Over-parameterized Student Model via Tensor Decomposition Boosted Knowledge Distillation

Yu-Liang Zhan[1], Zhong-Yi Lu[2], Hao Sun[1*], Ze-Feng Gao[2*]

[1] Gaoling School of Artificial Intelligence， Renmin University of China

[2] School of Physics，Renmin University of China

Background&Motivation

The limitations of Large-scale pre-trained model

- The substantial storage demands and high computational complexity hinder the practical deployment of Large-scale pre-trained models in real-world applications.

Limited capacity of small student models

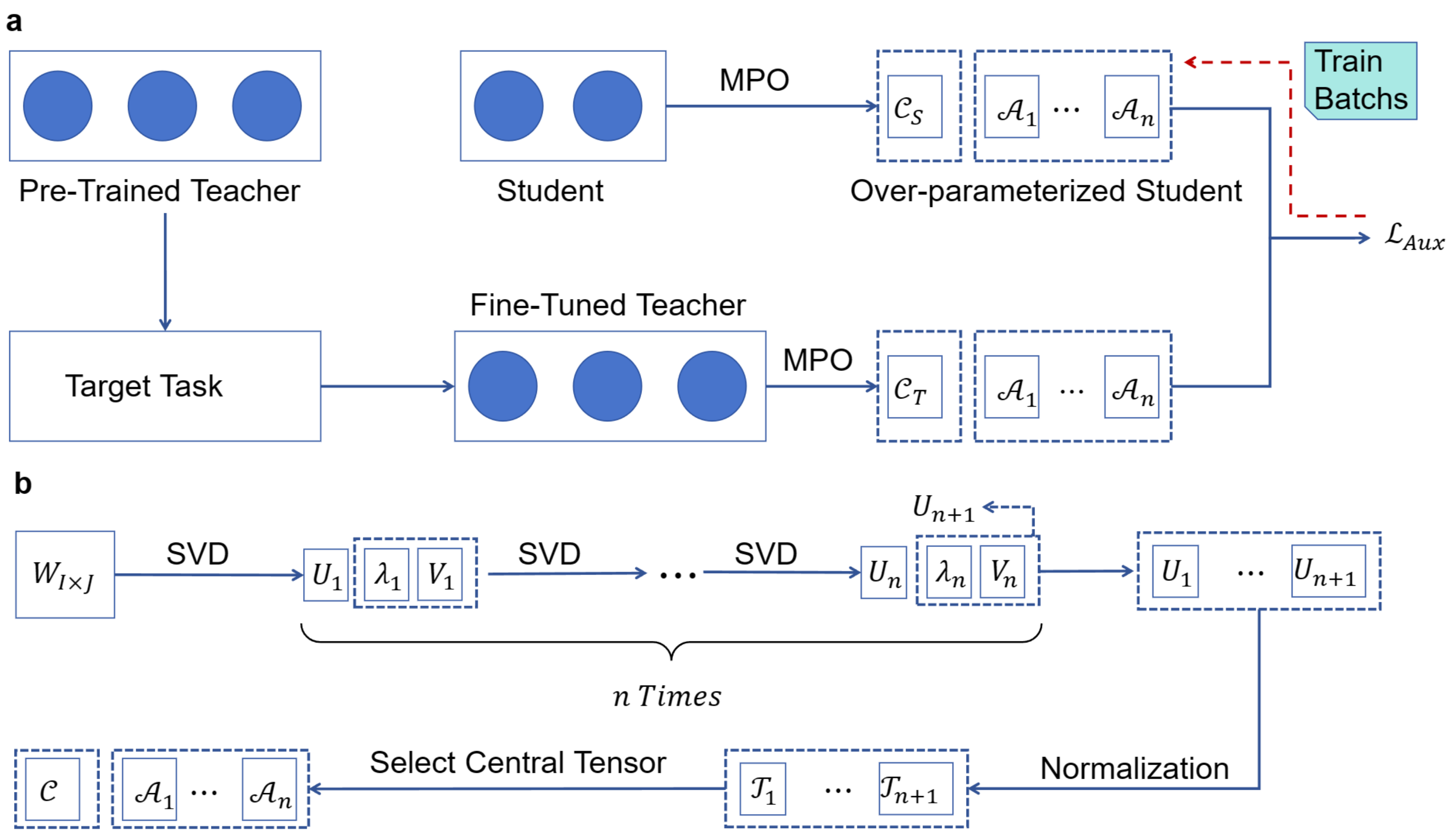
- Due to the lesser over-parameterization of small models compared to large ones, small student models' generalization capability often falls short, resulting in suboptimal fine-tuning performance on downstream tasks.

Major concerns of over-parameterizing student models

- The potential information loss caused by tensor decomposition should be minimized, as small computation errors may accumulate and propagate exponentially within the stacked layers of student models.

- There is no effective mechanism to ensure the consistency of information between student and teacher models in the over-parameterized student models.

➢ Method



- **Augmentation in parameter count :**

$$N_{add} = \sum_{k=1}^{m} i_k j_k d_{k-1} d_k - \prod_{k=1}^{m} i_k j_k$$

- **Aligning the auxiliary tensors:**

$$\mathcal{L}_{Aux} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{MSE}\left(\mathcal{A}_{s,k}, \mathcal{A}_{t,k}\right)$$

➢ Experiments

- NLP Tasks

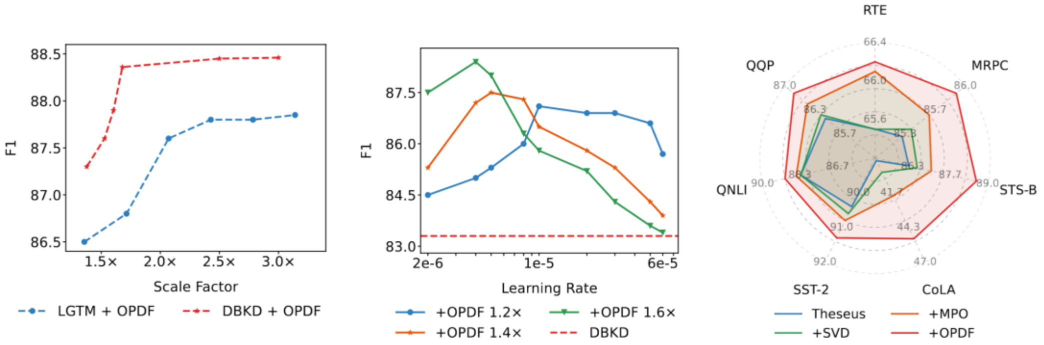
Datasets	RTE Acc.	MRPC F1/Acc.	STS-B Corr.	CoLA Mcc.	SST-2 F1/Acc.	QNLI Acc.	QQP F1/Acc.	MNLI Acc.	Avg.	# Train Params (M)	# Inference Params (M)
BERT-base [2]	70.5	86.5/81.8	86.6	54.2	92.0	91.2	88.0/91.0	84.2	83.4	110	110
BERT-of-Theseus [49]											
None	65.5	85.3/79.6	86.2	39.2*	90.4	88.7	86.1/89.6	**81.5**	79.2	66	66
+SVD	65.5	85.4/80.0	86.5	43.1	90.6	88.6	86.2/89.7	80.3	79.6	90	66
+OPDF (Ours)	**66.2**	**85.9/80.5**	**88.6**	**45.2**	**91.3**	**89.0**	**86.8/90.2**	81.4	**80.5**	160	66
LGTM [42]											
None	63.3	86.3/80.1	82.9*	33.9*	91.1	**89.3**	**88.0/91.1**	**82.2**	78.8	67	67
+SVD	64.7	86.8/81.9	83.1	37.4	91.2	88.6	86.5/89.4	79.3	78.9	91	67
+OPDF (Ours)	**66.9**	**87.8/82.4**	**83.3**	**38.9**	**91.5**	88.7	87.0/90.2	80.9	**79.8**	163	67
DBKD [43]											
None	61.2	83.3/75.5	/	25.2	88.1	86.1	85.3/88.7	76.1	74.4	53	53
+SVD	64.7	86.5/78.6	/	26.4	88.8	85.8	85.5/89.0	76.5	75.8	69	53
+OPDF (Ours)	**69.1**	**88.4/83.3**	/	**27.2**	**89.8**	**86.5**	**86.9/90.2**	**77.7**	**77.6**	83	53
AD-KD [44]											
None	68.8	88.7/84.3	**89.3**	53.1	**91.5**	90.8	85.9/89.5	81.7	82.4	67	67
+SVD	69.4	89.3/85.8	88.8	53.5	89.9	90.1	86.4/89.8	81.5	82.6	91	67
+OPDF (Ours)	**71.7**	**90.3/86.8**	88.9	**55.0**	91.3	**91.1**	**86.8/90.0**	**82.1**	**83.4**	182	67

- CV Tasks

Datasets	Imagenet-1k top-1	Imagenet-1k top-5	Imagenet Real top-1	Imagenet Real top-5	Imagenet V2 top-1	Imagenet V2 top-5	# Train Params (M)	# Inference Params (M)
CLIP-ViT-L/14[59]	84.8*	/	88.9*	/	75.1*	/	321	321
TinyVit-5M[58]								
None	77.4*	94.1*	86.1*	97.5*	66.8*	87.6*	5.4	5.4
+SVD	77.9	95.1	86.3	97.3	68.7	88.4	7.6	5.4
+OPDF	**80.0**	**96.7**	**87.4**	**98.1**	**69.4**	**88.9**	9.9	5.4
TinyVit-11M								
None	80.5*	95.6*	87.8*	98.0*	70.7*	90.4*	11	11
+SVD	82.0	96.7	88.4	97.9	71.7	91.4	17	11
+OPDF	**82.5**	**96.9**	**88.9**	**98.3**	**72.4**	**92.6**	23	11
TinyVit-21M								
None	82.3*	96.3*	88.9*	98.3*	73.0*	91.9*	21	21
+SVD	82.9	96.8	88.3	97.8	71.8	92.4	29	21
+OPDF	**84.0**	**97.5**	**89.4**	**98.4**	**74.9**	**93.4**	38	21

- ◆ OPDF can enhance the performance of the distillation model without increasing the inference latency .
- ◆ There are inherent limits to the benefits that can be achieved through over-parameterization in knowledge distillation models.
- ◆ The performance of the model with OPDF consistently remains at least as high as that of the original method.
- ◆ Learning rate decrease as the over-parameterization scale increases.
- ◆ All components in our approach are effective.

- Further Analysis



(a) Impact of scale factor (b) Impact of learning rate (c) Impact of OPDF components

Over-parameterized Student Model via Tensor Decomposition Boosted Knowledge Distillation

Thank You!

For further details, feel free to get in touch with us.

zhanyuliang@ruc.edu.cn