TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

SGD with Momentum

Momentum

The standard (PyTorch) momentum SGD equations are

$$v_t = \mu v_{t-1} + \eta * \hat{g}_t \quad \mu \text{ is typically .9 or .99}$$

$$\Phi_{t+1} = \Phi_t - v_t$$

Here v is velocity, $0 \leq \mu < 1$ represents friction drag and $\eta \hat{g}$ is the acceleration generated by the gradient force.

Momentum

The theory of momentum is generally given in terms of second order structure and total gradients (GD rather than SGD).

But second order analyses of are controversial for SDG in very large dimension.

Still, momentum is widely used in practice.

Momentum and Temperature

$$v_t = \mu v_{t-1} + \eta * \hat{g}_t \quad \mu \text{ is typically .9 or .99}$$

$$\Phi_{t+1} = \Phi_t - v_t$$

We will use a first order analysis to argue that by setting

$$\eta = (1 - \mu) B \eta_0$$

the temperature will be essentially determined by η_0 independent of the choice of the momentum parameter μ or the batch size B.

Momentum and Temperature

$$\eta = (1 - \mu)B\eta_0$$

Emprical evidence for this setting of η is given in

Don't Decay the Learning Rate, Increase the Batch Size, Smith et al., 2018

Momentum as a Running Average

Consider a sequence x_1, x_2, x_3, ...

For $t \geq N$, consider the average of the N most recent values.

$$\overline{x}_t = \frac{1}{N} \sum_{k=0}^{N-1} x_{t-k}$$

This can be approximated efficiently with

$$\tilde{x}_0 = 0$$

$$\tilde{x}_t = \left(1 - \frac{1}{N}\right) \tilde{x}_{t-1} + \left(\frac{1}{N}\right) x_t$$

Deep Learning Convention for Running Averages

In deep learning a running average

$$\tilde{x}_t = \left(1 - \frac{1}{N}\right)\tilde{x}_{t-1} + \left(\frac{1}{N}\right)x_t$$

is written as

$$\tilde{x}_t = \beta\tilde{x}_{t-1} + (1-\beta)x_t$$

where

$$\beta = 1 - 1/N$$

Typical values for β are .9, .99 or .999 corresponding to N being 10, 100 or 1000.

It will be convenient here to use N rather than β.

Momentum as a Running Average

$$
\begin{aligned}
v_t &= \mu v_{t-1} + \eta \hat{g}_t \\
&= \left(1 - \frac{1}{N}\right) v_{t-1} + \frac{1}{N}(N\eta\hat{g}_t)
\end{aligned}
$$

We see that v_t is a running average of $N\eta\hat{g}$.

Momentum as a Running Average

v_t is a running average of $N\eta\hat{g}$.

Alternatively, we can consider a direct running average of the gradient.

$$\tilde{g}_t = \left(1 - \frac{1}{N}\right)\tilde{g}_{t-1} + \left(\frac{1}{N}\right)\hat{g}_t$$

The running average of $N\eta\hat{g}$ is the same as $N\eta$ times the running average of $\hat{g}$. Hence

$$v_t = N\eta\tilde{g}_t$$

Momentum as a Running Average

We have now shown that the standard formulation of momentum can be written as

$$\tilde{g}_t = \left(1 - \frac{1}{N}\right)\tilde{g}_{t-1} + \left(\frac{1}{N}\right)\hat{g}_t$$

$$\Phi_{t+1} = \Phi_t - N\eta\tilde{g}_t$$

Total Effect Rule

We will adopt the rule of thumb that the temperature is determined by the total effect of a single training gradient $g_{t,b}$.

Also that "temperature" corresponds to the converged loss at fixed learning rate.

Total Effect Rule

The effect of $g_{t,b}$ on the batch average $\hat{g}_t$ is $\left(\frac{1}{B}\right) g_{t,b}$.

$$\text{Using} \qquad \sum_{i=0}^{\infty} \frac{1}{N}\left(1-\frac{1}{N}\right)^i = 1$$

we get that the effect of $\hat{g}_t$ on $\sum_{t=0}^{\infty} \tilde{g}_t$ equals $\hat{g}_t$.

So for $\Phi_{t+1} = \Phi_t - N\eta\tilde{g}_t$ the total effect of $g_{t,b}$ is $(N/B)\eta\ g_{t,b}$.

Total Effect Rule

For $\Phi_{t+1} = \Phi_t - N\eta\tilde{g}_t$ the total effect of $g_{t,b}$ is $(N/B)\eta\; g_{t,b}$.

By taking $\eta = \frac{B}{N}\eta_0$ we get that the total effect, and hence the temperature, is determined by η_0 independent of the choice of N and B.

For the standard momentum paramenter $\mu = (1 - 1/N)$ this becomes

$$\eta = (1 - \mu)B\eta_0$$

where η_0 determines temperature independent of μ and B.

END