

# 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



关注小夕并星标，解锁自然语言处理  
搜索、推荐与算法岗求职秘籍

文 | 苏剑林（追一科技，人称苏神）

美 | 人美心细小谨思密达

## 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

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

其中 $\beta_2$ 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

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

其中 $\beta_2$ 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  $\hat{m}_t = m_t / (1 - \beta_1^t)$ （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

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

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

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

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 $t \rightarrow \infty$ 时， $\hat{\beta}_{2,t} \rightarrow \beta_2$ ，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 $t \rightarrow \infty$ ，学习率最好恒定为常数（这时候相当于退化为SGD），这就要求 $t \rightarrow \infty$ 时， $\hat{\beta}_{2,t} \rightarrow 1$ 。

对于AdaX来说，当 $t = 0$ 时  $\hat{\beta}_{2,t} = 0$ ，当 $t \rightarrow \infty$ ， $\hat{\beta}_{2,t} \rightarrow 1$ ，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 $\hat{\beta}_{2,t} = 1 - \frac{1}{t^2}$ ，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。



### 萌屋公告

喜欢本文的小伙伴们，记得扫描下方二维码**关注并星标置顶**，我才能来到你面前哦。

卖萌屋妹子们的原创技术干货有 **ACL2020学术前沿系列**、**NLP综述系列**、**NLP论文清单系列**、**NLP基础入门系列**、**搜索与推荐系列**、**深度学习初/中/高级炼丹技巧**、**机器学习入门系列**、**算法岗offer收割系列**等。订阅号后台回复【**干货**】即可打包带走。

卖萌屋里有众多顶会审稿人、大厂研究员、知乎大V和美丽小姐姐（划掉）  
校招求职 高质量讨论群，订阅号后台回复【**入群**】即可上车。

自然语言处理知识图谱 / 深度学习 / 机器学习 /



### 夕小瑶的卖萌屋

关注&星标小夕，带你解锁AI秘籍

订阅号主页下方「撩一下」有惊喜哦



声明：pdf仅供学习使用，一切版权归原创公众号所有；建议持续关注原创公众号获取最新文章，学习愉快！