Adam

- Adaptive Moment Estimation (Adam) Optimizer
- The Adam optimizer is developed by Diederik P. Kingma and Jimmy Ba in 2014.
- The Adam optimizer, short for "Adaptive Moment Estimation," is an iterative optimization algorithm used to minimize the loss function during the training of neural networks.
- Adaptive Moment Estimation is an algorithm for optimization technique for gradient descent
- It is really efficient when working with large problem involving a lot of data or parameters
- It requires less memory and is efficient
- Adam has become a go-to choice for many machine learning practitioners.
- Adam can be looked at as a combination of RMSprop and SGD with momentum
- Here, we control the rate of gradient descent in such a way that there is minimum oscillation when it reaches the global minimum while taking big enough steps (step-size) so as to pass the local minima hurdles along the way.
- Hence, combining the features of the above methods to reach the global minimum efficiently

Adaptive Learning Rates:

- Adam adjusts the learning rates for each parameter individually.
- It calculates a moving average of the first-order moments (the mean of gradients) and the second-order moments (the uncentered variance of gradients) to scale the learning rates adaptively.
- This makes it well-suited for problems with sparse gradients or noisy data.

Bias Correction:

- To counteract the initialization bias in the first moments, Adam applies bias correction during the early iterations of training.
- This ensures faster convergence and stabilizes the training process.

Low Memory Requirements:

- Unlike some optimization algorithms that require storing a history of gradients for each parameter, Adam only needs to maintain two moving averages per parameter.

- This makes it memory-efficient, especially for large neural networks.

Learning Rate:

While Adam adapts the learning rates, choosing a reasonable initial learning rate is still essential. It often performs well with the default value of 0.001.

Epsilon Value:

The epsilon (ε) value is a small constant added for numerical stability.

Steps Involved in the Adam Optimization Algorithm:

- Step 1. Initialize the first and second moments' moving averages (m and v) to zero.

  $$m_0 = 0, v_0 = 0$$

- Step 2. Compute the gradient of the loss function to the model parameters.
- Step 3. Update the moving averages using exponentially decaying averages.

  This involves calculating m_t and v_t as weighted averages of the previous moments and the current gradient.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) \nabla w_t$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla w_t)^2$$

$\beta_1 = 0.9$ $\beta_2 = 0.99$

- Step 4. Apply bias correction to the moving averages, particularly during the early iterations.

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

- Step 5. Calculate the parameter update by dividing the bias-corrected first moment by the square root of the bias-corrected second moment, with an added small constant (epsilon) for numerical stability.

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} * \hat{m}_t$$

$$\eta = 0.001$$

- Step 6. Update the model parameters using the calculated updates.
- Step 7. Repeat steps 2-6 for a specified number of iterations or until convergence.