TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Stochastic Gradient Descent (SGD)

Decoupling the Learning Rate From the Batch Size

Decoupling η from B

For vanilla SGD with minibatching we have

$$
\begin{aligned}
\Phi_{t+1} \;\; &\texttt{-=} \;\; \eta \hat{g}_t \\
\hat{g}_t \;\; &= \;\; \frac{1}{B} \sum_b \hat{g}_{t,b}
\end{aligned}
$$

Where $\hat{g}_{t,b}$ is the gradient of the element b of the batch.

Decoupling η from B

We can compare batch size B to batch size 1.

For batch size 1 on the same sequence of data points with $b \in \{1, \ldots, B\}$ and with learning rate η_0 we have

$$\Phi_{t+b} = \Phi_{t+b-1} - \eta_0 \, \nabla_\Phi \, \mathcal{L}(b, \Phi_{t+b-1})$$

where $\mathcal{L}(b, \Phi_{t+b-1})$ is the gradient of the batch element b at parameter value Φ_{t+b-1}.

Decoupling η from B

$$\Phi_{t+b} = \Phi_{t+b-1} - \eta_0 \, \nabla_\Phi \, \mathcal{L}(b, \Phi_{t+b-1})$$

If the parameters are moving slowing we have

$$\nabla_\Phi \, \mathcal{L}(b, \Phi_{t+b-1}) \approx \nabla_\Phi \, \mathcal{L}(b, \Phi_t) = \hat{g}_{t,b}$$

So for Batch size 1 we get

$$\Phi_{t+B} \approx \Phi_t - \eta_0 \sum_b \hat{g}_{t,b}$$

Decoupling η from B

For batch size 1 we have

$$\Phi_{t+B} \approx \Phi_t - \eta_0 \sum_b \hat{g}_{t,b}$$

For batch size B we have

$$\Phi_{t+1} = \Phi_t - \eta \frac{1}{B} \sum_b \hat{g}_{t,b}$$

If η_0 causes convergence to a desirable loss at batch size 1 then

$$\eta = B\eta_0$$

will cause a similar convergence at batch size B.

Decoupling η from B

Recent work has show that using $\eta = B\eta_0$ leads to effective learning with very large (highly parallel) batches.

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., 2017.

END