Teacher Intervention: Improving Convergence of Quantization Aware Training for Ultra-Low Precision Transformers

[1]Minsoo Kim, [2]Kyuhong Shim, [1]Seongmin Park, [2]Wonyong Sung, and [1]Jungwook Choi*

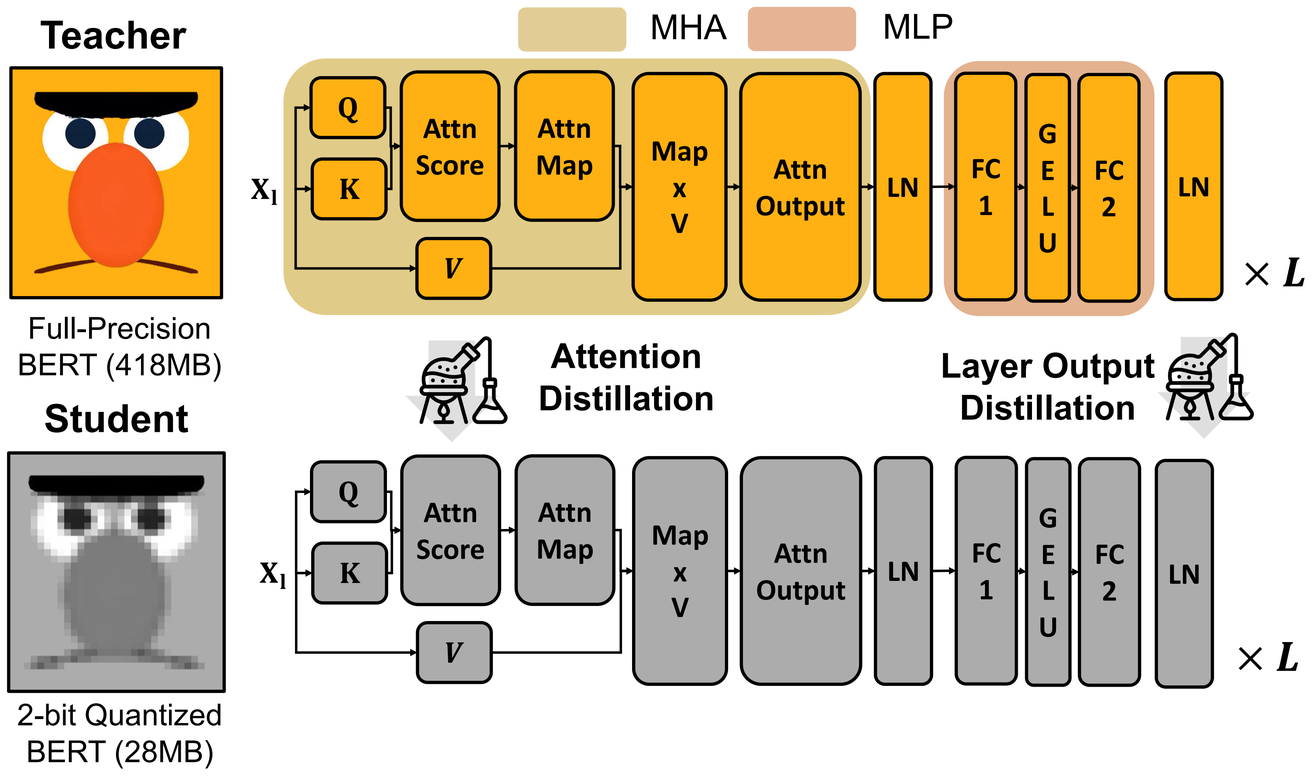
[1]Department of Electronic Engineering, Hanyang University
[2]Department of Electrical and Computer Engineering, Seoul National University
minsoo2333@hanyang.ac.kr

EACL 2023

Paper Code

1. Background and Motivation

- Quantization (QAT) & Knowledge Distillation (KD)
  : KD provides extra guidance for low precision (2-bit) quantization

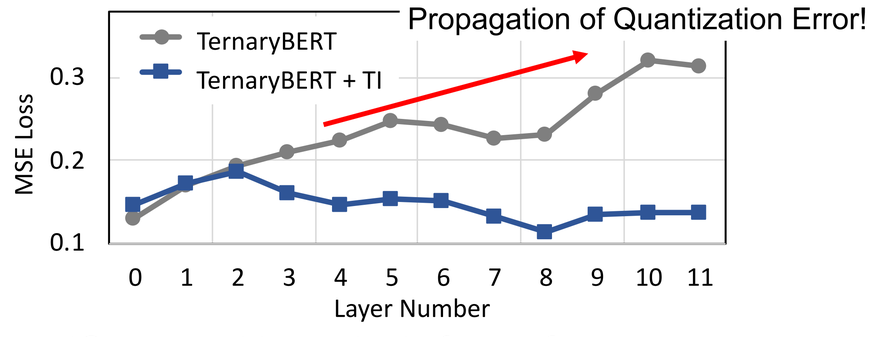


Quantization Aware Training (QAT) with **Knowledge Distillation**
[Zhang et al, TernaryBERT, EMNLP 2020]

- Prior works suffer noticeable **accuracy degradation** and require **increased iterations** for fine-tuning in few-sample tasks

Task (Num. Samples)	QQP (364K)	CoLA 8.5K	RTE (2.5K)
Full precision (Fine-tune iters.)	87.7 (34,113)	58.0 (1,650)	73.3 (234)
Ternary weight (Fine-tune iters.)	87.8 (34,113)	**49.6** (1,650)	**68.5** (234)
Ternary Weight (Fine-tune iters. w/ DA)	- (-)	58.29 (**20,862**)	73.3 (**1,654**)

QAT performance of TernaryBERT and number of iterations

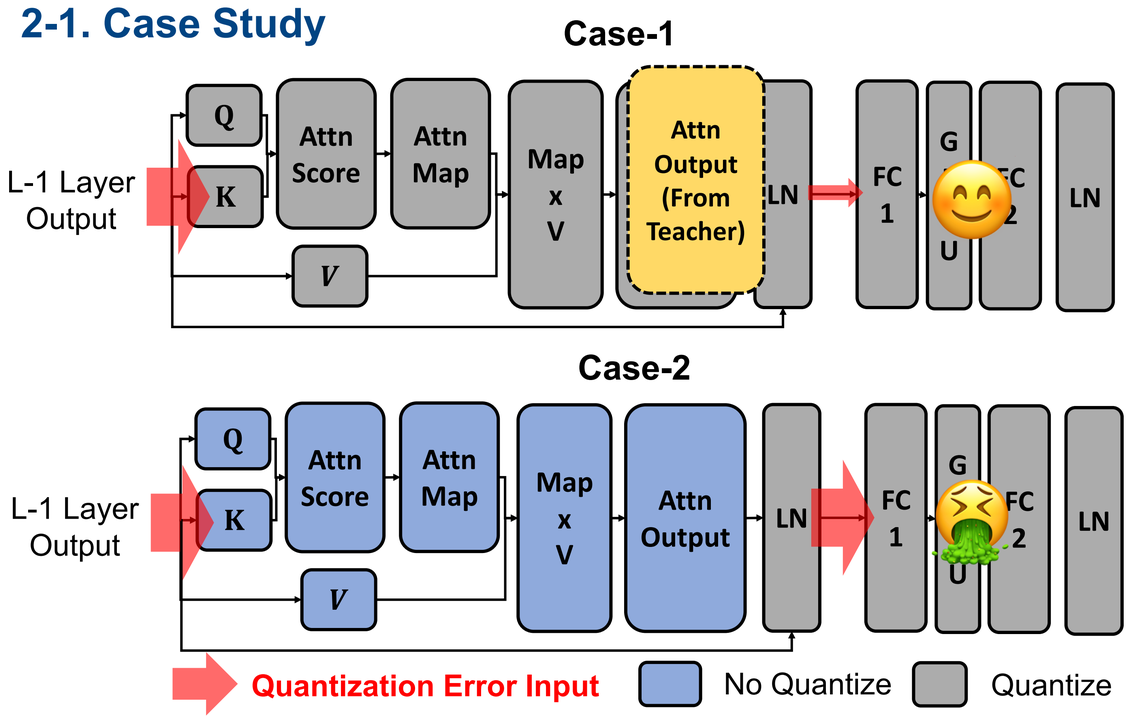


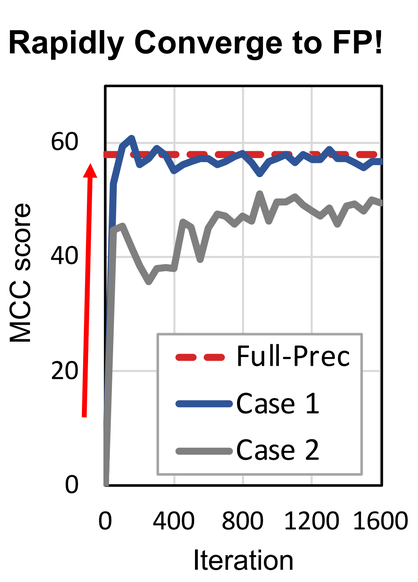
MSE Loss at the output of Transformer layers

💡: **Propagating impact of quantization error along the layers**

2. Methods

2-1. Case Study





Accuracy curve in QAT
(BERT-base, CoLA Task)

- Case 1: Intervene student's attention output with the teacher's (TI-O) + Quantize all
  ⇒ **Propagation of Q Error X** 🤗
- Case 2: No quantize attention sub-layers
  ⇒ **Propagation of Q Error O** 😔

- **Teacher Intervention (TI)** : step-by-step reconstruction of sub-layers of Transformer
  - Step 1: QAT with TI (Few-Steps)
  - Step 2: QAT with KD

2-2. Teacher Intervention

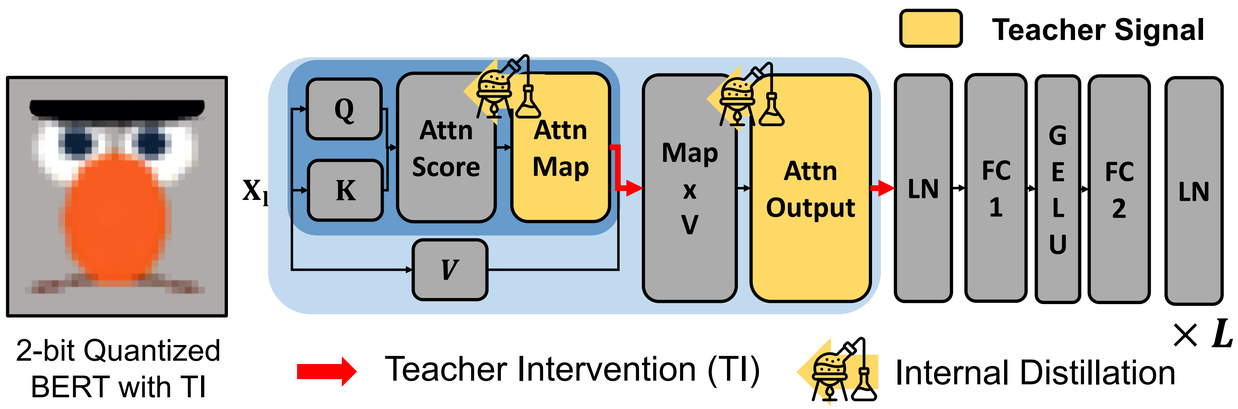
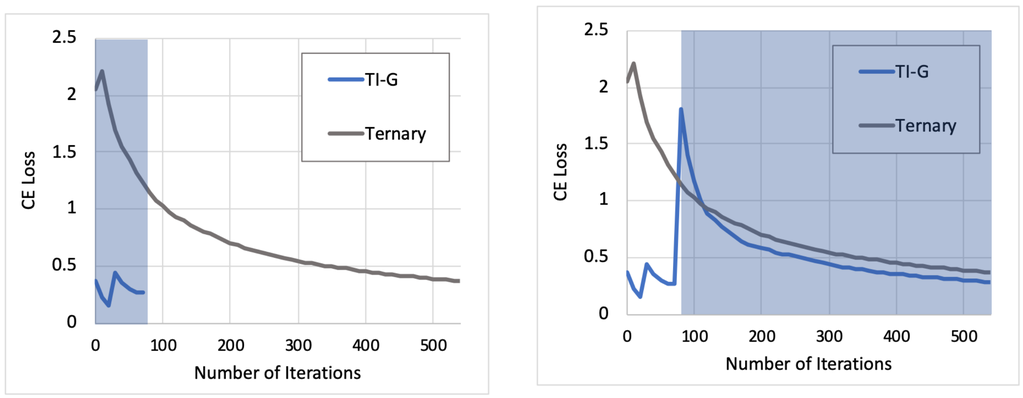
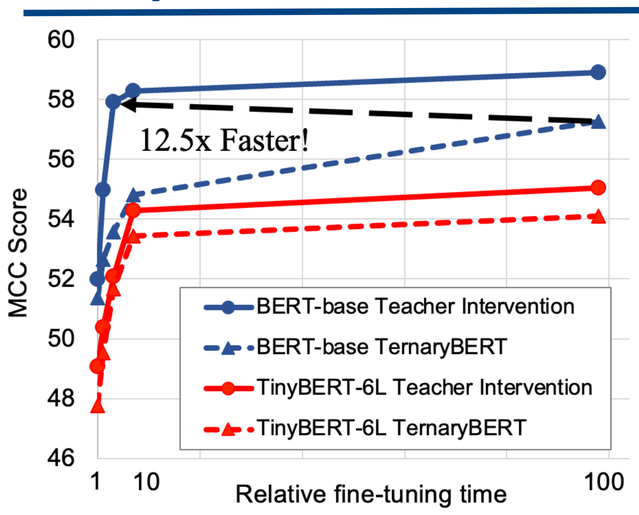


Illustration of Teacher Intervention (TI-M, TI-O)

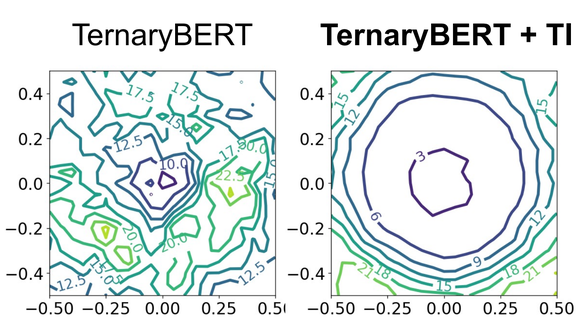


Two-Step QAT with Teacher Intervention
(Cross-Entropy Loss Curve)

3. Experiments



Fine-Tuning time vs Accuracy
of quantized BERT



Loss landscape visualization of a
fine-tuned Transformer

- TI achieves higher accuracy within shorter fine-tuning time
- TI flattens the loss surface of QAT

Summary

- Quantization below 2-bit -> considerable accuracy degradation due to unstable convergence in few-sample Fine-Tuning.
- ***Teacher Intervention*** **(TI)**: proactive knowledge distillation method for fast converging QAT of ultra-low precision (2 bit) Transformers.
- TI achieves **superior accuracy** with **significantly lower fine-tuning iterations** (up to x12.5) on Transformers of NLP (BERT) as well as computer vision (ViT) compared to SOTA QAT methods