Adam Optimization Algorithm

Adam (short for Adaptive Moment Estimation) is an optimization algorithm that computes adaptive learning rates for each parameter. It combines the advantages of two other extensions of stochastic gradient descent: AdaGrad and RMSProp.

The update rules for the Adam optimizer are as follows:

1. Compute the gradients:

$$g_t = \nabla_\theta J(\theta_t)$$

where g_t is the gradient of the loss function J with respect to the parameters θ at time step t.

2. Update the moment estimates:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

where m_t is the first moment (mean of gradients) and v_t is the second moment (uncentered variance of gradients). The parameters β_1 and β_2 are the decay rates for these moment estimates, typically set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

3. Correct the bias in the moment estimates:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected moment estimates.

4. Update the parameters:

$$\theta_{t+1} = \theta_t - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where α is the learning rate, and ϵ is a small constant (e.g., 10^{-8}) to prevent division by zero.

Adam adapts the learning rate for each parameter individually, making it robust to noisy gradients and sparse gradients. This results in faster convergence and better performance in many cases.

Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 10^{-3} or 5×10^{-4} is a great starting point for many models.