AdaX：一个比Adam更优秀，带”长期记忆“的优化器

原创 苏剑林 夕小瑶的卖萌屋 1周前

前言

这篇文章简单介绍一个叫做AdaX的优化器，来自《AdaX: Adaptive Gradient Descent with Exponential Long Term Memory》。介绍这个优化器的原因是它再次印证了之前在《硬核推导Google　AdaFactor：一个省显存的宝藏优化器 》一文中提到的一个结论，两篇文章可以对比着阅读。

Adam & AdaX

AdaX的更新格式是

$$\begin{cases}g_t = \nabla_\theta L(\theta_t) \\ m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t = (1+\beta_2) v_{t-1} + \beta_2 g_t^2 \\ \hat{v}_t = v_t \Big/ \Big((1+\beta_2)^t - 1\Big) \\ \theta_t = \theta_{t-1} - \alpha_t m_t \Big/ \sqrt{\hat{v}_t + \epsilon}\end{cases} \tag{1}$$

其中β_2的默认值是0.0001。对了，顺便附上自己的Keras实现：https://github.com/bojone/adax 作为比较，Adam的更新格式是

$$\begin{cases}g_t = \nabla_\theta L(\theta_t) \\ m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \hat{m}_t = m_t \big/ \big(1-\beta_1^t\big) \\ \hat{v}_t = v_t \big/ \big(1-\beta_2^t\big) \\ \theta_t = \theta_{t-1} - \alpha_t \hat{m}_t \Big/ \sqrt{\hat{v}_t + \epsilon}\end{cases} \tag{2}$$

其中β_2的默认值是0.999。

等价形式变换

可以看到，两者的第一个差别是AdaX去掉了动量的偏置校正 $\hat{m}_t = m_t/\left(1-\beta_1^t\right)$ （这一步），但这其实影响不大，AdaX最大的改动是在v_t处，本来$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$是滑动平均格式，而$v_t = (1+\beta_2) v_{t-1} + \beta_2 g_t^2$不像是滑动平均了，而且$1+\beta_2 > 1$，似乎有指数爆炸的风险?

原论文称之为“with Exponential Long Term Memory”，就是指 $1+\beta_2 > 1$导致历史累积梯度的比重不会越来越小，反而会越来越大，这就是它的长期记忆性。

事实上，学习率校正用的是$\hat{v}_t$，所以有没有爆炸我们要观察的是$\hat{v}_t$。对于Adam，我们有

$$\begin{aligned}\hat{v}_t =& v_t/\left(1-\beta_2^t\right)\\ =&\frac{\beta_2 v_{t-1} + (1-\beta_2) g_t^2}{1-\beta_2^t}\\ =&\frac{\beta_2 \hat{v}_{t-1}\left(1-\beta_2^{t-1}\right) + (1-\beta_2) g_t^2}{1-\beta_2^t}\\ =&\beta_2\frac{1-\beta_2^{t-1}}{1-\beta_2^t}\hat{v}_{t-1} + \left(1-\beta_2\frac{1-\beta_2^{t-1}}{1-\beta_2^t}\right) g_t^2\end{aligned} \tag{3}$$

所以如果设$\hat{\beta}_{2,t} = \beta_2\frac{1-\beta_2^{t-1}}{1-\beta_2^t}$，那么更新公式就是

$$\hat{v}_t = \hat{\beta}_{2,t}\hat{v}_{t-1} + \left(1-\hat{\beta}_{2,t}\right) g_t^2 \tag{4}$$

基于同样的道理，如果设$\hat{\beta}_{2,t} = 1 - \frac{\beta_2}{(1+\beta_2)^t-1}$，那么AdaX的$\hat{v}_t$的更新公式也可以写成上式。

衰减策略比较

所以，从真正用来校正梯度的$\hat{v}_t$来看，不管是Adam还是AdaX，其更新公式都是滑动平均的格式，只不过对应的衰减系数$\hat{\beta}_{2,t}$不一样。

对于Adam来说，当时t = 0， $\hat{\beta}_{2,t} = 0$，这时候$\hat{v}_t$就是g_t^2，也就是用实时梯度来校正学习率，这时候校正力度最大；当$t\to\infty$时，$\hat{\beta}_{2,t}\to\beta_2$，这时候$\hat{v}_t$是累积梯度平方与当前梯度平方的加权平均，由于$\beta_2 < 1$，所以意味着当前梯度的权重$1-\beta_2$不为0，这可能导致训练不稳定，因为训练后期梯度变小，训练本身趋于稳定，校正学习率的意义就不大了，因此学习率的校正力度应该变小，并且$t\to\infty$，学习率最好恒定为常数（这时候相当于退化为SGD），这就要求$t\to\infty$时，$\hat{\beta}_{2,t}\to 1$。

对于AdaX来说，当t = 0时 $\hat{\beta}_{2,t} = 0$，当$t\to\infty$，$\hat{\beta}_{2,t}\to 1$，满足上述的理想性质，因此，从这个角度来看，AdaX确实是Adam的一个改进。在AdaFactor中使用的则是$\hat{\beta}_{2,t} = 1 - \frac{1}{t^c}$，它也是从这个角度设计的。至于AdaX和AdaFactor的策略孰优孰劣，笔者认为就很难从理论上解释清楚了，估计只能靠实验。

就这样结束了

嗯，文章就到这儿结束了。开头就说了，本文只是简单介绍一下AdaX，因为它再次印证了之前的一个结论——$\hat{\beta}_{2,t}$应当满足条件“$\hat{\beta}_{2,0} = 0, \hat{\beta}_{2,\infty} = 1$”，这也许会成为日后优化器改进的基本条件之一。