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 $\eta_t$, update frequency $k$.

for $t \leftarrow 1$ to $T$
do
Sample minibatch $B = \{(x_i, y_i), \ldots, (x_B, y_B)\}$ from $X$.
Compute gradient $g = \nabla_w L_B(w)$ on minibatch $B$.
Compute $\epsilon^{(t)} = \rho \frac{||g^{(t)}||}{||g||} \cdot \nabla_w L_S(w)/||\nabla_w L_S(w)||$.
Compute gradient approximation for the SAM objective: $g_s = \nabla_w L_B(w)|_{w+\epsilon}$.
Update weights: $w_{t+1} = w_t - \eta_t \cdot g_s$
end for

Algorithm 3 Look-LayerSAM

Input: $x \in \mathbb{R}^d$, learning rate $\eta_t$, update frequency $k$.

for $t \leftarrow 1$ to $T$
do
Sample minibatch $B = \{(x_i, y_i), \ldots, (x_B, y_B)\}$ from $X$.
Compute gradient $g = \nabla_w L_B(w)$ on minibatch $B$.
if $t \% k = 0$ then
Compute $\epsilon^{(t)} = \rho \frac{||g^{(t)}||}{||g||} \cdot \nabla_w L_S(w)/||\nabla_w L_S(w)||$.
Compute SAM gradient: $g_s = \nabla_w L_B(w)|_{w+\epsilon}$.
else
$g_s = g + \alpha \cdot \frac{||g||}{||g||} \cdot g_v$.
end if
Update weights: $w_{t+1} = w_t - \eta_t \cdot g_s$
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 [15].

Theorem 1. With probability $1 - \delta$ over the choice the training set $S \sim D$, we have

$$
L_D(w) \leq \max_{||\epsilon||_p \leq \rho} \mathcal{L}_S(w + \epsilon') + \\
\frac{k \log(1 + ||w||^2 p^2 + \sqrt{\log(n)/k})^2 + 4 \log \frac{2}{\delta} + O(1)}{n - 1}
$$

wheren $n = |S|$ and $\rho^2 = \rho^2 + \rho_0^2$.

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

$$
\mathbb{E}_{w \sim L}[L_D(w)] \leq \mathbb{E}_{w \sim L}[L_S(w)] + \sqrt{KL(Q||P) + \log \frac{2}{\delta}}
$$

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 $\epsilon^0$ to indicate the difference between our imitated weight perturbation, $\epsilon'$, from LookSAM and the real weight perturbation, $\epsilon$, from SAM. As the optimization is in fact regarding the distribution of $\epsilon'$, we assume that $L_D(w) \leq \mathbb{E}_{\epsilon' \sim \mathcal{N}(0, \sigma')^T}[L_D(w + \epsilon')]$, which indicates adding Gaussian perturbation should not decrease the test error[13]. Following [13], the generalization bound can be written as follows:

$$
\mathbb{E}_{\epsilon' \sim \mathcal{N}(0, \sigma')^T}[L_D(w + \epsilon')] \leq \mathbb{E}_{\epsilon' \sim \mathcal{N}(0, \sigma')^T}[L_S(w + \epsilon')] + \\
\sqrt{\frac{1}{k} k \log(1 + \frac{||w||^2}{\kappa^2}) + \frac{1}{n} + \log \frac{2}{\delta} + 2 \log(6n + 3k)},
$$

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

In Equation (10), we assume that $\epsilon_i$ and $\epsilon_i^0$ are independent normal variables with mean 0, and corresponding variance $\sigma$ and $\sigma_0$ respectively. Let $\{\epsilon'_i\}$, where $\epsilon'_i = \epsilon_i + \epsilon^0_i$,
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, $\sigma_0^2$ becomes zero and $\sigma^2$ equals to $\sigma^2$. As $||e'||^2$ has chi-square distribution in this case and based on concentration inequality from Lemma 1 in [27], we obtain the following for any positive $x$:

$$P(||\epsilon + \epsilon_0||^2) \leq \sqrt{k n}$$

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

$$P(||\epsilon + \epsilon_0||^2) \leq \sqrt{k n}$$

With probability of $(1 - \frac{1}{\sqrt{n}})$, we have,
Table 10. Architectures of ViTs

<table>
<thead>
<tr>
<th>Model</th>
<th>Params</th>
<th>Patch Resolution</th>
<th>Sequence Length</th>
<th>Hidden Size</th>
<th>Heads</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B-16</td>
<td>87M</td>
<td>16 × 16</td>
<td>196</td>
<td>768</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>ViT-B-32</td>
<td>88M</td>
<td>32 × 32</td>
<td>49</td>
<td>768</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>ViT-S-16</td>
<td>22M</td>
<td>16 × 16</td>
<td>196</td>
<td>384</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>ViT-S-32</td>
<td>23M</td>
<td>32 × 32</td>
<td>49</td>
<td>384</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

\[
\left\| \epsilon' \right\|^2 = \left| \epsilon + \epsilon_0 \right|^2 \\
\leq (\sigma^2 + \sigma_0^2)k + 2\sqrt{k \ln n + 2 \ln n} \\
\leq (\sigma^2 + \sigma_0^2)k(1 + \sqrt{\frac{\ln n}{k}})^2 \\
\leq \rho^2 + \rho_0^2,
\]

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

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

\[
\mathcal{L}_D(w) \leq (1 - \frac{1}{\sqrt{n}}) \max_{||\epsilon'||_{p=\rho'}} \mathcal{L}_S(w + \epsilon') + \frac{1}{\sqrt{n}} \\
+ \frac{k \log (1 + ||w||_2^2 (1 + \sqrt{\frac{\log(n)}{k}})^2 + \log \frac{\rho^2}{\rho^2} + 2 \log(6n + 3k))}{n - 1} \\
\leq \max_{||\epsilon'||_{p=\rho'}} \mathcal{L}_S(w + \epsilon') \\
+ \frac{k \log (1 + ||w||_2^2 (1 + \sqrt{\frac{\log(n)}{k}})^2 + 4 \log \frac{\rho^2}{\rho^2} + 8 \log(6n + 3k))}{n - 1}
\]

where \( \rho^2 = \rho^2 + \rho_0^2 \).