Understanding and Improving Knowledge Distillation for Quantization-Aware Training of Large Transformer Encoders

[1]Minsoo Kim, [2]Sihwa Lee, [3]Sukjin Hong, [3]Du-Seong Chang, and [1,2]Jungwook Choi*

[1]Department of Electronic Engineering, Hanyang University
[2]Department of Artificial Intelligence, Hanyang University
[3]KT

Paper Code

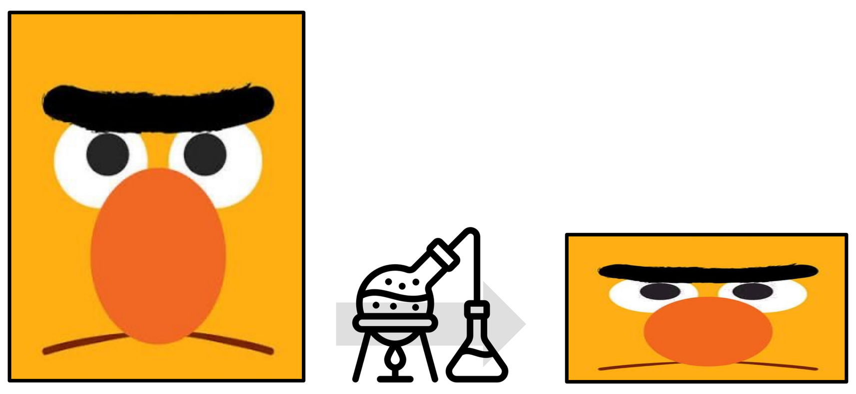
[1,2]{minsoo2333, macto94, choij}@hanyang.ac.kr
[3]{sukjin.hong, dschang}@kt.com

1. Summary

- Analyze prior Knowledge Distillation (KD) techniques for Quantization-Aware Training (QAT).
- Revealing task-dependent attention characteristics from weight quantization of large Transformer encoder.
- Propose new KD methods for QAT on Large Transformer Encoders.

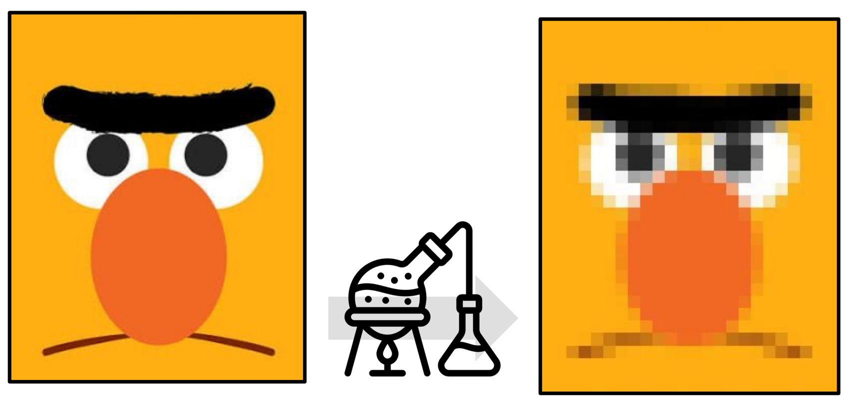
2. Motivation

Teacher Student



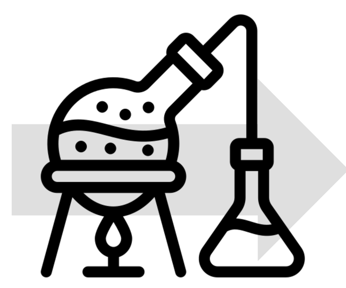
Model Compression

Teacher Student



Model Quantization

DistillBERT,
PKD,
MobileBERT,
MiniLM,
MiniLMv2,
TinyBERT
…



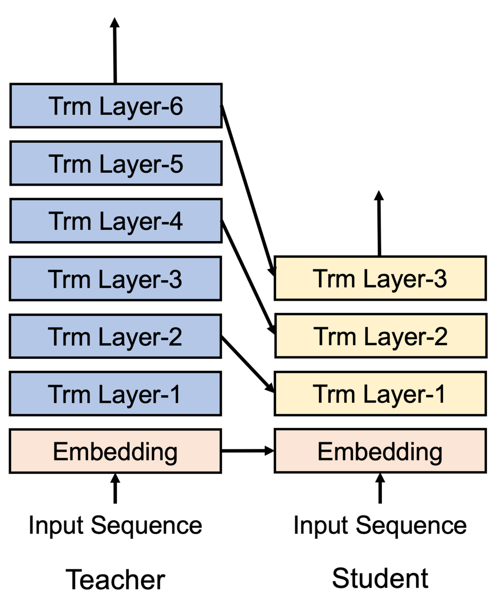
Knowledge Distillation

TernaryBERT
(≈TinyBERT)

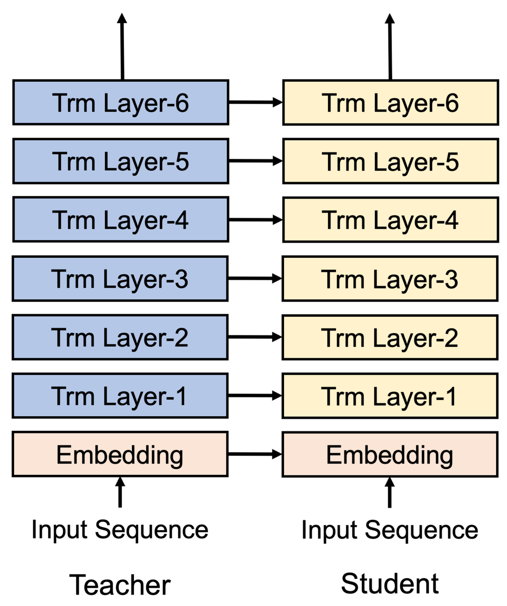
Same Strategy?

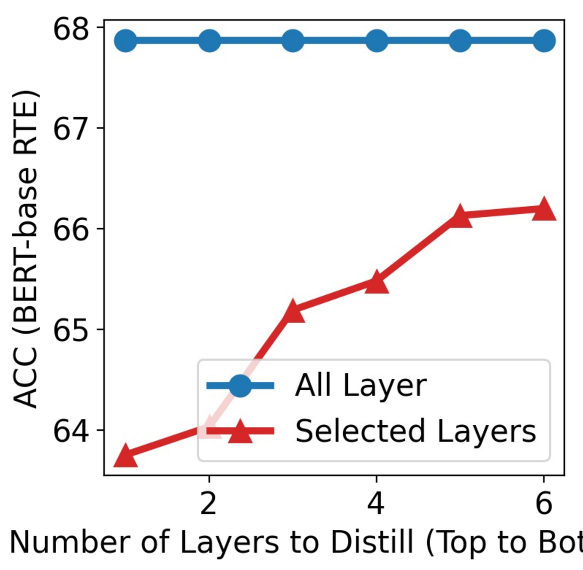
3. Prior KD Techniques for QAT

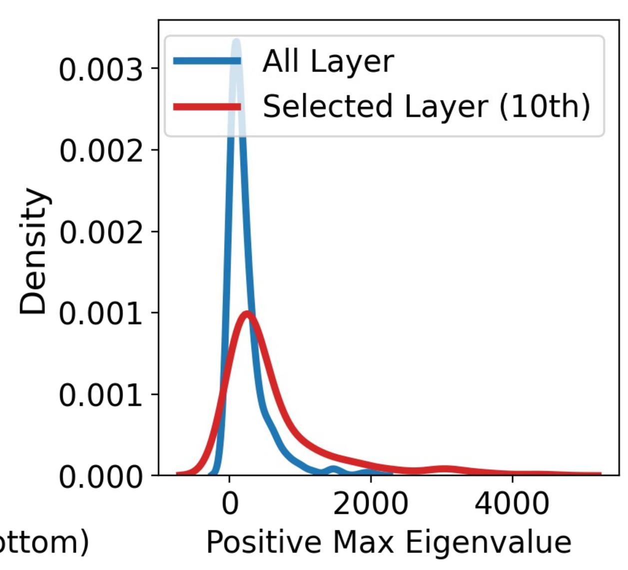
3-1. All-Layer Distillation for QAT



Selected Layer Distillation (**Compression**) vs All Layer Distillation (**QAT**)

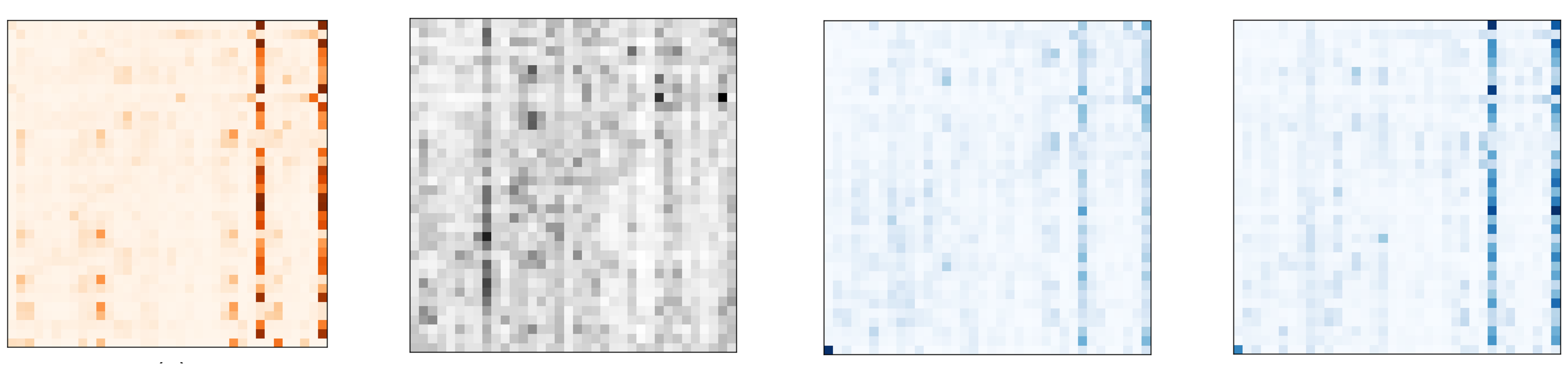






- Layer-wise Distillation helps QAT of quantized student model.

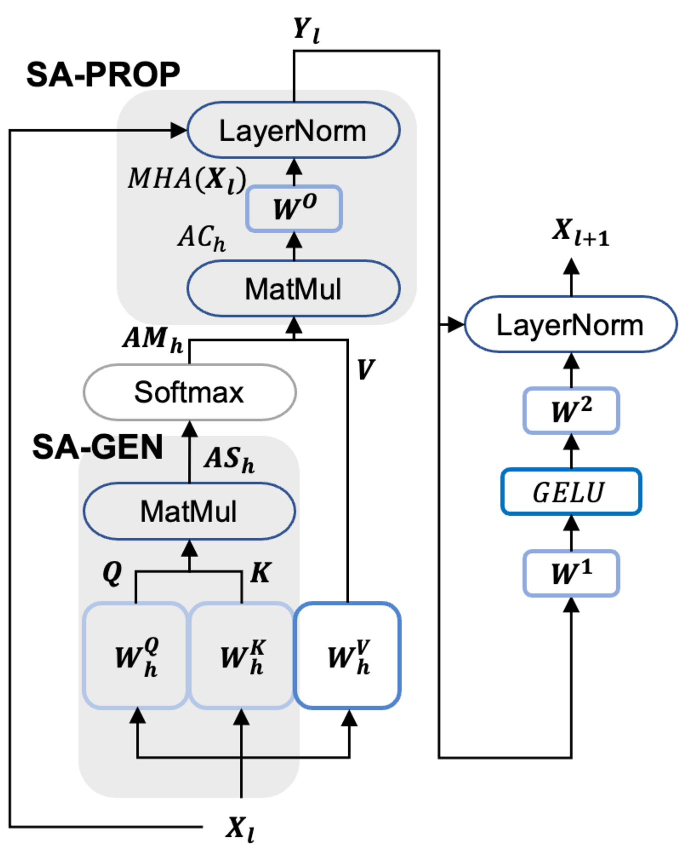
3-2. Improve KD on Self-Attention Generation



FP SA-Map w/o QAT Baseline **Ours**

- KL-Div loss function with self-attention map maintain the relative importance of attention across tokens (attention-map loss).

4. KD for QAT on Large Transformers



SA-GEN
(attention-map loss)

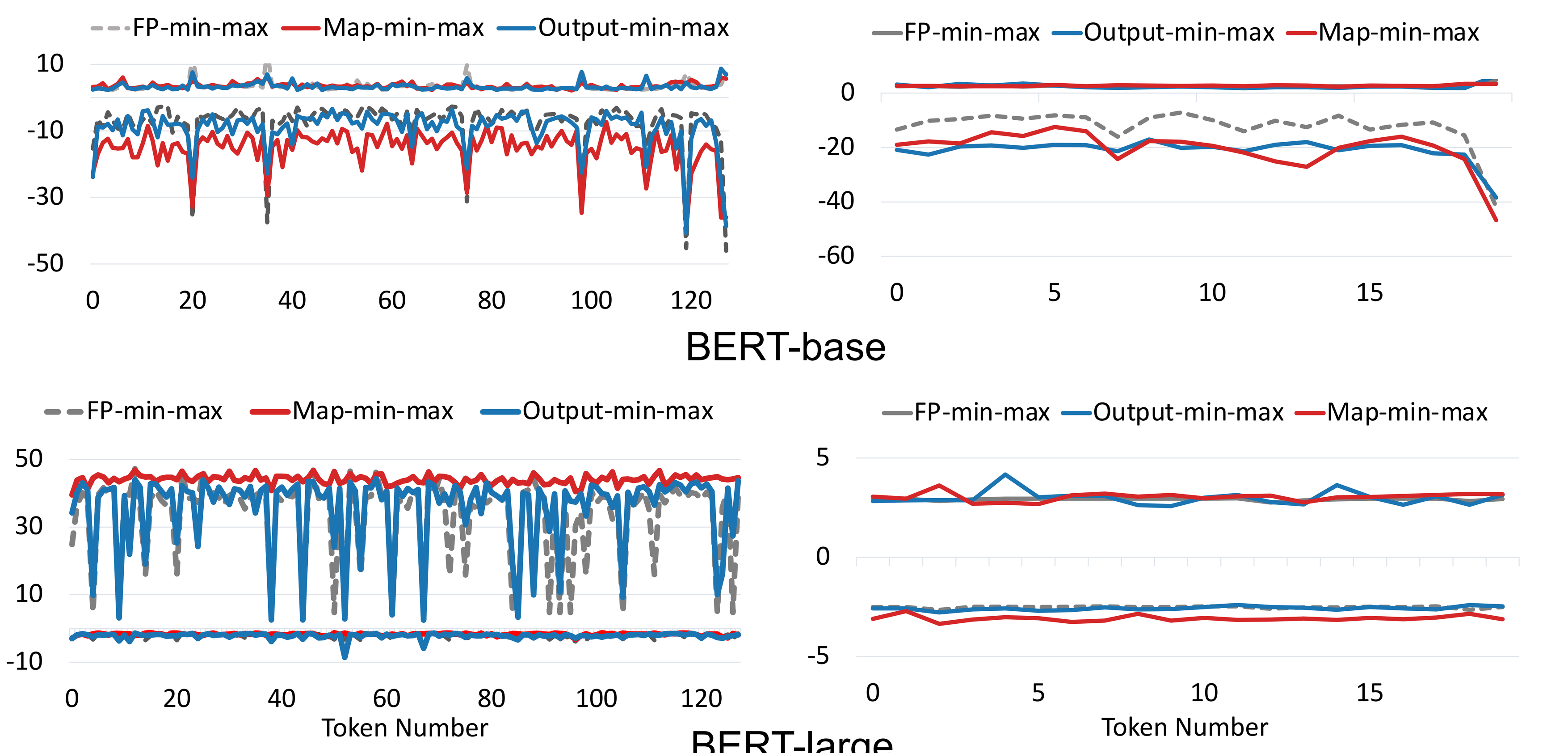
$$\mathcal{L}_{map} = \frac{1}{N_h n}\sum_{h=1}^{N_h}\sum_{t=1}^{n} D_{KL}(\boldsymbol{AM}^T_{l,h,t} \,||\, \boldsymbol{AM}^S_{l,h,t})$$

SA-PROP
(attention-output loss)

$$\mathcal{L}_{output} = \sum_{l=0}^{L-1} MSE(Y_l^T, Y_l^S)$$

Task Dependent Charateristics

SA-PROP Min-Max Range Comparison (Teacher vs Student)



- SA-PROP show distinct features depending on NLU tasks.
- Task-dependent attention characteristics are intensified when the model size increases.

$$\mathcal{L}_{unified_1} = \mathcal{L}_{map} + \gamma\mathcal{L}_{output}$$
$$\mathcal{L}_{unified_2} = \gamma\mathcal{L}_{map} + \mathcal{L}_{output}$$
$$where\ \gamma \in \{0.1, 0.2, 0.3, \ldots, 0.9\}$$

5. Experimental Results

KD-QAT Results on GLUE benchmark (8-bit Activation, 2-bit Weight)

GLUE Task (Dataset)	RTE† (2.5k)	CoLA† (8.5k)	STS-B† (5.7k)	SST-2* (67k)	QNLI* (108k)	MNLI* (393k)	QQP* (364k)	MRPC (3.5k)	AVG
Full-Prec	73.28	58.04	89.24	92.09	91.32	84.37	89.30	87.77	83.39
Baseline	68.53 ±1.69	49.61 ±0.79	87.55 ±0.14	92.01 ±0.29	90.65 ±0.05	84.21 ±0.10	89.06 ±0.40	**88.58** ±0.40	81.28
Map	70.39 ±0.78	50.40 ±1.03	**87.78** ±0.15	92.13 ±0.22	**90.98** ±0.17	84.31 ±0.10	89.22 ±0.40	88.07 ±0.40	81.66
Output	70.65 ±1.27	49.05 ±0.50	87.77 ±0.14	92.13 ±0.22	90.58 ±0.07	84.24 ±0.01	89.17 ±0.20	87.01 ±0.43	81.33
Map+Output	**71.68** ±1.19	**50.50** ±0.45	87.73 ±0.16	**92.39** ±0.18	90.91 ±0.14	**84.33** ±0.06	**89.28** ±0.10	88.18 ±0.53	**81.87**

BERT-base (110M param, Compression rate is 14.9x)

GLUE Task (Dataset)	RTE† (2.5k)	CoLA† (8.5k)	STS-B† (5.7k)	SST-2* (67k)	QNLI* (108k)	MNLI* (393k)	QQP* (364k)	MRPC (3.5k)	AVG
Full-Prec	70.39	60.31	89.83	92.32	92.29	86.49	89.55	88.43	83.70
Baseline	65.02 ±1.40	52.87 ±0.99	88.75 ±0.09	91.82 ±0.22	91.87 ±0.15	85.70 ±0.17	89.29 ±0.07	**89.26** ±0.54	81.84
Map	66.42 ±0.75	53.16 ±0.53	88.65 ±0.11	92.20 ±0.30	91.93 ±0.13	86.10 ±0.13	**89.53** ±0.07	88.67 ±0.37	82.08
Output	**69.50** ±1.20	**54.71** ±0.71	**89.10** ±0.08	92.13 ±0.26	91.92 ±0.13	86.22 ±0.05	89.44 ±0.09	88.75 ±0.71	**82.72**
Map+Output	68.83 ±1.45	54.69 ±1.08	88.85 ±0.15	**92.30** ±0.11	**92.16** ±0.15	**86.36** ±0.06	89.48 ±0.06	88.64 ±0.79	82.66

BERT-large (340M param, Compression rate is 15.4x)

Task (Dataset)	KLUE-TC (45k)	KLUE-STS (11k)	NSMC (150k)	AVG
Full-Prec	85.76	92.11	91.87	89.91
Baseline	85.56 ±0.08	91.04 ±0.10	91.13 ±0.04	89.24
Map	85.41 ±0.10	**91.44** ±0.23	91.24 ±0.10	89.36
Output	**85.63** ±0.23	91.03 ±0.11	91.39 ±0.15	89.35
Map + Output	85.57 ±0.21	91.11 ±0.14	**91.65** ±0.12	**89.44**

ULM-large (280M param, Compression rate is 15.9x)

- Case-1 (†): RTE, CoLA, STS-B
- Case-2 (⋆): SST-2, QNLI, MNLI, QQP
- Baseline: TernaryBERT

- In the BERT-base, attention-map loss benefits all the tasks in Case-1 and Case-2.
- In the BERT-large, attention-output loss significantly boosts the accuracy of Case-1.
- Overall, the unified loss facilitates QAT accuracy in every tasks.