DATA PARALLELISM: HOW TO TRAIN DEEP LEARNING MODELS ON MULTIPLE GPUS

LAB 3, PART 1: SCALING THE BATCH SIZE

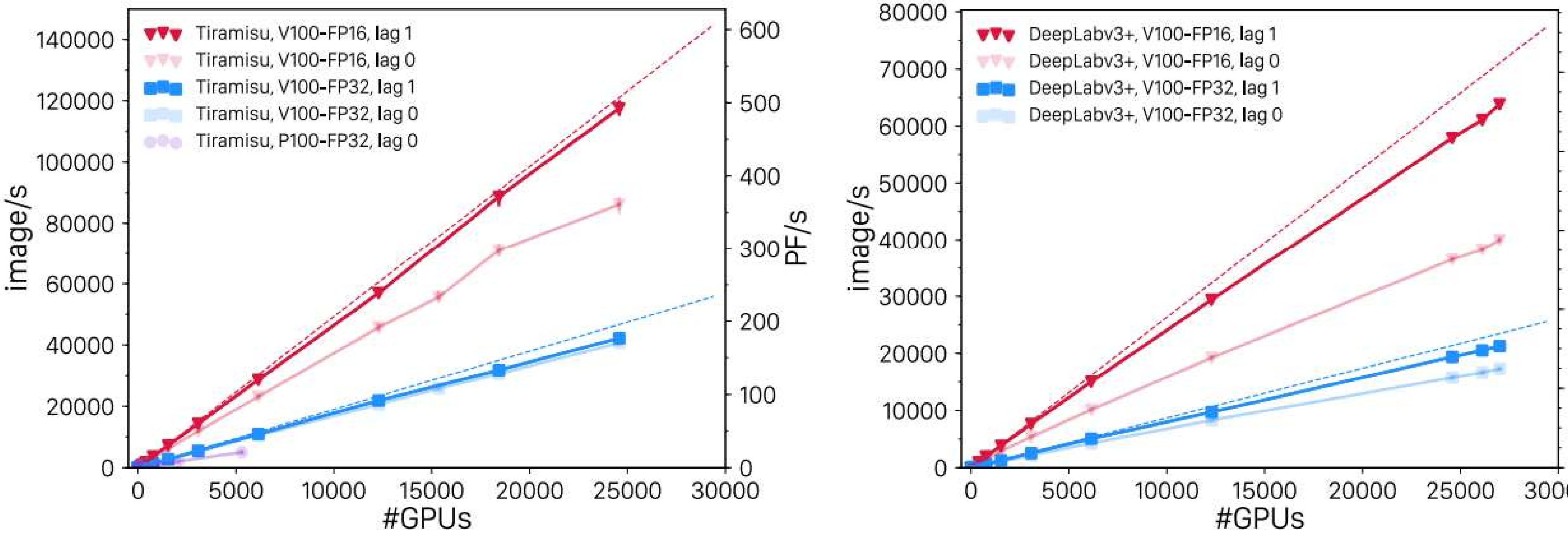
NVIDIA

DEEP
LEARNING
INSTITUTE

CAN WE INCREASE THE BATCH SIZE INDEFINITELY?

IN TERMS OF IMAGES / SECOND?

Yes



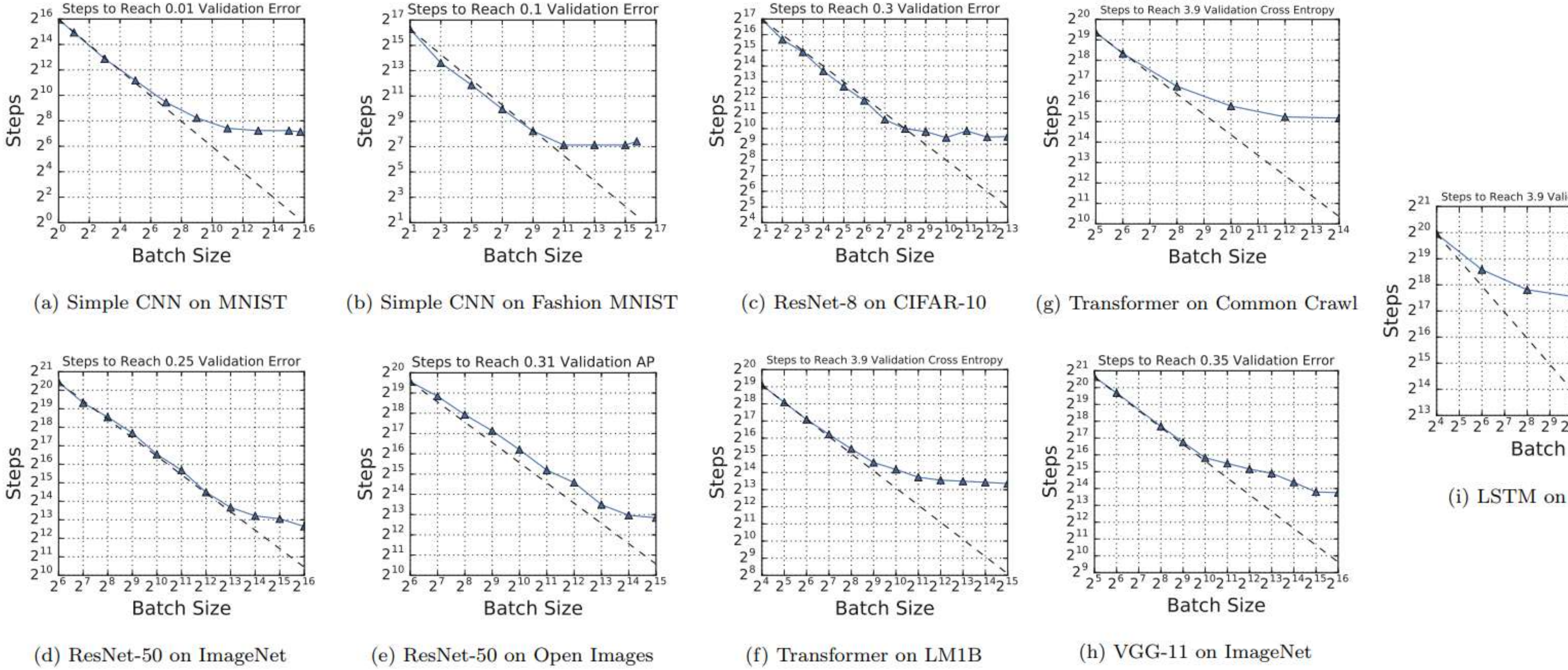
(a) Tiramisu

(b) DeepLabv3+

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., ... & Houston, M. (2018, November). Exascale deep learning for climate analytics. In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (p. 51). IEEE Press. arXiv:1810.01993

IN TERMS OF STEPS TO CONVERGENCE?

There are limits



(a) Simple CNN on MNIST

(b) Simple CNN on Fashion MNIST

(c) ResNet-8 on CIFAR-10

(g) Transformer on Common Crawl

(i) LSTM on

(d) ResNet-50 on ImageNet

(e) ResNet-50 on Open Images

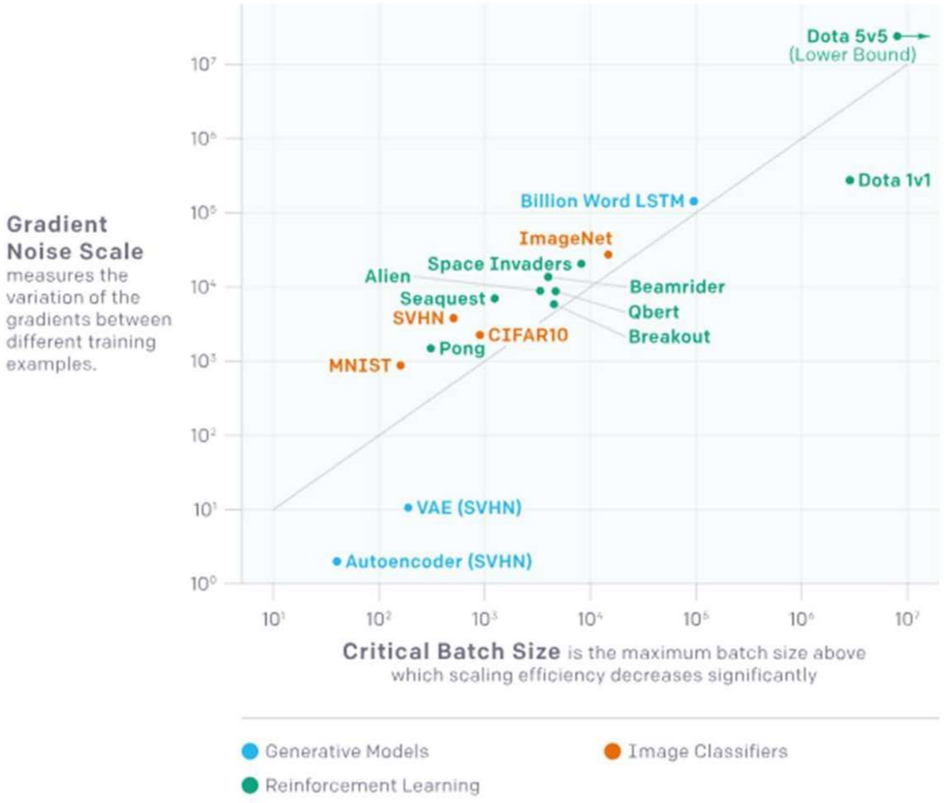
(f) Transformer on LM1B

(h) VGG-11 on ImageNet

Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., & Dahl, G. E. (2018). Measuring the effects of data parallelism on neural network training. arXiv:1811.03600

IN TERMS OF STEPS TO CONVERGENCE?

There are limits

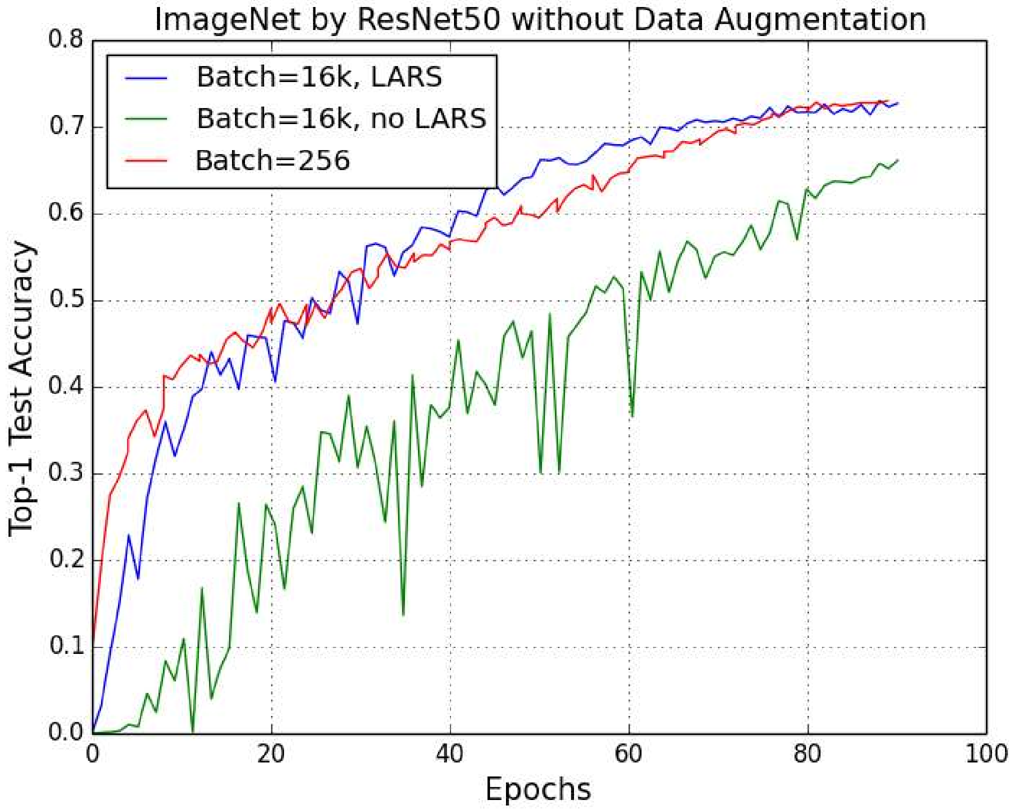


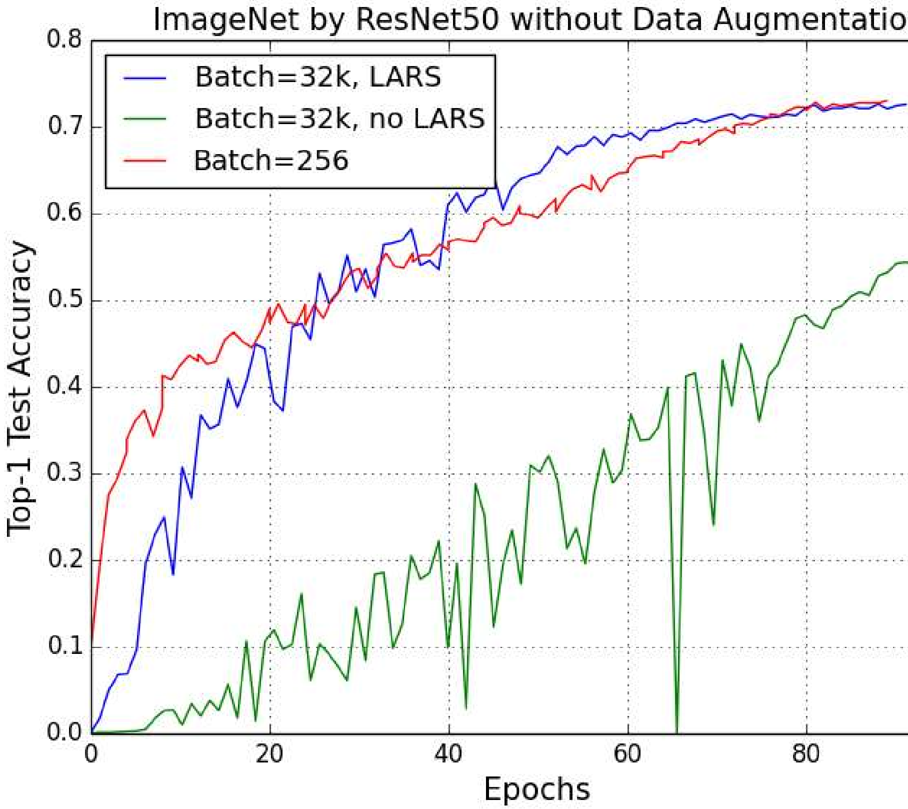
https://blog.openai.com/science-of-ai/

LARGE MINIBATCH AND ITS IMPACT ON ACCURACY

IMPACT ON ACCURACY

Naïve approaches lead to degraded accuracy

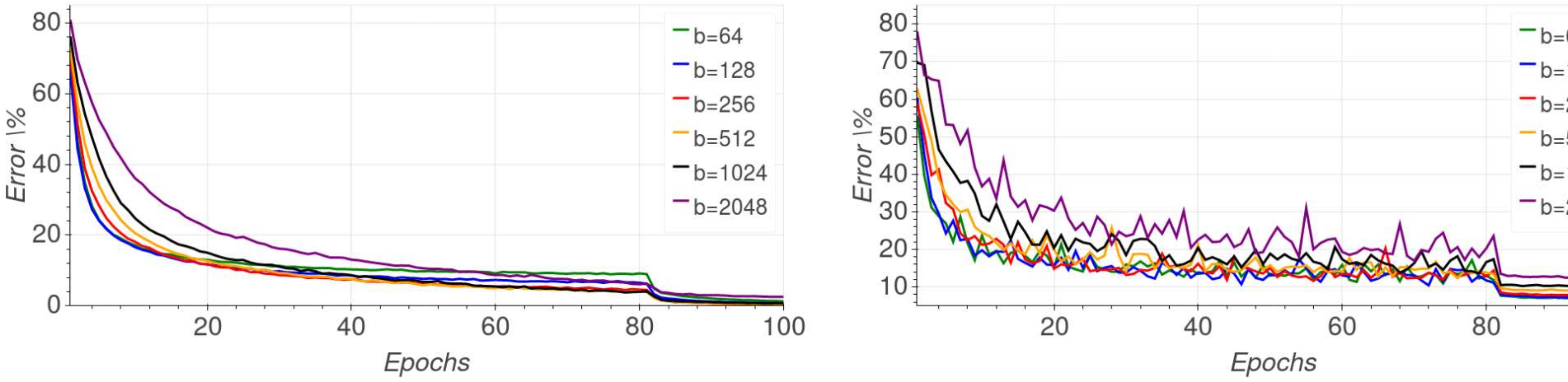




You, Y., Zhang, Z., Hsieh, C., Demmel, J., & Keutzer, K. (2017). ImageNet training in minutes. arXiv: 1709.05011

IMPACT ON ACCURACY

Naïve approaches lead to degraded accuracy



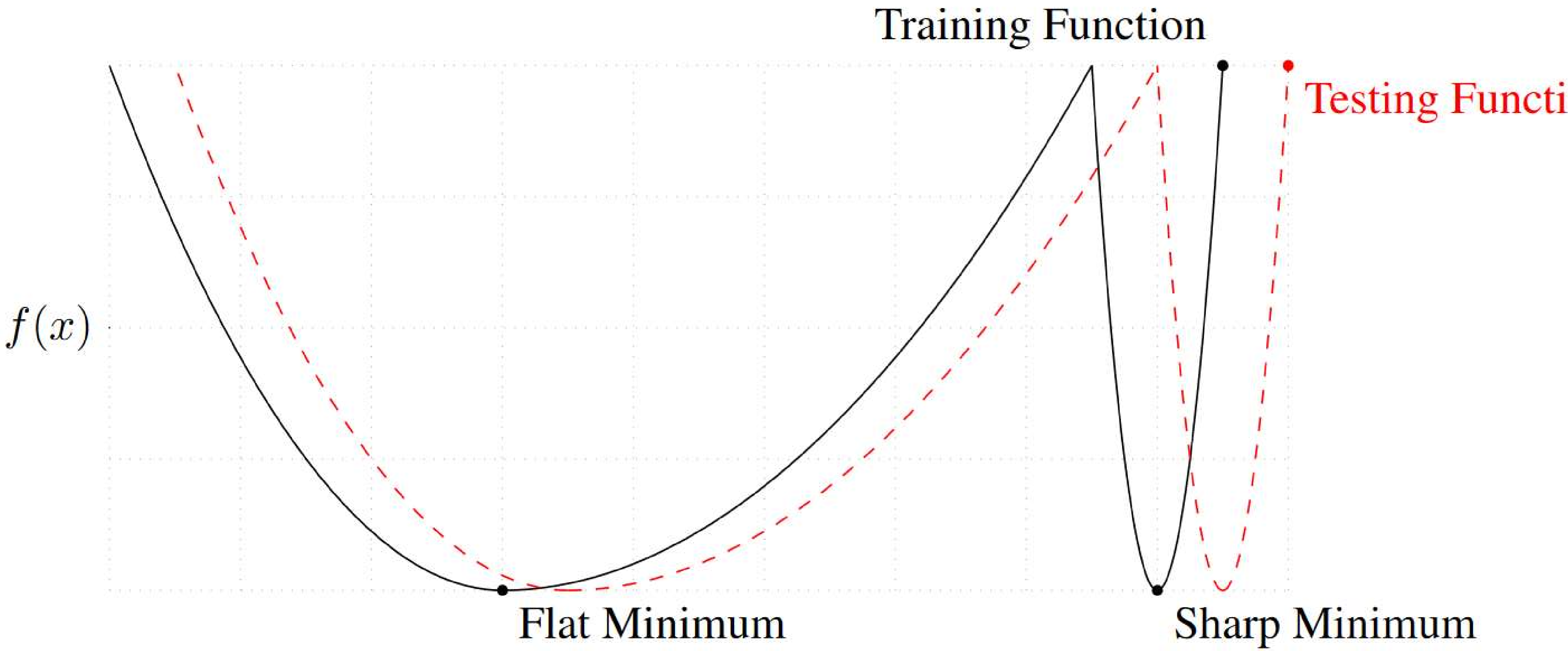
(a) Training error

(b) Validation error

Hoffer, E., Hubara, I., & Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. arXiv:1705.08741

IMPACT ON ACCURACY

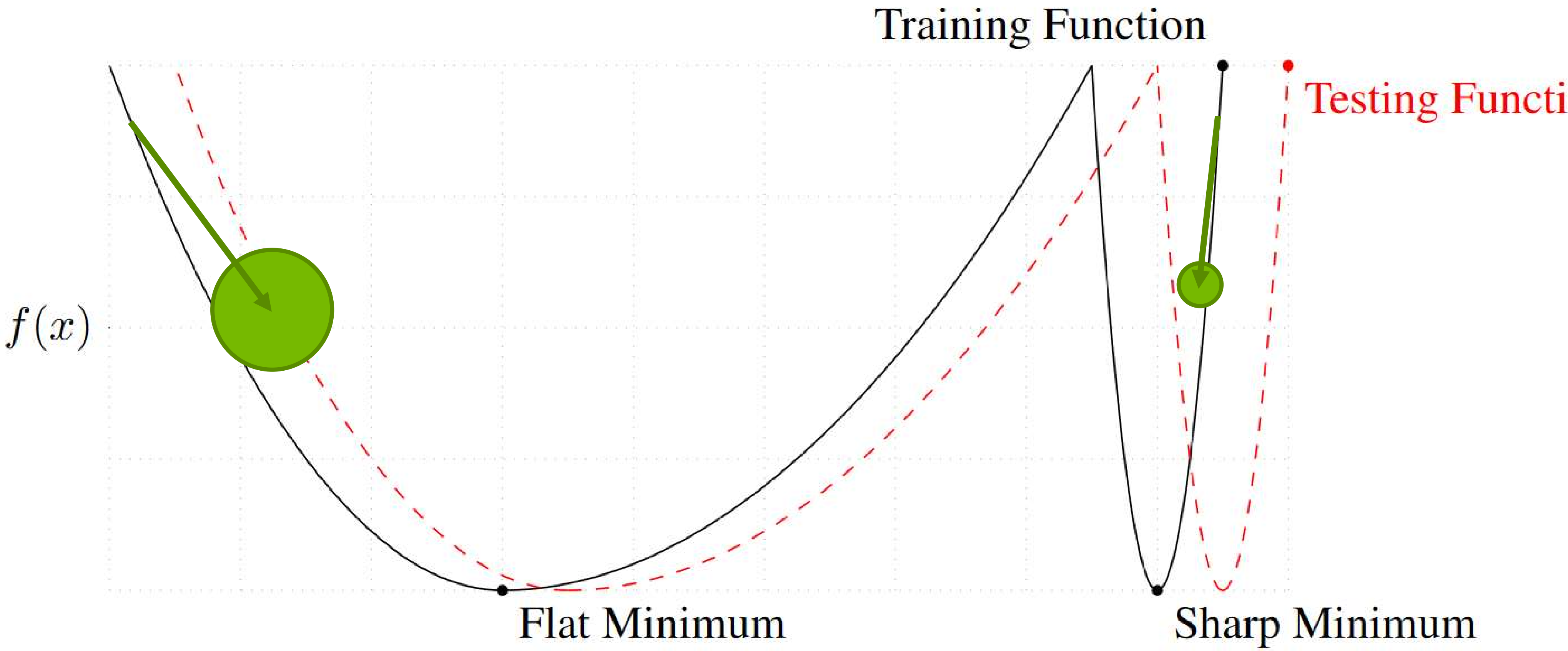
Why? Generalization and flatness of minima?



Keskar, N. S., et al. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv:1609.04836

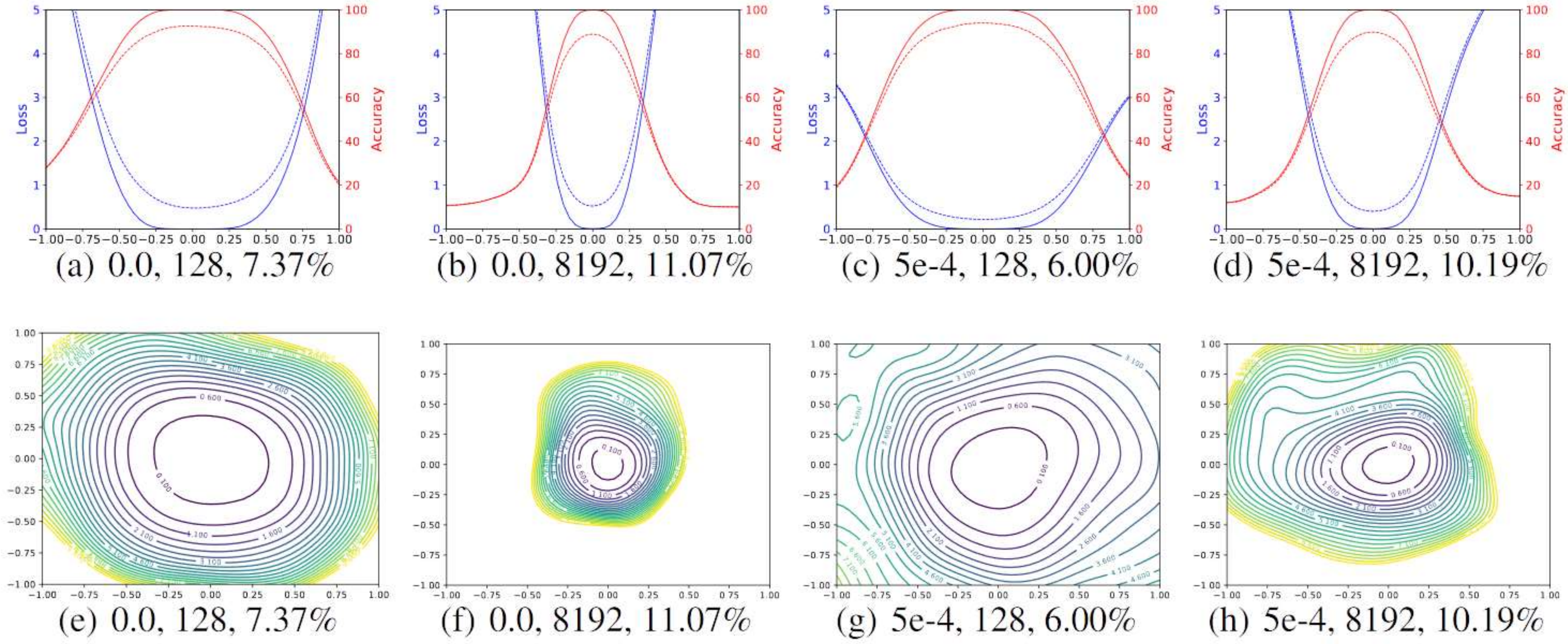
IMPACT ON ACCURACY

Why does it happen? Noise in the gradient update.



Keskar, N. S., et al. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv:1609.04836

IMPACT ON ACCURACY



(a) 0.0, 128, 7.37%

(b) 0.0, 8192, 11.07%

(c) 5e-4, 128, 6.00%

(d) 5e-4, 8192, 10.19%

(e) 0.0, 128, 7.37%

(f) 0.0, 8192, 11.07%

(g) 5e-4, 128, 6.00%

(h) 5e-4, 8192, 10.19%

Figure 3: The 1D and 2D visualization of solutions obtained using SGD with different weight decay and batch size. The title of each subfigure contains the weight decay, batch size, and test error.

Li, H., Xu, Z., Taylor, G., & Goldstein, T. (2017). Visualizing the Loss Landscape of Neural Nets. arXiv:1712.09913

NVIDIA

DEEP
LEARNING
INSTITUTE

www.nvidia.com/dli