High Efficiency Online Data Generation to Improve Pretraining Scaling Laws of Deep Neural Networks

Sushmit Chakma [1], Sike Ogieva [2], Randall Balestriero [3]

[1] Haverford College, [2] Amherst College, [3] Brown University

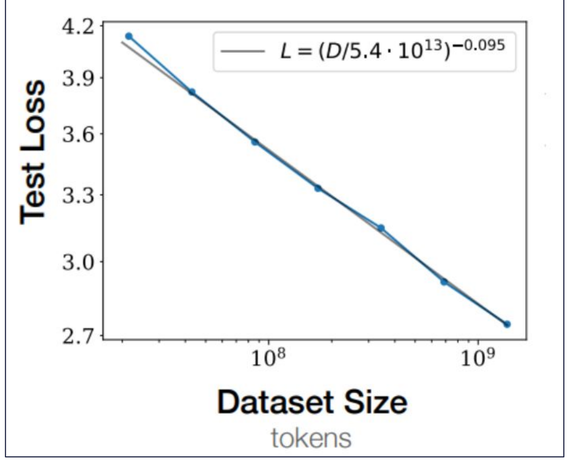
Abstract

Training deep neural networks at scale is increasingly constrained by fundamental problems of data scarcity. In this project, we propose a fast and cheap pipeline for online data generation in training supervised learning models.

Our pipeline introduces a warm-up phase where data is procedurally generated on the fly, eliminating disk-based data loading overhead. This synthetic data can be adapted to model weaknesses through an adversarial reinforcement learning framework.

Introduction

Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute used for training.



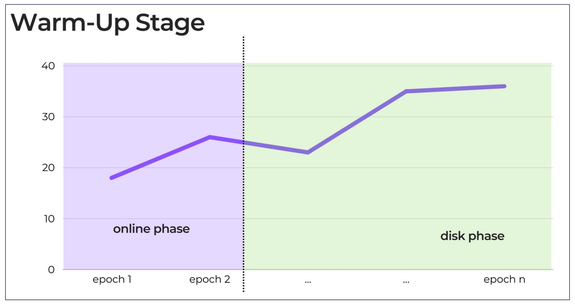
Kaplan et al, Scaling Laws for Neural Language Models, 2020.

Scaling laws for neural networks refers to the empirical fact that model performance scales as a power law with dataset size (model size, compute budget). We seek to achieve similar performances with smaller amounts of data and compute.

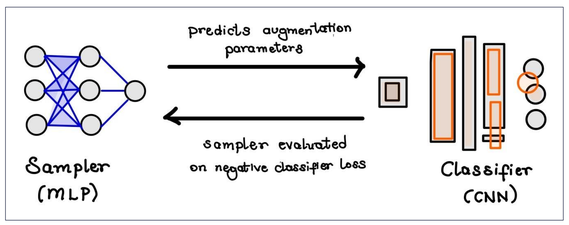
Methodology

Our warm-up phase precedes classic supervised training and is done on quickly generated synthetic data. We run our experiments on the MNIST classification task with a small, custom CNN.

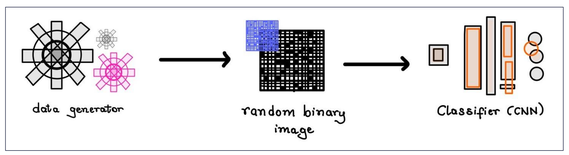
We fix 500 training steps and vary what proportion of these steps make up the warm-up phase.



Experiment 1 We collect single samples per class and the warm-up training is done entirely on versions of these augmented by a model trained on the classifier's loss.



Experiment 2 Then we try basing the warm-up on a different, but related task— generating random binary images, and counting the number of black pixels.



Results

Synthetic training reduces program running time. This can be explained by the decrease in data loading times and overhead.

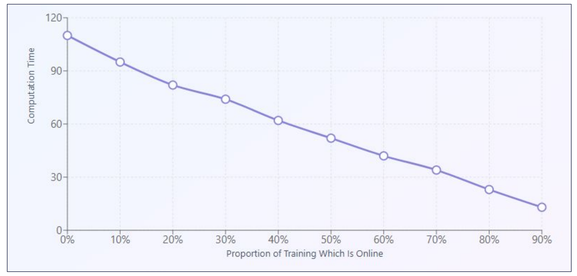


Figure 5: Program running time (in seconds) plotted against online proportion of training (Experiment 2).

Replacing natural data with generated data preserves final accuracies at first but then these accuracies plunge after proportion of training time which is done on synthetic data surpasses 30%.

The 30% threshold is constant across both experiments. And the pattern of improvement then deterioration also holds for the final loss values.

Figure 6: Final accuracy scores plotted against online proportion of training (Experiment 1).

Conclusion

This work demonstrates that incorporating a warm-up phase reduces training times while maintaining final accuracies up to an extent. Specifically, we observe that warm-up proportions up to approximately 30% preserve performance in both experiments, despite their warm-up tasks being structurally different, which presents the question of whether this threshold is a constant function of the dataset.

Experiment 2 validates that even if the early training of the classifier is done on a task which is different from the primary task, the warm-up benefits are still observed. This implies that there is considerable value in researching smarter (yet cheap) generation techniques.

Acknowledgements

Sike and Sushmit would like to thank our Principal Investigator, Dr. Randall Balestriero for his mentorship and expertise.

This research was done under the exploreCSR (program funded by Brown Computer Science and Google Research.

Code: https://github.com/sike25/online_data_generation