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

Minsoo Kim¹, Kyuhong Shim², Seongmin Park¹, Wonyong Sung² and Jungwook Choi¹*

¹Department of Electronic Engineering, Hanyang University
²Department of Electrical and Computer Engineering, Seoul National University

(minsoo2333, skstjdals, choij)@hanyang.ac.kr
(sku20, wysung)@snu.ac.kr

Abstract

Pre-trained Transformer models such as BERT have shown great success in a wide range of applications, but at the cost of substantial increases in model complexity. Quantization-aware training (QAT) is a promising method to lower the implementation cost and energy consumption. However, aggressive quantization below 2-bit causes considerable accuracy degradation due to unstable convergence, 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 of QAT and expediting the convergence. Furthermore, we propose a gradual intervention mechanism to stabilize the recovery of subsections of Transformer layers from quantization. The proposed schemes enable fast convergence of QAT and improve the model accuracy regardless of the diverse characteristics of downstream fine-tuning tasks. We demonstrate that TI consistently achieves superior accuracy with significantly lower fine-tuning iterations on well-known Transformers of natural language processing as well as computer vision compared to the state-of-the-art QAT methods.

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., 2019; Raffel et al., 2020; Brown et al., 2020) and computer vision (CV) (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021a). 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 have been increasing needs 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., 2019) contain a few hundred million parameters, incurring profound memory and computation overhead for resource-constrained devices with limited memory and computing fabric. Therefore, seminal research efforts attempted to reduce this burden via model compression. (Behnke and 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...
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., 2021). 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), efforts on understanding and improving QAT on small dataset tasks are limited. XTC (Wu et al., 2022) recently revealed that a simple expansion in fine-tuning iterations could heal the QAT accuracy. Still, it caused an order-of-magnitude increase in QAT time, hindering the broad deployment of quantized Transformers.

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 the difficulty of quantization on few-sample fine-tuning originates from disruption of loss surface due to quantization error propagation. To mitigate this undesirable phenomenon, we propose TI to intervene layer-wise signal propagation with the intact signal from the teacher. TI removes the interference of propagated quantization errors to smooth out the loss surface and expedite the convergence. We further discover that subsections of Transformer layers exhibit different susceptibility to quantization error for diverse downstream tasks. Thus, we propose a gradual intervention mechanism that first intervenes at the attention output for stable tuning of the feed-forward network, followed by self-attention map intervention for its recovery from quantization. The proposed gradual intervention along the subsections of Transformer layers enables fast convergence of QAT and improves the model accuracy regardless of the diverse characteristics of downstream fine-tuning tasks. We perform an extensive evaluation on various fine-tuned Transformers (BERT-base/large (Devlin et al., 2019), TinyBERT-4L/6L (Jiao et al., 2020), and SkipBERT-6L (Wu et al., 2022) for NLP, and ViT (Dosovitskiy et al., 2020) for CV) and demonstrate that TI consistently achieves superior accuracy with lower fine-tuning iterations compared to the state-of-the-art QAT methods. In particular, TI outperforms TernaryBERT (Zhang et al., 2020b) on GLUE tasks with $12.5 \times$ savings in fine-tuning hours, as shown in Fig.1.

2 Related Work

2.1 Knowledge Distillation for BERT Compression

Knowledge distillation (KD) (Hinton et al., 2015) is a transfer learning framework that passes on knowledge of a large model (teacher) to a smaller one (student). Since KD provides extra guidance on how the student should behave, it can help mitigate accuracy degradation for model compression. Therefore, KD has been widely employed in training 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 shown in DistilBERT (Sanh et al., 2019). In addition to the distillation loss at the model output, PKD (Sun et al., 2019) suggested loss on intermediate output that matched the normalized output of the teacher and the student at each Transformer layer. MobileBERT (Sun et al., 2020) also employed per-head attention map transfer along with the customized network structure for constructing efficient Transformers. 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 above mentioned studies focused on the task-agnostic BERT, there have been several efforts (Tang et al., 2019; Aguilar et al., 2020) to train tiny task-specific students. 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., 2020) proposed a two-step KD, the first step for general distillation, followed by task-specific distillation.

Although extensive research has been conducted to utilize KD for BERT compression, there are
limited efforts in investigating and developing KD techniques for model quantization, an orthogonal model compression method. In this work, we develop a new KD technique primarily focusing on improving QAT of pre-trained Transformers. In particular, we reveal that more aggressive intervention of the teacher on the subsections of each Transformer 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., 2021) 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., 2019) is built with Transformer layers (Vaswani et al., 2017), which include 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 \(H\) be the number of attention heads and \(d_h = d/H\). \(W^Q_h, W^K_h, W^V_h \in \mathbb{R}^{d \times d_h}\) are the weight parameters converting \(X_l\) into Query (\(Q = X_l W^Q_h\)), Key (\(K = X_l W^K_h\)), and Value (\(V = X_l W^V_h\)), respectively. Then, attention score (\(SA_h = Q K^\top\)), self-attention map (\(SA_h = \text{Softmax}_h(\frac{SA_h}{\sqrt{d}})\)), and attention context (\(AC = SA_h V\)) can be defined.

The attention output (AO) is defined as \(AO = \text{MHA}(X_l) = \text{Concat}(AC_1, \ldots, AC_{NH}) \times W^O\). Motivated by (Kobayashi et al., 2020), the attention output can be re-written per each token \(i\):

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

where \(f(x) := (x^W + b^V)W^O\) and \(\alpha_{i,j}\) is \(j\)-th value of \(i\)-th token in \(SA_h\). Therefore, the computation of MHA consists of two parts: self-attention generation (SA-GEN) corresponding to the attention map (\(\alpha\)) and self-attention propagation (SA-PROP) corresponding to \(f(x)\). Fig. 3 illustrates the structure of a Transformer layer indicating regions corresponding to SA-GEN and SA-PROP. FFN consists of two fully-connected layers with weight parameters \(W^1\) and \(W^2\):

\[
\text{FFN}(Y_l) = \text{GeLU}(X_l W^1 + b^1)W^2 + b^2. \tag{2}
\]
Therefore, an output of Transformer layer $X_{l+1}$ is defined as:
\[
Y_l = \text{LayerNorm}(X_l + \text{MHA}(X_l)),
\]
\[
X_{l+1} = \text{LayerNorm}(Y_l + \text{FFN}(Y_l)).
\]

3.2 Quantization-Aware Training
Quantization-aware training (QAT) emulates inference-time quantization during fine-tuning of the full-precision model to adjust the parameters to be robust to the quantization error (Hwang and Sung, 2014) (Hubara et al., 2016). In particular, ternary quantization represents all the weight parameters ($W^Q, W^K, W^V, W^O, W^2$) into ternary values $t \in \{+1, 0, -1\}$ along with a scale factor $s$ for sub-2bit inference. In this work, we follow the approach of TWN (Zhu et al., 2016) that analytically estimates the optimal scale factor $s$ and the ternary weight $W_t = \{t\}^{|W|}$ to minimize $\|W - s \cdot W_t\|$.

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 $X_{l+1}$ as well as the attention score (AS) with mean squared error (MSE) loss:
\[
L_{\text{trm}} = \sum_{l=1}^{L+1} \text{MSE}(X^S_l, X^T_l) + \sum_{l=1}^{L} \text{MSE}(AS^S_l, AS^T_l),
\]
where $S$ and $T$ represent the student and teacher models, respectively. Also, the output logits of the student ($P^S$) and the teacher ($P^T$) are used in TernaryBERT to compute the cross-entropy (CE) loss:
\[
L_{\text{pred}} = \text{CE}(P^S, P^T).
\]

We follow the settings of TernaryBERT as our baseline QAT method.

3.3 Challenges
Despite attempts to bridge the accuracy gap, prior works on ultra-low precision Transformers (Bai et al., 2021; 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 unstable. But we observed that QAT often fails even if it fine-tunes from the successfully trained model.

4 Method
4.1 Teacher Intervention
Teacher intervention (TI) is a KD method that aggressively intervenes in the student’s signal propagation along the Transformer layers to suppress propagation of quantization error. Fig. 3(b) illus-
Figure 3: (a) Architecture of Transformer layer. (b) Locations of Teacher Intervention: output intervention (TI-O) and map intervention (TI-M). \( \text{AO}^T \) and \( \text{SA}^T \) are teacher’s attention outputs. (c) Accuracy curves of controlled experiments. Group-A: Quantize all sub-layers with TI-O (Case 1) or do not quantize MHA (Case 2), Group-B: Quantize all sub-layers with TI-M (Case 3) or do not quantize SA-GEN (Case 4). Note rapid convergence for the cases with TI.

The development of TI is motivated by the previous observation of the aggravating impact of quantization error along the layers (cf. Fig. 2(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 and TI-M with two groups of quantization cases:

- Group-A: Quantize all sub-layers with TI-O (Case 1) or do not quantize MHA (Case 2).
- Group-B: Quantize all sub-layers with TI-M (Case 3) or do not quantize SA-GEN (Case 4).

The key difference between the two groups is that Case2/4 propagates the quantization error through the sub-layers while Case1/3 does not, thanks to TI. Fig. 3(c) shows the convergence curves of the four cases on CoLA. As shown in the figure, Case1/3 converges rapidly to full-precision accuracy despite the ultra-low bit quantization in all the sub-layers. Whereas Case2/4 converges slowly to the sub-optimal point with noticeable accuracy degradation. Although Case2/4’s MHA/SA-GEN computations are in full-precision, the error propagated from its preceding quantized FFN sub-layers still affects it, corrupting the attention output/self-attention map. On the other hand, TI-O and TI-M interrupt this error propagation to stabilize QAT on FFN sub-layers and MHA outputs, respectively. Therefore, Fig. 2(b) confirms that the error propagation disappears when TI is applied.

Motivated by this insightful observation, we devise a new QAT method that employs TI for the step-by-step reconstruction of sub-layers of Transformers. In the first step, TI is used to fine-tune the quantized weights of either FFN (TI-O) or SA-PROP (TI-M) sub-layers. Note that the convergence in this step is quick, as shown in Fig. 3(c). Therefore, Step1 takes only a fraction of fine-tuning epoch. In the second step, quantization is applied to the entire weights of Transformer layers for QAT. Since part of the parameters is already trained in Step1 with the guidance of TI, Step2 converges faster to a superior local minimum. In Sec. 5, we demonstrate this improved convergence leads to boosts in accuracy for ultra-low bit Transformers.
Figure 4: (a) Cross-entropy loss of QAT with different TI methods, and (b) Layer-wise MSE loss for different TI methods (Left: STS-B, Right: CoLA). Note that TI-G converges to lower loss than TI-M and TI-O on both tasks.

### 4.2 Gradual Teacher Intervention

<table>
<thead>
<tr>
<th>Type</th>
<th>TI-M</th>
<th>TI-O</th>
<th>TI-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-layer</td>
<td>GEN</td>
<td>PROP</td>
<td>FFN</td>
</tr>
<tr>
<td>Step1-Phase1</td>
<td>Q+TI</td>
<td>Q</td>
<td>Q+TI</td>
</tr>
<tr>
<td>Step1-Phase2</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
</tr>
<tr>
<td>Step2</td>
<td>Q</td>
<td>Q</td>
<td>Q</td>
</tr>
</tbody>
</table>

Table 1: Quantization settings for teacher intervention. Different from TI-M and TI-O, gradual teacher intervention (TI-G) applies TI to quantization (Q) from large (GEN+PROP) to small (GEN) scope in Step1.

Given multiple teacher intervention options, which would achieve the best performance? In this section, we propose a unified approach that gradually applies the output intervention followed by the map intervention 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, developing a unified solution that utilizes teacher intervention in various scopes is beneficial. As a natural combination, we propose a gradual teacher intervention mechanism that applies TI-O first to tune the FFN sub-layers (Step1-Phase1), followed by TI-M to recover the self-attention map (Step1-Phase2). The proposed method, called **gradual intervention** (TI-G), has shown practical success in most ultra-low bit Transformers studied in this work. (We conducted an ablation study for the other possibilities of a unified solution for teacher intervention in \ref{sec:A3}.) Table 1 summarizes the TI settings. Note that the two phases of TI-G in Step1 do not increase the total number of iterations thanks to fast convergence.

Fig. 4(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. In particular, on STS-B, TI-G benefits from the output intervention in the first phase to favorably warm up the FFN sub-layers, and thus the map intervention in the next phase can reduce the loss more than TI-M. A similar situation happens in the case of CoLA.

Fig. 4(b) investigates the internal behavior of Transformers via the layer-wise MSE at the output of MHA. As discussed earlier, the fine-tuned Transformers exhibit distinct characteristics depending on the downstream tasks. For example, in the case of STS-B, TI-M is more effective in reducing MSE since the recovery of the self-attention map is essential. On the other hand, CoLA has preferred TI-O for its focus on tuning FFN sub-layers. In both cases, TI-G’s gradual intervention with decreasing scopes from the attention output to the self-attention map helps achieve the smallest MSE. Therefore, this investigation suggests that the proposed unified intervention mechanism can manage diverse characteristics of fine-tuned Transformers for various downstream tasks. We demonstrate the general applicability of TI-G in the next section.

## 5 Experiments

In this section, we evaluate TI for QAT of fine-tuned BERT and vision Transformer with ternary weight quantization. We demonstrate that TI significantly boosts the convergence of QAT to achieve higher model accuracy within shorter fine-tuning time compared to the state-of-the-art QAT methods
(TernaryBERT (Zhang et al., 2020b) and XTC (Wu et al., 2022)) on various Transformer tasks and datasets. More extensive evaluation results on a wide range of BERT models as well as the ablation studies on various aspects of TI can be found in Appendix A.3. Due to limited space, detailed experiment settings are also summarized in Appendix A.1.

<table>
<thead>
<tr>
<th>Iterations</th>
<th>CoLA</th>
<th>RTE</th>
<th>MRPC</th>
<th>STS-B</th>
<th>Avg. Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Budget-O</td>
<td>1,603</td>
<td>233</td>
<td>343</td>
<td>538</td>
<td>1</td>
</tr>
<tr>
<td>Budget-A</td>
<td>13,325</td>
<td>4,468</td>
<td>7,050</td>
<td>10,066</td>
<td>16</td>
</tr>
<tr>
<td>Budget-C</td>
<td>159,909</td>
<td>53,621</td>
<td>84,611</td>
<td>120,795</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 2: Different budgets for fine-tuning iterations for GLUE tasks.

5.1 QAT Accuracy on Few-Sample BERT

First, we perform an extensive performance comparison of the proposed teacher intervention methods with the state-of-the-art QAT methods: TernaryBERT (Zhang et al., 2020b) and XTC (Wu et al., 2022). We evaluate TI on fine-tuned BERT\textsubscript{base} (12 layer), BERT\textsubscript{large} (24 layer) (Devlin et al., 2019), TinyBERT (Jiao et al., 2020) (6 layer), and SkipBERT (Wu et al., 2022) (6 layer) for GLUE tasks (Wang et al., 2018). We report the experimental results for the small dataset tasks (less than 10k of dataset size) here; the results on the other models and tasks are reported in Appendix A.2.

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**: 2× iterations of Budget-O.
- **Budget-O4**: 4× iterations of Budget-O.
- **Budget-A/C**: prolonged fine-tuning budgets employed in XTC (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 16× larger than Budget-O. 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).

**Scenario-1: Few-Sample Fine-tuning**: Table 3 summarizes the experimental results of few-sample fine-tuning. Consistent with the prior observations (Wu et al., 2022), the accuracy of TernaryBERT increased as the fine-tuning budget grew from Budget-O to Budget-O4. But all teacher intervention options have improved these baseline accuracies. Note that the preference between TI-M and TI-O varies across the tasks and models. For example, TI-M is mainly preferred on CoLA and RTE, while there is a marginally higher preference for TI-O on STS-B. For most cases, however, TI-G outperforms the other TI options. Although the accuracies of TI-M and TI-O are similar, TI-G’s accuracy significantly exceeds 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 4 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 on both TinyBERT and SkipBERT. As expected, TI-G’s accuracy is significantly increased from Budget-O to Budget-A/C than Budget-O4. Interestingly, 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 Budget-O4 (for TinyBERT-6L), highlighting the superior convergence of TI-G compared to XTC.

5.2 QAT Accuracy on Vision Transformers

We further evaluate the proposed teacher intervention method on 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 Table, 5, TI-G outperforms TernaryBERT on all the QAT cases for ternary weight quantization. It is noteworthy that TI-G is exceptionally effective for fine-tuned ImageNet.
5.3 Convergence Analysis

In Sec. 4, we discussed that TI impedes propagation of quantization errors and flattens the loss surface for favorable convergence as shown in Fig. 2. To quantitatively analyze the convergence of QAT, we conducted the Hessian eigenvalue analysis (Park and Kim, 2022) on $BERT_{base}$ over CoLA task. As shown in Fig. 5, TernaryBERT suffers large magnitude (positive and negative) eigenvalues. The large magnitude Hessian eigenvalues indicate sharp loss surfaces, unfriendly for quantization (Shen et al., 2020; Dong et al., 2019). A prolonged fine-tuning would suppress the large Hessian eigenvalues at the expense of significantly increased fine-tuning computations. In contrast, TI significantly reduces the magnitude with an order of magnitude smaller fine-tuning iteration. From this analysis, we can conclude that TI helps flatten the loss surface to improve the convergence of QAT.

5.4 Further Analysis of TI-G: Self-Attention Map Recovery

In this section, we provide an analysis of TI-M’s effectiveness in the use of TI-G. Since the self-attention map encodes the relational representation of symbols within a sequence, quantization on SAGEN would disrupt this relational encoding. Therefore, TI on the self-attention map (TI-M) is crucial for maintaining relative order within the self-attention map. As shown in Sec. 4.2, TI-O lacks consideration of the self-attention map recovery, requiring gradual teacher intervention mechanism.
that applies TI-M after TI-O for recovering the self-attention map. Fig. 6 shows two empirical evidences that TI-M helps to recover the self-attention map with TI-O.

Fig. 6(a) shows the ranking loss computed on the self-attention map of BERT$_{base}$ after QAT on CoLA. Note that the ranking loss quantitatively indicates how much the relative order in the quantized self-attention map has been skewed from the teacher’s self-attention map (Liu et al., 2021b). As can be seen, QAT without TI-M leads to significant disruption of the self-attention map. In contrast, QAT with TI-M maintains the original ranking of cross-symbol correlation, successfully suppressing the ranking loss. We further visualize each quantized model’s resulting self-attention maps in Fig. 6(b). Contrary to using TI-O only, TI-O with TI-M successfully recovers the teacher self-attention map’s distinctive feature. From this observation, we can conclude that TI-M approach used with TI-O is particularly beneficial for preserving the self-attention map under quantization.

6 Conclusion

In this work, we proposed a proactive knowledge distillation method for improving the 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 QAT methods in achieving higher accuracy on various fine-tuned Transformers with smaller fine-tuning budgets. Our code is available at https://github.com/MarsJacobs/ti-kd-qat.

7 Limitations

This work proposes a proactive knowledge distillation method called teacher intervention for improving the convergence of QAT with small fine-tuning budgets. Although our analysis reveals that quantization error propagation is one of the main causes of sub-optimal convergence of QAT, a more in-depth investigation of the detailed recovery mechanism of QAT from quantization error would be interesting. Also, extending the discovery of proactive knowledge distillation to diverse Transformer architectures, including encoder-decoder and decoder-only models, would be a promising future research direction.

Acknowledgements

This work was partly supported by Samsung Advanced Institute of Technology, Samsung Electronics Co., Ltd, Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (2020-0-01297, Development of Ultra-Low Power Deep Learning Processor Technology using Advanced Data Reuse for Edge Applications) and National Research Foundation of Korea (NRF) grant funded by Korea government (MSIT) (No. 2021R1A2C1013513).

References


Yihuan Mao, Yujing Wang, Chufan Wu, Chen Zhang, Yang Wang, Quanlu Zhang, Yaming Yang, Yunhai Tong, and Jing Bai. 2020. LadaBERT: Lightweight


A Appendix

In this section, we provide experimental setup and TI implementation in Sec. A.1. We also provide evaluation results of GLUE (Wang et al., 2018) large datasets (SST-2, QNLI, MNLI, QQP) in Sec. A.2 and ablation studies in Sec. A.3. Further analysis of TI-G’s effectiveness is provided in 5.4.

A.1 Experimental Settings

Models Description We use task-specific, fine-tuned BERT family models to evaluate our method. Further information on each model is as follows:

1. BERT$_{base}$ $L=12$, $d=768$ $N_H=12$, Contains about 110M parameters.

2. BERT$_{large}$ $L=24$, $d=1024$ $N_H=16$, Contains about 340M parameters.

3. TinyBERT-4L $L=4$, $d=312$ $N_H=12$, Contains about 14M parameters. This compressed model is fine-tuned using task-specific distillation method (Jiao et al., 2020) initialized from general pre-trained TinyBERT model file. ²

4. TinyBERT-6L $L=6$, $d=768$ $N_H=12$, Contains about 67M parameters. This compressed model is fine-tuned using task-specific distillation method (Jiao et al., 2020) initialized from general pre-trained TinyBERT model file. ³

5. SkipBERT-6L $L=6$, $d=768$ $N_H=12$, Contains about 67M parameters. This compressed model parameters are initialized from XTC (Wu et al., 2022) model initialization method: using every other layer of the fine-tuned BERT$_{base}$’s parameter to initialize the layer-reduced model.

Datasets and Settings We evaluate our method on GLUE benchmark (Wang et al., 2018), which is a collection of resources for training, evaluating, and analyzing natural language understanding systems. For the training setting, we use the batch size of 16 for CoLA and 32 for other tasks. The learning rate starts from zero and gradually increases to 2e-5 during the warm-up stage and decays linearly to 2e-9 for total training epochs.

²https://huggingface.co/huawei-noah/TinyBERT_4L_zh
³https://huggingface.co/huawei-noah/TinyBERT_6L_zh

<table>
<thead>
<tr>
<th></th>
<th>(Zhang et al., 2020b)</th>
<th>(Wu et al., 2022)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Budget-O</td>
<td>Budget-O2</td>
</tr>
<tr>
<td>DA</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 6: Training budgets for the GLUE tasks. DA is data augmentation.

<table>
<thead>
<tr>
<th></th>
<th>CoLA</th>
<th>RTE</th>
<th>MRPC</th>
<th>STS-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train (noDA)</td>
<td>8,551</td>
<td>2,490</td>
<td>3,668</td>
<td>5,749</td>
</tr>
<tr>
<td>Train (DA)</td>
<td>213,212</td>
<td>142,991</td>
<td>225,630</td>
<td>322,121</td>
</tr>
<tr>
<td>DA / noDA</td>
<td>24.9</td>
<td>57.4</td>
<td>61.5</td>
<td>56.0</td>
</tr>
<tr>
<td>Budget-O Time</td>
<td>259</td>
<td>97</td>
<td>72</td>
<td>223</td>
</tr>
<tr>
<td>Budget-A Time</td>
<td>2,149</td>
<td>1,862</td>
<td>1,484</td>
<td>4,160</td>
</tr>
<tr>
<td>Budget-C Time</td>
<td>25,792</td>
<td>22,342</td>
<td>17,813</td>
<td>49,915</td>
</tr>
</tbody>
</table>

Table 7: Training details for the GLUE tasks. DA is data augmentation.

Comparison of Training Cost We performed an extensive performance comparison in Sec. 5.1 along the training budget to investigate the convergence of QAT methods. Table. 6 shows the training setting of each budget with the Data Augmentation (DA) option, which is first proposed by (Zhang et al., 2020b) (Jiao et al., 2020) artificially expanding the training sample for prolonged fine-tuning, causing a blow-up in fine-tuning costs. To investigate how much the DA magnifies the QAT overhead, we compare the number of training samples of GLUE small datasets and the number of data-augmented training samples in Table. 7. Furthermore, we measured the total QAT time for each Budget over GLUE small datasets with the A6000 single GPU, indicating that prolonged fine-tuning with DA would become profoundly painful, increasing model deployment costs and time.

Teacher Intervention Implementation Our experiments were performed on A6000 GPUs. We use Pytorch 1.10.2 for the implementation of TI and this implementation is based on the TernaryBERT Pytorch code base. ⁴ We would like to emphasize that TI has strong advantages of simplicity for implementation. To implement TI operation, we modified the teacher model forward operation, enforcing the teacher model to return layer-wise MHA outputs and SA-GEN outputs which will be used in TI-O and TI-M respectively. When the teacher model’s layer-wise required outputs are returned,
we delve deeper into evaluating the efficacy of TI two step QAT and output loss In this section, we derive deeper into evaluating the efficacy of the TI method. As TI-O employs one additive loss for internal distillation (Attention Output MSE loss, called output loss) and TI QAT is performed in two steps, conducting an ablation study on TI (exploiting output loss and two-step QAT separately) provides a deeper understanding of our method. Table. 10 demonstrates the effect of output loss and the use of a two-step method with TI independently with BERT family models over GLUE small datasets. As can be seen, both factors act positively on the performance of quantized

**A.3 Ablation Study**

**TI two step QAT and output loss** In this section, we delve deeper into evaluating the efficacy of the TI method. As TI-O employs one additive loss for internal distillation (Attention Output MSE loss, called output loss) and TI QAT is performed in two steps, conducting an ablation study on TI (exploiting output loss and two-step QAT separately) provides a deeper understanding of our method. Table. 10 demonstrates the effect of output loss and the use of a two-step method with TI independently with BERT family models over GLUE small datasets. As can be seen, both factors act positively on the performance of quantized

**Table 8**: Accuracy comparison of QAT methods on BERT family (large-sample GLUE tasks) without and with TI for regular fine-tuning iterations (Budget-O). Each experiment is repeated 3 times.

<table>
<thead>
<tr>
<th>Method</th>
<th>CoLA</th>
<th>STS-B</th>
<th>MRPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-Prec</td>
<td>73.28</td>
<td>58.04</td>
<td>89.24</td>
</tr>
<tr>
<td>TernaryBERT Budget-O</td>
<td>73.16</td>
<td>54.81</td>
<td>89.90</td>
</tr>
<tr>
<td>TI-G</td>
<td>74.24</td>
<td>58.29</td>
<td>89.01</td>
</tr>
<tr>
<td>TernaryBERT Budget-C</td>
<td>73.40</td>
<td>58.25</td>
<td>89.23</td>
</tr>
<tr>
<td>TI-G</td>
<td>75.93</td>
<td>58.91</td>
<td>89.39</td>
</tr>
</tbody>
</table>

**Table 9**: Accuracy comparison of QAT methods on BERT family (few-sample GLUE tasks) without and with TI for regular fine-tuning iterations (Budget-A/C). Each experiment is repeated 10 times.

<table>
<thead>
<tr>
<th>Method</th>
<th>CoLA</th>
<th>STS-B</th>
<th>MRPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-Prec</td>
<td>73.28</td>
<td>58.04</td>
<td>89.24</td>
</tr>
<tr>
<td>TernaryBERT Budget-A</td>
<td>73.16</td>
<td>54.81</td>
<td>89.90</td>
</tr>
<tr>
<td>TI-G</td>
<td>74.24</td>
<td>58.29</td>
<td>89.01</td>
</tr>
<tr>
<td>TernaryBERT Budget-C</td>
<td>73.40</td>
<td>58.25</td>
<td>89.23</td>
</tr>
<tr>
<td>TI-G</td>
<td>75.93</td>
<td>58.91</td>
<td>89.39</td>
</tr>
</tbody>
</table>

**Table 10**: Ablation study of Teacher Intervention. Each result is averaged score over GLUE small datasets (RTE, CoLA, STS-B, MRPC)

<table>
<thead>
<tr>
<th>Method</th>
<th>CoLA</th>
<th>STS-B</th>
<th>MRPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-Prec</td>
<td>72.51</td>
<td>65.88</td>
<td>72.17</td>
</tr>
<tr>
<td>TernaryBERT Budget-O</td>
<td>73.92</td>
<td>66.85</td>
<td>72.77</td>
</tr>
<tr>
<td>TI-G</td>
<td>74.13</td>
<td>67.13</td>
<td>73.02</td>
</tr>
</tbody>
</table>

**Table 11**: Comparison of unified TI strategies. Each result is averaged score over GLUE small datasets (RTE, CoLA, STS-B, MRPC)
Interestingly, two-step QAT factor provides a powerful impact (Row 2) in compressed models (TinyBERT-4L and TinyBERT-6L), and output loss shows a distinct impact (Row 3) in the BERT\textsubscript{large} model. When both factors are implied together (Row 4), performance is boosted further.

**Exploration of unified teacher intervention methods** In this section, we investigate the best unified teacher intervention strategy. We proposed a unified intervention approach (TI-G) in Sec. 4.2, applying TI with output intervention (TI-O) followed by map intervention (TI-M). As more scheduling choices are left when using TI-O and TI-M together, we propose additional two scheduling methods, Inverted TI, and Stochastic TI. Inverted TI applies TI-O and TI-M in the opposite order of TI-G, and stochastic TI means choosing TI-O and TI-M options randomly in every training iteration, ensembling each TI option’s effect in a stochastic way. Table 11 shows the comparison between unified TI scheduling methods with BERT family models over GLUE small datasets. TI-G outperforms other unified TI scheduling methods in almost every model, which shows our proposed teacher intervention unification approach’s efficacy.