On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima

Nitish Shirish Keskar¹ Dheevatsa Mudigere² Jorge Nocedal¹ Mikhail Smelyanskiy² Ping Tak Peter Tang²

¹Northwestern University
²Intel Corporation

ICLR, 2017
Presenter: Tianlu Wang
Outline

1 Introduction
   - Batch Size of Stochastic Gradient Methods

2 Drawbacks of Large-Batch Methods
   - Main Observation
   - Numerical Results
   - Parametric Plots
   - Sharpness of Minima

3 Success of Small-Batch Methods
   - Deterioration along Increasing of Batch-Size
   - Warm-started Large Batch experiments

4 Summary
Outline

1 Introduction
   - Batch Size of Stochastic Gradient Methods

2 Drawbacks of Large-Batch Methods
   - Main Observation
   - Numerical Results
   - Parametric Plots
   - Sharpness of Minima

3 Success of Small-Batch Methods
   - Deterioration along Increasing of Batch-Size
   - Warm-started Large Batch experiments

4 Summary
Batch Size of Stochastic Gradient Methods

- Non-convex optimization in deep learning:
  \[ \min_{x \in \mathbb{R}^n} f(x) := \frac{1}{M} \sum_{i=1}^{M} f_i(x) \]

- Stochastic Gradient Methods and its variants:
  \[ |B_k| \in \{32, 64, \ldots, 512\} \]

- Increase batch size to improve parallelism leads to a loss in generalization performance
Non-convex optimization in deep learning:
\[
\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{M} \sum_{i=1}^{M} f_i(x)
\]

Stochastic Gradient Methods and its variants:
\[|B_k| \in \{32, 64, \ldots, 512\}\]

Increase batch size to improve parallelism leads to a loss in generalization performance

---

Figure 2: Training and testing accuracy for SB and LB methods as a function of epochs.
Outline

1 Introduction
   - Batch Size of Stochastic Gradient Methods

2 Drawbacks of Large-Batch Methods
   - Main Observation
   - Numerical Results
   - Parametric Plots
   - Sharpness of Minima

3 Success of Small-Batch Methods
   - Deterioration along Increasing of Batch-Size
   - Warm-started Large Batch experiments

4 Summary
Main Observations

- Large-batch methods tend to converge to **sharp minimizers** of the training function and tend to generalize less well. Small-batch methods converge to **flat minimizers** and are able to escape basins of attraction of sharp minimizers.
Main Observations

- Large-batch methods tend to converge to sharp minimizers of the training function and tend to generalize less well. Small-batch methods converge to flat minimizers and are able to escape basins of attraction of sharp minimizers.
- Sharp Minimizer $\hat{x}$: function increases rapidly in a small neighborhood of $\hat{x}$
- Flat Minimizer $\bar{x}$: function varies slowly in a large neighborhood of $\bar{x}$

![Diagram showing flat and sharp minima](image)

Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)
Outline

1. Introduction
   • Batch Size of Stochastic Gradient Methods

2. Drawbacks of Large-Batch Methods
   • Main Observation
   • Numerical Results
   • Parametric Plots
   • Sharpness of Minima

3. Success of Small-Batch Methods
   • Deterioration along Increasing of Batch-Size
   • Warm-started Large Batch experiments

4. Summary
Numerical Results

- 6 multi-class classification networks, mean cross entropy, ADAM optimizer, LB: 10% of training data, SB: 256 data points
Numerical Results

- 6 multi-class classification networks, mean cross entropy, ADAM optimizer, LB: 10% of training data, SB: 256 data points

<table>
<thead>
<tr>
<th>Name</th>
<th>Network Type</th>
<th>Architecture</th>
<th>Data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>Fully Connected</td>
<td>Section B.1</td>
<td>MNIST (LeCun et al., 1998a)</td>
</tr>
<tr>
<td>$F_2$</td>
<td>Fully Connected</td>
<td>Section B.2</td>
<td>TIMIT (Garofolo et al., 1993)</td>
</tr>
<tr>
<td>$C_1$</td>
<td>(Shallow) Convolutional</td>
<td>Section B.3</td>
<td>CIFAR-10 (Krizhevsky &amp; Hinton, 2009)</td>
</tr>
<tr>
<td>$C_2$</td>
<td>(Deep) Convolutional</td>
<td>Section B.4</td>
<td>CIFAR-10</td>
</tr>
<tr>
<td>$C_3$</td>
<td>(Shallow) Convolutional</td>
<td>Section B.3</td>
<td>CIFAR-100 (Krizhevsky &amp; Hinton, 2009)</td>
</tr>
<tr>
<td>$C_4$</td>
<td>(Deep) Convolutional</td>
<td>Section B.4</td>
<td>CIFAR-100</td>
</tr>
</tbody>
</table>
6 multi-class classification networks, mean cross entropy, ADAM optimizer, LB: 10% of training data, SB: 256 data points

Table 1: Network Configurations

<table>
<thead>
<tr>
<th>Name</th>
<th>Network Type</th>
<th>Architecture</th>
<th>Data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>Fully Connected</td>
<td>Section B.1</td>
<td>MNIST (LeCun et al., 1998a)</td>
</tr>
<tr>
<td>$F_2$</td>
<td>Fully Connected</td>
<td>Section B.2</td>
<td>TIMIT (Garofolo et al., 1993)</td>
</tr>
<tr>
<td>$C_1$</td>
<td>(Shallow) Convolutional</td>
<td>Section B.3</td>
<td>CIFAR-10 (Krizhevsky &amp; Hinton, 2009)</td>
</tr>
<tr>
<td>$C_2$</td>
<td>(Deep) Convolutional</td>
<td>Section B.4</td>
<td>CIFAR-10</td>
</tr>
<tr>
<td>$C_3$</td>
<td>(Shallow) Convolutional</td>
<td>Section B.3</td>
<td>CIFAR-100 (Krizhevsky &amp; Hinton, 2009)</td>
</tr>
<tr>
<td>$C_4$</td>
<td>(Deep) Convolutional</td>
<td>Section B.4</td>
<td>CIFAR-100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SB</td>
<td>LB</td>
</tr>
<tr>
<td>$F_1$</td>
<td>$99.66% \pm 0.05%$</td>
<td>$99.92% \pm 0.01%$</td>
</tr>
<tr>
<td>$F_2$</td>
<td>$99.99% \pm 0.03%$</td>
<td>$98.35% \pm 2.08%$</td>
</tr>
<tr>
<td>$C_1$</td>
<td>$99.89% \pm 0.02%$</td>
<td>$99.66% \pm 0.2%$</td>
</tr>
<tr>
<td>$C_2$</td>
<td>$99.99% \pm 0.04%$</td>
<td>$99.99% \pm 0.01%$</td>
</tr>
<tr>
<td>$C_3$</td>
<td>$99.56% \pm 0.44%$</td>
<td>$99.88% \pm 0.30%$</td>
</tr>
<tr>
<td>$C_4$</td>
<td>$99.10% \pm 1.23%$</td>
<td>$99.57% \pm 1.84%$</td>
</tr>
</tbody>
</table>
Generalization gap is not due to over-fitting or over-training ???
Generalization gap is not due to *over-fitting* or *over-training* ???

**Figure 2:** Training and testing accuracy for SB and LB methods as a function of epochs.
Outline

1. Introduction
   - Batch Size of Stochastic Gradient Methods

2. Drawbacks of Large-Batch Methods
   - Main Observation
   - Numerical Results
   - Parametric Plots
   - Sharpness of Minima

3. Success of Small-Batch Methods
   - Deterioration along Increasing of Batch-Size
   - Warm-started Large Batch experiments

4. Summary
Parametric Plots

- $x^*_s$ and $x^*_l$: solutions obtained by SB and LB
- plot $f(\alpha x^*_l + (1 - \alpha)x^*_s)$:
- \( x_s^* \) and \( x_l^* \): solutions obtained by SB and LB
- plot \( f(\alpha x_l^* + (1 - \alpha)x_s^*) \):
Outline

1 Introduction
   - Batch Size of Stochastic Gradient Methods

2 Drawbacks of Large-Batch Methods
   - Main Observation
   - Numerical Results
   - Parametric Plots
   - Sharpness of Minima

3 Success of Small-Batch Methods
   - Deterioration along Increasing of Batch-Size
   - Warm-started Large Batch experiments

4 Summary
Motivation: Measure the sensitivity of training function at the given local minimizer, so we want to explore a small neighborhood of a minimizer and compute the largest value that $f$ can attain in this neighborhood.

Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)
Sharpness of Minima

- Small neighborhood:
  \( p \): dimension of manifold
  \( A \): \( n \times p \) matrix, columns are randomly generated
  \( A^+ \): pseudo-inverse of \( A \)

\[
C_\varepsilon = \{ z \in \mathbb{R}^n : -\varepsilon(|x_i| + 1) \leq z_i \leq \varepsilon(|x_i| + 1) \} \\
\forall i \in \{1, 2, \ldots, n\}
\]

\[
C_\varepsilon = \{ z \in \mathbb{R}^p : -\varepsilon(|(A^+ x)_i| + 1) \leq z_i \leq \varepsilon(|(A^+ x)_i| + 1) \} \\
\forall i \in \{1, 2, \ldots, p\}
\]

- **Metric 2.1.** Given \( x \in \mathbb{R}^n \), \( \varepsilon > 0 \) and \( A \in \mathbb{R}^{n \times p} \), the sharpness of \( f \) at \( x \):

\[
\phi_{x,f}(\varepsilon, A) := \frac{(\max_{y \in C_\varepsilon} f(x + Ay)) - f(x)}{1 + f(x)} \times 100
\]  

- \( A \) can be the identity matrix \( I_n \)
Sharpness of Minima in Full Space ($A$ is the identity matrix):

<table>
<thead>
<tr>
<th></th>
<th>$\epsilon = 10^{-3}$</th>
<th></th>
<th>$\epsilon = 5 \cdot 10^{-4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SB</td>
<td>LB</td>
<td>SB</td>
</tr>
<tr>
<td>$F_1$</td>
<td>$1.23 \pm 0.83$</td>
<td>$205.14 \pm 69.52$</td>
<td>$0.61 \pm 0.27$</td>
</tr>
<tr>
<td>$F_2$</td>
<td>$1.39 \pm 0.02$</td>
<td>$310.64 \pm 38.46$</td>
<td>$0.90 \pm 0.05$</td>
</tr>
<tr>
<td>$C_1$</td>
<td>$28.58 \pm 3.13$</td>
<td>$707.23 \pm 43.04$</td>
<td>$7.08 \pm 0.88$</td>
</tr>
<tr>
<td>$C_2$</td>
<td>$8.68 \pm 1.32$</td>
<td>$925.32 \pm 38.29$</td>
<td>$2.07 \pm 0.86$</td>
</tr>
<tr>
<td>$C_3$</td>
<td>$29.85 \pm 5.98$</td>
<td>$258.75 \pm 8.96$</td>
<td>$8.56 \pm 0.99$</td>
</tr>
<tr>
<td>$C_4$</td>
<td>$12.83 \pm 3.84$</td>
<td>$421.84 \pm 36.97$</td>
<td>$4.07 \pm 0.87$</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   - Batch Size of Stochastic Gradient Methods

2. Drawbacks of Large-Batch Methods
   - Main Observation
   - Numerical Results
   - Parametric Plots
   - Sharpness of Minima

3. Success of Small-Batch Methods
   - Deterioration along Increasing of Batch-Size
   - Warm-started Large Batch experiments

4. Summary
Deterioration along Increasing of Batch-Size

- Note batch-size $\approx 15000$ for $F_2$ and batch-size $\approx 500$ for $C_1$

- There exists a threshold after which there is a deterioration in the quality of the model.
Outline

1 Introduction
   - Batch Size of Stochastic Gradient Methods

2 Drawbacks of Large-Batch Methods
   - Main Observation
   - Numerical Results
   - Parametric Plots
   - Sharpness of Minima

3 Success of Small-Batch Methods
   - Deterioration along Increasing of Batch-Size
   - Warm-started Large Batch experiments

4 Summary
Warm-started Large Batch experiments

- Train network for 100 epochs with batch-size=256 and use these 100 epochs as starting points.

- The SB method needs some epochs to explore and discover a flat minimizer.
Numerical experiments that support the view that convergence to sharp minimizers gives rise to the poor generalization of large-batch methods for deep learning.

SB methods have an exploration phase followed by convergence to a flat minimizer.

Attempts to remedy the problem:
- Data augmentation
- Conservative training
- Adversarial training
- Robust optimization