Adam Optimizer

Adaptive Moment Estimation

It is currently the most powerfull Optimization technique

Adam Optimizer is Most Famous & almost most used Optimizer in the World

Adam is combination of Learning Decay & Momentum,

Adam, which stands for Adaptive Moment Estimation, is an optimization algorithm widely used in the training of deep neural networks. It combines ideas from both RMSprop and momentum to provide an adaptive learning rate for each parameter, along with momentum-like terms for handling different scales and accelerating convergence. Adam has been successful in practice and is considered one of the state-of-the-art optimization algorithms for deep learning.

he Adam optimizer is designed to provide an adaptive and efficient method for updating the parameters during the training of neural networks. Here's an intuitive understanding of how Adam works:

1. **Adaptive Learning Rates:**
   - One key feature of Adam is its adaptive learning rates. Instead of using a fixed learning rate for all parameters, Adam adjusts the learning rate for each parameter individually. This is achieved by maintaining separate moving averages for the first and second moments of the gradients.
2. **Exponential Moving Averages:**
   - Adam maintains two exponentially decaying moving averages. The first moment ⬚m_t is an exponentially weighted average of past gradients, similar to the momentum term in other optimization algorithms. The second moment ⬚v_t is an exponentially weighted average of the squared gradients, similar to the moving average used in RMSprop.
3. **Bias Correction:**
   - The moving averages ⬚m_t and ⬚v_t are biased towards zero, especially during the early steps of optimization. Adam corrects these biases by dividing ⬚m_t and ⬚v_t by $1-⬚1⬚1-\beta_{1t}$ and $1-⬚2⬚1-\beta_{2t}$, respectively, where ⬚t is the current time step.
4. **Parameter Update:**

- The parameters are updated based on the adaptive learning rates and the bias-corrected estimates of the first and second moments. The update rule for each parameter ⬚ is: ⬚+1=⬚−⬚^+⋅^ where is the learning rate, ^ and ^ are the bias-corrected estimates, and is a small constant added for numerical stability.

5. **Benefits:**
   - **Adaptability:** Adam adapts the learning rates based on the characteristics of each parameter. Parameters with sparse gradients or infrequent updates receive larger effective learning rates, while parameters with frequent updates receive smaller effective learning rates.
   - **Momentum-Like Term:** The inclusion of the ⬚ term provides momentum-like effects, helping to accelerate convergence, especially in the presence of noisy gradients or flat regions in the loss landscape.
   - **Numerical Stability:** The small constant is added to the denominator to prevent division by zero and enhance numerical stability.
6. **Hyperparameters:**
   - The hyperparameters 1, 2, and need to be chosen based on the specific task and dataset. Common default values are 1=0.9, 2=0.999, and =1−8.

In summary, Adam combines the advantages of adaptive learning rates, momentum, and bias correction to provide an effective and widely used optimization algorithm for training neural networks. Its adaptability to different scales and robust performance on various tasks make it a popular choice in practice.

Verdict:

Last 4-5 Years mai Adam is most used on every different dataset,

If Adam doesn’t satisfy the needs Try RMSProp once,