Adam(Adaptive Moment Estimation)本质上是带有动量项的RMSprop，它利用梯度的一阶矩估计和二阶矩估计动态调整每个参数的学习率。Adam的优点主要在于经过偏置校正后，每一次迭代学习率都有个确定范围，使得参数比较平稳。公式如下：

$$m_t = \mu * m_{t-1} + (1 - \mu) * g_t$$

$$n_t = \nu * n_{t-1} + (1 - \nu) * g_t^2$$

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

$$\hat{n}_t = \frac{n_t}{1 - \nu^t}$$

$$\Delta\theta_t = -\frac{\hat{m}_t}{\sqrt{\hat{n}_t} + \epsilon} * \eta$$

其中，

m_t

，

n_t

分别是对梯度的一阶矩估计和二阶矩估计，可以看作对期望

$E|g_t|$

，

$E|g_t^2|$

的估计；

$\hat{m}_t$

，

$\hat{n}_t$

是对

m_t

，

n_t

的校正，这样可以近似为对期望的无偏估计。 可以看出，直接对梯度的矩估计对内存没有额外的要求，而且可以根据梯度进行动态调整，而

$$-\frac{\hat{m}_t}{\sqrt{\hat{n}_t}+\epsilon}$$

对学习率形成一个动态约束，而且有明确的范围。

特点：

- 结合了Adagrad善于处理稀疏梯度和RMSprop善于处理非平稳目标的优点
- 对内存需求较小
- 为不同的参数计算不同的自适应学习率
- 也适用于大多非凸优化 - 适用于大数据集和高维空间