Welford's Online Algorithm for the computation of the Running Variance

In the weights' gradients stats tracker class (WeightsGradientsStatsTracker) we have two mains statistics, the running absolute mean and the running absolute mean variance. The first is the mean of the absolute value of all the gradient for every weights per layer. The second is the variance of the observed absolute mean of all the gradient for every weights per layer. To compute those statistic, we use Welford's online algorithm [Welford(1962)].

Computing the Running Absolute Mean of the Weights' Gradient Per Layer

The n-th running absolute mean for the i-th layer gradients mean weights' is

$$\bar{x}_{n,i} = \bar{x}_{n-1,i} + \frac{x_{n,i} - \bar{x}_{n-1,i}}{n}$$

where $\bar{x}_{n-1,i}$ is the previous running absolute mean for the layer i, n is the batch number (i.e. how many times we have updated the variance) and $x_{n,i}$ is the absolute mean of the weights' gradients of the layer i. Also, when $n = 1$, $\bar{x}_{n-1,i} = 0$.

Computing the Running Variance of the Absolute Mean Weights Gradient Per Layer

The n-th running variance for the i-th layer weights' gradients absolute mean is

$$s^2_{n,i} = \frac{M_{2,n,i}}{n-1}$$

where

$$M_{2,n,i} = M_{2,n-1,i} + (x_{n,i} - \bar{x}_{n-1,i}) \times (x_{n,i} - \bar{x}_{n,i})$$

Also, when $n = 1$, $M_{2,n,i} = 0$ and $s^2_{n,i} = 0$.

Example of Computation

Having the following two layers gradients weights' update

$$\begin{aligned}
\text{layer}_1 &= [0.24, 0.00, -0.15] \\
\text{layer}_2 &= [-0.16, 0.25, 0.00]
\end{aligned}$$

Thus, if $n = 1$

$$\begin{aligned}
\bar{x}_{1,1} &= 0 + \frac{0.13 - 0}{1} = 0.13 \\
\bar{x}_{1,2} &= 0 + \frac{0.13\bar{6} - 0}{1} = 0.13\bar{6} \\
s^2_{1,1} &= 0 \\
s^2_{1,2} &= 0
\end{aligned}$$

For $n = 2$, assuming the updated weights' gradients vectors are

$$\begin{aligned}
\text{layer}_1 &= [0.24, 0.00, -0.15] \times 2 = [0.48, 0.00, -0.30] \\
\text{layer}_2 &= [-0.16, 0.25, 0.00] \times 2 = [-0.32, 0.50, 0.00]
\end{aligned}$$

the running means and variances are

$$\begin{aligned}
\bar{x}_{2,1} &= 0.13 + \frac{0.26 - 0.13}{2} = 0.195 \\
\bar{x}_{2,2} &= 0.13\bar{6} + \frac{0.27\bar{3} - 0.13\bar{6}}{2} = 0.205 \\
s^2_{2,1} &= \frac{0 + (0.26 - 0.13) \times (0.26 - 0.195)}{2 - 1} = 0.00845 \\
s^2_{2,2} &= \frac{0 + (0.27\bar{3} - 0.13\bar{6}) \times (0.27\bar{3} - 0.205)}{2 - 1} = 0.00933889
\end{aligned}$$

For $n = 3$, assuming the updated weights' gradients vectors are

$$\begin{aligned}
\text{layer}_1 &= [0.24, 0.00, -0.15] \times 3 = [0.72, 0.00, -0.45] \\
\text{layer}_2 &= [-0.16, 0.25, 0.00] \times 3 = [-0.48, 0.75, 0.00]
\end{aligned}$$

the running means and variances are

$$\begin{aligned}
\bar{x}_{3,1} &= 0.195 + \frac{0.39 - 0.195}{3} = 0.26 \\
\bar{x}_{3,2} &= 0.205 + \frac{0.41 - 0.205}{3} = 0.27\bar{3} \\
s^2_{3,1} &= \frac{0.00845 + (0.39 - 0.195) \times (0.39 - 0.26)}{3 - 1} = 0.0169 \\
s^2_{3,2} &= \frac{0.00933889 + (0.41 - 0.205) \times (0.41 - 0.27\bar{3})}{3 - 1} = 0.018677778
\end{aligned}$$

References

[Welford(1962)] B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4 (3):419–420, 1962. doi: 10.1080/00401706.1962.10490022. URL `https://www.tandfonline.com/doi/abs/10.1080/00401706.1962.10490022`.