Supplementary Material - Augment Your Batch: Improving Generalization Through Instance Repetition

A. Hypothesis: large batch training issues

Previous works [5, 7, 10] suggested that large-batch training issues may result from an implicit bias in the SGD training process: with large batch sizes, SGD selects minima with worse generalization. We examine the dynamics of SGD to find how such a selection mechanism might work, and suggest why BA has less of these issues, in comparison to standard large batch.

Consider the optimization of non-augmented datasets, using loss functions of the form

$$f\left(\mathbf{w}\right)=\frac{1}{N}\sum_{n=1}^{N}\ell\left(\mathbf{w},\mathbf{x}_{n},\mathbf{y}_{n}\right)\ , \tag{1}$$

where we recall $\left\{\mathbf{x}_{n},\mathbf{y}_{n}\right\}_{n=1}^{N}$ is a dataset of N data sample-target pairs and ℓ is the loss function. We use SGD with batch of size B, where the update rule is given by

$$\mathbf{w}_{t+1}=\mathbf{w}_{t}-\eta\frac{1}{B}\sum_{n\in\mathcal{B}(k(t))}\nabla_{\mathbf{w}}\ell\left(\mathbf{w}_{t},\mathbf{x}_{n},\mathbf{y}_{n}\right)\ . \tag{2}$$

Here, we assume for simplicity that the indices are sampled with replacement, B divides N, and that $k\left(t\right)$ is sampled uniformly from $\{1,\ldots,N/B\}$. When our model is sufficiently rich and over-parameterized (e.g., deep networks), we typically converge to a minimum $\mathbf{w}^{*}$ which is a global minimum on all data points in the training set [11, 8]. This means that $\forall n:\nabla_{\mathbf{w}}\ell\left(\mathbf{w}^{*},\mathbf{x}_{n},\mathbf{y}_{n}\right)=0$. We linearize the dynamics of Eq. 2 near $\mathbf{w}^{*}$ to obtain

$$\mathbf{w}_{t+1}=\mathbf{w}_{t}-\eta\frac{1}{B}\sum_{n\in\mathcal{B}(k(t))}\mathbf{H}_{n}\mathbf{w}_{t}\ , \tag{3}$$

where we assume (without loss of generality) that $\mathbf{w}^{*}=0$, and denote $\mathbf{H}_{n}\triangleq\nabla_{\mathbf{w}}^{2}\ell\left(\mathbf{w},\mathbf{x}_{n},\mathbf{y}_{n}\right)$ as the per-sample Hessian. Since we are at a global minimum, all $\mathbf{H}_{n}$ are symmetric PSD (there are no descent directions). However, recall that there can be many different global minima (on the training set). SGD selects only certain minima. As we shall see this selection depends on the batch sizes and learning rate, through the following quantities: the averaged Hessian over batch k

$$\left\langle\mathbf{H}\right\rangle_{k}\triangleq\frac{1}{B}\sum_{n\in\mathcal{B}(k)}\mathbf{H}_{n}$$

and the maximum over the maximal eigenvalues of $\{\left\langle\mathbf{H}\right\rangle_{k}\}_{k=1}^{N/B}$

$$\lambda_{\max}=\max_{k\in[N/B]}\max_{\forall\mathbf{v}:\|\mathbf{v}\|=1}\mathbf{v}^{\top}\left\langle\mathbf{H}\right\rangle_{k}\mathbf{v}. \tag{4}$$

This $\lambda_{\max}$ affects SGD through the following Theorem (proof in Appendix A.1):

The iterates of SGD (Eq. 3) will converge if

$$\lambda_{\max}<\frac{2}{\eta}\,. \tag{5}$$

In addition, this bound is tight in the sense that it is also a necessary condition for certain datasets.

According to the Theorem, SGD with high learning rate will prefer to converge to minima with low $\lambda_{\max}$, thus selecting them from all (global) minima. Such minima, with low $\lambda_{\max}$, tend to have low variability of $\mathbf{H}_{n}$ (as high variability usually results in larger maximal values).

Next, when increasing the batch size, we typically *decrease* $\lambda_{\max}$, as we *decrease* the variability of $\left\langle\mathbf{H}\right\rangle_{k}$ and replace max operations with averaging. Therefore, certain minima with high variability in $\mathbf{H}_{n}$ will thus become accessible to SGD. Now SGD may converge to these high variability minima, which were suggested to exhibit worse generalization performance than the original minima [10].

This issue can be partially mitigated by increasing the learning rate [3, 1], in a way which will make these new minima inaccessible again, while keeping the original minima accessible. However, merely changing the learning rate may not be sufficient for very large batch sizes, when some minima with high variability and low variability will eventually have similar $\lambda_{\max}$, so SGD will not be able to discriminate between these minima. For example, in the limit of full batch (GD), the variability of $\mathbf{H}_{n}$ will not affect $\lambda_{\max}$ (only their mean).

Now, recall BA can achieve variance reduction that is significantly lower than the $1/B$ reduction, which may occur with an uncorrelated sum of B samples. This implies that the $\lambda_{\max}$ (Eq. 5) would change less in BA than standard large-batch training, allowing the model to exhibit less of the aforementioned SGD convergence issues.

A.1. Proof of Theorem A

We examine the first moment dynamics of Eq. 3, by taking its expectation

$$\mathbf{w}_{t+1} = \left(\mathbf{I} - \eta \left\langle \mathbf{H} \right\rangle\right) \mathbf{w}_t \,, \tag{6}$$

where

$$\left\langle \mathbf{H} \right\rangle \triangleq \frac{1}{N} \sum_{n=1}^{N} \mathbf{H}_n$$

it is easy to see that a necessary and sufficient condition for convergence of Eq. 6

$$\bar{\lambda}_{\max} < \frac{2}{\eta} \,, \tag{7}$$

where $\bar{\lambda}_{\max}$ is the maximal eigenvalue of $\left\langle \mathbf{H} \right\rangle$. This is the standard convergence condition for full batch SGD, i.e., gradient descent.

First, to see Eq. 5 is a necessary condition for certain datasets, suppose we have $\mathbf{H}_n = 0$ in all samples, except, in a single batch k, for which we have

$$\lambda_{\max} = \max_{\forall \mathbf{v}:\|\mathbf{v}\|=1} \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle_k \mathbf{v} \,,$$

In this case, the weights are updated only when we are at batch k. Therefore, ignoring all the batches, the dynamics are equivalent to full batch gradient descent with the dataset restricted to batch k. Therefore, $\bar{\lambda}_{\max} = \lambda_{\max}$, and we only have first order dynamics (with no noise). Thus, the necessary and sufficient condition for stability is Eq. 7 with $\bar{\lambda}_{\max} = \lambda_{\max}$, which is Eq. 5.

Next, to show Eq. 5 is also a sufficient condition (for all data sets) we examine the second moment dynamics. First we observe that

$$\begin{aligned}\mathbf{w}_{t+1}^\top \mathbf{w}_{t+1} &= \mathbf{w}_t^\top \left(\mathbf{I} - \eta \left\langle \mathbf{H} \right\rangle_{k(t)}\right)^\top \left(\mathbf{I} - \eta \left\langle \mathbf{H} \right\rangle_{k(t)}\right) \mathbf{w}_t \,. \\ &= \mathbf{w}_t^\top \left(\mathbf{I} - 2\eta \left\langle \mathbf{H} \right\rangle_{k(t)} + \eta^2 \left\langle \mathbf{H} \right\rangle_{k(t)} \left\langle \mathbf{H} \right\rangle_{k(t)}\right) \mathbf{w}_t \,.\end{aligned}$$

Denoting

$$\left\langle \mathbf{H}^2 \right\rangle \triangleq \frac{1}{N/B} \sum_{k=0}^{N/B} \left\langle \mathbf{H} \right\rangle_k \left\langle \mathbf{H} \right\rangle_k \,.$$

Thus, we obtain

$$\left\|\mathbf{w}_{t+1}\right\|^2 = \left[\mathbf{w}_{t+1}^\top \left(\mathbf{I} - 2\eta \left\langle \mathbf{H} \right\rangle + \eta^2 \left\langle \mathbf{H}^2 \right\rangle\right) \mathbf{w}_t\right] \,. \tag{8}$$

Since $\mathbf{H}_n$ are all PSDs it is easy to see that if $\mathbf{z}$ is a zero eigenvector of $\left\langle \mathbf{H} \right\rangle$ or $\left\langle \mathbf{H}^2 \right\rangle$ then it must be a zero vector eigenvector of other matrix, and also of all $\mathbf{H}_n$, $\forall n$. We denote the null space

$$\mathcal{V} \triangleq \left\{\mathbf{v} \in \mathbb{R}^d \middle| \left\|\mathbf{v}\right\| = 1, \left\langle \mathbf{H} \right\rangle \mathbf{z} = 0\right\}$$

and its complement $\bar{\mathcal{V}}$. From Eq. 8 a necessary and sufficient condition for convergence of this equation is

$$\max_{\mathbf{v} \in \bar{\mathcal{V}}} \mathbf{v}^\top \left(\mathbf{I} - 2\eta \left\langle \mathbf{H} \right\rangle + \eta^2 \left\langle \mathbf{H}^2 \right\rangle\right) \mathbf{v} < 1 \,. \tag{9}$$

To complete the proof we will show that Eq. 5 also implies Eq. 9, for any B.

First we notice that Eq. 4 implies that $\forall \mathbf{v} \in \bar{\mathcal{V}}$:

$$\begin{aligned}\mathbf{v}^\top \left\langle \mathbf{H}^2 \right\rangle \mathbf{v} &= \frac{1}{N} \sum_{k=0}^{N/B} \sum_{n \in \mathcal{B}(k)} \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle_k \mathbf{H}_m \mathbf{v} \\ &\leq \frac{1}{N} \sum_{n=1}^{N} \lambda_{\max} \mathbf{v}^\top \mathbf{H}_n \mathbf{v} \\ &= \lambda_{\max} \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle \mathbf{v} \,.\end{aligned} \tag{10}$$

Also, since $\lambda_{\max} > \bar{\lambda}_{\max}$, we have

$$\mathbf{v}^\top \left\langle \mathbf{H} \right\rangle^2 \mathbf{v} \leq \lambda_{\max} \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle \mathbf{v} \,. \tag{11}$$

We combine the above results to prove the Lemma, and $\forall \mathbf{v} \in \bar{\mathcal{V}}$:

$$\begin{aligned}&\mathbf{v}^\top \left[\left(\mathbf{I} - 2\eta \left\langle \mathbf{H} \right\rangle\right) + \eta^2 \left\langle \mathbf{H}^2 \right\rangle\right] \mathbf{v} \\ =&1 - 2\eta \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle \mathbf{v} + \eta^2 \mathbf{v}^\top \left\langle \mathbf{H}^2 \right\rangle \mathbf{v} \\ \overset{(1)}{\leq}&1 - 2\eta \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle \mathbf{v} + \eta^2 \lambda_{\max} \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle \mathbf{v} \\ =&1 - \eta \left(2 - \eta \lambda_{\max}\right) \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle \mathbf{v} \,,\end{aligned}$$

where in (1) we used Eqs. 10 and 11. Given the condition in Eq. 5 this is smaller than 1, so Eq. 9 holds, so this proves the Theorem.

As a side note, we can bound the convergence rate using the last equation. To see this, we denote $\mathcal{P}_{\bar{\mathcal{V}}}$ as the projection to $\bar{\mathcal{V}}$, and

$$\lambda_{\min} \triangleq \min_{\forall \mathbf{v} \in \bar{\mathcal{V}}} \mathbf{v}^\top \left\langle \mathbf{H} \right\rangle \mathbf{v}$$

as the smallest non-zero eigenvalue of $\left\langle \mathbf{H} \right\rangle$. iterating the recursion we obtain that the convergence rate is linear

$$\left\|\mathcal{P}_{\bar{\mathcal{V}}} \mathbf{w}_t\right\|^2 \leq \left(1 - \eta \left(2 - \eta \lambda_{\max}\right) \lambda_{\min}\right)^t \left\|\mathcal{P}_{\bar{\mathcal{V}}} \mathbf{w}_0\right\|^2 \,. \tag{12}$$

However, note this bound is not necessarily tight.

B. ImageNet Experiments Details

For ResNet50 [2], we used the data augmentation method advocated by [9] that employed various sized patches of the image with size distributed evenly between 8% and 100% and aspect ratio constrained to the interval $[3/4, 4/3]$. The images were also flipped horizontally with $p = 0.5$, and no additional color jitter was performed. For the MobileNet model [4], we used a less aggressive augmentation method, as described in the original paper. In the AlexNet model [6], we used the original augmentation regime.

For all ImageNet models, we followed the training regime by [1] in which an initial learning rate of 0.1 is decreased by a factor of 10 in epochs $30, 60$, and 80 for a total of 90 epochs. We applied a weight decay factor of 10^{-4} to every parameter in the network except for those of batch-norm layers.

The ResNet50 model was trained using multiple feed-forwards and gradient accumulations, creating a ”Ghost batch normalization” [3] effect, where subsets of 64 images in the batch are normalized separately

References

[1] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[3] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *NIPS*, pages 1731–1741, 2017. 1, 3

[4] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[5] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*, 2017. 1

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[7] Kamil Nar and S Shankar Sastry. Step size matters tep size matters in deep learning deep learning. In *NIPS*, 2018. 1

[8] Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017. 1

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[10] Lei Wu, Chao Ma, and Weinan E. How SGD Selects the Global Minima in Over-parameterized Learning : A Dynamical Stability Perspective. In *NeurIPS*, 2018. 1

[11] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 1