Sparse Teachers Can Be Dense with Knowledge

Yi Yang#, Chen Zhang#, Dawei Song*

Beijing Institute of Technology, #equal contributions, *corresponding author

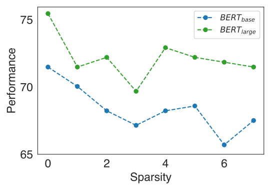
BACKGROUND & MOTIVATION

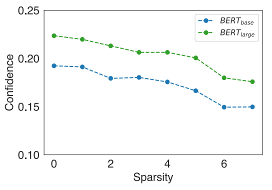
Background

- Recent advances in distilling pretrained language models have discovered that, the student-friendliness should be taken into consideration besides the expressiveness of knowledge to realize a truly knowledgeable teacher.

Motivation

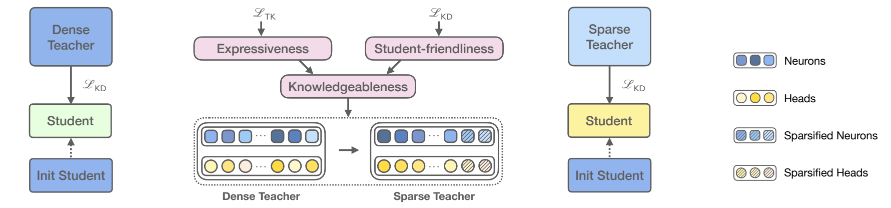
- From a pilot study, we find that LMs of large scale tend to have a good performance and high confidence, and that both performance and confidence can be degraded through randomly sparsifying a small portion of parameters. This indicates that some parameters resulting in student-unfriendliness can be rather removed, to improve student-friendliness of the teacher without sacrificing too much its expressiveness.
- Motivated by this finding, we propose a sparse teacher trick (in short, **STARK** 🤖) under the guidance of an overall knowledgeable score for each teacher parameter, which accords not only with the expressiveness but also the student-friendliness of the parameter by interpolation.





SPARSE TEACHER TRICK

Our trick involves three stages in the student learning procedure. First, we distil a trial student from the dense teacher on a specific task (***trial distillation***). Then, we sparsify the parameters of the dense teacher that are associated with adequately low knowledgeable scores (***parameter sparsification***). Finally, rewinding is applied, where the student is set to the initialization exactly used in the trial distillation stage and is learned from the sparse teacher during (***actual distillation***).



(a) Trial Distillation (b) Parameter Sparsification (c) Actual Distillation

Trial and Actual Distillation

- ***Trial distillation*** and ***actual distillation*** share the same distillation regime. We employ the widely used logits distillation as the distillation objective. The trial distillation and actual distillation also reuse the initialization of the student for better convergence.

$$\mathscr{L}_{\mathsf{KD}} = -\operatorname{softmax}(\mathbf{z}^t/\tau)\log\operatorname{softmax}(\mathbf{z}^s/\tau), \qquad \mathscr{L}_{\mathsf{TK}} = -\mathbf{y}\log\mathbf{y}^s, \qquad \mathscr{L} = \mathscr{L}_{\mathsf{KD}} + \alpha\cdot\mathscr{L}_{\mathsf{TK}}$$

$\mathscr{L}_{\mathsf{KD}}$ -- distillation loss $\mathscr{L}_{\mathsf{TK}}$ -- task loss $\mathscr{L}$ -- distillation objective

$\mathbf{y}^s, \mathbf{y}$ --prediction normalized probabilities of the student and ground-truth one-hot probabilities.

$\mathbf{z}^t, \mathbf{z}^s$ -- logits of the teacher and student τ -- temperature controlling the smoothness the logits

Parameter Sparsification

- We mainly sparsify the attention heads of MHA blocks and intermediate neurons of FFN blocks in the teacher. We attach a set of variables $\boldsymbol{\xi}^{(i)}$ and $\boldsymbol{\nu}$ to the attention heads and the intermediate neurons, to record the parameter sensitivities for a specific task through accumulated absolute gradients.

$$\mathrm{MHA}^{\circ}(\mathbf{X}) = \sum_{i=1}^{A} \xi^{(i)}\mathrm{Attn}(\mathbf{X}, \mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)})\mathbf{W}_O^{(i)}, \qquad \mathrm{FFN}^{\circ}(\mathbf{X}) = \mathrm{GELU}(\mathbf{X}\mathbf{W}_1)\mathrm{diag}(\nu)\mathbf{W}$$

- **Expressiveness**. The expressiveness of the teacher is tied to the expressiveness score. A higher expressiveness score indicates that the corresponding parameter has bigger contribution towards the performance. The expressiveness scores of the attention heads in MHA and the intermediate neurons in FFN can be depicted as:

$$\mathbb{P}_{\mathsf{head}}^{(i)} = \mathbb{E}_{\mathscr{D}}\left|\frac{\partial\mathscr{L}_{\mathsf{TK}}}{\partial\xi^{(i)}}\right|, \qquad \mathbb{P}_{\mathsf{neuron}}^{(i)} = \mathbb{E}_{\mathscr{D}}\left|\frac{\partial\mathscr{L}_{\mathsf{TK}}}{\partial\mathrm{diag}(\nu)}\right|$$

- **Student-friendliness**. Likewise, the student-friendliness of the teacher can be described as student-friendliness scores, which are approximated from distillation loss of the trial distillation.

$$\mathbb{Q}_{\mathsf{head}}^{(i)} = \mathbb{E}_{\mathscr{D}}\left|\frac{\partial\mathscr{L}_{\mathsf{KD}}}{\partial\xi^{(i)}}\right|, \qquad \mathbb{Q}_{\mathsf{neuron}}^{(i)} = \mathbb{E}_{\mathscr{D}}\left|\frac{\partial\mathscr{L}_{\mathsf{KD}}}{\partial\mathrm{diag}(\nu)}\right|$$

- **Knowledgeable scores.** we normalize the expressiveness and student-friendliness scores with ℓ2 norm. We introduce λ to quantify the tradeoff to balance the expressiveness and student-friendliness of teacher.

$$\mathbb{I}_{\mathsf{head}}^{(i)} = \lambda\mathbb{P}_{\mathsf{head}}^{(i)} + (1-\lambda)\mathbb{Q}_{\mathsf{head}}^{(i)}, \qquad \mathbb{I}_{\mathsf{neuron}} = \lambda\mathbb{P}_{\mathsf{neuron}} + (1-\lambda)\mathbb{Q}_{\mathsf{neuron}}$$

- ***Parameter sparsification*** sparsifies the parameters in the teacher with adequately low knowledgeable scores. The adequacy is met by enumerating diverse sparsity levels and obtaining the one leading to the best student during the actual distillation.

EXPERIMENTS

Datasets

- Experiments on GLUE benchmark that contains a collection of NLU tasks.
- We exclude CoLA on which general model distillation methods transfer knowledge poorly.

Baselines

- ***Student*** model is initialized by dropping 2/3 layers or pruning 70% parameters of the ***teacher*** model BERT
- **FT** which directly finetune the student, **KD**, **PKD**, **CKD** and **DynaBERT**.
- Student-friendly baselines: **TAKD** that employs a reasonable assistant, **MetaKD** that adapts the teacher with the student feedback, and **DKD** hat amplifies the student-friendly knowledge.

Main Comparison

- The best results on datasets are **boldfaced**. § is the optimal sparsity on each dataset. $*_4$ and $*_{30\%}$ mean the student is initialized by dropping 2/3 layers or pruning 70\% parameters of the teacher. **STARK₄** and **STARK₃₀%** exactly mean KD₄ and KD30\% w/STARK.

EXPERIMENTS

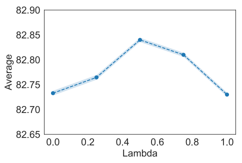
- We only report MetaKD on small datasets due to limited resources, and DynaBERT without data augmentation due to unavailable augmented data.

Method	MNLI-m Acc	MNLI-mm Acc	MRPC F1	QNLI Acc	QQP F1	RTE Acc	STSB SpCorr	SST-2 Acc	Average
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	93.8	86.8
layer-dropped student									
FT$_4$	77.5	77.7	86.0	85.3	86.1	65.0	86.5	89.5	81.7
KD$_4$	77.7	77.7	86.9	85.1	86.1	65.3	86.4	89.6	81.8
PKD$_4$	77.7	77.7	**87.6**	85.0	86.0	65.3	86.4	89.9	82.0
CKD$_4$	77.7	77.9	87.2	85.0	86.2	64.6	86.4	89.6	81.8
MetaKD$_4$	\	\	85.1	\	\	63.9	86.5	89.5	\
DKD$_4$	77.9	78.0	86.9	84.8	86.0	66.3	86.5	88.8	81.9
TAKD$_4$	77.1	77.3	87.2	84.5	86.3	**67.9**	86.7	89.9	82.1
STARK$_4$	**78.8**	**79.0**	87.4	**85.7**	**86.5**	67.5	**87.2**	**90.6**	**82.8**
§	40%	50%	50%	50%	30%	60%	40%	50%	46%
parameter-pruned student									
FT$_{30\%}$	82.0	82.6	88.5	89.5	87.7	69.0	87.2	91.9	84.8
KD$_{30\%}$	82.5	82.4	89.1	89.5	87.8	69.3	87.0	91.9	84.9
PKD$_{30\%}$	82.5	82.8	**89.5**	89.9	**88.0**	68.6	86.4	91.9	84.9
DynaBERT$_{30\%}$	81.5	82.8	87.4	89.1	86.6	68.1	87.2	90.3	84.1
DKD$_{30\%}$	82.4	82.4	88.4	89.6	87.7	**70.4**	87.0	91.9	85.0
TAKD$_{30\%}$	82.7	82.3	89.1	89.8	87.8	68.6	87.6	91.9	85.0
STARK$_{30\%}$	**82.8**	**82.9**	89.4	**90.0**	87.8	69.7	**87.9**	**92.2**	**85.3**
§	30%	20%	30%	70%	40%	20%	30%	40%	35%

- **Scalability**: the results of scalability to larger teachers and smaller students.

Method	MNLI-m Acc	MNLI-mm Acc	MRPC F1	QNLI Acc	QQP F1	RTE Acc	STSB SpCorr	SST-2 Acc	Average
BERT$_{\text{large}}$	86.6	86.1	92.3	92.2	89.0	75.5	89.9	93.9	88.2
KD$_8$	78.9	79.5	84.9	86.1	86.4	63.9	85.6	90.5	82.0
STARK$_8$	**79.4**	**80.5**	**85.0**	**86.3**	**87.0**	**65.7**	**88.7**	**90.9**	**82.9**
§	30%	20%	9%	10%	30%	60%	20%	20%	35%
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	93.8	86.8
KD$_{30\%}$	73.2	72.8	82.9	78.9	83.5	**58.5**	46.5	86.8	72.9
STARK$_{30\%}$	**73.9**	**74.3**	**83.1**	**80.4**	**83.8**	57.8	**48.6**	**88.1**	**73.7**
§	50%	50%	30%	50%	30%	40%	30%	40%	40%

- **Knowledgeableness Tradeoff** :the performance variation along with the change of λ.
- **Training Efficiency**: the training time consumed during trial distillation and actual distillation stages



Stage	Train time on MNLI
trial distillation	~2.5h
actual distillation	~7h

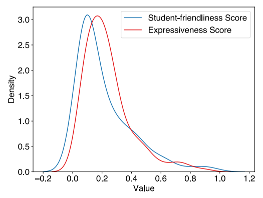
- **Pluggability: STARK** is pluggable to any distillation methods since it is orthogonal to existing paradigms.

Method	MNLI-m Acc	MNLI-mm Acc	MRPC F1	QNLI Acc	QQP F1	RTE Acc	STSB SpCorr	SST-2 Acc	Average
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	93.8	86.8
KD$_4$	77.7	77.7	86.9	85.1	86.1	65.3	86.4	89.6	81.8
w/ STARK	**78.8**	**79.0**	**87.4**	**85.7**	**86.5**	**67.5**	**87.2**	**90.6**	**82.8**
PKD$_4$	77.7	77.7	87.6	85.0	86.0	65.3	86.4	89.9	82.0
w/ STARK	**78.8**	**79.1**	**87.7**	**85.9**	**86.6**	**66.8**	**87.2**	**90.1**	**82.8**
CKD$_4$	77.7	77.9	87.2	85.0	86.2	64.6	86.4	89.6	81.8
w/ STARK	**78.8**	**79.0**	**87.6**	**86.4**	**86.5**	**66.4**	**87.2**	**90.4**	**82.8**

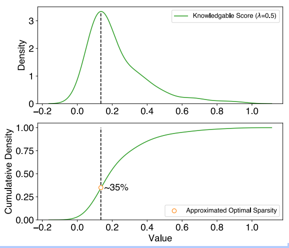
- **Unstructured Pruning: STARK** is capable of unstructured pruning. * indicates that unstructured pruning is otherwise used.

Method	MNLI-m Acc	MNLI-mm Acc	MRPC F1	QNLI Acc	QQP F1	RTE Acc	STSB SpCorr	SST-2 Acc	Average
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	93.8	86.8
KD$_4$	77.7	77.7	86.9	85.1	86.1	65.3	86.4	89.6	81.8
STARK$_4$	78.8	**79.0**	**87.4**	**85.7**	86.5	**67.5**	87.2	**90.6**	**82.8**
STARK$_4$*	**79.0**	**79.0**	**87.4**	85.3	**86.8**	66.1	**87.3**	89.8	82.6

- **Automatic STARK:** We explore an alternative algorithm to get the obtain optimal sparsity more efficiently. attentive solution is proposed based on a surprising observation that a sparse teacher under the guidance of randomness can achieve a promising Average score of 82.5%. This weird phenomenon drives us to put forward a proposition.
- **Assumption 1.** *Both expressiveness and student-friendliness scores are densely located at their clusters, where the cluster center of student-friendliness scores owns a smaller magnitude than that of expressiveness scores.* Density of expressiveness and student-friendliness scores of BERT$_{\text{base}}$ attention heads finetuned on MRPC



- **Assumption 2.** *An optimal sparsity is positively correlated to the first density peak of a sparsification sequence.*



Method	MNLI-m Acc	MNLI- mm	MRPC F1	QNLI Acc	QQP F1	RTE Acc	STSB SpCor	SST-2 Acc	Average
STARK$_4$	**78.8**	**79.0**	87.4	**85.7**	**86.5**	67.5	**87.2**	**90.6**	**82.8**
§	40%	50%	50%	50%	30%	60%	40%	50%	46%
STARK-AUTO$_4$	78.1	79.0	86.6	**85.7**	**86.0**	67.5	87.2	90.0	82.6
§	47%	51%	35%	46%	44%	42%	38%	38%	43%

CONCLUSIONS & LIMITATIONS

Conclusions

- We validate that sparse teachers can be dense with knowledge under the guidance of our designed knowledgeable score.
- The knowledgeable score is carefully crafted to make sure that the student-unfriendly knowledge can be reduced without hurting too much the expressive knowledge.
- Extensive experimental results on the GLUE benchmark support our claim to a large degree.

Limitations

- **STARK** can be further explored under two additional settings: 1) in a task-agnostic setting (e.g., MiniLM) and 2) on large LMs (e.g., BERT$_{\text{large}}$)
- Our attentive automatic solution for **STARK** can be enhanced.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers and chairs for their constructive suggestions. This research was supported in part by Natural Science Foundation of Beijing (grant number: 4222036) and Huawei Technologies (grant number: TC20201228005).

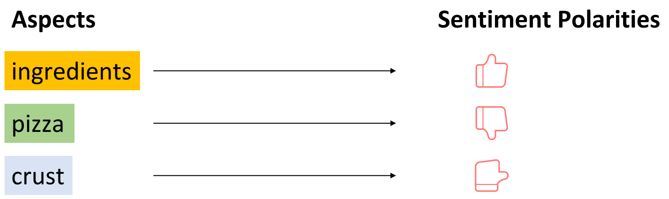
Syntax-Aware Aspect-Level Sentiment Classification with Proximity-Weighted Convolution Network

Chen Zhang[1], Qiuchi Li[2], Dawei Song[1]*

[1]Beijing Institute of Technology & Zhejiang Lab, [2]University of Padua. *Corresponding author

BACKGROUND

"They use fancy ingredients but even fancy ingredients don't make for good pizza unless someone knows how to get the crust right."



MOTIVATION

Limitations of the State of the Art

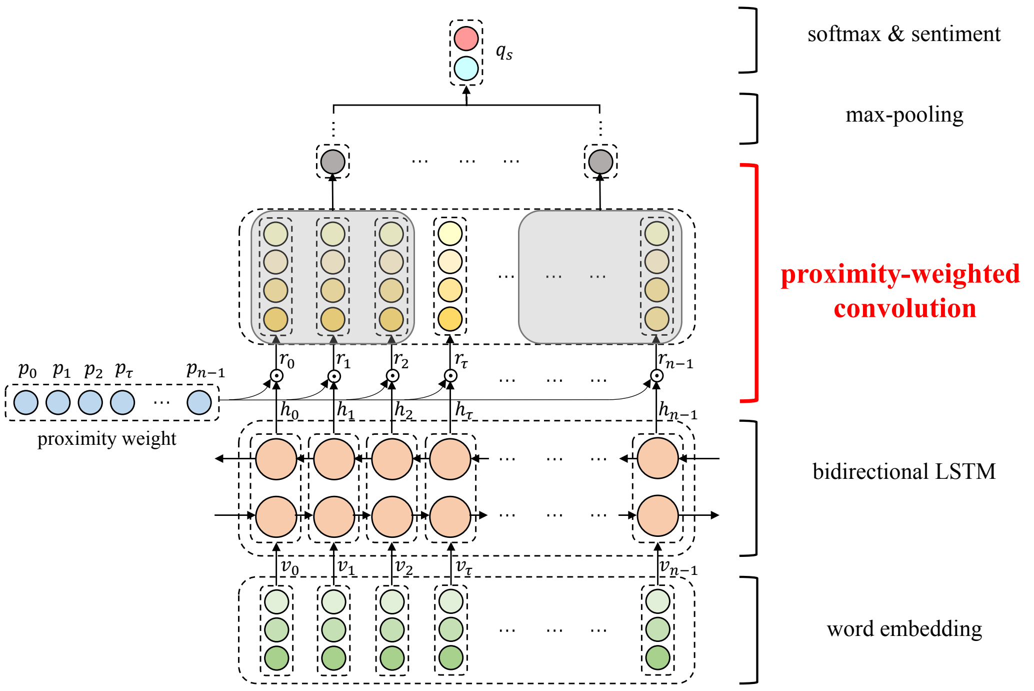
- **Syntax** has been generally neglected in aspect-level sentiment classification.
- The sentiment polarity of an aspect needs to be determined by key **phrases** instead of single **words**.

Research Questions

- How to capture syntactic information and n-gram level features in relation to aspects in a unified framework?
- Can syntactic information help improve aspect-level sentiment classification?
- Will n-gram level features work better than word level ones?

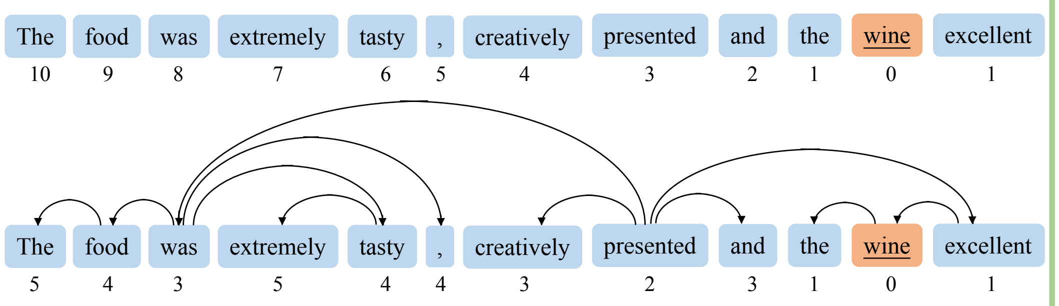
MODEL

Proximity-**W**eighted **C**onvolution **N**etwork (**PWCN**)


q_s
softmax & sentiment
max-pooling
proximity-weighted convolution
p_0 p_1 p_2 p_τ p_{n-1}
proximity weight
r_0 r_1 r_2 r_τ r_{n-1}
h_0 h_1 h_2 h_τ h_{n-1}
bidirectional LSTM
v_0 v_1 v_2 v_τ v_{n-1}
word embedding

Proximity weight -- Capturing Syntactic Information

- Position / dependency distance


The food was extremely tasty , creatively presented and the wine excellent
10 9 8 7 6 5 4 3 2 1 0 1
The food was extremely tasty , creatively presented and the wine excellent
5 4 3 5 4 4 3 2 3 1 0 1

- Position / dependency proximity weight

$$p_i = 1 - \frac{d_i}{n}$$

p_i -- proximity weight, d_i -- position / dependency distance, n -- sentence length.

Accordingly, we have two variants of **PWCN**: **PWCN-Pos** and **PWCN-Dep**

Proximity-weighted convolution -- Capturing n-gram level features

- Proximity weight assigning

$$r_i = h_i \cdot p_i$$

r_i -- proximity-weighted representation, h_i -- hidden states output by bidirectional LSTM.

- Convolution

EXPERIMEMTS

Datasets

- Experiments on two benchmarking datasets from **SemEval 2014**.
- The datasets consist of reviews and comments from two categories: **laptop** and **restaurant**, respectively.

Baselines

- **LSTM** only uses the last hidden state vector to predict sentiment polarity.
- **RAM** considers hidden state vectors of context as external memory and applies Gated Recurrent Unit (GRU) structure to multi-hop attention. The top-most representation is used for predicting polarity.
- **IAN** models attention between aspect and its context interactively with two LSTMs.
- **TNet-LF** leverages Context-Preserving Transformation to preserve and strengthen the informative part of context. It also benefits from a multi-layer architecture.

Variants of PWCN-Pos

- **Att-PWCN-Pos:** the proximity weight is **multiplied by the normalized attention weight**, to integrate semantic relatedness and syntax relationship.
- **Point-PWCN-Pos**: n-gram level convolution is set to **1-gram level**, which degrades the convolution process to point-wise (word level) feed-forward network.

Main results

Average accuracy and macro-F1 score over 3 runs with random initialization. The best results are in bold. The marker † refers to *p*-value < 0.05 when comparing with IAN, while the marker ‡ refers to *p*-value < 0.05 when comparing with TNet-LF. The relative increase over the LSTM baseline is given in bracket.

Model	Laptop		Restaurant	
	Acc	Macro-F1	Acc	Macro-F1
LSTM	69.63	63.51	77.99	66.91
RAM	72.81(+4.57%)	68.59(+8.00%)	79.89(+2.44%)	69.49(+3.86%)
IAN	71.63(+2.87%)	65.94(+3.83%)	78.59(+0.77%)	68.41(+2.24%)
TNet-LF	75.16(+7.94%)	71.10(+11.95%)	80.20(+2.83%)	70.78(+5.78%)
Att-PWCN-Pos	72.92(+4.72%)	68.14(+7.29%)	80.15(+2.77%)	70.17(+4.87%)
Point-PWCN-Pos	74.45(+6.92%)	69.47(+9.38%)	80.00(+2.58%)	69.93(+4.51%)
PWCN-Pos	75.23†(+8.17%)	70.71†‡(+11.34%)	**81.12**†‡(4.01%)	71.81†(+7.32%)
PWCN-Dep	**76.12**†‡(+9.32%)	**72.12**†(+13.56%)	80.96†(+3.81%)	**72.21**†(+7.92%)

Impact of Syntax

Visualization of a case with respect to *food*

Method	Visualization	Prediction
Att.	great food but the service was dreadful !	Negative
Pos.	great food but the service was dreadful !	Positive
Dep.	great food but the service was dreadful !	Positive

CONCLUSIONS AND FUTURE WORK

Conclusions

- We have proposed a framework that leverages n-gram information and syntactic dependency between aspect and contextual terms into an applicable model.
- Experimental results have demonstrated the effectiveness of our proposed models and suggested that syntactic dependency is more beneficial to aspect-level sentiment classification than semantic relatedness.
- N-gram level features are more significant than word level ones in aspect-level sentiment classification.

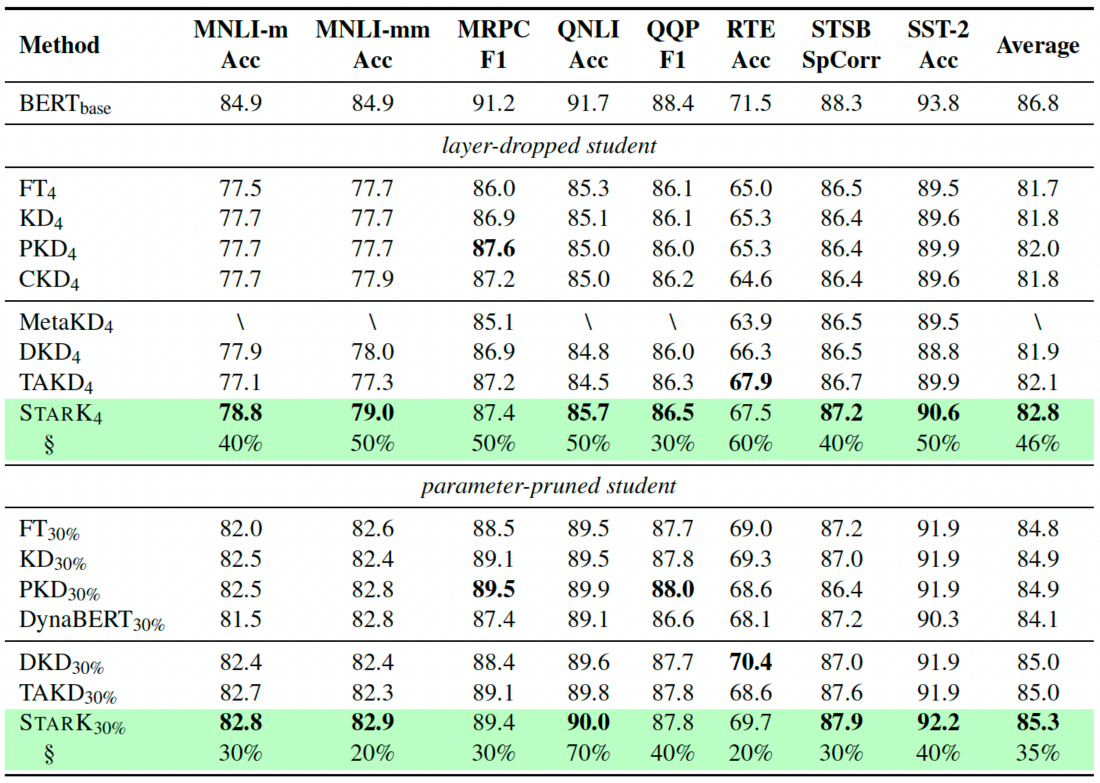
Future work

- In-depth analysis of the difference between PWCN models and attention-based models to achieve a deep understanding of where the syntactical dependencies overwhelm semantic relatedness.

ACKNOWLEDGEMENTS

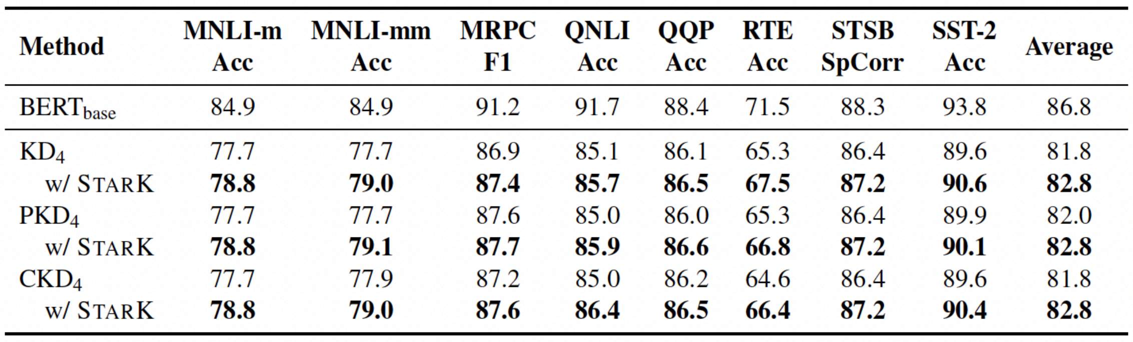
- This work is supported by The National Key Research and Development Program of China (grant No. 2018YFC0831700), Natural Science Foundation of China (grant No. U1636203), and Major Project Program of Zhejiang Lab (grant No. 2019DH0ZX01).
- Co-author Chen Zhang has received Travel Grants from ACM SIGIR.

42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19) dwsong@bit.edu.cn



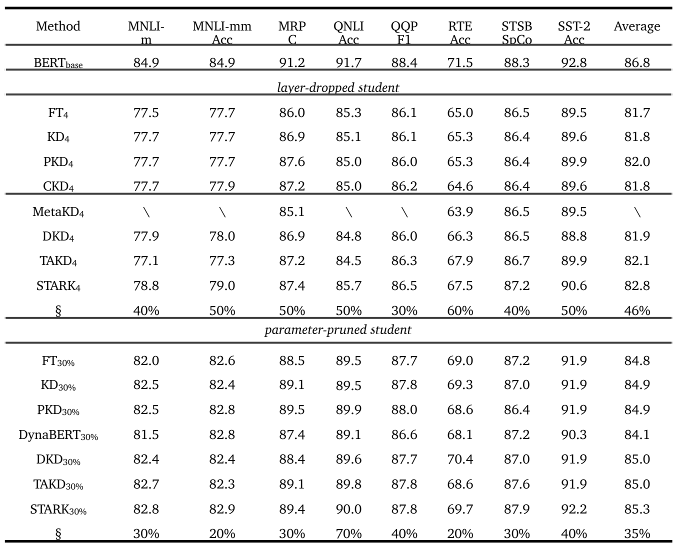
Method	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	STSB	SST-2	Average
	Acc	Acc	F1	Acc	F1	Acc	SpCorr	Acc	
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	93.8	86.8
layer-dropped student									
FT$_4$	77.5	77.7	86.0	85.3	86.1	65.0	86.5	89.5	81.7
KD$_4$	77.7	77.7	86.9	85.1	86.1	65.3	86.4	89.6	81.8
PKD$_4$	77.7	77.7	**87.6**	85.0	86.0	65.3	86.4	89.9	82.0
CKD$_4$	77.7	77.9	87.2	85.0	86.2	64.6	86.4	89.6	81.8
MetaKD$_4$	\	\	85.1	\	\	63.9	86.5	89.5	\
DKD$_4$	77.9	78.0	86.9	84.8	86.0	66.3	86.5	88.8	81.9
TAKD$_4$	77.1	77.3	87.2	84.5	86.3	**67.9**	86.7	89.9	82.1
STARK$_4$	**78.8**	**79.0**	87.4	**85.7**	**86.5**	67.5	**87.2**	**90.6**	**82.8**
§	40%	50%	50%	50%	30%	60%	40%	50%	46%
parameter-pruned student									
FT$_{30\%}$	82.0	82.6	88.5	89.5	87.7	69.0	87.2	91.9	84.8
KD$_{30\%}$	82.5	82.4	89.1	89.5	87.8	69.3	87.0	91.9	84.9
PKD$_{30\%}$	82.5	82.8	**89.5**	89.9	**88.0**	68.6	86.4	91.9	84.9
DynaBERT$_{30\%}$	81.5	82.8	87.4	89.1	86.6	68.1	87.2	90.3	84.1
DKD$_{30\%}$	82.4	82.4	88.4	89.6	87.7	**70.4**	87.0	91.9	85.0
TAKD$_{30\%}$	82.7	82.3	89.1	89.8	87.8	68.6	87.6	91.9	85.0
STARK$_{30\%}$	**82.8**	**82.9**	89.4	**90.0**	87.8	69.7	**87.9**	**92.2**	**85.3**
§	30%	20%	30%	70%	40%	20%	30%	40%	35%

Table 2: The results of main comparison on GLUE development set. The best results on datasets are **boldfaced**. § is
udent is initialized by dropping 2/3 layers or pruning

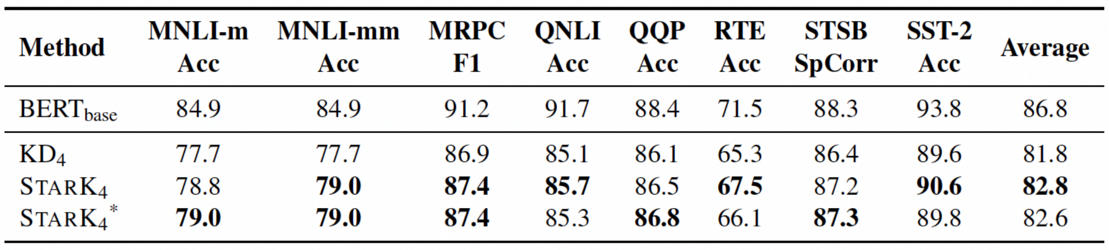


Method	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	STSB	SST-2	Average
	Acc	Acc	F1	Acc	Acc	Acc	SpCorr	Acc	
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	93.8	86.8
KD$_4$	77.7	77.7	86.9	85.1	86.1	65.3	86.4	89.6	81.8
w/ STARK	**78.8**	**79.0**	**87.4**	**85.7**	**86.5**	**67.5**	**87.2**	**90.6**	**82.8**
PKD$_4$	77.7	77.7	87.6	85.0	86.0	65.3	86.4	89.9	82.0
w/ STARK	**78.8**	**79.1**	**87.7**	**85.9**	**86.6**	**66.8**	**87.2**	**90.1**	**82.8**
CKD$_4$	77.7	77.9	87.2	85.0	86.2	64.6	86.4	89.6	81.8
w/ STARK	**78.8**	**79.0**	**87.6**	**86.4**	**86.5**	**66.4**	**87.2**	**90.4**	**82.8**

Table 5: The results of pluggability to baselines.

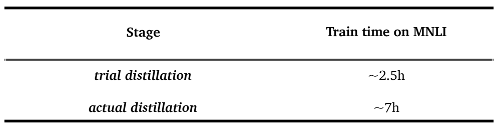


Method	MNLI-	MNLI-mm	MRP	QNLI	QQP	RTE	STSB	SST-2	Average
	m	Acc	C	Acc	F1	Acc	SpCo	Acc	
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	92.8	86.8
layer-dropped student									
FT$_4$	77.5	77.7	86.0	85.3	86.1	65.0	86.5	89.5	81.7
KD$_4$	77.7	77.7	86.9	85.1	86.1	65.3	86.4	89.6	81.8
PKD$_4$	77.7	77.7	87.6	85.0	86.0	65.3	86.4	89.9	82.0
CKD$_4$	77.7	77.9	87.2	85.0	86.2	64.6	86.4	89.6	81.8
MetaKD$_4$	\	\	85.1	\	\	63.9	86.5	89.5	\
DKD$_4$	77.9	78.0	86.9	84.8	86.0	66.3	86.5	88.8	81.9
TAKD$_4$	77.1	77.3	87.2	84.5	86.3	67.9	86.7	89.9	82.1
STARK$_4$	78.8	79.0	87.4	85.7	86.5	67.5	87.2	90.6	82.8
§	40%	50%	50%	50%	30%	60%	40%	50%	46%
parameter-pruned student									
FT$_{30\%}$	82.0	82.6	88.5	89.5	87.7	69.0	87.2	91.9	84.8
KD$_{30\%}$	82.5	82.4	89.1	89.5	87.8	69.3	87.0	91.9	84.9
PKD$_{30\%}$	82.5	82.8	89.5	89.9	88.0	68.6	86.4	91.9	84.9
DynaBERT$_{30\%}$	81.5	82.8	87.4	89.1	86.6	68.1	87.2	90.3	84.1
DKD$_{30\%}$	82.4	82.4	88.4	89.6	87.7	70.4	87.0	91.9	85.0
TAKD$_{30\%}$	82.7	82.3	89.1	89.8	87.8	68.6	87.6	91.9	85.0
STARK$_{30\%}$	82.8	82.9	89.4	90.0	87.8	69.7	87.9	92.2	85.3
§	30%	20%	30%	70%	40%	20%	30%	40%	35%

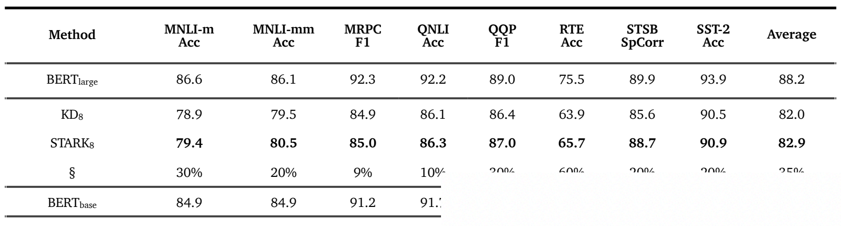


Method	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	STSB	SST-2	Average
	Acc	Acc	F1	Acc	Acc	Acc	SpCorr	Acc	
BERT$_{\text{base}}$	84.9	84.9	91.2	91.7	88.4	71.5	88.3	93.8	86.8
KD$_4$	77.7	77.7	86.9	85.1	86.1	65.3	86.4	89.6	81.8
STARK$_4$	78.8	**79.0**	**87.4**	**85.7**	86.5	**67.5**	87.2	**90.6**	**82.8**
STARK$_4^*$	**79.0**	**79.0**	**87.4**	85.3	**86.8**	66.1	**87.3**	89.8	82.6

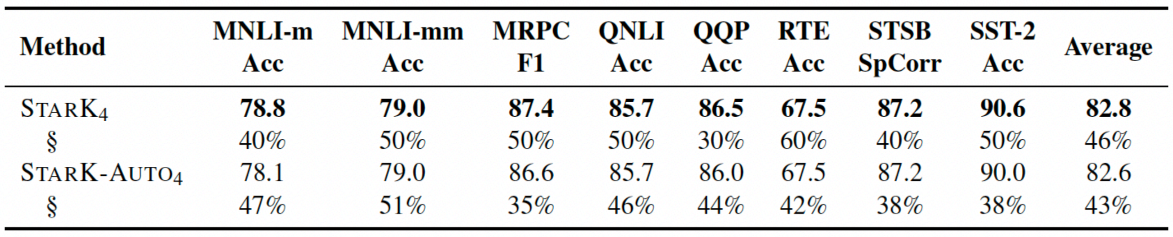
Table 6: The results of compatibility with unstructured pruning. * indicates that unstructured pruning is otherwise used.



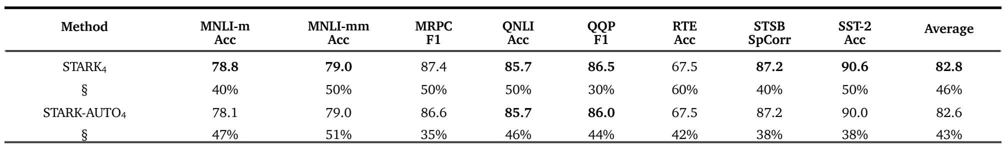
Stage	Train time on MNLI
trial distillation	~2.5h
actual distillation	~7h



Method	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	STSB	SST-2	Average
	Acc	Acc	F1	Acc	F1	Acc	SpCorr	Acc	
BERT$_{\text{large}}$	86.6	86.1	92.3	92.2	89.0	75.5	89.9	93.9	88.2
KD$_8$	78.9	79.5	84.9	86.1	86.4	63.9	85.6	90.5	82.0
STARK$_8$	**79.4**	**80.5**	**85.0**	**86.3**	**87.0**	**65.7**	**88.7**	**90.9**	**82.9**
§	30%	20%	9%	10%	30%	60%	20%	20%	35%
BERT$_{\text{base}}$	84.9	84.9	91.2	91.					
KD$_{30\%}$	73.2	72.8	82.9	78.					
STARK$_{30\%}$	**73.9**	**74.3**	**83.1**	**80.**					
§	50%	50%	30%	50%					



Method	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	STSB	SST-2	Average
	Acc	Acc	F1	Acc	Acc	Acc	SpCorr	Acc	
STARK$_4$	**78.8**	**79.0**	**87.4**	**85.7**	**86.5**	**67.5**	**87.2**	**90.6**	**82.8**
§	40%	50%	50%	50%	30%	60%	40%	50%	46%
STARK-AUTO$_4$	78.1	79.0	86.6	85.7	86.0	67.5	87.2	90.0	82.6
§	47%	51%	35%	46%	44%	42%	38%	38%	43%



Method	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	STSB	SST-2	Average
	Acc	Acc	F1	Acc	F1	Acc	SpCorr	Acc	
STARK$_4$	**78.8**	**79.0**	87.4	**85.7**	**86.5**	67.5	**87.2**	**90.6**	**82.8**
§	40%	50%	50%	50%	30%	60%	40%	50%	46%
STARK-AUTO$_4$	78.1	79.0	86.6	**85.7**	**86.0**	67.5	87.2	90.0	82.6
§	47%	51%	35%	46%	44%	42%	38%	38%	43%