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

N. Keskar†, D. Mudigere‡, J. Nocedal*, M. Smelyanskiy†, P. T. P. Tang†

*Northwestern University, †Intel

Abstract

1. Observation: when using a larger batch (LB) methods there is a degradation in the quality of the model, as measured by its ability to generalize.
2. We investigate the cause for this generalization drop and present numerical evidence that LB methods tend to converge to sharp minimizers of the training and testing functions.
3. SB methods converge to flat minimizers, and this is due to the inherent noise in the gradient estimation.

1 Motivation - Why LB Methods ?

• Small Batch (SB) SGD — Simple, effective but limited scaling (100s of nodes).
• Large Batch (LB) methods — Improved concurrency, potential to scale to 1000s+ nodes, faster time-to-train, larger problems.

Generalization Gap with LB methods

<table>
<thead>
<tr>
<th></th>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SB</td>
<td>LB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parametric Plots

Plot for \( \alpha \in [-1,2]: f(\alpha) + (1-\alpha)2\)

2 Evidence for Sharpness

Conventional Wisdom

• LB methods lose noise/explorative properties.
  “Why is this bad?”
• Minimum number of iterations are required for convergence.
  “Not true: gap exists even if run for 1000s of epochs.”
• LB methods “xoverfit”.
  “Once model is specified, unclear what this means. Surely not over-training.”

Our Observation

1. The lack of generalization ability is due to the fact that LB methods tend to converge to sharp minimizers of the training function.
2. These minimizers are characterized by large positive eigenvalues in \( \nabla^2 f(x) \).
3. SB methods converge to flat minimizers characterized by small positive eigenvalues of \( \nabla^2 f(x) \).

2.0

3 Success of SB Methods

Evolution

3.1 Conclusions

• LB methods → sharp minimizers and these minimizers correlate with poorer generalization.
• SB methods avoid sharp minimizers due to noise.
• Our attempts at data augmentation, conservative training, robust optimization, adversarial training etc. did not consistently close the gap.

7 Open Questions

• Rigorous relationship between training algorithm, minimizer properties and generalization performance.
• Steering to flat minimizers: scalable LB method with SOTA performance.