Towards Efficient and Scalable Sharpness-Aware Minimization

Supplementary Material

A. Appendix

A.1. LayerSAM & LookLayerSAM

Algorithm 2 Layer-wise SAM (LayerSAM)

Input: $x \in \mathbb{R}^d$, learning rate η_t, update frequency k.
for $t \leftarrow 1$ **to** T **do**
  Sample Minibatch $\mathcal{B} = \{(x_i, y_i), \cdots, (x_{|\mathcal{B}|}, y_{|\mathcal{B}|})\}$ from X.
  Compute gradient $g = \nabla_{\boldsymbol{w}} L_B(\boldsymbol{w})$ on minibatch $\mathcal{B}$.
  Compute $\boldsymbol{\epsilon^{(i)}} = \rho \frac{||\boldsymbol{w}^{(i)}||}{||\boldsymbol{g}^{(i)}||} \cdot \nabla_{\boldsymbol{w}} \mathcal{L}_S(\boldsymbol{w}) / \|\nabla_{\boldsymbol{w}} \mathcal{L}_S(\boldsymbol{w})\|$
  Compute gradient approximation for the SAM objective: $\boldsymbol{g_s} = \nabla_{\boldsymbol{w}} L_{\mathcal{B}}(\boldsymbol{w})|_{\boldsymbol{w}+\boldsymbol{\epsilon}}$
  Update weights: $\boldsymbol{w^{(i)}_{t+1}} = \boldsymbol{w^{(i)}_t} - \eta^{(i)}_t \cdot \boldsymbol{g^{(i)}_s}$
end for

Algorithm 3 Look-LayerSAM

Input: $x \in \mathbb{R}^d$, learning rate η_t, update frequency k.
for $t \leftarrow 1$ **to** T **do**
  Sample Minibatch $\mathcal{B} = \{(x_i, y_i), \cdots, (x_{|\mathcal{B}|}, y_{|\mathcal{B}|})\}$ from X.
  Compute gradient $\boldsymbol{g} = \nabla_{\boldsymbol{w}} L_B(\boldsymbol{w})$ on minibatch $\mathcal{B}$.
  if t%k = 0 **then**
    $\epsilon(\boldsymbol{w})^{(i)} = \rho \frac{\|\boldsymbol{w}^{(i)}\|}{\|\boldsymbol{g}^{(i)}\|} \cdot \nabla_{\boldsymbol{w}} \mathcal{L}_S(\boldsymbol{w}) / \|\nabla_{\boldsymbol{w}} \mathcal{L}_S(\boldsymbol{w})\|$
    Compute SAM gradient: $\boldsymbol{g_s} = \nabla_{\boldsymbol{w}} L_B(\boldsymbol{w})|_{\boldsymbol{w}+\epsilon(\boldsymbol{w})}$
    $\boldsymbol{g_v} = \boldsymbol{g_s} - \|\boldsymbol{g_s}\| \cos(\theta) \cdot \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|}$, where $\cos(\theta) = \frac{\boldsymbol{g} \cdot \boldsymbol{g_s}}{\|\boldsymbol{g}\| \|\boldsymbol{g_s}\|}$
  else
    $\boldsymbol{g_s} = \boldsymbol{g} + \alpha \cdot \frac{\|\boldsymbol{g}\|}{\|\boldsymbol{g_v}\|} \cdot \boldsymbol{g_v}$
  end if
  Update weights: $\boldsymbol{w_{t+1}}^{(i)} = \boldsymbol{w_t}^{(i)} - \eta^{(i)}_t \cdot \boldsymbol{g_s}^{(i)}$
end for

A.2. Parameter Settings

In this section, we will introduce the architectures of ViTs in this paper (Table 10). Next, we provide the hyperparameters in Table 8 for ViT training, including learning rate, warmup, optimizer, gradient clipping, epoch, etc. In addition, Table 9 gives us the parameter settings of ViT for large-batch training in this paper.

A.3. Generalization bound

We firstly introduce Theorem 1 regarding generalization bound based on sharpness of LookSAM and then give a proof for it. Note that a similar bound was also established in the original SAM paper [13].

Theorem 1. *With probability 1 - δ over the choice the training set $\mathcal{S} \sim \mathcal{D}$, we have*

$$\mathcal{L}_D(\boldsymbol{w}) \leq \max_{||\boldsymbol{\epsilon'}||_p \leq \rho'} \mathcal{L}_S(\boldsymbol{w} + \boldsymbol{\epsilon'}) + \sqrt{\frac{k \log(1 + \frac{||\boldsymbol{w}||_2^2}{\rho'^2}(1 + \sqrt{\frac{\log(n)}{k}})^2) + 4 \log \frac{n}{\delta} + \tilde{O}(1)}{n-1}} \tag{8}$$

where $n = |\mathcal{S}|$ and $\rho'^2 = \rho^2 + \rho_0^2$.

Proof. We start by illustrating the PAC-Bayesian Generation Bound theorem, which gives a bound on the generalization error of any posterior distribution $\mathcal{Q}$ on parameters that can be achieved using a selected prior distribution $\mathcal{P}$ over parameters training with data set $\mathcal{S}$. Let $KL(\mathcal{Q}||\mathcal{P})$ denote the KL divergence between two Bernoulli distributions $\mathcal{P}$ and $\mathcal{Q}$, we have:

$$\mathbb{E}_{\boldsymbol{w} \sim \mathcal{L}}[L_{\mathcal{D}}(\boldsymbol{w})] \leq \mathbb{E}_{\boldsymbol{w} \sim \mathcal{L}}[L_{\mathcal{S}}(\boldsymbol{w})] + \sqrt{\frac{KL(\mathcal{Q}||\mathcal{P}) + \log \frac{n}{\delta}}{2(n-1)}} \tag{9}$$

In order to accelerate the training process, LookSAM calculate the SAM gradient only at every k step and try to reuse the projected components to imitate the weight perturbations introduced from SAM procedure in the subsequent steps. We use $\boldsymbol{\epsilon^0}$ to indicate the difference between our imitated weight perturbation, $\boldsymbol{\epsilon'}$, from LookSAM and the real weight perturbation, $\boldsymbol{\epsilon}$, from SAM. As the optimization is in fact regarding the distribution of $\boldsymbol{\epsilon'}$, we assume that $\mathcal{L}_{\mathcal{D}}(\boldsymbol{w}) \leq \mathbb{E}_{\epsilon'_i \sim \mathcal{N}(0,\rho')}[L_{\mathcal{D}}(\boldsymbol{w} + \boldsymbol{\epsilon'})]$, which indicates adding Gaussian perturbation should not decrease the test error[13]. Following [13], the generalization bound can be written as follows:

$$\mathbb{E}_{\epsilon'_i \sim \mathcal{N}(0,\sigma')}[L_{\mathcal{D}}(\boldsymbol{w} + \boldsymbol{\epsilon'})] \leq \mathbb{E}_{\epsilon'_i \sim \mathcal{N}(0,\sigma')}[L_{\mathcal{S}}(\boldsymbol{w} + \boldsymbol{\epsilon'})] + \sqrt{\frac{\frac{1}{4} k \log(1 + \frac{||\boldsymbol{w}||_2^2}{k\sigma'^2}) + \frac{1}{4} + log \frac{n}{\delta} + 2log(6n + 3k)}{n-1}}, \tag{10}$$

where $\epsilon'_i = \epsilon_i + \epsilon^0_i$

In Equation (10), we assume that ϵ_i and ϵ^0_i are independent normal variables with mean 0, and corresponding variance σ and σ_0 respectively. Let $\{\epsilon'_i\}$, where $\epsilon'_i = \epsilon_i + \epsilon^0_i$,

Table 8. Parameter Settings of ViT for Vanilla Training

Model	Input Resolution	Batch Size	Epoch	Warmup Steps	Peak LR	LR Decay	Optimizer	ρ	Weight Decay	Gradient Clipping
ViT-B-16	224	4096	300	10000	3e-3	cosine	AdamW	/	0.3	1.0
ViT-B-32	224	4096	300	10000	3e-3	cosine	AdamW	/	0.3	1.0
ViT-S-16	224	4096	300	10000	3e-3	cosine	AdamW	/	0.3	1.0
ViT-S-32	224	4096	300	10000	3e-3	cosine	AdamW	/	0.3	1.0
ViT-B-16 + SAM	224	4096	300	10000	3e-3	cosine	AdamW	0.18	0.3	1.0
ViT-B-32 + SAM	224	4096	300	10000	3e-3	cosine	AdamW	0.15	0.3	1.0
ViT-S-16 + SAM	224	4096	300	10000	3e-3	cosine	AdamW	0.1	0.3	1.0
ViT-S-32 + SAM	224	4096	300	10000	3e-3	cosine	AdamW	0.05	0.3	1.0
ViT-B-16 + LookSAM	224	4096	300	10000	3e-3	cosine	AdamW	0.18	0.3	1.0
ViT-B-32 + LookSAM	224	4096	300	10000	3e-3	cosine	AdamW	0.15	0.3	1.0
ViT-S-16 + LookSAM	224	4096	300	10000	3e-3	cosine	AdamW	0.1	0.3	1.0
ViT-S-32 + LookSAM	224	4096	300	10000	3e-3	cosine	AdamW	0.05	0.3	1.0

Table 9. Parameter Settings of ViT for Large-Batch Training

Model	Batch Size	Epoch	Warmup Steps	Peak LR	LR Decay	Optimizer	ρ	α	Weight Decay	Gradient Clipping
ViT-B-16 + SAM	4096	300	10000	1e-2	linear	LAMB	0.18	/	0.1	1.0
ViT-B-16 + SAM	8192	300	10000	1.7e-2	linear	LAMB	0.18	/	0.1	1.0
ViT-B-16 + SAM	16834	300	7000	1.8e-2	linear	LAMB	0.18	/	0.1	1.0
ViT-B-16 + SAM	32768	300	6000	1.8e-2	linear	LAMB	0.18	/	0.1	1.0
ViT-B-16 + LayerSAM	4096	300	10000	1e-2	linear	LAMB	1.0	/	0.1	1.0
ViT-B-16 + LayerSAM	8192	300	10000	1.7e-2	linear	LAMB	1.0	/	0.1	1.0
ViT-B-16 + LayerSAM	16384	300	7000	1.8e-2	linear	LAMB	1.0	/	0.1	1.0
ViT-B-16 + LayerSAM	32768	300	6000	1.8e-2	linear	LAMB	1.0	/	0.1	1.0
ViT-B-16 + LayerSAM	65536	300	3500	2e-2	linear	LAMB	1.0	/	0.2	1.0
ViT-B-16 + Look-LayerSAM	4096	300	10000	1e-2	linear	LAMB	1.0	0.7	0.1	1.0
ViT-B-16 + Look-LayerSAM	8192	300	10000	1.7e-2	linear	LAMB	1.0	0.7	0.1	1.0
ViT-B-16 + Look-LayerSAM	16384	300	7000	1.8e-2	linear	LAMB	1.0	0.7	0.1	1.0
ViT-B-16 + Look-LayerSAM	32768	300	6000	1.8e-2	linear	LAMB	1.0	0.7	0.1	1.0
ViT-B-16 + Look-LayerSAM	65536	300	3500	2e-2	linear	LAMB	1.0	0.7	0.2	1.0

be the independent normal variable with mean 0 and variance $\sigma'^2 = \sigma^2 + \sigma_0^2$. In particular, at the time when LookSAM can perfectly imitate the SAM procedure by reusing the projected gradient, σ_0^2 becomes zero and σ'^2 equals to σ^2. As $||\boldsymbol{\epsilon'}||_2^2$ has chi-square distribution in this case and based on concentration inequality from Lemma 1 in [27], we obtain the following for any positive x:

$$
\begin{aligned}
&P(||\boldsymbol{\epsilon} + \boldsymbol{\epsilon_0}||_2^2 - k(\sigma^2 + \sigma_0^2) \\
&\quad \geq 2(\sigma^2 + \sigma_0^2)\sqrt{kx} + 2x(\sigma^2 + \sigma_0^2)) \\
&\quad \leq exp(-x)
\end{aligned} \tag{11}
$$

Let $x = \ln \sqrt{n}$, then we have that

$$
\begin{aligned}
&P(||\boldsymbol{\epsilon} + \boldsymbol{\epsilon_0}||_2^2 \\
&\quad \geq (\sigma^2 + \sigma_0^2)(k + 2\sqrt{k \ln \sqrt{n}} + 2 \ln \sqrt{n})) \\
&\quad \leq \frac{1}{\sqrt{n}}
\end{aligned} \tag{12}
$$

With probability of $(1 - \frac{1}{\sqrt{n}})$, we have,

Table 10. Architectures of ViTs

Model	Params	Patch Resolution	Sequence Length	Hidden Size	Heads	Layers
ViT-B-16	87M	16×16	196	768	12	12
ViT-B-32	88M	32×32	49	768	12	12
ViT-S-16	22M	16×16	196	384	6	12
ViT-S-32	23M	32×32	49	384	6	12

$$\begin{aligned} ||\boldsymbol{\epsilon'}||_2^2 &= ||\boldsymbol{\epsilon} + \boldsymbol{\epsilon_0}||_2^2 \\ &\leq (\sigma^2 + \sigma_0^2)(k + 2\sqrt{k \ln \sqrt{n}} + 2 \ln \sqrt{n}) \\ &\leq (\sigma^2 + \sigma_0^2) k (1 + \sqrt{\frac{\ln n}{k}})^2 \\ &\leq \rho^2 + \rho_0^2, \end{aligned} \tag{13}$$

where $\rho_0^2 = \sigma_0^2 k (1 + \sqrt{\frac{\ln n}{k}})^2$.

After substituting the value for σ' back to Equation (10), we can generate the following bounds:

$$\begin{aligned} &\mathcal{L}_{\mathcal{D}}(\boldsymbol{w}) \leq (1 - \frac{1}{\sqrt{n}}) \max_{||\boldsymbol{\epsilon'}||_{p \leq \rho'}} \mathcal{L}_S(\boldsymbol{w} + \boldsymbol{\epsilon'}) + \frac{1}{\sqrt{n}} \\ &+ \sqrt{\frac{\frac{1}{4} k \log(1 + \frac{||\boldsymbol{w}||_2^2}{\rho^2}(1 + \sqrt{\frac{\log(n)}{k}}))^2 + \log \frac{n}{\delta} + 2 \log(6n + 3k))}{n - 1}} \\ &\leq \max_{||\boldsymbol{\epsilon'}||_{p \leq \rho'}} \mathcal{L}_S(\boldsymbol{w} + \boldsymbol{\epsilon'}) \\ &+ \sqrt{\frac{k \log(1 + \frac{||\boldsymbol{w}||_2^2}{\rho'^2}(1 + \sqrt{\frac{\log(n)}{k}})^2) + 4 \log \frac{n}{\delta} + 8 \log(6n + 3k))}{n - 1}} \end{aligned} \tag{14}$$

where $\rho'^2 = \rho^2 + \rho_0^2$.