SGD、Momentum、RMSprop、Adam区别与联系

徐小贱民
互联网，Data Scientist，快手

说到优化算法，不由得想起六个月前刚毕业那段时间。我们班连我一共三个去搞了算法，另外两个分别去了滴滴和京东，于是我们决定抱团学习深度学习，制定了一系列的计划，从哪一块开始？那就从优化算法吧。然后.....就没有然后了。

深度学习优化算法经历了 BGD -> SGD -> MBGD -> SGDM -> NAG ->AdaGrad -> AdaDelta/RMSprop -> Adam -> AdaMax -> Nadam 这样的发展历程，本文简单来梳理这些优化算法是如何一步一步演变而来的。

优化算法框架：

计算目标函数关于当前参数的梯度：

$$g_t = \nabla f(w_t)$$

根据历史梯度计算一阶动量和二阶动量：

$$m_t = \phi(g_1, g_2, \dots, g_t)$$
$$V_t = \psi(g_1, g_2, \dots, g_t)$$

计算当前时刻的下降梯度：

$$\eta_t = \frac{\alpha}{\sqrt{V_t}} \cdot m_t$$

根据下降梯度进行更新：

$$w_{t+1} = w_t - \eta_t$$

最核心的区别就是第三步所执行的下降方向，在这个式子中，前半部分是实际的学习率（也即下降步长），后半部分是实际的下降方向。不同优化算法也就是不断地在这两部分上做文章。

最朴素的优化算法就是SGD了，没有动量和自适应学习率的概念，但还是有很多人在用着。

SGD

梯度更新规则:

SGD的形式最简单

$$\eta_t = \alpha \cdot g_t$$

存在问题:

因为更新比较频繁，会造成 cost function 有严重的震荡，最终停留在Local Minima或者Saddle Point处。

为了能够跳出Local Minima和Saddle Point，提出了动量的概念。

SGD with Momentum

梯度更新规则:

Momentum在梯度下降的过程中加入了惯性，使得梯度方向不变的维度上速度变快，梯度方向有所改变的维度上的更新速度变慢，这样就可以加快收敛并减小震荡。

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$$

一阶动量是移动平均值，这里 β_1 的经验值为0.9，也就是说时刻t的主要下降方向是由t-1时刻的下降方向再加一点点t时刻的偏向决定的。

存在问题:

不具备一些先知，例如快要上坡时，就知道需要减速了，适应性会更好；
不能根据参数的重要性而对不同的参数进行不同程度的更新。

SGD with Nesterov Acceleration

NAG改进了SGDM的第一个存在问题，在计算梯度时，不是在当前位置，而是未来的位置上。因此这里的梯度是跟着累积动量走了一步后的梯度，即

$$g_t = \nabla f(w_t - \frac{\alpha}{\sqrt{V_{t-1}}} \cdot m_{t-1})$$

我们希望能够根据参数的重要性而对不同的参数进行不同程度的更新，学习率是自适应的。对于经常更新的参数，我们已经积累了大量关于它的知识，不希望被单个样本影响太大，希望学习速率慢一些；对于偶尔更新的参数，我们了解的信息太少，希望能从每个偶然出现的样本身上多学一些，即学习速率大一些。

AdaGrad

梯度更新规则:

二阶动量为该维度上迄今为止所有梯度值的平方和

$$V_t = \sum_{\tau=1}^{t} g_\tau^2$$

为了避免分母为0，加了一项随机扰动

$$\eta_{ti} = \frac{\alpha}{\sqrt{V_{ti} + \varepsilon}} \cdot m_{ti}$$

存在问题:

分母会不断积累，这样学习率就会收缩并最终会变得非常小。

RMSProp

梯度更新规则:

解决Adagrad学习率急剧下降的问题，RMSProp改变了二阶动量计算方法，即用窗口滑动加权平均值计算二阶动量。

$$V_t = \beta_2 \cdot V_{t-1} + (1-\beta_2) g_t^2$$

Hinton 建议设定 β_2 为 0.9, 学习率 α 为 0.001。

集成动量+自适应学习率的优化算法应该是目前最好的。

Adam

梯度更新规则:

Adam = Adaptive + Momentum，顾名思义Adam集成了SGD的一阶动量和RMSProp的二阶动量。

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$$
$$V_t = \beta_2 \cdot V_{t-1} + (1-\beta_2) g_t^2$$

优化算法里最常见的两个超参数 β_1, β_2 就都在这里了，前者控制一阶动量，后者控制二阶动量。

若在Adam基础上再加一个Nesterov加速，是不是更牛逼了，这就是Nadam。

参考资料:

https://www.youtube.com/watch?v=h7iBpEHGVNc&index=3...
www.youtube.com

如何选择优化器 optimizer
www.jianshu.com

https://arxiv.org/pdf/1609.04747.pdf
link.jianshu.com

Juliuszh：Adam那么棒，为什么还对SGD念念不忘 (1) —— 一个框架看懂优...
zhuanlan.zhihu.com