不同优化器的比较

> https://medium.com/%E9%9B%9E%E9%9B%9E%E8%88%87%E5%85%94%E5%85%94%E7%9A%84%E5%B7%A5%E7%A8%8B%E4%B8%96%E7%95%8C/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92ml-note-sgd-momentum-adagrad-adam-optimizer-f20568c968db

优化器决定权重更新的方式，调整权重值，让损失函数变小！

1. SGD-随机梯度下降法 (stochastic gradient decent)

$$W \leftarrow W - \eta \frac{\partial L}{\partial W}$$

SGD Weight update equation

W 為權重(weight)參數，**L** 為損失函數(loss function)， **η** 是學習率(learning rate)， **∂L/∂W** 是損失函數對參數的梯度(微分)

沿着梯度下降的方向更新权重值。

2. Momentum

此優化器為模擬物理動量的概念，在同方向的維度上學習速度會變快，方向改變的時候學習速度會變慢。

$$V_t \leftarrow \beta V_{t-1} - \eta \frac{\partial L}{\partial W}$$

$$W \leftarrow W + V_t$$

Momentum Weight update equation

這裡多了一個 **Vt** 的參數，可以將他想像成「方向速度」，會跟上一次的更新有關，如果上一次的梯度跟這次同方向的話，**|Vt|**(速度)會越來越大(代表梯度增強)，**W**參數的更新梯度便會越來越快，如果方向不同，**|Vt|**便會比上次更小(梯度減弱)，**W**參數的更新梯度便會變小， **β** 可以想像成空氣阻力或是地面摩擦力，通常設定成0.9

3. AdaGrad

對於Optimizer來說，learning rate(學習率) **η** 相當的重要，太小會花費太多時間學習，太大有可能會造成overfitting，無法正確學習，前面幾種Optimizer的學習率 **η**，都為固定值，而AdaGrad就是會依照梯度去調整 learning rate **η** 的優化器，Ada對我來說就是Adaptive的意思。

$$W \leftarrow W - \eta \frac{1}{\sqrt{n+\epsilon}} \frac{\partial L}{\partial W}$$

$$n = \sum_{r=1}^{t} (\frac{\partial L_r}{\partial W_r})^2$$

$$W \leftarrow W - \eta \frac{1}{\sqrt{\sum_{r=1}^{t} (\frac{\partial L_r}{\partial W_r})^2 + \epsilon}} \frac{\partial L}{\partial W}$$

AdaGrad Weight update equation

在AdaGrad Optimizer 中，**η** 乘上 **1/√(n+ϵ)** 再做參數更新，出現了一個**n**的參數，**n**為前面所有梯度值的平方和，利用前面學習的梯度值**平方和**來調整learning rate ，**ϵ** 為平滑值加上 **ϵ** 的原因是為了不讓分母為0，**ϵ** 一般值為1e-8

- 前期梯度較小的時候，n較小，能夠放大學習率
- 後期梯度較大的時候，n較大，能夠約束學習率，但分母上梯度平方的累加會越來越大，會使梯度趨近於0，訓練便會結束，為了防止這個情況，後面有開發出 `RMSprop Optimizer` ，主要就是把n變成RMS(均方根)。

4. Adam

Adam Optimizer 其實可以說就是把前面介紹的Momentum 跟 AdaGrad這二種Optimizer做結合，

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) \frac{\partial L_t}{\partial W_t}$$

$$v_t = \beta_1 v_{t-1} + (1-\beta_2) (\frac{\partial L_t}{\partial W_t})^2$$

像Momentum一樣保持了過去梯度的指數衰減平均值，像Adam一樣存了過去梯度的平方衰減平均值

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

對mt跟vt做偏離校正

$$W \leftarrow W - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Adam Weight update equation

Adam 保留了 Momentum 對過去梯度的方向做梯度速度調整與Adam對過去梯度的平方值做learning rate的調整，再加上Adam有做參數的”**偏離校正”**，使得每一次的學習率都會有個確定的範圍，會讓參數的更新較為平穩。

5. Pytorch 优化器应用

寻找方程的最优解