When Will Gradient Regularization Be Harmful?

Yang Zhao [1] **Hao Zhang** [1] **Xiuyuan Hu** [1]

Abstract

Gradient regularization (GR), which aims to penalize the gradient norm atop the loss function, has shown promising results in training modern over-parameterized deep neural networks. However, can we trust this powerful technique? This paper reveals that GR can cause performance degeneration in adaptive optimization scenarios, particularly with learning rate warmup. Our empirical and theoretical analyses suggest this is due to GR inducing instability and divergence in gradient statistics of adaptive optimizers at the initial training stage. Inspired by the warmup heuristic, we propose three GR warmup strategies, each relaxing the regularization effect to a certain extent during the warmup course to ensure the accurate and stable accumulation of gradients. With experiments on Vision Transformer family, we confirm the three GR warmup strategies can effectively circumvent these issues, thereby largely improving the model performance. Meanwhile, we note that scalable models tend to rely more on the GR warmup, where the performance can be improved by up to 3% on Cifar10 compared to baseline GR. Code is available at https://github.com/zhaoyang-0204/gnp.

1. Introduction

Advancements in computational hardware have catalyzed the design of modern deep neural networks such as the Transformers (Dosovitskiy et al., 2021; Vaswani et al., 2017b; Brown et al., 2020; Liu et al., 2021; Touvron et al., 2021), characterized by an extraordinarily vast number of parameters, far exceeding their predecessors. In this context, regularization techniques emerge as a more pivotal role to resist overfitting in training these over-parameterized net-
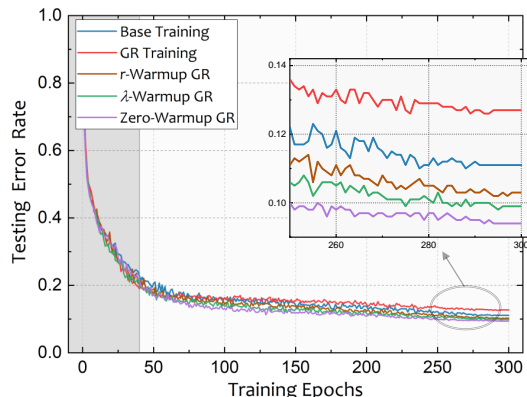

Figure 1. Comparison of test error rates (lower values are preferable) of the ViT-B model on the Cifar10 dataset under Base training, GR, and our three proposed GR warmup strategies. All the training instances have also applied the LR warmup. Notably, the performance with LR warmup and normal GR (red line) can be worse compared to training with only LR warmup (blue line).

works (Srivastava et al., 2014; Ioffe & Szegedy, 2015; Ba et al., 2016; Foret et al., 2021).

Recent studies highlight the gradient regularization (GR) as an effective regularization strategy (Barrett & Dherin, 2020; Smith et al., 2021; Zhao et al., 2022b; Karakida et al., 2023; Reizinger & Huszár, 2023). By imposing an additional penalty concerning gradient norm atop the loss function, this technique deliberately biases the optimization process towards the attainment of flat minima, fostering better generalization. Meanwhile, it has been revealed that a strong association exists between GR and Sharpness-Aware Minimization (SAM) family (Foret et al., 2021), which posits SAM as a special parameter configuration in the first-order solution of GR (Zhao et al., 2022b; Karakida et al., 2023).

However, despite its practical utility, the scope and limitations of GR are yet to be fully understood, particularly in terms of establishing when it can be beneficial or safe to apply this technique. We find that GR can lead to serious performance degeneration (see Figure 1) in the specific scenarios of adaptive optimization such as Adam (Kingma & Ba, 2015) and RMSProp (Hinton et al., 2012). With both our empirical observations and theoretical analysis, we find that the biased estimation introduced in GR can induce the

---

[1]Department of Electronic Engineering, Tsinghua University. Correspondence to: Hao Zhang <haozhang@tsinghua.edu.cn>, Yang Zhao <zhao-yang@tsinghua.edu.cn>.

instability and divergence in gradient statistics of adaptive optimizers at the initial stage of training, especially with a learning rate warmup technique which originally aims to benefit gradient statistics (Vaswani et al., 2017a; Popel & Bojar, 2018; Liu et al., 2020). Notably, this issue tends to become more severe as the complexity of the model increases.

To mitigate this issue, we draw inspirations from the idea of warmup techniques, and propose three GR warmup strategies: λ-warmup, r-warmup and zero-warmup GR. Each of the three strategies can relax the GR effect during warmup course in certain ways to ensure the accuracy of gradient statistics. Then, training reverts to normal GR for the rest of training, allowing the optimization to fully enjoy the performance gain derived from the GR. Finally, we empirically confirm that all the three GR warmup strategies can not only successfully avoid this issue but further enhance the performance for almost all training cases. Of these strategies, the zero-warmup GR can give the best improvements, significant outperforming the baseline.

2. Background

2.1. Gradient Regularization: an Overview

Gradient regularization typically aims to impose an additional gradient norm penalty on top of the loss function (Barrett & Dherin, 2020; Smith et al., 2021; Zhao et al., 2022b; Karakida et al., 2023),

$$L^{(gr)}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \lambda ||\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})||_2 \qquad (1)$$

where $\boldsymbol{\theta}$ is the model parameter and λ denotes the regularization degree, effectively controlling the extent to which this regularization influences the overall process. Note that some research choose to penalize the $||\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})||_2^2$, which actually results in the same effect as $||\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})||_2$.

In practical applications, Hessian-free techniques are employed to approximate the involved gradient $L^{(gr)}(\boldsymbol{\theta})$,

$$g^{(gr)} = (1 - \frac{\lambda}{r})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) + \frac{\lambda}{r}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta} + \boldsymbol{\epsilon}) \qquad (2)$$

with

$$\boldsymbol{\epsilon} = r \cdot \frac{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})}{||\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})||_2} \qquad (3)$$

where the parameter r denotes as a small, positive scalar representing neighborhood perturbation, which is used for Taylor approximation.

Notably, the gradients in SAM (Foret et al., 2021) can be expressed as,

$$g^{(sam)} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta} + \boldsymbol{\epsilon}) \qquad (4)$$

which clearly means that SAM essentially regularizes gradient norm with a special parameter configuration in GR where $\lambda = r$ for any case.

2.2. Related Works

Recent research has been burgeoning in the field of GR with a specific emphasis on its applications and effects in deep learning. Barrett & Dherin (2020) along with Smith et al. (2021) have conducted studies first exploring the aspects of GR in deep learning. They have revealed that the stepped update of standard gradient descent implicitly regularizes the gradient norm through its inherent dynamics, especially when correlated with its continuous equivalent, which is identified as implicit GR. When considering explicit GR, they suggest that applying GR explicitly can yield even better results. Recently, Zhao et al. (2022b) and Karakida et al. (2023) proposes to directly penalize the gradient norm atop the loss and shows that such techniques can potentially give state-of-the-art performance. Meanwhile, it has been highlighted that GR exhibits a strong correlation with SAM (Zhao et al., 2022b; Karakida et al., 2023; Reizinger & Huszár, 2023).

Considering the close connections between GR and SAM family, we would also like to discuss works associated with flat minima. In (Hochreiter & Schmidhuber, 1997), the authors are the first to point out that the flatness of minima could be associated with the model generalization, where models with better generalization should converge to flat minima. And such claim has been supported extensively by both empirical evidences and theoretical demonstrations (Keskar et al., 2017; Dinh et al., 2017). In the meantime, researchers are also fascinating by how to implement practical algorithms to force the models to converge to such flat minima. By summarizing this problem to a specific minimax optimization, (Foret et al., 2021) introduce the SAM training scheme, which successfully guides optimizers to converge to flat minima. Since then, many SAM variants are proposed, such as improved approaches (Kwon et al., 2021; Zhuang et al., 2022) and efficient approaches (Du et al., 2021; Zhao et al., 2022a), in expectation to contributing the original SAM from various perspectives.

3. The Incompatibility Between Gradient Regularization and Adaptive Optimization

In this section, we will explore the effect of employing gradient regularization in the context of adaptive optimization. Firstly, we would present empirical results derived from practical training scenarios.

3.1. Practical Training with Gradient Regularization in Adaptive Optimization

Vanilla Adaptive Optimization Adaptive optimization is an advanced optimization algorithm in the training of DNNs. By dynamically adjusting learning rates during training, such a technique can largely improve the training

Table 1. Testing error rate of ViT models using Adam and RMSProp optimizers on Cifar-{10, 100} when employing a cross-utilization of LR warmup and standard GR techniques.

Model Architecture	Base Optimizer	LR Warmup	Cifar-10 Error [%]		Cifar-100 Error [%]	
			Vanilla	GR	Vanilla	GR
ViT-Ti	Adam	✗	$15.94_{\pm0.33}$	$14.37_{\pm0.31}$	$41.19_{\pm0.21}$	$39.99_{\pm0.39}$
		✓	$14.82_{\pm0.64}$	$13.92_{\pm0.19}$	$39.25_{\pm0.41}$	$39.28_{\pm0.50}$
	RMSProp	✗	$17.05_{\pm0.45}$	$16.32_{\pm0.30}$	$41.62_{\pm0.16}$	$41.06_{\pm0.90}$
		✓	$16.47_{\pm0.39}$	$15.79_{\pm0.25}$	$40.31_{\pm0.50}$	$40.02_{\pm0.22}$
ViT-S	Adam	✗	$15.43_{\pm0.52}$	$14.41_{\pm0.42}$	$42.83_{\pm0.34}$	$39.16_{\pm0.27}$
		✓	$12.07_{\pm0.53}$	$\mathbf{12.40_{\pm0.17}}$	$37.17_{\pm0.24}$	$36.65_{\pm0.23}$
	RMSProp	✗	$19.94_{\pm0.44}$	$17.74_{\pm0.23}$	$43.09_{\pm0.34}$	$40.63_{\pm0.30}$
		✓	$15.47_{\pm0.15}$	$14.77_{\pm0.35}$	$38.55_{\pm0.22}$	$37.90_{\pm0.19}$
ViT-B	Adam	✗	$18.72_{\pm0.21}$	$16.11_{\pm0.43}$	$44.16_{\pm0.50}$	$41.69_{\pm0.42}$
		✓	$10.83_{\pm0.28}$	$\mathbf{12.36_{\pm0.33}}$	$36.70_{\pm0.17}$	$36.65_{\pm0.26}$
	RMSProp	✗	$20.19_{\pm0.17}$	$19.06_{\pm0.67}$	$44.47_{\pm0.47}$	$42.28_{\pm0.65}$
		✓	$15.76_{\pm0.25}$	$14.97_{\pm0.25}$	$39.03_{\pm0.34}$	$39.04_{\pm0.35}$

efficiency and model performance, especially on complex, non-stationary problems. Unlike fixed learning rate methods, adaptive optimizers such as Adam and RMSProp can expediently navigate the model towards optimal solutions by tuning learning rates per parameter based on estimations of first and second moments of the gradients. For Adam, its gradient momentum $\phi(g; t)$ and adaptive learning rate $\psi(g; t)$ at step t can be expressed as,

$$\phi(g_0, \cdots, g_t; t) = \frac{(1 - \beta_1) \sum_{i=1}^{t} \beta_1^{t-i} g_i}{1 - \beta_1^t}$$
$$\psi(g_0, \cdots, g_t; t) = \sqrt{\frac{1 - \beta_2^t}{(1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} g_i^2}} \quad (5)$$

where g_t represents the gradient at step t.

Nonetheless, the initial stages of training often involve inaccurate estimation of gradients. Compared to other optimizers, the intrinsic reliance of adaptive optimizers on gradient moments makes it more difficult to get rid of these misleading estimations later in the process. As a consequence, this can potentially steer the optimization towards converging to suboptimal local minima.

A commonly adopted strategy to mitigate this issue is the implementation of learning rate (LR) warmup techniques. With this method, a lower learning rate than the base rate is applied during the initial stages of training, before being increased to the standard learning rate for the rest of the training period. Specifically, there are two typical LR warmup implementations, constant LR warmup and gradual LR warmup. For constant LR warmup, a low constant learning rate is maintained throughout the warmup period. Whereas for gradual LR warmup, the learning rate will ramp up progressively from a very low value to the normal level, commonly in a linear manner,

$$\eta(t) = \min\{\frac{t}{T_w}, 1\}\eta_0 \quad (6)$$

where T_w denotes the warmup steps and η denotes the base learning rate.

In practice, the gradual LR warmup method is often preferred over its constant warmup. Due to its smoother transition in learning rates, this method often yields better model performance. Accordingly, we will adopt the gradual LR warmup methodology in our subsequent investigations.

Training Implementation Firstly, we aim to empirically present the effect of cross-utilizing LR warmup and gradient regularization techniques in the context of adaptive optimization, with a particular emphasis on checking their compatibility in the training process. As for adaptive optimizers, we will choose the Adam optimizer and RMSProp optimizers.

For the network models, we will focus on the prevalent Vision Transformer (Dosovitskiy et al., 2021) models, considering that the conventional CNN models are commonly trained via the SGD optimizer to give better performance. We select to train ViT-Ti, ViT-S and ViT-B model architectures from scratch on Cifar-{10, 100} dataset. Notably, the model complexity gradually increases from ViT-Ti to ViT-B.

To ensure optimal model performance, we will leverage established training recipes recommended in the contemporary works (Dosovitskiy et al., 2021; Foret et al., 2021;
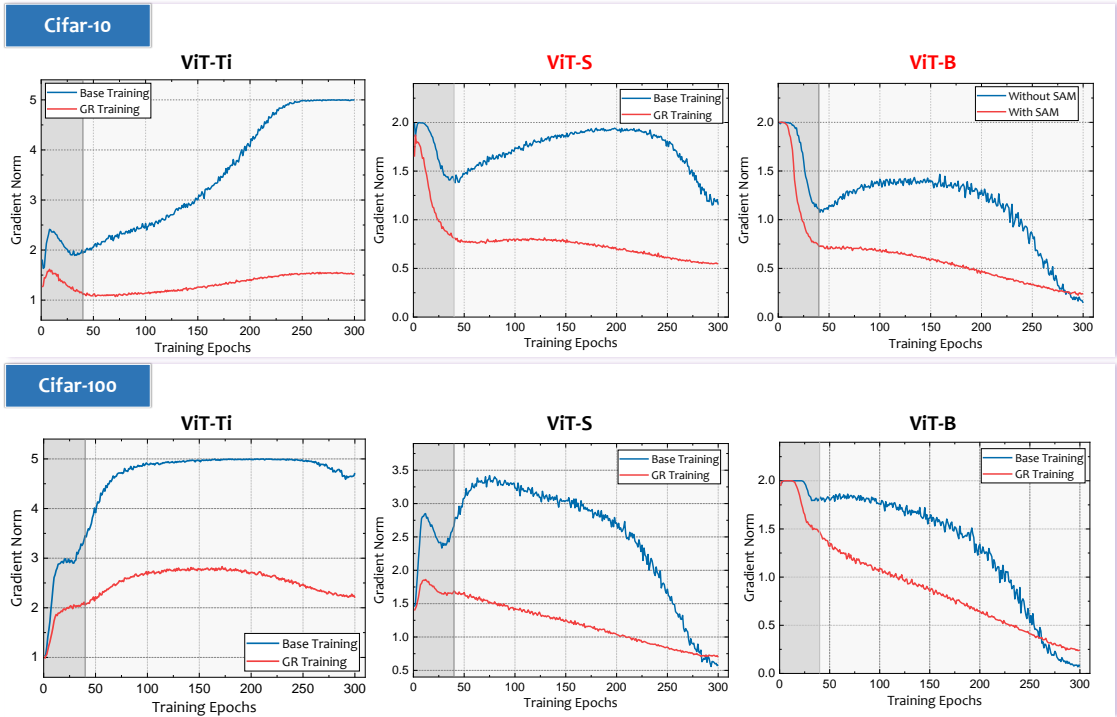
Figure 2. Evolution of gradient norm during training for ViT-Ti, ViT-S and ViT-B models on Cifar-{10, 100}, comparing training with GR (red line) and without GR (blue line).

Zhao et al., 2022b; Karakida et al., 2023) and additionally conduct searches on essential hyperparameters. Details on the training process can be found in the Appendix. Table 1 shows the final training results, reporting both the mean and standard deviation of testing error rates under five different run seeds.

3.2. Performance Degradation When Using GR and Adaptive Optimization Simultaneously

LR warmup can improve training performance. From the table, we can clearly see that for all the training comparison pairs in vertical (training instances with and without the implementation of a LR warmup technique), training with LR warmup will universally give better model performance for both vanilla training and gradient regularization training. Concretely, for both Adam and RMSProp optimizers, leveraging a LR warmup technique generally reduce the testing error rate on Cifar10 and Cifar100 datasets. Notably, in the absence of the LR warmup technique, the performance of ViT-B models, which possess the highest model capacity among the three architectures, degenerates to a catastrophic level, significantly inferior to even the simplest ViT-Ti models. However, with LR warmup, ViT-B models can outperform the other two architectures. The observed enhancement in model performance progressively intensifies from ViT-Ti to ViT-B architectures, suggesting that more

complicated models are considerably dependent on the LR warmup technique. Therefore, the LR warmup technique emerges as a indispensable component for adaptive learning algorithms.

Additionally, we can also find that for these ViT models, training with Adam optimizers noticeably excels over RMSProp optimizers, particularly when LR warmup techniques are adopted for more complex models. For instance, when applied to ViT-B models on Cifar10 in vanilla training, the testing error rate when utilizing the Adam optimizer can be approximately 5% lower compared to the usage of the RMSProp optimizer. Meanwhile, from the improvements, Adam optimizers derive greater benefits from the LR warmup techniques as compared to RMSProp optimizers, especially in training without gradient regularization.

Gradient regularization can have side-effect with LR warmup techniques. The horizontal training paris in the table (with and without the implementation of GR) elucidates a nuanced relationship between GR and LR warmup techniques. When training without LR warmup, GR consistently enhances model performance across all instances, just as its effectiveness observed in CNN training. However, when GR is combined with LR warmup, GR can be less helpful, as some performances can be far from behind those achieved via vanilla training with LR warmup techniques.

When GR training with LR warmup, for the RMSProp optimizer, GR can marginally improve model performance for most instances. Nevertheless, RMSProp fails to match the competitive performance demonstrated by the Adam optimizer. For the latter, a noticeable degradation in model performance can be observed for certain training instances. In these cases, models trained with both GR and LR warmup yield higher testing error rates compared to their counterparts trained with LR warmup but without GR. Notably, this degeneration becomes more significant as model complexity increases. This suggests that GR and warmup can be incompatible when employed simultaneously, especially for large models. Coupled with our previous observation that larger models rely more on the LR warmup technique, we can reason that larger architectural models like ViT-B are more sensitive to gradient statistics at the inception of training phase.

Figure 3. Plot of the Equation 9

4. Understanding the Incompatibility Between GR and Adaptive Optimization

In this section, we delve deeper into the conditions under which this compatibility issue between GR and warmup techniques occur, and perform a theoretical analysis to understand their underlying causes.

Compatibility issues arise with an initial decline in gradient norm. Recall that GR essentially regularizes the norm of loss gradient with a degree λ. It implies that once training starts, GR will bias the gradient towards regions where the norm is smaller. Such biased estimation may potentially undermine the stable and precise accumulation of gradient statistics at the initial stage of training. Consequently, this motivates us to check the evolution of the gradient norm during training. Figure 2 shows this evolution across epochs when employing the Adam optimizer for training with a LR warmup duration of 40 epochs, where the gradient norm is averaged over each individual step within a single epoch.

From the figure, we can clearly see that the gradient norm while training with GR is significantly smaller than that observed without GR across all training instances. Combined with previous incompatible training instances in Table 1, we can find that these instances all originally exhibit a decrease in gradient norm at the initial stages of training and the use of GR further exacerbates this decline. In other words, if the original gradient norm experiences a decline, then imposing gradient regularization can trigger compatibility issues.

GR can lead to large adaptive leraning rate variance in theory Next, we will give theoretical analysis regarding this compatibility from the perspective of adaptive LR variance.

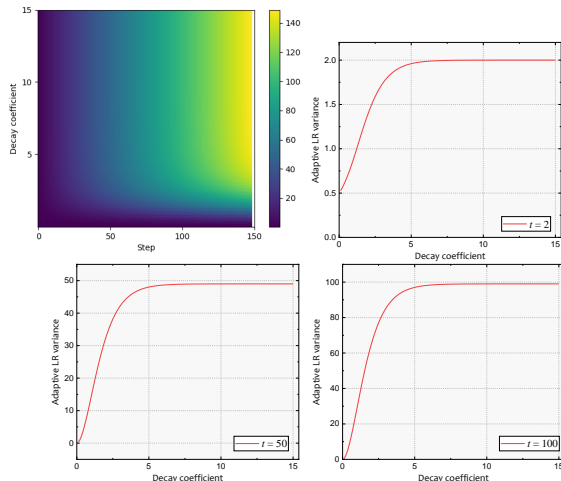Consider gradients $g_1, \cdots, g_t$, where each element g_k is independent and conforms $g_k \sim \mathcal{N}(0, \sigma_k)$. We target to investigate the variance of the adaptive LR $\psi(\cdot)$, in the case where gradient norm decreases as t increases. It is important to note that this scenario differs from the one proposed in (Liu et al., 2020), where $g_k \sim \mathcal{N}(0, \sigma)$ consistently.

A decrease in gradient norm generally indicates a contraction in the overall magnitude of the gradients. Given that these gradients are from a Gaussian distribution $g_k \sim \mathcal{N}(0, \sigma_i)$, a decreased gradient norm implies a corresponding drop in variance since variance quantifies the extent of deviation around the mean. Thereby, we assume the variance σ_k decays exponentially,

$$\sigma_k = e^{-\gamma k} \sigma_1 \qquad (7)$$

Now, the distribution of $\psi(\cdot)$ follows a nonstandard and intricate distribution. Perform a straightforward analysis can be challenging. Similar to that in (Liu et al., 2020), we adopt the approximation proposed by (Nau, 2014), which is $p(\psi(\cdot; \gamma)) \approx p(\sqrt{\frac{t}{\sum_{k=1}^{t} g_k^2}})$.

Lemma 4.1. *The variance of the adaptive learning rate can be evaluated via the Taylor series method (Benaroya et al., 2005),*

$$\mathrm{Var}(\psi(\cdot; \gamma)) \approx \frac{\mathrm{Var}(Y)}{4\mathbb{E}(Y)^3} \qquad (8)$$

where $Y = \frac{1}{t} \sum_{k=1}^{t} g_k^2$

By leveraging the Taylor approximation, Lemma 4.1 provides a valid way for calculating the adaptive LR variance. The proof can be found in the Appendix.

Theorem 4.2. *Given g_k are independent, and each follows $g_k \sim \mathcal{N}(0, \sigma_k)$, where $\sigma_k = e^{-\gamma k} \sigma_1$, $\gamma > 0$. $\mathrm{Var}(\psi(\cdot; \gamma))$ is given by,*

$$\mathrm{Var}(\psi(\cdot; \gamma)) = \frac{t}{2\sigma_1} \frac{(1 + e^{-\gamma t})(1 - e^{-\gamma})^2}{(1 + e^{-\gamma})(1 - e^{-\gamma t})^2} \qquad (9)$$

Here, we provide a plot for the Equation 9, as shown in Figure 3. In the figure, x-axis denotes the step t while y-axis denotes the decay coefficient γ. We can clearly see that the adaptive LR variance $\mathrm{Var}(\psi(\cdot;\gamma))$ increases along with the decay coefficient γ.

Theorem 4.3. *For $te^{-\gamma(t-1)} \leq 1$, $\mathrm{Var}(\psi(\cdot;\gamma))$ monotonically increases as γ increases.*

Also, the proof can be found in the Appendix. And actually, $te^{-\gamma(t-1)} \leq 1$ would hold almost, as t keeps increasing.

Theorem 4.3 theoretically confirms that $\mathrm{Var}(\psi(\cdot;\gamma))$ increases as γ increase. In conclusion, this indicates that GR will lead to a larger adaptive LR variance as compared to vanilla training in scenarios involving a decline in gradient norm. This is the fundamental reason for the observed incompatibility between GR and adaptive optimization techniques.

5. GR Warmup Strategies

As indicated by previous demonstrations, the current GR will impose excessive regularization on gradient norm during warmup course. This results in a substantial increase in adaptive LR variance, especially when the gradient norm originally experiences a drop. To mitigate this issue, it is motivated to reduce the regularization degree on gradient norm at the beginning of training, analogous to the LR warmup techniques.

Here, we will introduce three GR warmup strategies: r-warmup, λ-warmup and zero-warmup. Algorithm 1 shows the general framework of training with our warmup strategies. Table 2 shows the summary of the model performance and Figure 4 shows the corresponding evolution of gradient norm when applying the three warmup strategies.

All the training details can be found in the Appendix. Additionally, we have included further results on various model architectures and datasets in the Appendix.

5.1. r-warmup Gradient Regularization

The r-warmup GR is designed with due consideration to not only GR but also SAM. It is important to recall that the regularization degree $r^{(sam)}$ in SAM corresponds to the parameter setting $r = \lambda$ in GR. As such, our first strategy involves concurrently ramping up parameters r and λ in GR (which equates to $r^{(sam)}$ in SAM), following the same pattern as that of the learning rate warmup schedule.

$$r_t = \min\{\frac{t}{T_w}, 1\} \cdot r_0 \qquad (10)$$

Algorithm 1 Gradient Regularization Warmup Strategies

Input: Training set $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=0}^{N}$; loss function $L(\cdot)$; batch size B; learning rate η_0; total iterations T; warmup iterations T_w; GR coefficient λ_0 and r_0; GR warmup strategies $\mathcal{P}$.
Parameter: Model weights $\boldsymbol{\theta}$.
Output: Optimized model weights $\hat{\boldsymbol{\theta}}$.
Algorithm:
  1: Parameter initialization $\boldsymbol{\theta}_0$.
  2: **for** step $t = 1$ **to** T **do**
  3:     Get batch data pairs $\mathcal{B} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=0}^{B}$ sampled from training set $\mathcal{S}$.
  4:     **if** $t < T_w$ **then**
  5:         $\eta(t) = \min\{\frac{t}{T_w}, 1\}\eta_0$
  6:         Perform $\mathcal{P} \in \{r, \lambda, \text{zero}\}$-warmup GR strategies
  7:     **else**
  8:         Training with η_0, λ_0 and r_0.
  9:     **end if**
 10:     Compute the gradient $\boldsymbol{g}^{(1)} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$.
 11:     **if** $r_t > 0$ **then**
 12:         $\boldsymbol{g}^{(2)} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$ at $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t + r_t \frac{\boldsymbol{g}^{(1)}}{||\boldsymbol{g}^{(1)}||}$.
 13:     **else**
 14:         $\boldsymbol{g}^{(2)} = 0$
 15:     **end if**
 16:     $\boldsymbol{g}_t = (1 - \frac{\lambda_t}{r_t})\boldsymbol{g}^{(1)} + \frac{\lambda_t}{r_t}\boldsymbol{g}^{(2)}$
 17:     Update weight $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \cdot \boldsymbol{g}_t$
 18: **end for**
 19: **return** Final optimization $\hat{\boldsymbol{\theta}}$.

---

where r_0 denotes the base value we set. Correspondingly, the gradient becomes,

$$g_t = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t + r_t \cdot \frac{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)}{||\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)||}) \qquad (11)$$

This approach results in a considerably milder regularization effect initially, which then linearly scales up to its standard strength by the end of the warmup course.

As illustrated in Figure 4, the gradient norm is regularized at a slower rate during the initial stages of training, situating the gradient norm between those trained with and without GR. As we can see from Table 2, when using the r-warmup GR during the warmup course, all training cases exhibit improved model performance compared to standard GR implementation. The previously observed degradation effect on model performance can no longer be seen. Particularly for ViT-B models, performance significantly surpasses that of models trained with GR, and slightly outperform those trained without GR. However, although r-warmup GR leads to some improvement in model performance, this enhanced performance remains relatively close to that before degradation, i.e., training without GR.
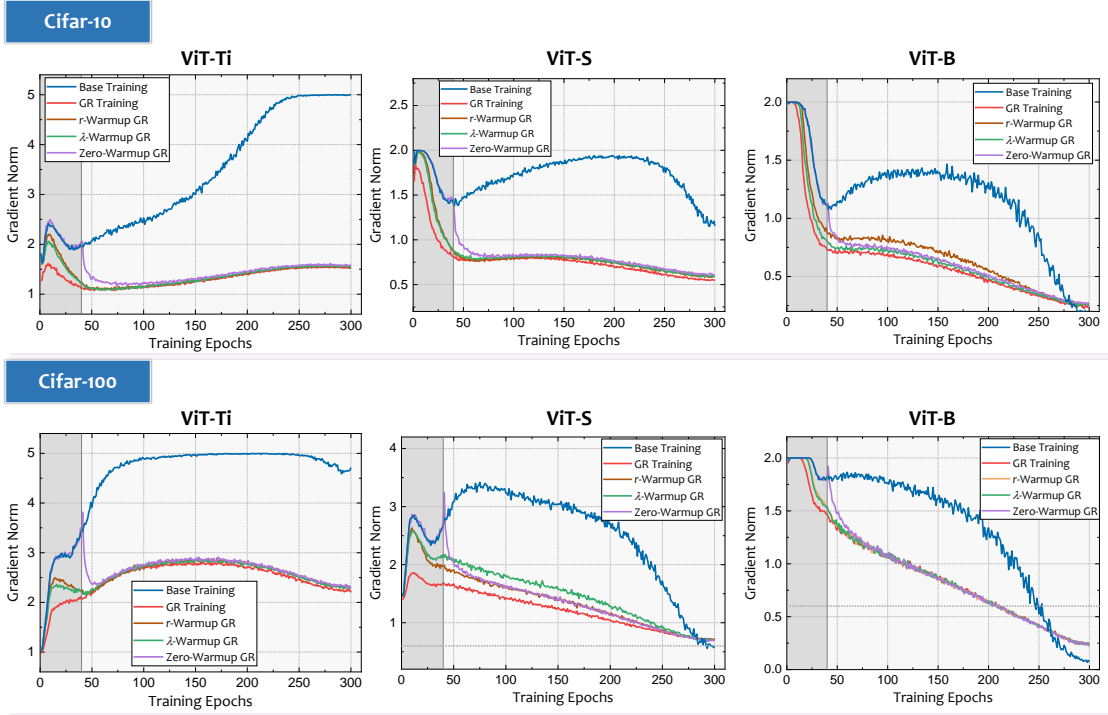
6

Figure 4. Evolution of gradient norm during training for ViT-Ti, ViT-S and ViT-B models on Cifar-{10, 100}, comparing training with r-warmup GR strategy (dark brown line), λ-Warmup GR strategy (green line) and zero-warmup GR strategy (purple line).

5.2. λ-warmup Gradient Regularization

We note that according to (Zhao et al., 2022b), a small r in GR can cause loss of precision since the perturbed gradient term $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t + r \cdot \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)/||\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)||)$ will become exceedingly close to the reference gradient $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$. Therefore, given that r in SAM corresponds to $\lambda = r$ in GR, r-warmup GR can suffer from this issue.

To this end, in our second strategy, we will choose to keep the parameter r fixed, and gradually ramp up the regularization degree λ in Equation 2 instead, which denotes that

$$\lambda_t = \min\{\frac{t}{T_w}, 1\} \cdot \lambda_0 \qquad (12)$$

where λ_0 denotes the base regularization degree. Now, the gradient corresponds to,

$$g_t = (1 - \frac{\lambda_t}{r_0})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) + \frac{\lambda_t}{r_0}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t + r_0 \frac{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)}{||\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)||}) \qquad (13)$$

Note that $\lambda_t < r_0$ for $t < T_w$, suggesting the term $\lambda_t/r_0 < 1$ during the warmup course. Contrary to the first r-warmup GR, this approach mitigates the regularization of gradient norm by interpolating the vanilla base (in this case, Adam) gradient and progressively diminishing its ratio in this interpolation. In other words, for λ-warmup GR, the corresponding gradient starts from the vanilla base gradient at at the beginning of the warmup course and eventually

evolve to the normal GR effect at the end of the warmup course.

From the Table 2 and Figure 4, we can find that although the gradients between adopting r-warmup GR and λ-warmup GR are quite close, λ-warmup GR can provide relatively better results. This suggests that performing warmup on λ in GR, rather than r in SAM, can effectively mitigate the precision loss problems caused by a small r value. Furthermore, we can find in the figure that while the regularization effects of both the r-warmup and λ-warmup GR are somewhat weaker than the normal GR implementation, they ultimately converge to an equivalent degree of regularization. Consequently, these observations motivate the development of our third strategy.

5.3. Zero-warmup Gradient Regularization

Our third strategy will further reduce the gradient norm regularization to a zero effect. In other words, we will use the vanilla base learning algorithm during warmup course, which will not impose any regularization on gradient norm. Subsequently, we switch to the normal GR for the remaining training iterations,

$$\lambda_t = \begin{cases} 0 & t \leq T_w \\ \lambda_0 & t > T_w \end{cases} \qquad (14)$$

Table 2. Test error rate of ViT models using Adam and RMSProp optimizers on the Cifar-{10, 100} datasets when trained with our three proposed GR warmup strategies

MODEL ARCHITECTURE	BASE OPTIMIZER	GR WARMUP POLICY	CIFAR-10 ERROR[%]	CIFAR-100 ERROR[%]
VIT-TI	ADAM	GR (BASELINE)	$13.92_{\pm 0.19}$	$39.38_{\pm 0.50}$
		r-WARMUP	$13.80_{\pm 0.17}$	$39.24_{\pm 0.37}$
		λ-WARMUP	$13.86_{\pm 0.35}$	$\mathbf{39.12_{\pm 0.37}}$
		ZERO-WARMUP	$\mathbf{13.61_{\pm 0.20}}$	$38.53_{\pm 0.38}$
	RMSPROP	GR (BASELINE)	$15.79_{\pm 0.25}$	$40.02_{\pm 0.22}$
		r-WARMUP	$15.32_{\pm 0.21}$	$39.80_{\pm 0.22}$
		λ-WARMUP	$15.47_{\pm 0.27}$	$39.55_{\pm 0.19}$
		ZERO-WARMUP	$\mathbf{15.06_{\pm 0.40}}$	$\mathbf{39.25_{\pm 0.21}}$
VIT-S	ADAM	GR (BASELINE)	$12.40_{\pm 0.17}$	$36.65_{\pm 0.23}$
		r-WARMUP	$11.72_{\pm 0.38}$	$36.53_{\pm 0.18}$
		λ-WARMUP	$11.47_{\pm 0.04}$	$36.48_{\pm 0.23}$
		ZERO-WARMUP	$\mathbf{10.68_{\pm 0.49}}$	$\mathbf{35.41_{\pm 0.25}}$
	RMSPROP	GR (BASELINE)	$14.77_{\pm 0.35}$	$37.90_{\pm 0.19}$
		r-WARMUP	$14.73_{\pm 0.18}$	$37.65_{\pm 0.30}$
		λ-WARMUP	$14.65_{\pm 0.39}$	$37.74_{\pm 0.27}$
		ZERO-WARMUP	$\mathbf{14.02_{\pm 0.29}}$	$\mathbf{37.37_{\pm 0.31}}$
VIT-B	ADAM	GR (BASELINE)	$12.36_{\pm 0.33}$	$36.65_{\pm 0.26}$
		r-WARMUP	$10.22_{\pm 0.26}$	$36.18_{\pm 0.37}$
		λ-WARMUP	$9.98_{\pm 0.23}$	$36.03_{\pm 0.30}$
		ZERO-WARMUP	$\mathbf{9.42_{\pm 0.15}}$	$\mathbf{35.75_{\pm 0.20}}$
	RMSPROP	GR (BASELINE)	$14.97_{\pm 0.25}$	$39.04_{\pm 0.35}$
		r-WARMUP	$14.42_{\pm 0.11}$	$38.36_{\pm 0.22}$
		λ-WARMUP	$14.55_{\pm 0.39}$	$38.50_{\pm 0.15}$
		ZERO-WARMUP	$\mathbf{14.34_{\pm 0.24}}$	$\mathbf{38.24_{\pm 0.21}}$

The corresponding gradient can be expressed as,

$$g_t = \begin{cases} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) & t \leq T_w \\ g_t^{(gr)} & t > T_w \end{cases} \quad (15)$$

In this strategy, warmup technique and GR are implemented separately across two distinct training periods. During the first period, we conduct warmup without imposing any regularization, resulting in a gradient norm identical to that obtained when training with the vanilla learning algorithm. This allows the optimizer to accumulate sufficient accurate gradients in statistics by the completion of the warmup phase. After that, the training process enters into the second period, which starts to regularize the gradient norm with normal GR method. Given the accumulation of gradients during the initial period, biasing the gradient towards convergence to regions with smaller gradient norm becomes considerably safer. It is worth noting that since zero-warmup employs base learning instead of sharpness-aware learning, it can further reduce computation overhead to some extent, dependent upon the number of warmup epochs.

As illustrated in the Figure 4, as expected, the evolution of gradient norms during the warmup phase when trained with the zero-warmup strategy matches precisely with that in vanilla training. But after that, as the zero-warmup GR switch to normal GR implementation, the gradient norm can be further reduced. For all three strategies, the corresponding gradient norms gradually become quite close, and eventually terminate at nearly the same values. This is because, after the warmup course, training is set to regularize the gradient norm to the same degree in all three warmup GR strategies, regardless of the level of gradient norm regularization applied during the warmup course.

We can see from the Table 2 that the zero-warmup GR yields superior performance compared to the other two strategies, significantly surpassing the benchmark method which involves training with normal GR method. For ViT-B models, this enhancement approximates nearly a 3% percent notch up on the Cifar10 dataset. These findings highlight that it is best to train with the base optimizer during the warmup course such that the variance of adaptive learning rate can be limited. Once the training stabilizes, we can safely bias the training with GR to achieve better model performance.

6. Conclusion

We show that GR and adaptive optimization can be incompatible when applied simultaneously, resulting in performance degeneration. Our empirical and theoretical evidences suggest that this incompatibility is caused by the excessive GR at the initial stage of training. To solve this issue, we propose three GR warmup strategies: r-warmup, λ-warmup, zero-warmup. We confirm the three strategies can successfully avoid this issue and give better performance.

Impact Statement

We believe the impacts of our work is positive, can be contributive to training modern complex neural networks. We claim our paper presents no potentially negative implications concerning ethical considerations or future societal consequences.

References

Ba, L. J., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXivPreprint*, abs/1607.06450, 2016.

Barrett, D. G. and Dherin, B. Implicit gradient regularization. *arXiv preprint arXiv:2009.11162*, 2020.

Benaroya, H., Han, S. M., and Nagurka, M. *Probability models in engineering and science*, volume 192. CRC press, 2005.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70, pp. 1019–1028, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. F. Efficient sharpness-aware minimization for improved training of neural networks. *arXivPreprint*, abs/2110.03141, 2021.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

Hinton, G., Srivastava, N., and Swersky, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Comput.*, 9(1):1–42, 1997.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37, pp. 448–456, 2015.

Karakida, R., Takase, T., Hayase, T., and Osawa, K. Understanding gradient regularization in deep learning: Efficient finite-difference computation and implicit bias. In *International Conference on Machine Learning*, pp. 15809–15827. PMLR, 2023.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kwon, J., Kim, J., Park, H., and Choi, I. K. ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914, 2021.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021,*

Montreal, QC, Canada, October 10-17, 2021, pp. 9992–10002. IEEE, 2021.

Nau, R. Forecasting with moving averages. *Fuqua School of Business, Duke University*, pp. 1–3, 2014.

Popel, M. and Bojar, O. Training tips for the transformer model. *Prague Bull. Math. Linguistics*, 110:43–70, 2018.

Reizinger, P. and Huszár, F. Samba: Regularized autoencoders perform sharpness-aware minimization. In *Fifth Symposium on Advances in Approximate Bayesian Inference*, 2023.

Smith, S. L., Dherin, B., Barrett, D. G., and De, S. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H. Going deeper with image transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 32–42. IEEE, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017a.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b.

Zhao, Y., Zhang, H., and Hu, X. Randomized sharpness-aware training for boosting computational efficiency in deep learning. *arXivPreprint*, abs/2203.09962, 2022a. doi: 10.48550/arXiv.2203.09962.

Zhao, Y., Zhang, H., and Hu, X. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 2022b.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N. C., Tatikonda, S., Duncan, J. S., and Liu, T. Surrogate gap minimization improves sharpness-aware training. In *The Tenth International Conference on Learning Representations*, 2022.

A. Proof of Theorems

Lemma A.1. *The variance of the adaptive learning rate can be evaluated via the Taylor series method ([Benaroya et al., 2005]),*

$$\text{Var}(\psi(\cdot;\gamma)) \approx \frac{\text{Var}(Y)}{4\mathbb{E}(Y)^3} \tag{16}$$

where $Y = \frac{1}{t}\sum_{k=1}^{t} g_k^2$

Proof:

We have $\text{Var}(\psi(\cdot;\gamma)) = \sqrt{\frac{1}{Y}}$, where $Y = \frac{1}{t}\sum_{k=1}^{t} g_k^2$. Assume $g(X) = \sqrt{\frac{1}{X}}$, where X is a random variable with mean μ. Based on the Taylor series,

$$g(X) = g(\mu) + (X-\mu)g'(\mu) + \frac{(X-\mu)^2}{2}g''(\mu) + \cdots \tag{17}$$

So, the variance can be,

$$\begin{aligned}
\text{Var}[g(X)] &= \text{Var}\left[g(\mu) + (X-\mu)g'(\mu) + \frac{(X-\mu)^2}{2}g''(\mu) + \cdots\right] \\
&= \text{Var}\left[(X-\mu)g' + \frac{(X-\mu)^2}{2}g''(\mu) + \cdots\right]
\end{aligned} \tag{18}$$

Often the first item is taken,

$$\text{Var}[g(X)] \approx g'(\mu)^2\text{Var}(X) \tag{19}$$

So, considering that $\text{Var}(\psi(\cdot;\gamma)) = g(Y) = \sqrt{\frac{1}{Y}}$, we have,

$$\begin{aligned}
\text{Var}[g(Y)] &\approx g'(\mu)^2\text{Var}(Y) \\
&= \left(-\frac{1}{2\mathbb{E}(Y)^{\frac{3}{2}}}\right)^2 \text{Var}(Y) \\
&= \frac{\text{Var}(Y)}{4\mathbb{E}(Y)^3}
\end{aligned} \tag{20}$$

Theorem A.2. *Given g_k are independent, and each follows $g_k \sim \mathcal{N}(0, \sigma_k)$, where $\sigma_k = e^{-\gamma k}\sigma_1$, $\gamma > 0$. $\text{Var}(\psi(\cdot;\gamma))$ is given by,*

$$\text{Var}(\psi(\cdot;\gamma)) = \frac{t}{2\sigma_1} \frac{(1+e^{-\gamma t})(1-e^{-\gamma})^2}{(1+e^{-\gamma})(1-e^{-\gamma t})^2} \tag{21}$$

Proof:

Assume $\sigma_k = e^{-\gamma k}\sigma_1$, then $\mathbb{E}(g_k^2) = \sigma_k$, $\text{Var}(g_k^2) = 2\sigma_k^2$.

So,

$$\begin{aligned}
\mathbb{E}(tY) &= \mathbb{E}(\sum_{k=1}^{t} g_k^2) \\
&= \sum_{k=1}^{t}\mathbb{E}(g_k^2) = \sum_{k=1}^{t}\sigma_k = \sum_{k=1}^{t}e^{-\gamma}\sigma_1 \\
&= \sigma_1\frac{1-e^{-\gamma t}}{1-e^{-\gamma}}
\end{aligned} \tag{22}$$

$$\text{Var}(tY) = \text{Var}(\sum_{k=1}^{t} g_k^2)$$

$$= \sum_{k=1}^{t} \text{Var}(g_k^2) = \sum_{k=1}^{t} 2\sigma_k^2 = \sum_{k=1}^{t} 2(e^{-\gamma}\sigma_1)^2 \tag{23}$$

$$= 2\sigma_1^2 \frac{1 - (e^{-\gamma})^{2t}}{1 - (e^{-\gamma})^2}$$

$$= 2\sigma_1^2 \frac{1 - e^{-2\gamma t}}{1 - e^{-2\gamma}}$$

This lead to,

$$\text{Var}(\psi(\cdot;\gamma)) \approx \frac{\text{Var}(Y)}{4\mathbb{E}(Y)^3}$$

$$= \frac{2\frac{1}{t^2}\sigma_1^2 \frac{1-e^{-2\gamma t}}{1-e^{-2\gamma}}}{4\left(\frac{1}{t}\sigma_1 \frac{1-e^{-\gamma t}}{1-e^{-\gamma}}\right)^3} \tag{24}$$

$$= \frac{t}{2\sigma_1} \frac{(1 + e^{-\gamma t})(1 - e^{-\gamma})^2}{(1 + e^{-\gamma})(1 - e^{-\gamma t})^2}$$

Theorem A.3. *For $te^{-\gamma(t-1)} \leq 1$, $\text{Var}(\psi(\cdot;\gamma))$ monotonically increases as γ increases.*

Proof:

Assume $x = e^{-\gamma}$. So,

$$\text{Var}(\psi(\cdot;\gamma)) \approx (h'(\mathbb{E}(Y)))^2 \text{Var}(Y)$$

$$= \frac{t}{2\sigma_0} \frac{(1 + x^t)(1 - x)^2}{(1 + x)(1 - x^t)^2} \tag{25}$$

Set

$$\beta(x) = \log \text{Var}(\psi(\cdot;\gamma)) = \log t + \log(1 + x^t) + 2\log(1 - x) - \log(1 + x) - 2\log(1 - x^t) \tag{26}$$

$$\frac{d\beta(x)}{dx} = \frac{tx^{t-1}}{x^t + 1} + \frac{2tx^{t-1}}{1 - x^t} - \frac{1}{1 + x} - \frac{2}{1 - x}$$

$$= tx^{t-1}\left(\frac{1}{1 + x^t} + \frac{2}{1 - x^t}\right) - \left(\frac{1}{1 + x} + \frac{2}{1 - x}\right) \tag{27}$$

Set

$$\zeta(a) = \frac{1}{1 + a} + \frac{2}{1 - a}$$

$$\frac{d\zeta(a)}{da} = \frac{a^2 + 6a + 1}{(1 - a^2)^2} > 0 \tag{28}$$

Given $a = x^t$ decrease with t. $\zeta(t)$ decreases. Namely, $\zeta(t) < \zeta(1), t > 1$

$$\frac{1}{1 + x^t} + \frac{2}{1 - x^t} < \frac{1}{1 + x} + \frac{2}{1 - x} \tag{29}$$

Assume $tx^{t-1} \leq 1$

$$tx^{t-1}\left(\frac{1}{1+x^t} + \frac{2}{1-x^t}\right) < \frac{1}{1+x} + \frac{2}{1-x} \tag{30}$$

That is $\frac{d\beta(x)}{dx} < 0$, $\text{Var}(\psi(\cdot;\gamma))$ increase as x decrease. Considering $x = e^{-\gamma}$, we can conclude that $\text{Var}(\psi(\cdot;\gamma))$ increase as γ increase.

B. Implementation Details

Our training task will utilize ViT models to address image classification problems on the CIFAR dataset. For data argumentation, the data will be subjected to an augmentation process following a standard strategy that has been widely adopted in contemporary works (Foret et al., 2021). This approach involves sequentially processing images through a series of operations, starting with four-pixel padding, followed by random flipping and concluding with random cropping.

To get good model performance, we will partially consult the modern ViT training recipes (Dosovitskiy et al., 2021), and meanwhile perform a slight more searching on core hyperparameters.

Table 3. The basic hyperparameters for training ViTs.

	Adam/RMSProp Training	GR Training
Epoch	300	300
LR/GR Warmup epoch	{20, 40, 80}	{20, 40, 80}
Batch size	256	256
Gradient Clipping	True	True
Basic learning rate	{5e-3, 1e-3, 5e-4}	{5e-3, 1e-3, 5e-4}
Learning rate schedule	cosine	cosine
Weight decay	0.03	0.03
r in GR	-	{0.05, 0.1}
$\frac{\lambda}{r}$ in GR	-	{0.8, 1}

C. Additional Results

In additional to the classic ViT architecture family, we have also evaluated our methods with Swin and CaiT architecture families. Table 4 shows the corresponds results on Cifar dataset. It should be mentioned that the common training hyper-parameters are the same as those of training ViT models.

As evidenced by the Table 4, our methods demonstrate improved performance across different Transformer architectures, which means that warmup GR can be essential for Transformer-based architecture. In fact, despite these models featuring distinct structural elements, their core mechanisms remain consistent with the conventional Transformer. It should be noted that among the three GR warmup strategy, zero-warmup strategy yield the best results for the most cases.

Meanwhile, we have also extended our evaluation to include TinyImageNet as well as ImageNet, as shown at Table 5. From the results, our methods shows the effectiveness, as evidenced by a consistent reduction in error rates across all models tested.

Table 4. Test error rate of Swin and CaiT models using Adam optimizers on the Cifar-{10, 100} datasets when trained with our three proposed GR warmup strategies

Model Architecture	Base Optimizer	GR Warmup Policy	Cifar-10 Error[%]	Cifar-100 Error[%]
Swin-Tiny	Adam	No GR	$12.39_{\pm 0.21}$	$42.36_{\pm 0.42}$
		Vanilla GR	$11.34_{\pm 0.14}$	$40.65_{\pm 0.33}$
		r-Warmup	$10.54_{\pm 0.19}$	$39.44_{\pm 0.29}$
		λ-Warmup	$10.70_{\pm 0.29}$	$\mathbf{39.13_{\pm 0.36}}$
		Zero-Warmup	$\mathbf{10.52_{\pm 0.16}}$	$39.51_{\pm 0.32}$
Swin-Small	Adam	No GR	$12.72_{\pm 0.17}$	$43.14_{\pm 0.44}$
		Vanilla GR	$11.85_{\pm 0.13}$	$42.20_{\pm 0.42}$
		r-Warmup	$11.68_{\pm 0.22}$	$40.82_{\pm 0.34}$
		λ-Warmup	$11.69_{\pm 0.19}$	$40.69_{\pm 0.40}$
		Zero-Warmup	$\mathbf{11.24_{\pm 0.17}}$	$\mathbf{40.12_{\pm 0.26}}$
Swin-Base	Adam	No GR	$12.26_{\pm 0.16}$	$41.63_{\pm 0.53}$
		Vanilla GR	$11.88_{\pm 0.20}$	$38.89_{\pm 0.44}$
		r-Warmup	$11.44_{\pm 0.17}$	$\mathbf{39.53_{\pm 0.37}}$
		λ-Warmup	$11.34_{\pm 0.27}$	$38.92_{\pm 0.30}$
		Zero-Warmup	$\mathbf{10.90_{\pm 0.15}}$	$38.69_{\pm 0.42}$
CaiT-XXS12	Adam	No GR	$13.25_{\pm 0.14}$	$41.74_{\pm 0.46}$
		Vanilla GR	$13.62_{\pm 0.23}$	$39.14_{\pm 0.17}$
		r-Warmup	$11.31_{\pm 0.24}$	$38.74_{\pm 0.33}$
		λ-Warmup	$12.06_{\pm 0.31}$	$\mathbf{39.31_{\pm 0.43}}$
		Zero-Warmup	$\mathbf{11.07_{\pm 0.18}}$	$38.51_{\pm 0.36}$
CaiT-XXS24	Adam	No GR	$11.89_{\pm 0.19}$	$39.42_{\pm 0.53}$
		Vanilla GR	$12.77_{\pm 0.15}$	$38.02_{\pm 0.36}$
		r-Warmup	$10.47_{\pm 0.13}$	$\mathbf{37.04_{\pm 0.41}}$
		λ-Warmup	$11.77_{\pm 0.24}$	$37.52_{\pm 0.29}$
		Zero-Warmup	$\mathbf{9.87_{\pm 0.20}}$	$37.17_{\pm 0.41}$
CaiT-XS12	Adam	No GR	$11.06_{\pm 0.16}$	$39.15_{\pm 0.47}$
		Vanilla GR	$12.52_{\pm 0.19}$	$37.65_{\pm 0.32}$
		r-Warmup	$\mathbf{9.48_{\pm 0.24}}$	$37.83_{\pm 0.36}$
		λ-Warmup	$10.23_{\pm 0.21}$	$37.78_{\pm 0.33}$
		Zero-Warmup	$9.59_{\pm 0.25}$	$\mathbf{37.46_{\pm 0.49}}$
CaiT-S12	Adam	No GR	$10.59_{\pm 0.18}$	$39.15_{\pm 0.45}$
		Vanilla GR	$12.34_{\pm 0.25}$	$36.41_{\pm 0.41}$
		r-Warmup	$9.03_{\pm 0.29}$	$35.70_{\pm 0.33}$
		λ-Warmup	$9.20_{\pm 0.31}$	$36.12_{\pm 0.43}$
		Zero-Warmup	$\mathbf{8.94_{\pm 0.14}}$	$\mathbf{35.16_{\pm 0.39}}$

Table 5. Test error rate of ViT models using Adam optimizers on the TinyImageNet and ImageNet datasets when trained with our three proposed GR warmup strategies

Model Architecture	Base Optimizer	GR Warmup Policy	Testing Error[%]
ViT-Ti	TinyImageNet	No GR	$52.15_{\pm 0.16}$
		Vanilla GR	$51.08_{\pm 0.19}$
		r-Warmup	$50.41_{\pm 0.21}$
		λ-Warmup	$50.54_{\pm 0.14}$
		Zero-Warmup	$\mathbf{50.19_{\pm 0.20}}$
ViT-S	TinyImageNet	No GR	$52.28_{\pm 0.23}$
		Vanilla GR	$50.64_{\pm 0.18}$
		r-Warmup	$50.07_{\pm 0.24}$
		λ-Warmup	$50.25_{\pm 0.26}$
		Zero-Warmup	$\mathbf{49.91_{\pm 0.19}}$
ViT-B	TinyImageNet	No GR	$62.23_{\pm 0.21}$
		Vanilla GR	$49.96_{\pm 0.23}$
		r-Warmup	$49.30_{\pm 0.14}$
		λ-Warmup	$48.73_{\pm 0.15}$
		Zero-Warmup	$\mathbf{48.75_{\pm 0.20}}$
Swin-Tiny	ImageNet	No GR	$23.43_{\pm 0.24}$
		Vanilla GR	$22.34_{\pm 0.15}$
		r-Warmup	$21.81_{\pm 0.28}$
		λ-Warmup	$22.04_{\pm 0.22}$
		Zero-Warmup	$\mathbf{21.58_{\pm 0.34}}$