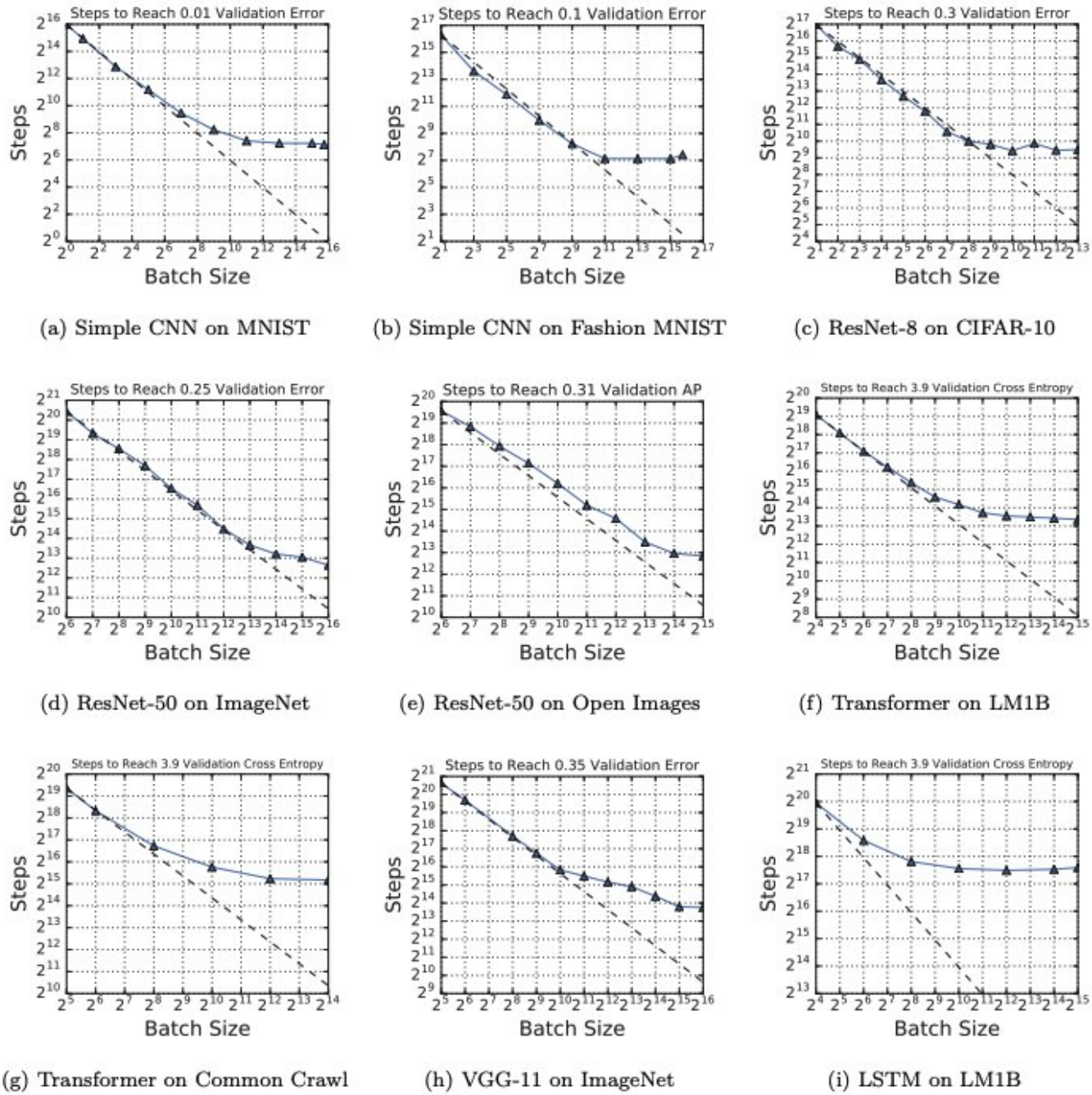
Twitter Thread by Roger Grosse

Roger Grosse
@RogerGrosse

Important paper from Google on large batch optimization. They do impressively careful experiments measuring # iterations needed to achieve target validation error at various batch sizes. The main "surprise" is the lack of surprises. [thread]

https://t.co/7QIx5CFdfJ



(a) Simple CNN on MNIST (b) Simple CNN on Fashion MNIST (c) ResNet-8 on CIFAR-10

(d) ResNet-50 on ImageNet (e) ResNet-50 on Open Images (f) Transformer on LM1B

(g) Transformer on Common Crawl (h) VGG-11 on ImageNet (i) LSTM on LM1B

Figure 1: **The relationship between steps to result and batch size has the same characteristic form for all problems.** In all cases, as the batch size grows, there is an initial period of **perfect scaling** (indicated with a dashed line) where the steps needed to achieve the error goal halves for each doubling of the batch size. Then there is a region of **diminishing returns** that eventually leads to a region of **maximal data parallelism** where additional parallelism provides no benefit whatsoever. AP denotes average precision (see Appendix A).

The paper is a good example of lots of elements of good experimental design. They validate their metric by showing lots of variants give consistent results. They tune hyperparamters separately for each condition, check that optimum isn't at the endpoints, and measure sensitivity.

They have separate experiments where the hold fixed # iterations and # epochs, which (as they explain) measure very different things. They avoid confounds, such as batch norm's artificial dependence between batch size and regularization strength.

When the experiments are done carefully enough, the results are remarkably consistent between different datasets and architectures. Qualitatively, MNIST behaves just like ImageNet.

Importantly, they don't find any evidence for a "sharp/flat optima" effect whereby better optimization leads to worse final results. They have a good discussion of experimental artifacts/confounds in past papers where such effects were reported.

The time-to-target-validation is explained purely by optimization considerations. There's a regime where variance dominates, and you get linear speedups w/ batch size. Then there's a regime where curvature dominates and larger batches don't help. As theory would predict.

Incidentally, this paper must have been absurdly expensive, even by Google's standards. Doing careful empirical work on optimizers requires many, many runs of the algorithm. (I think surprising phenomena on ImageNet are often due to the difficulty of running proper experiments.)