Adam Optimization Algorithm:

Adam 作为一个泛用性极强的优化算法，集 Gradient Descent with Momentum 和 RMSprop 的优点为一体，其基本实现如下所示，分别由 momentum 和 RMSprop 两个 portions 组成：

Adam: Adaptive momentum estimation

On interaction t, compute $d\omega, db$ on current mini batch, the learning rate is α:

$$\text{Initial with } V_{d\omega} = V_{db} = 0 \text{ and } S_{d\omega} = S_{db} = 0$$

$$\textbf{Momentum Portion}: V_{d\omega} = \beta_1 V_{d\omega} + (1-\beta_1)d\omega, \qquad V_{\alpha b} = \beta_1 V_{db} + (1-\beta_1)db$$

$$\textbf{RMSprop Portion}: S_{d\omega} = \beta_2 S_{d\omega} + (1-\beta_2)(d\omega)^2, \qquad S_{db} = \beta_2 S_{dl} + (1-\beta_2)(db)^2$$

Adding **bias correction, introduced in** (https://github.com/GuoJiaqi-1020/Jacky-s-ML-notebook/blob/main/pdf%20note/Exponentially%20Weighted%20Averages.pdf) :

$$V_{d\omega}^{\text{corrected}} = \frac{V_{d\omega}}{(1-\beta_1^t)}, \qquad V_{db}^{\text{corrected}} = \frac{V_{db}}{(1-\beta_1^t)}$$

$$S_{d\omega}^{\text{corrected}} = \frac{S_{d\omega}}{(1-\beta_2^t)}, \qquad S_{db}^{\text{corrected}} = \frac{S_{db}}{(1-\beta_2^t)}$$

Then, the position(parameters) will be updated using the above two equations:

$$\omega := \omega - \alpha \frac{V_{d\omega}^{\text{corrected}}}{\sqrt{S_{d\omega}^{\text{corrected}} + \epsilon}} \qquad b := b - \alpha \frac{V_{db}^{\text{corrected}}}{\sqrt{S_{db}^{\text{corrected}} + \epsilon}}$$

Usually, $\beta_1 = 0.9 \rightarrow (d\omega)$, $\beta_2 = 0.999 \rightarrow (d\omega^2)$ which are the default values. β_1 is for computing the mean of the derivatives, called the first momentum, and β_2 is used to compute exponentially weighted average of the squares (in RMSprop), called the second momentum.