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

In the weights' gradients stats tracker class (WeightsGradientsStatsTracker), the goal is to estimate, for each set of weights updated over an epoch, the mean of the means of the absolute values of the weights' gradient and the variance of the means of the absolute values of the weights' gradient.

Formally, for gradient $g \in \mathbb{R}^n$ of a given set of weights, let $x_i = \frac{1}{n}\sum_{j=1}^{n}|g_j|$ be the the mean of the elements in g of the i-th batch of the epoch. Then, we wish to estimate the absolute mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and the variance $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ over the x_is. To compute those statistic in a numerically stable way, we use Welford's online algorithm [Welford(1962)].

In the remainder of this document, we denote by $\bar{x}_i = \frac{1}{i}\sum_{j=1}^{i} x_j$ and $s^2 = \frac{1}{i-1}\sum_{i=1}^{j}(x_j - \bar{x})^2$ the absolute mean and the variance of the absolute mean of the i-th first elements or otherwise called running mean and running variance respectively.

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

The i-th running absolute mean for the means of the weights' gradient is

$$\bar{x}_i = \bar{x}_{i-1} + \frac{x_i - \bar{x}_{i-1}}{i}$$

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

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

The i-th running variance for the means of the weights' gradient is

$$s_i^2 = \frac{M_{2,i}}{i-1}$$

where

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

Also, when $i = 1$, $M_{2,i} = 0$ and $s_i^2 = 0$.

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 $n = 1$, for layer_1

$$\begin{aligned}\bar{x}_1 &= 0 + \frac{0.13 - 0}{1} = 0.13\\ s_1^2 &= 0\end{aligned}$$

and for layer_2

$$\begin{aligned}\bar{x}_1 &= 0 + \frac{0.13\bar{6} - 0}{1} = 0.13\bar{6}\\ s_1^2 &= 0\end{aligned}$$

For $i = 2$, 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 mean and variance for layer_1 are

$$\begin{aligned}\bar{x}_2 &= 0.13 + \frac{0.26 - 0.13}{2} = 0.195 \\ s_2^2 &= \frac{0 + (0.26 - 0.13) \times (0.26 - 0.195)}{2 - 1} = 0.00845\end{aligned}$$

and for layer_2

$$\begin{aligned}\bar{x}_2 &= 0.13\bar{6} + \frac{0.27\bar{3} - 0.13\bar{6}}{2} = 0.205 \\ s_2^2 &= \frac{0 + (0.27\bar{3} - 0.13\bar{6}) \times (0.27\bar{3} - 0.205)}{2 - 1} = 0.00933889\end{aligned}$$

For $i = 3$, 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 mean and variance for layer_1 are

$$\begin{aligned}\bar{x}_3 &= 0.195 + \frac{0.39 - 0.195}{3} = 0.26 \\ s_3^2 &= \frac{0.00845 + (0.39 - 0.195) \times (0.39 - 0.26)}{3 - 1} = 0.0169\end{aligned}$$

and for layer_2

$$\begin{aligned}\bar{x}_3 &= 0.205 + \frac{0.41 - 0.205}{3} = 0.27\bar{3} \\ s_3^2 &= \frac{0.00933889 + (0.41 - 0.205) \times (0.41 - 0.27\bar{3})}{3 - 1} = 0.018677778\end{aligned}$$

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`.