2 = 1 - \frac{u_x}{U}$$

$$\mathcal{L}_{KD} = w_1 \mathcal{L}_{KL}^{T_1} + w_2 \mathcal{L}_{KL}^{T_2}$$

Method	RTE	IMDB	CoLA	Avg.
$\text{BERT}_{\text{BASE}}$	67.8	89.1	54.2	70.4
$\text{BERT}_{\text{LARGE}}$	72.6	90.4	60.1	74.4
No KD	63.7	86.3	39.0	63.0
KD w/ $\text{BERT}_{\text{BASE}}$	64.9	86.9	39.4	63.7
KD w/ $\text{BERT}_{\text{LARGE}}$	64.5	86.5	38.2	63.1
KD w/ Ensemble	64.9	86.7	39.9	63.8
Uncertainty-Hard	**66.9***	86.3	**42.7***	**65.3**
Uncertainty-Soft	66.4*	**87.1***	41.0	64.8

Table 1: We find that bigger teacher with better performance raises a worse student model. Results are average of 3 seeds on the validation set. * denotes statistically significant improvement over the best performing baseline with $p < 0.05$.

Dynamic Data Selection

3 types of uncertainty score

- choose the top $N \times r$ instances to query the **teacher model**
  (r = selection ratio controlling the number to query)

- Entropy

$$u_x = -\sum_y P(y \mid x) \log P(y \mid x) .$$

- Margin

$$u_x = P(y_1^* \mid x) - P(y_2^* \mid x) .$$

- Least-Confidence(LC)

$$u_x = 1 - P(\hat{y} \mid x)$$

Dataset	# Train	# Aug Train	# Dev	# Test	# Class
SST-5	8.8k	176k	1.1k	2.2k	5
IMDB	20k	400k	5k	25k	2
MNLI	393k	786,0k	20k	20k	3
MRPC	3.7k	74k	0.4k	1.7k	2
RTE	2.5k	50k	0.3k	3k	2
CoLA	8.5k	170k	1k	1k	2

Table 3: Statistics of datasets. **# Aug Train** denotes the number of the augmented training dataset following Jiao et al. (2020).

Method	#FLOPs	SST-5	IMDB	MRPC	MNLI-m / mm
$\text{BERT}_{\text{BASE}}$ (Teacher)	-	52.0	89.1	86.8	84.0 / 84.4
Vanilla KD	45.1B	47.4	86.8	80.2	81.7 / 82.0
Random	22.6B	46.8	86.4	79.7	81.4 / 81.6
Uncertainty-Entropy	28.2B	46.7	86.8	79.4	81.5 / 82.0
Uncertainty-Margin	28.2B	46.6	86.8	79.4	81.4 / 81.9
Uncertainty-LC	28.2B	46.5	86.8	79.4	81.4 / 81.9
Δ	-	- 0.6	0.0	- 0.5	- 0.2 / 0.0

Table 2: Dynamic data selection results with r set to 0.5. Results are averaged of 3 seeds on the validation set. Δ denotes the minimal performance degradation of different selection strategies compares to vanilla KD.

But no performance enhancement
→ Tiny dataset need augmentation

Dynamic Data Selection

Augment data

- Can sufficiently cover the possible data space

uncertainty-based selection strategy

- can **maintain the superior performance** while saving

the **computational cost**

Method	#FLOPs	SST-5	IMDB	MRPC	MNLI-m/ mm	Avg. (↑)	Δ (↓)
$BERT_{BASE}$ (Teacher)	-	53.7	88.8	87.5	83.9/ 83.4	79.5	-
TinyBERT†	24.9B	-	-	86.4	82.5/ 81.8	-	-
TinyBERT	24.9B	51.4	87.6	86.2	82.6/ 82.0	78.0	0.0
Random	2.49B	51.1	87.0	83.3	80.8/ 80.5	76.5	1.5
Uncertainty-Entropy	4.65B	51.5	**87.7**	**86.5**	**81.8**/ 81.0	**77.7**	**0.3**
Uncertainty-Margin	4.65B	**51.6**	**87.7**	**86.5**	81.6/ **81.1**	**77.7**	**0.3**
Uncertainty-LC	4.65B	51.2	**87.7**	**86.5**	81.4/ 80.8	77.5	0.5

Table 4: Test results when the selection ratio $r = 0.1$ for dynamic data selection on various tasks. #FLOPs denotes the average computational cost of KD for each instance. † denotes results from Jiao et al. (2020).

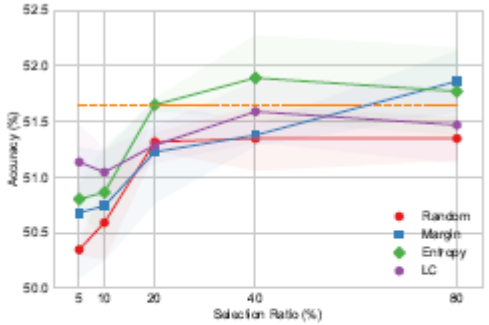


Figure 3: We plot the mean accuracy on the validation set of 3 seeds (± one standard deviation) under different selection ratios of various strategies. Orange dashed line denotes the performance of vanilla KD.

Dynamic Supervision Adjustment

$$\mathcal{L}_{KD} = \lambda_{KL} * \mathcal{L}_{KL} + \lambda_{PT} * \mathcal{L}_{PT}$$

$$\mathcal{L}_{PT} = \sum_{i=1}^{M} \left\| \frac{\mathbf{h}_i^s}{\|\mathbf{h}_i^s\|_2} - \frac{\mathbf{h}_{I_{pt}(j)}^t}{\left\|\mathbf{h}_{I_{pt}(j)}^t\right\|_2} \right\|_2^2$$

$$\lambda_{KL} = \lambda_{KL}^*(1 - \frac{u_x}{U}), \quad \lambda_{PT} = \lambda_{PT}^* \frac{u_x}{U}$$

L_{pt} *PaTient loss*
measures the alignment between normalized internal representations

M
num of student layer

$I_{pt}(i)$
corresponding alignment of teacher layer for the student i-th layer

h_i^s, h_i^t
i-th layer represenatation of student and teacher

λ_{kl}^*, U
Predefined weights by parameter search for each objective
Normalization factor

Method	SST-5	MRPC	RTE	Avg.
$\text{BERT}_{\text{BASE}}$ (Teacher)	52.0	86.8	67.8	68.9
Vanilla KD	47.4	80.2	64.9	64.2
BERT-PKD	46.6	80.8	65.1	64.2
Uncertainty	**48.1**	**81.5***	**66.4***	**65.3**

Table 5: Results of dynamic adjusting the supervision weights, showing the uncertainty-based adjustment is effective. * denote results are statistically significant with $p < 0.05$.

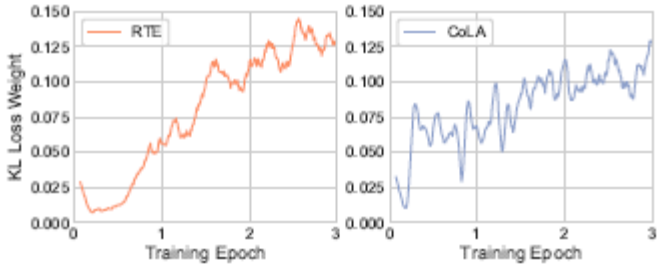
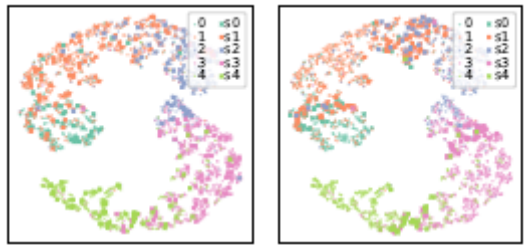


Figure 6: Evolution of the dynamically adjusted weight of KL-divergence loss weight.

corresponding weight
of the prediction probability alignment objective is increasing as the **student becomes more confident** about its predictions, thus paying more attention to **matching the output distribution** with the teacher model

Dynamic Data Selection



(a) Random (b) Margin

Figure 5: The t-SNE visualization of instance representations. Uncertainty-based strategies select the instances close to the class boundary, which is useful for the learning of the student model. Best viewed in color.

?

감사합니다