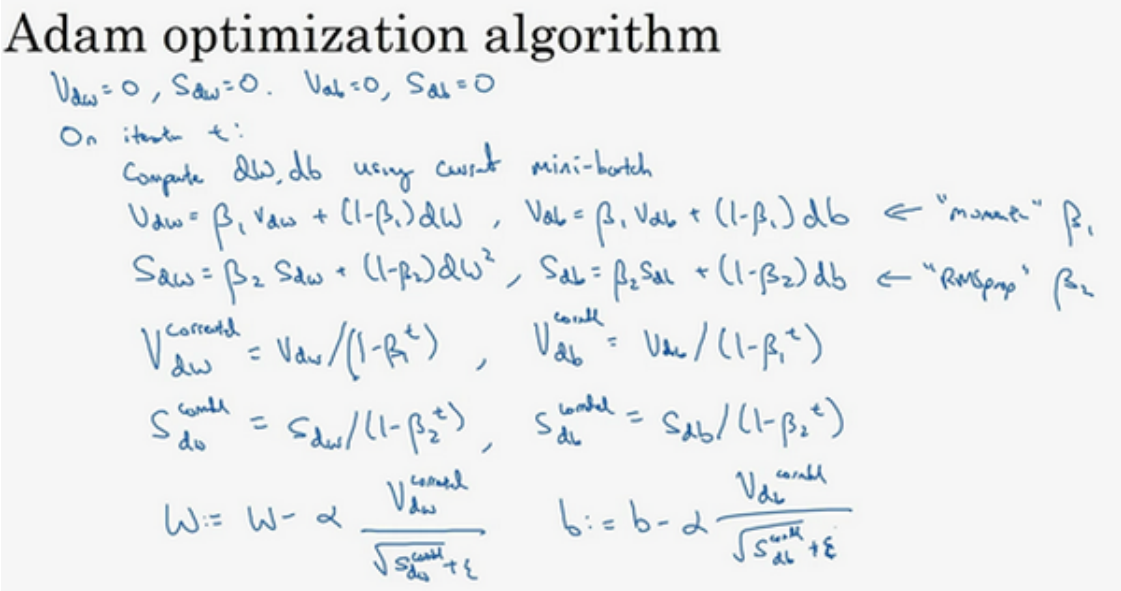
2.8 Adam 优化算法(Adam optimization algorithm)

在深度学习的历史上，包括许多知名研究者在内，提出了优化算法，并很好地解决了一些问题，但随后这些优化算法被指出并不能一般化，并不适用于多种神经网络，时间久了，深度学习圈子里的人开始多少有些质疑全新的优化算法，很多人都觉得动量（**Momentum**）梯度下降法很好用，很难再想出更好的优化算法。所以 **RMSprop** 以及 **Adam** 优化算法（**Adam** 优化算法也是本视频的内容），就是少有的经受住人们考验的两种算法，已被证明适用于不同的深度学习结构，这个算法我会毫不犹豫地推荐给你，因为很多人都试过，并且用它很好地解决了许多问题。

Adam 优化算法基本上就是将 **Momentum** 和 **RMSprop** 结合在一起，那么来看看如何使用 **Adam** 算法。



使用 **Adam** 算法，首先你要初始化，$v_{dW}=0$，$S_{dW}=0$，$v_{db}=0$，$S_{db}=0$，在第t次迭代中，你要计算微分，用当前的 **mini-batch** 计算dW，db，一般你会用 **mini-batch** 梯度下降法。接下来计算 **Momentum** 指数加权平均数，所以$v_{dW}=\beta_1 v_{dW}+(1-\beta_1)dW$（使用$\beta_1$，这样就不会跟超参数$\beta_2$混淆，因为后面 **RMSprop** 要用到$\beta_2$），使用 **Momentum** 时我们肯定会用这个公式，但现在不叫它β，而叫它β_1。同样$v_{db}=\beta_1 v_{db}+(1-\beta_1)db$。

接着你用 **RMSprop** 进行更新，即用不同的超参数β_2，$S_{dW}=\beta_2 S_{dW}+(1-\beta_2)(dW)^2$，再说一次，这里是对整个微分$dW$进行平方处理，$S_{db}=\beta_2 S_{db}+(1-\beta_2)(db)^2$。

相当于 **Momentum** 更新了超参数β_1，**RMSprop** 更新了超参数β_2。一般使用 **Adam** 算法

的时候，要计算偏差修正，$v_{dW}^{\text{corrected}}$，修正也就是在偏差修正之后，

$$v_{dW}^{\text{corrected}} = \frac{v_{dW}}{1-\beta_1^t},$$

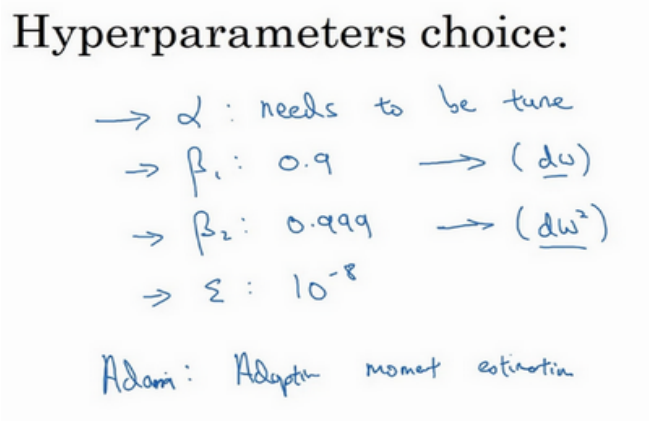
同样$v_{db}^{\text{corrected}} = \frac{v_{db}}{1-\beta_1^t}$，

S也使用偏差修正，也就是$S_{dW}^{\text{corrected}} = \frac{S_{dW}}{1-\beta_2^t}$，$S_{db}^{\text{corrected}} = \frac{S_{db}}{1-\beta_2^t}$。

最后更新权重，所以W更新后是$W := W - \frac{\alpha v_{dW}^{\text{corrected}}}{\sqrt{S_{dW}^{\text{corrected}}}+\varepsilon}$（如果你只是用 **Momentum**，使用$v_{dW}$或者修正后的$v_{dW}$，但现在我们加入了 **RMSprop** 的部分，所以我们要除以修正后S_{dW}的平方根加上ε）。

根据类似的公式更新b值，$b := b - \frac{\alpha v_{\text{db}}^{\text{corrected}}}{\sqrt{S_{\text{db}}^{\text{corrected}}}+\varepsilon}$。

所以 **Adam** 算法结合了 **Momentum** 和 **RMSprop** 梯度下降法，并且是一种极其常用的学习算法，被证明能有效适用于不同神经网络，适用于广泛的结构。



本算法中有很多超参数，超参数学习率α很重要，也经常需要调试，你可以尝试一系列值，然后看哪个有效。β_1常用的缺省值为 0.9，这是 dW 的移动平均数，也就是dW的加权平均数，这是 **Momentum** 涉及的项。至于超参数β_2，**Adam** 论文作者，也就是 **Adam** 算法的发明者，推荐使用 0.999，这是在计算$(dW)^2$以及$(db)^2$的移动加权平均值，关于ε的选择其实没那么重要，**Adam** 论文的作者建议ε为10^{-8}，但你并不需要设置它，因为它并不会影响算法表现。但是在使用 **Adam** 的时候，人们往往使用缺省值即可，β_1，β_2和ε都是如此，我觉得没人会去调整ε，然后尝试不同的α值，看看哪个效果最好。你也可以调整β_1和β_2，但我认识的业内人士很少这么干。

为什么这个算法叫做 **Adam**？**Adam** 代表的是 **Adaptive Moment Estimation**，β_1用于计算

这个微分（dW），叫做第一矩，β_2用来计算平方数的指数加权平均数（$(dW)^2$），叫做第二矩，所以 **Adam** 的名字由此而来，但是大家都简称 **Adam** 权威算法。