TEACHER INTERVENTION: IMPROVING CONVERGENCE OF QUANTIZATION AWARE TRAINING FOR ULTRA-LOW PRECISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-trained Transformer models such as BERT have shown great success in a wide range of applications, but with the cost of substantial increases in model complexity. Quantization-aware training (QAT) is a promising way to lower the implementation cost and energy consumption. However, aggressive quantization below 2-bit causes considerable accuracy degradation, especially when the downstream dataset is not abundant. This work proposes a proactive knowledge distillation method called *Teacher Intervention* (TI) for fast converging QAT of ultra-low precision pre-trained Transformers. TI intervenes layer-wise signal propagation with the intact signal from the teacher to remove the interference of propagated quantization errors, smoothing loss surface and expediting the convergence. We further propose a gradual intervention mechanism to stabilize tuning of the feed-forward network and recover the self-attention map in steps. The proposed scheme enables fast convergence of QAT and improves the model accuracy regardless of diverse characteristics of downstream fine-tuning tasks. We demonstrate that TI consistently achieves superior accuracy with lower fine-tuning budget.

1 INTRODUCTION

The Transformer-based pre-trained neural networks have significantly improved the performance of various applications of artificial intelligence, including natural language processing (NLP) (Devlin et al., 2018; Raffel et al., 2020; Brown et al., 2020) and computer vision (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021). The self-attention mechanism represents these models (Vaswani et al., 2017), which links different symbols within a sequence to obtain a relational representation. Thanks to the exceptional performance of the pre-trained Transformer models, there has been an increasing need for their efficient deployment. However, the gigantic size of the pre-trained Transformer models hinders straightforward implementation. Even relatively small models like BERT-base (Devlin et al., 2018) contain a few hundred million parameters, incurring profound memory and computation overhead for resource-constrained devices with limited memory and computing fabric. Seminal research efforts attempted to reduce this burden via model compression. Behnke & Heafield (2020) and Gordon et al. (2020) pruned unimportant weights to reduce the number of parameters, while Mao et al. (2020) further employed low-rank matrix factorization. In addition, Knowledge Distillation (KD) Hinton et al. (2015) was employed in (Sanh et al., 2019; Sun et al., 2019; 2020; Wang et al., 2020) to transfer knowledge of the original model (*teacher*) to the compressed one (*student*) by mimicking the Teacher's behavior.

Among many model compression techniques, quantization-aware training (QAT) stands out for its recent success in reducing computational complexity and memory requirements of Transformer models (Bhandare et al., 2019; Zafrir et al., 2019; Kim et al., 2021). QAT reflects quantization errors during the forward pass computation of stochastic gradient descent to train a more accurate quantized model. However, quantizing weight parameters of Transformers to a precision lower than 2-bits degrades the accuracy, especially when the dataset size for the target downstream tasks is not large enough (Zhang et al., 2020b; Bai et al., 2020). Although few-sample fine-tuning of Transformer models has been reported to be highly unstable (Grießhaber et al., 2020; Yu et al., 2021; Zhang et al., 2020a; Dodge et al., 2020; Mosbach et al., 2020), there is limited understanding on why QAT is susceptible for small dataset tasks and how to improve the QAT accuracy. Recently, XTC (Wu et al., 2022) claimed that the few-sample fine-tuning is under-trained, and thus augmented the QAT method with data augmentation and significantly increased fine-tuning iterations. While XTC was successful in recovering accuracy degradation, it took an order of magnitude higher fine-tuning time. The increased fine-tuning cost for QAT would become a significant hurdle for the broad deployment of quantized Transformers.

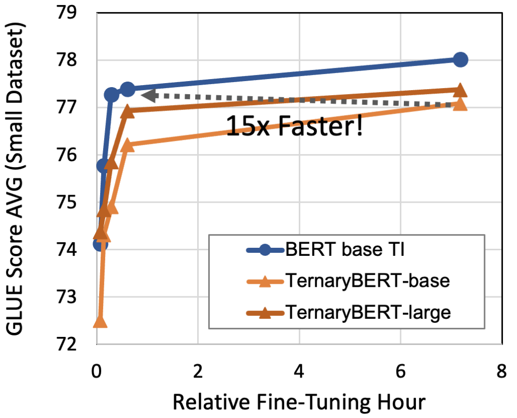


Figure 1: The training time and accuracy comparison between ours and other SOTA results on CoLA task

This work proposes a proactive KD method called *Teacher Intervention* (TI) for fast converging QAT of ultra-low precision pre-trained Transformers. We reveal that difficulty of quantization on few-sample fine-tuning originates from disruption of loss surface due to quantization error propagation. To mitigation this undesirable phenomena, we propose TI to intervene layer-wise signal propagation with the intact signal from the teacher. TI removes the interference of propagated quantization errors, smoothing loss surface and expediting the convergence. We further discover that self-attention map is particularly susceptible to the quantization error. Thus, we propose a gradual intervention mechanism that first intervenes attention output for stable tuning of the feed-forward network, followed by self-attention map intervention for its recovery. The proposed gradual intervention scheme enables fast convergence of QAT and improves the model accuracy regardless of diverse characteristics of downstream fine-tuning tasks. We perform extensive evaluation on various fine-tuned Transformers (BERT-base/large, TinyBERT-4L/6L, and SkipBERT-6L for NLP, and ViT for CV), and demonstrate that TI consistently achieves superior accuracy with lower fine-tuning budget. In particular, TI outperforms TernaryBERT on GLUE tasks with 15× savings in fine-tuning hours, as shown in Fig.1.

We summarize our contributions as follows:

- We reveal that root cause of QAT's failure is the disruption of loss surface due to propagation of quantization error.
- We propose a proactive KD method called Teacher Intervention (TI), which removes the interference of propagated quantization errors. TI works as a quick warm-up phase to stabilize ultra-low precision QAT and improve convergence.
- We propose a gradual intervention mechanism that first intervene attention output to tune the feed-forward network followed by self-attention map intervention for its enhanced recovery. The proposed gradual intervention improves QAT convergence regardless of diverse characteristics of downstream fine-tuning tasks.
- We demonstrate that TI significantly enhances convergence of state-of-the-art QAT methods on various fine-tuned Transformers in NLP and CV by achieving higher accuracy within smaller fine-tuning budgets. This improved fine-tuning efficiency will facilitate rapid deployment of ultra-low precision fine-tuned Transformers.

2 Related Work

2.1 Knowledge Distillation for BERT Compression

Knowledge distillation (KD) Hinton et al. (2015) is a transfer learning framework to pass on knowledge of a large model (*teacher*) to a smaller one (*student*) by mimicking the teacher's behavior. Because KD provides fruitful information not contained in the output label, KD can compress the model size without significant performance degradation. Therefore, KD has been widely studied to train smaller BERT models for various application domains.

The most common distillation approach is to match the probability distribution from the final output softmax between the Teacher and Student for the same input, as in DistilBERT Sanh et al. (2019). In addition to the distillation loss at the model output, PKD Sun et al. (2019) suggested matching the normalized output of each Transformer layer claiming that loss on intermediate output can be beneficial for the Student. MobileBERT Sun et al. (2020) also employed per-head attention map transfer along with the customized architecture for efficient Transformer computations. MiniLM Wang et al. (2020) further transferred knowledge from the self-attention map as well as the value-relation. Considering the structural mismatch between the Teacher and Student models, MiniLM performed distillation only at the last Transformer layer.

While the abovementioned studies focused on the task-agnostic BERT, there have been several efforts Tang et al. (2019); Aguilar et al. (2020) to train a tiny task-specific Student. In this line of study, the task-specific, downstream fine-tuned BERT is first prepared, and the Student is trained with KD by utilizing this fine-tuned model as a teacher. As a hybrid approach, TinyBERT Jiao et al. (2019) proposed a two-step KD, the first step for general distillation, followed by task-specific distillation.

Although these KD techniques for BERT compression have developed efficient BERT structures with a reduced number of parameters and computations, there has been limited understanding of KD on the model quantization. In this work, we reveal that more aggressive intervention of the teacher on the self-attention map of each layer helps the ultra-low precision model regain the model accuracy.

2.2 Quantization for Ultra-Low Precision BERT

Quantization is a promising technique for reducing the high inference cost of large-scale models without changing the model structure. Instead of representing numbers with the 32-bit floating-point (FP32) format, employing fixed-point representation, such as 8-bit integer (INT8) quantization, has achieved significant speedup and storage savings for BERT Zafrir et al. (2019); Kim et al. (2021). However, direct quantization of weight parameters would suffer accuracy degradation of the original model accuracy when the quantization bit-precision is low. Therefore, quantization-aware training (QAT) is commonly applied for ultra-low precision model quantization.

Recently, QAT has been applied for compressing BERT with precision lower than 2-bit. TernaryBERT Zhang et al. (2020b) represents each weight element into one of three values $\{-1, 0, 1\}$. TernaryBERT actively incorporates KD into QAT for improving accuracy degradation. Especially, KD with the MSE loss on the attention score (before taking Softmax) and the output of each Transformer layer is employed for QAT. To further reduce the bit-precision, BinaryBERT Bai et al. (2020) suggested a modified QAT procedure that initializes the weights for binary quantization. However, ternarizing or binarizing weight parameters significantly degrades the model accuracy, especially when the dataset size for the target downstream tasks is not large enough.

In fact, it has been reported that finetuning BERT on downstream tasks with insufficient data is highly unstable Grießhaber et al. (2020); Yu et al. (2021). As a result, several works proposed modified finetuning procedures for improving the stability (Zhang et al., 2020a; Dodge et al., 2020; Mosbach et al., 2020). Still, the proposed approaches do not address the sensitivity of Transformer models on QAT for small datasets. XTC (Wu et al., 2022) recently proposed a QAT method with significantly increased iterations and data augmentation to improve quantization accuracy of ultra-low bit precision Transformers. As illustrated in Fig. 1, however, this prolonged fine-tuning results in sizable deployment overhead, let alone costly data augmentation. In this work, we discover that quantization significantly disrupts the propagation of self-attention in Transformer layers hindering the optimization process of QAT. Therefore, we propose a new KD-based method that proactively intervene the error propagation to improve convergence of QAT methods.

3 Background and Motivation

3.1 Transformer Layer

The BERT model Devlin et al. (2018) is built with Transformer layers Vaswani et al. (2017). A standard Transformer layer includes two main sub-modules: Multi-Head Attention (MHA) and Feed-Forward Network (FFN). Input to the l-th Transformer layer is $X_l \in \mathbb{R}^{n \times d}$ where n and d are the sequence length and hidden state size, respectively. Let N_H be the number of attention heads and $d_h = d/N_H$. $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times d_n}$ are the weight parameters converting X_l into Querry ($Q = X_l W_h^Q$), Key ($K = X_l W_h^K$), and Value ($V = X_l W_h^V$), respectively. Then, attention score (AS), self-attention map (SA), and attention context (AC) are defined as follows:

$$\text{AS}_h = QK^\top; \text{SA}_h = \text{Softmax}_h(\frac{\text{AS}_h}{\sqrt{d}}); \text{AC}_h = \text{SA}_h V. \tag{1}$$

Then, attention output (= MHA) is defined as $\text{MHA}(X_l) = \text{Concat}(\text{AC}_1, \text{AC}_2, ...\text{AC}_{N_H})W^O$. Motivated by Kobayashi et al. (2020), attention output can be re-written per each token i:

$$\text{MHA}(X_l)(i) = \sum_{j=1}^{n} \alpha_{i,j} f(X_l(j)), \tag{2}$$

where $f(x) := (xW^V + b^V)W^O$ and $\alpha_{i,j}$ is j'th value of i'th token in AM_h. Therefore, MHA can be decomposed into two parts: self-attention generation (SA-GEN) corresponding to the attention map (α), and self-attention propagation (SA-PROP) corresponding to $f(x)$. Fig. **??** shows which part is SA-GEN and SA-PROP respectively. FFN consists of two fully-connected layers with weight parameters W^1 and W^2:

$$\text{FFN}(Y_l) = \text{GeLU}(X_l W^1 + b^1)W^2 + b^2. \tag{3}$$

Therefore, a Transformer layer X_l is defined as:

$$\begin{aligned} Y_l &= \text{LayerNorm}(X_l + \text{MHA}(X_l)), \\ X_{l+1} &= \text{LayerNorm}(Y_l + \text{FFN}(Y_l)). \end{aligned} \tag{4}$$

3.2 Quantization-Aware Training

Quantization-aware training (QAT) emulates inference-time quantization during training to learn parameters robust to the quantization error. In particular, ternary quantization represents all the weight parameters ($W^Q, W^K, W^V, W^O, W^1, W^2$) into ternary values $\{+1, 0, -1\}$ along with a scale factor α for sub-2bit inference at deployment. In this work, we follow the approach of TWN Zhu et al. (2016) that analytically estimates the optimal α and T to minimize $\|W - \alpha T\|$.

Due to aggressive bit-reduction, ternary quantization causes significant accuracy loss. KD can help compensate for accuracy degradation, where the original full-precision model works as a teacher to guide the training of the quantized model as a student. In case of Transformer models, TernaryBERT Zhang et al. (2020b) applied KD on every output activation X^l as well as attention scores AS with mean squared error (MSE) loss:

$$L_{trm} = \sum_{l=1}^{L+1} \text{MSE}(X_l^S, X_l^T) + \sum_{l=1}^{L} \text{MSE}(A_l^S, A_l^T), \tag{5}$$

where S and T represent the student and teacher models, respectively.

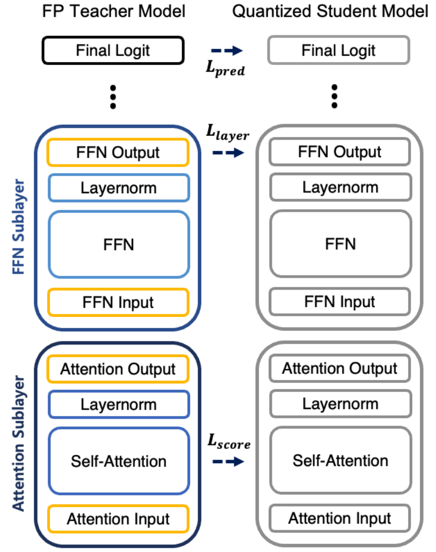
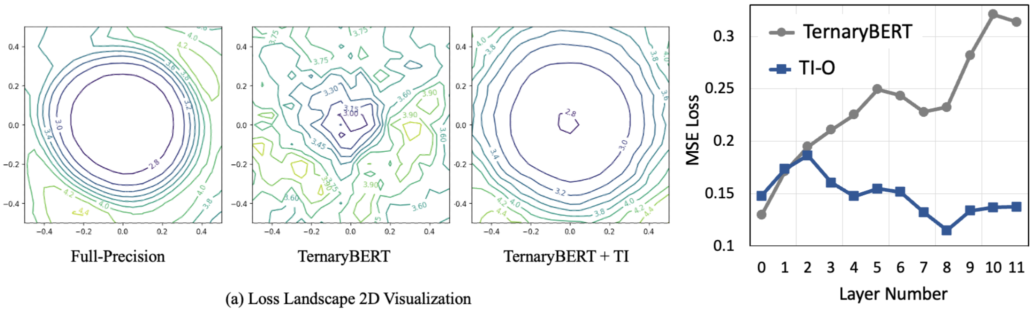


Figure 2: Illustration of three knowledge distillation locations (attention score-L_{score}, layer output-L_{layer}, and the final prediction-L_{pred}) for quantization-aware training of Transformer models.

Also, the output logits of the student (P^S) and the teacher (P^T) are used in TernaryBERT Zhang et al. (2020b) to compute the cross-entropy loss:

$$L_{pred} = \text{CE}(P^S, P^T). \tag{6}$$

We follow the settings of TernaryBERT Zhang et al. (2020b) as our baseline QAT method. Fig. 2 shows the KD locations utilized for Transformer models.



(a) Loss Landscape 2D Visualization

(b) Per Layer Transformer Layer output MSE loss

Figure 3: (a) Loss landscape visualization of different QAT methods over CoLA task (b) Hessian max eigenvalue spectra of different QAT methods

3.3 Challenges

Despite attempts to bridge the accuracy gap, prior works on ultra-low precision Transformers Bai et al. (2020); Zhang et al. (2020b) still suffer noticeable accuracy degradation, especially when the dataset size is small. Recall Mosbach et al. (2020) that the few-sample fine-tuning is susceptible to bad local minima. However, we observed that QAT often fails even if it fine-tunes from the successfully trained model. Therefore, we visualize the loss landscape of the quantized Transformers. Fig. 3(a) illustrates the loss surfaces of the full-precision BERT-base model successfully fine-tuned on CoLA of GLUE tasks, along with the models quantized with TernaryBERT. As expected, the loss surface of TernaryBERT exhibits sharp curvatures with many bad valleys, contrasting with the smooth loss surface of the full-precision model. To gain further intuition of the internal behavior of Transformer layers under quantization, we measured the mean-square error (MSE) of the output of each Transformer layer between the full-precision baseline (teacher) and TernaryBERT (student). As shown in Fig. 3(b), there is a trend of growing MSE over the layers. (Similar trends can be observed in the other GLUE tasks.) This exaggerated error along the layers would degrade the model's accuracy.

In this work, we focus on alleviating this aggravating impact of quantization errors on Transformers with a proactive knowledge distillation called Teacher Intervention (TI). Interestingly, as shown in Fig. 3, TI could successfully suppress the error propagation and flatten the loss surface for favorable convergence without precipitously increased fine-tuning iterations. We describe the detail of TI in the next section.

4 Method

4.1 Teacher Intervention

Teacher intervention (TI) is a KD method that aggressively intervenes in the student's signal propagation to guide QAT for more rapid convergence. Fig. 4(a) illustrates the options for teacher intervention. First, intervention on the attention output (a.k.a. output intervention, TI-O) replaces the student's attention output (= MHA) with the teacher's. In this case, the FFN sub-layers are trained with ultra-low precision quantization without concerns of erroneous input from MHA. Meanwhile, the computation within the MHA sub-layer is quantized for internal distillation. Similarly, intervention on the self-attention map (a.k.a. map intervention, TI-M) replaces the student's SA-GEN output with the teacher's.

The development of TI is motivated by the previous observation of the aggravating impact of quantization error along the layers (cf. Fig. 3(b)). We conjecture that the root cause of this phenomenon is error propagation instead of the quantization error itself. To confirm our hypothesis, we conducted controlled experiments for TI-O with two quantization cases (and similar experiments for TI-M):

- Case-1: Quantize all except MHA (SA-GEN for TI-M).
- Case-2: Quantize all + TI.

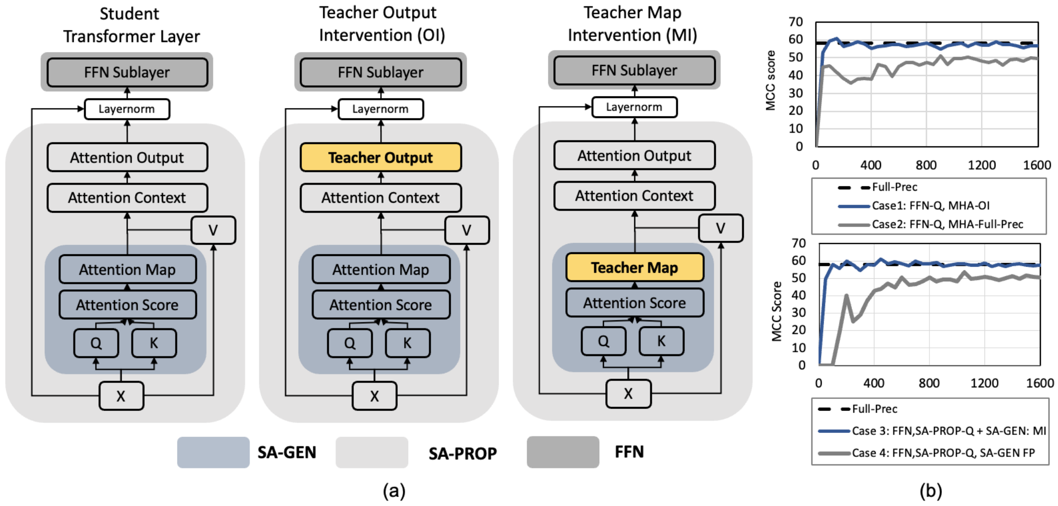


Figure 4: (a) Illustration of Teacher Intervention TI (b) Accuracy curves during QAT with Teacher intervention (TI) and MSA Full-Precision model QAT in CoLA task (c) Layer-wise MSE loss of FFN output with Zhang et al. (2020b) and QAT with TI

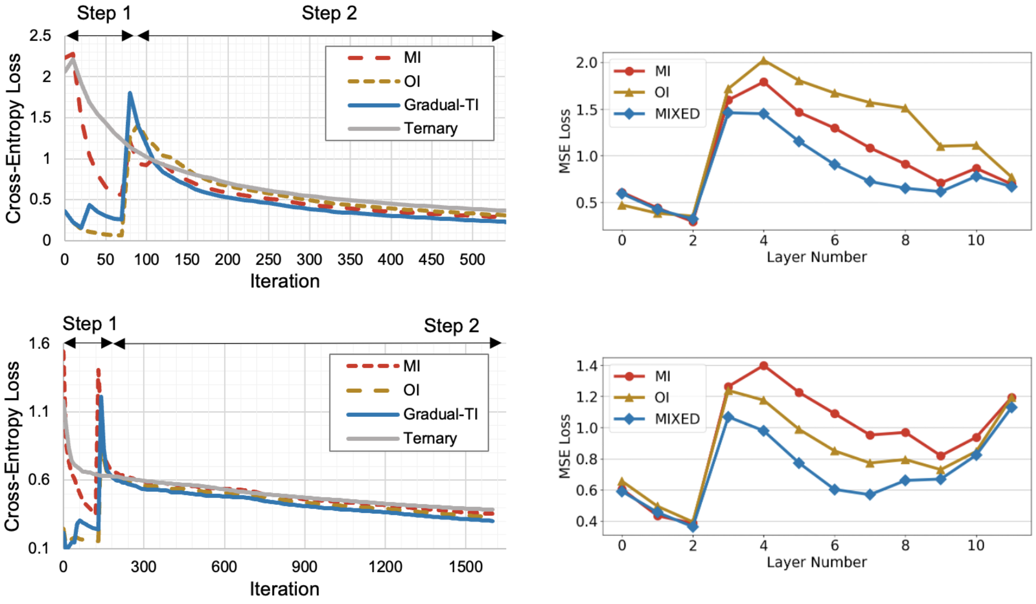
	Map Intervention			Output Intervention			Gradual Intervention		
	SA-GEN	PROP	FFN	SA-GEN	PROP	FFN	SA-GEN	PROP	FFN
Step-1-Phase1	T	S	S	T		S	T		S
Step-1-Phase2							T	S	S
Step-2	S			S			S		

Table 1: Detailed TI settings. S: Stduent Model, T: Teacher Model

The key difference is that Case-2 propagates the quantization error through the Transformer layers while Case-1 does not. Fig. 4(b) shows the convergence curves of the two cases on CoLA for TI-O and TI-M. As shown in the figure, Case-1 converges rapidly to full-precision accuracy despite the ultra-low bit quantization in the quantized sub-layers. Whereas, Case-2 converges slowly to the sub-optimal point with noticeable accuracy degradation. For example (in the setting of TI-O), although Case-2's MHA computation is in full-precision, the error propagated from its preceding FFN sub-layers still affects it, corrupting the attention output. On the other hand, TI interrupts this error propagation to stabilize QAT on FFN sub-layers. Therefore, Fig. 3(b) shows that error propagation disappears when TI is applied.

4.2 Gradual Teacher Intervention

Given that there are two options for teacher intervention, how can we choose one for achieving the best performance? In this section, we propose a unified approach that gradually applies the output intervention followed by the map intervention. Note that the two TI options have strengths and weaknesses. For example, TI-O focuses on tuning the FFN sub-layers, but it lacks consideration of the self-attention map recovery. On the other hand, TI-M is best suited for recovering the self-attention map, but it does not protect signal propagation through SA-PROP. Interestingly, we empirically discover that different downstream tasks of the pre-trained Transformers have diverse preferences; e.g., BERT-base fine-tuned on STS-B is sensitive to disruption in the self-attention map while the model fine-tuned on CoLA prefers careful tuning of the FFN sub-layers. Therefore, it is helpful to develop a unified solution for teacher intervention. As a natural combination, we propose a gradual teacher intervention mechanism that applies TI-O first to tune the FFN sub-layers, followed by TI-M to recover the self-attention map. (We conducted an ablation study for the other possibilities.) The proposed gradual intervention method (TI-G) has shown practical success in most cases studied in this work. Table **??** summarizes the TI settings.



(a)

(b)

Figure 5: Left: Cross-entropy loss in QAT with different TI methods Right: Layer-wise MSE loss with different TI methods

Fig. 5(a) shows the convergence curves of different TI options for STS-B and CoLA. As discussed earlier, TI-M shows better convergences than TI-O on STS-B, but the opposite trend is shown on CoLA. Nevertheless, TI-G always shows superior convergence compared to the other options, demonstrating its universal applicability. The MSE error of MHA reported in Fig. 5(b) explains the reasoning behind TI-G's superiority. As discussed earlier, BERT-base trained on STS-B and CoLA exhibit distinct characteristics. In the case of STS-B, TI-M is more effective in reducing the error at the MHA output than TI-O. Whereas, TI-G takes advantage of the superior tuning of the FFN sub-layers in the first phase, and thus map intervention in its second phase reduces more error. A similar situation happens in the CoLA case. Therefore, we can conclude that the proposed gradual intervention can be a great fit for managing diverse characteristics of various downstream tasks for QAT of fine-tuned Transformers.

5 Experiments

In this section, we evaluate SARQ on fine-tuned BERT with ternary quantization. We empirically demonstrate that 1) SARQ outperforms the state-of-the-art (TernaryBERT) on various tasks and datasets and 2) SARQ significantly boosts the convergence of QAT.

5.1 Experiment Settings

We use the task-specific, fine-tuned BERT-Base Devlin et al. (2018) (12 layer) and TinyBERT Jiao et al. (2019) (4 layer) for evaluation of SARQ. We evaluate our method on the GLUE Wang et al. (2018) and SQuAD Rajpurkar et al. (2016). TernaryBERT and SARQ-1step perform training for three epochs, the same as the full-precision finetuning. In the case of SARQ, the first step QAT with Teacher intervention is performed until convergence (up to one epoch), then the second step QAT is performed for two epochs. Note that the QAT with Teacher intervention converges much faster than one epoch.

Table 2: Different training budgets and training details for the GLUE tasks.

	Zhang et al. (2020b)			Wu et al. (2022)	
	Budget-O	Budget-O2	Budget-O4	Budget-A	Budget-C
DA	✗	✗	✗	✓	✓
Epoch	3	6	12	1	12

	CoLA	RTE	MRPC	STS-B
Train batch-size	16	32	32	32
Train (noDA)	8551	2490	3668	5749
Train (DA)	213212	142991	225630	322121
DA / noDA	24.9	57.4	61.5	56.0
Budget-O Time	259	97	72	223
Budget-A Time	2149	1862	1484	4160
Budget-C Time	25792	22342	17813	49915

		$\text{BERT}_{\text{BASE}}$				$\text{BERT}_{\text{LARGE}}$			
GLUE Task	Cost	RTE	CoLA	STS-B	MRPC	RTE	CoLA	STS-B	MRPC
	Full-Prec	73.28	58.04	89.24	87.77	70.39	60.31	89.83	88.43
TernaryBERT	Budget-O	67.44 ±1.30	49.44 ±1.11	87.58 ±0.09	85.58 ±0.58	63.36 ±1.01	53.25 ±1.20	88.65 ±0.16	88.31 ±0.20
TI-Map	Budget-O	69.60 ±0.92	51.37 ±1.23	87.75 ±0.12	86.25 ±1.03	66.13 ±1.12	52.40 ±1.65	88.61 ±0.16	88.67 ±0.32
TI-Output	Budget-O	69.31 ±0.57	50.91 ±0.94	87.76 ±0.22	86.04 ±0.61	65.20 ±0.94	52.66 ±1.27	88.56 ±0.16	88.68 ±0.51
TI-Gradual	Budget-O	**70.32** ±0.72	**51.98** ±1.35	**87.77** ±0.29	**86.44** ±0.49	**66.27** ±0.79	**54.12** ±1.13	**88.66** ±0.05	**88.80** ±0.41
TernaryBERT	Budget-O2	70.51 ±0.41	52.65 ±0.77	88.04 ±0.14	86.00 ±0.38	66.42 ±0.62	55.72 ±1.26	89.00 ±0.09	88.22 ±0.82
TI-Gradual	Budget-O2	**71.48** ±0.36	**54.98** ±0.66	**88.04** ±0.18	**88.63** ±0.55	**68.11** ±0.75	**57.55** ±1.69	**89.12** ±0.04	**88.25** ±0.46
TernaryBERT	Budget-O4	71.23 ±0.42	53.57 ±0.80	88.22 ±0.04	86.58 ±0.32	67.50 ±0.95	57.70 ±0.64	89.12 ±0.01	**89.12** ±0.84
TI-Gradual	Budget-O4	**73.16** ±0.36	**57.92** ±1.19	**88.48** ±0.61	**89.56** ±0.52	**69.67** ±0.72	**59.89** ±1.07	**89.33** ±0.10	88.74 ±0.76

		TinyBERT-4L				TinyBERT-6L			
GLUE Task	Cost	RTE	CoLA	STS-B	MRPC	RTE	CoLA	STS-B	MRPC
	Full-Prec	68.23	43.06	87.07	87.76	74.00	57.78	88.74	87.35
TernaryBERT	Budget-O	63.15 ±0.50	32.15 ±1.43	83.33 ±0.36	84.90 ±0.38	68.74 ±1.42	47.77 ±0.35	87.29±0.12	84.89 ±0.53
TI-Map	Budget-O	64.25 ±0.83	35.59 ±1.24	83.41 ±0.21	85.22 ±0.29	68.92 ±1.25	48.30 ±1.10	87.23 ±0.12	86.16 ±0.91
TI-Output	Budget-O	63.89 ±0.68	35.06 ±1.43	83.57 ±0.29	85.40 ±0.25	68.08 ±1.04	48.15 ±0.50	87.31 ±0.10	86.04 ±0.61
TI-Gradual	Budget-O	**64.29** ±0.72	**35.17** ±1.35	**83.58** ±0.29	**85.48** ±0.49	**69.38** ±0.78	**49.08** ±1.24	**87.31** ±0.11	**86.30** ±0.63
TernaryBERT	Budget-O2	64.74 ±0.76	34.49 ±1.89	84.10 ±0.34	85.73 ±0.05	69.67 ±0.95	49.54 ±0.22	87.51 ±0.04	86.61 ±0.41
TI-Gradual	Budget-O2	**64.98** ±1.45	**36.30** ±0.24	**84.27** ±0.22	**86.44** ±0.31	**71.12** ±0.63	**50.38** ±0.56	**87.56** ±0.09	**86.80** ±0.42
TernaryBERT	Budget-O4	65.10 ±0.55	35.91 ±0.31	84.36 ±0.26	86.70 ±0.13	70.75 ±0.36	51.66 ±0.51	87.75 ±0.04	86.76 ±0.23
TI-Gradual	Budget-O4	**65.22** ±0.55	**37.40** ±1.30	**84.39** ±0.27	**87.22** ±0.13	**72.13** ±0.12	**52.08** ±0.01	**87.90** ±0.14	**87.15** ±0.39

Table 3: Evaluation Budget-O/O2/O4 results of BERT family models on GLUE benchmark (Small dataset). Each task are repeated 10 times.

5.2 QAT Accuracy on GLUE Benchmark

First, we perform an extensive performance comparison of the proposed teacher intervention methods with the basic and the state-of-the-art QAT methods (TernaryBERT (Zhang et al., 2020b) and XTC (Wu et al., 2022), respectively). To investigate the convergence of these QAT methods, we consider the following fine-tuning budgets:

- Budget-O: The number of iterations reported in the original paper (Zhang et al., 2020b).
- Budget-O2: 2x number of iterations from Budget-O.
- Budget-O4: 4x number of iterations from Budget-O.
- Budget-A/C: the budget reported in (Wu et al., 2022).

A summary of fine-tuning budgets is shown in Table 2. Note that the number of fine-tuning iterations of Budge-A is roughly 15× larger than Budget-O4. In the following subsections, we categorize the experiments to two scenarios: Few-Sample Fine-tuning (Scenario-1: Budget O/O2/O4) and Prolonged Fine-tuning (Scenario-2: Budget A/C)[1].

Scenario-1: Few-Sample Fine-tuning Table 3 summarizes the experimental results of few-sample fine-tuning. The following lessons can be observed:

- Consistent with the prior observations (), the QAT accuracy increases as the fine-tining budget grows from Budget-O to Budget-O4.

[1] We report the experimental results for the small dataset tasks here; the experimental results on the other GLUE tasks are reported in Appendix.

		TinyBERT-6L					SkipBERT-6L			
Method	Cost	RTE	CoLA	STS-B	MRPC	Method	RTE	CoLA	STS-B	MRPC
	Full-Prec	74.00	57.78	88.74	87.35		74.72	55.37	89.27	86.11
TernaryBERT	Budget-A	72.02 ±0.21	53.44 ±1.11	88.43 ±0.07	88.14 ±0.31	XTC	69.91 ±0.41	53.74 ±0.77	88.77 ±0.03	86.29 ±0.57
TI-Gradual	Budget-A	**72.92** ±0.72	**54.29** ±1.35	**88.45** ±0.29	**88.36** ±0.49	TI-Gradual	**70.87** ±0.20	**56.46** ±0.68	**88.94** ±0.04	**86.98** ±0.44
TernaryBERT	Budget-C	73.40 ±1.30	54.11 ±1.11	**88.60** ±0.02	88.43 ±0.58	XTC	73.76 ±0.54	56.30 ±0.67	88.91 ±0.03	**87.38** ±0.19
TI-Gradual	Budget-C	**73.82** ±0.41	**55.05** ±1.13	**88.60** ±0.01	**88.62** ±0.02	TI-Gradual	**74.48** ±0.79	**56.32** ±1.13	**88.92** ±0.03	87.34 ±0.41

Table 4: Evaluation results of BERT family models on GLUE benchmark (Small dataset) with Budget-A/C.

	ViT-B with Short Fine-Tuning (1K steps)				ViT-B with Fine-Tuning (10K/20K steps)			
Dataset	CIFAR100		CIFAR10		CIFAR100		ImageNet	
Method	2bit	4bit	2bit	4bit	2bit	4bit	2bit	4bit
Full-Prec	92.78		99.1		92.78		82.65	
Baseline	84.61 ±0.12	89.80 ±0.09	97.32 ±0.02	98.46 ±0.02	89.57 ±0.04	91.82 ±0.07	75.40 ±0.12	79.92 ±0.12
Ours	**85.28** ±0.04	**90.19** ±0.12	**97.59** ±0.05	**98.59** ±0.05	**90.07** ±0.04	**91.98** ±0.10	**76.66** ±0.04	**80.40** ±0.14

Table 5: Evaluation QAT performance of ImageNet-21K pre-trained ViT-B on Vision benchmarks with shorter train and longer train.

- All teacher intervention options achieve higher accuracy than the baseline (TernaryBERT).
- The preference between TI-M and TI-O varies across the tasks and models. For example, TI-M is mostly preferred on CoLA and RTE, while there is a marginal preference for TI-O for STS-B.
- For all the cases (except one corner case for BERT-Large on MRPC), TI-G outperforms the other TI options. Although the accuracies of TI-M and TI-O are similar, TI-G significantly outperforms both.

From these observations, we can conclude that the proposed gradual intervention significantly improves the QAT convergence for regaining model accuracy.

Scenario-2: Prolonged Fine-tuning Table 3 summarizes the experimental results of prolonged fine-tuning. For fair comparisons, we followed the instructions of XTC to match the fine-tuning budgets, learning rates, and the model compression mechanism (including both TinyBERT and SkipBERT). As expected, TI-G achieves higher accuracy with Budget-A/C than Budget-O4. But TI-G's accuracy is higher than TernaryBERT and XTC for these prolonged fine-tuning budgets with noticeable margins. In fact, XTC's average accuracy on Budget-A (for SkipBERT-6L) is lower than TI-G's average accuracy on Budge-O, highlighting superior convergence of TI-G compared to XTC.

5.3 QAT Accuracy on ViT Benchmarks

We further evaluate the proposed teacher intervention method on vision Transformer (ViT). Table 5 summarizes the QAT accuracies of ViT fine-tuned for CIFAR10, CIFAR100, and ImageNet with the fine-tuning budgets following the original paper Dosovitskiy et al. (2020). As shown in the figure, TI-G outperforms TernaryBERT on all the test cases for both ternary (2-bit) and 4-bit quantization-aware training. It is noteworthy that TI-G is exceptionally effective for fine-tuned ImageNet.

6 Conclusion

In this work, we proposed a proactive knowledge distillation method for improving convergence of QAT for ultra-low precision Transformers called teacher intervention. The proposed method intervenes in the propagation of quantization error to suppress accuracy degradation and improve QAT's convergence speed. We demonstrate that the proposed method outperforms the state-of-the-art in achieving higher QAT accuracy on various fine-tuned Transformers.

REFERENCES

Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7350–7357, 2020.

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. Binarybert: Pushing the limit of bert quantization. *arXiv preprint arXiv:2012.15701*, 2020.

Maximiliana Behnke and Kenneth Heafield. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2664–2674, 2020.

Aishwarya Bhandare, Vamsi Sripathi, Deepthi Karkada, Vivek Menon, Sun Choi, Kushal Datta, and Vikram Saletore. Efficient 8-bit quantization of transformer neural machine language translation model. *arXiv preprint arXiv:1906.00532*, 2019.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Mitchell A Gordon, Kevin Duh, and Nicholas Andrews. Compressing bert: Studying the effects of weight pruning on transfer learning. *arXiv preprint arXiv:2002.08307*, 2020.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. Fine-tuning bert for low-resource natural language understanding via active learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1158–1171, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. *arXiv preprint arXiv:2101.01321*, 2021.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020. URL `https://www.aclweb.org/anthology/2020.emnlp-main.574`.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Yihuan Mao, Yujing Wang, Chufan Wu, Chen Zhang, Yang Wang, Quanlu Zhang, Yaming Yang, Yunhai Tong, and Jing Bai. Ladabert: Lightweight adaptation of bert through hybrid model compression. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3225–3234, 2020.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. doi: 10.18653/v1/d16-1264.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4323–4332, 2019.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2158–2170, 2020.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*, 2020.

Xiaoxia Wu, Zhewei Yao, Minjia Zhang, Conglong Li, and Yuxiong He. Extreme compression for pre-trained transformers made simple and efficient, 2022.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1063–1077, 2021.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit bert. *arXiv preprint arXiv:1910.06188*, 2019.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning. In *International Conference on Learning Representations*, 2020a.

Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. Ternarybert: Distillation-aware ultra-low bit bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 509–521, 2020b.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.