復旦大學

FUDAN UNIVERSITY

ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION

Tan Wang

2021-11-01

About this paper

- The name Adam is derived from adaptive moment
- Diederik P. Kingma: Senior Research Scientist at Google Brain, Contributions: the Variational Auto-Encoder (VAE), the Adam method
- Jimmy Ba: Assistant Professor in University of Toronto, CIFAR AI chair, completed PhD under the supervision of Geoffrey Hinton, Published as a conference paper at ICLR 2015
- The Adam optimization paper is the world's #1 most cited scientific paper of the past five years (2015-2019)
- Google Scholar citations: 88891

Problem

- The focus of this paper is on the optimization of stochastic objectives with high-dimensional parameters
- Higher-order optimization methods are ill-suited, and discussion is restricted to first-order methods (memory constriction of GPU)
- Efficient stochastic optimization techniques are required for a noisy objective
- An extension to Stochastic Gradient Decent

$$\min E\left[f(\theta)\right]$$

Algorithm

Momentum

(Polyak, 1964)

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

$$\theta_t = \theta_{t-1} - \alpha m_t$$

AdaGrad

(Duchi et al., 2011)

$$v_t = v_{t-1} + g_t^2$$

$$\theta_t = \theta_{t-1} - \alpha \frac{g_t}{\sqrt{v_t} + \varepsilon}$$

RMSProp

(Tieleman & Hinton, 2012)

$$v_t = \beta v_{t-1} + (1-\beta) g_t^2$$

$$\theta_t = \theta_{t-1} - \alpha \frac{g_t}{\sqrt{v_t} + \varepsilon}$$

RMSProp + Momentum

(Graves, 2013)

$$v_t = \beta v_{t-1} + (1-\beta) g_t^2$$

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

$$\Delta\theta_t = \gamma \Delta\theta_{t-1} - \alpha \frac{g_t}{\sqrt{v_t - m_t^2} + \varepsilon}$$

$$\theta_t = \theta_{t-1} + \Delta\theta_t$$

Algorithm

- Adam ≈ *RMSProp* + *Momentum*

$$m_t = \beta_1 m_{t-1} + \left(1 - \beta_1\right) g_t$$

$$v_t = \beta_2 v_{t-1} + \left(1 - \beta_2\right) g_t^2$$

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

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

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

Adam’s Update Rule

- Assuming ε = 0, the effective step taken in parameter space at time step *t is* $\Delta_t = a \cdot \hat{m}_t / \sqrt{\hat{v}_t}$
- Two Upper Bounds $\left|\Delta_t\right| \le a \cdot (1-\beta_1)/\sqrt{1-\beta_2}\ \ when\ \ (1-\beta_1) > \sqrt{1-\beta_2}$

$$\left|\Delta_t\right| \le a\ \ otherwise$$

- Proof

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

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

$$\Rightarrow \hat{m}_t = \frac{1-\beta_1}{1-\beta_1^t} \sum_{j=1}^{t} \beta_1^{t-j} g_j$$

$$\hat{v}_t = \frac{1-\beta_2}{1-\beta_2^t} \sum_{j=1}^{t} \beta_2^{t-j} g_j^2$$

$$\beta_i \in [0,1)$$

$$(1-\beta_1) > \sqrt{1-\beta_2}$$

$$\Rightarrow \beta_2 > 2\beta_1 - \beta_1^2 \Rightarrow \beta_1 < \beta_2$$

Adam’s Update Rule

- This can be understood as establishing a trust region around the current parameter value, beyond which the current gradient estimate does not provide sufficient information

- This typically makes it relatively easy to know the right scale of a in advance.

- The signal-to-noise ratio (SNR): $\hat{m}_t / \sqrt{\hat{v}_t}$

- A smaller SNR means that there is greater uncertainty about whether the direction of $\overset{?}{m}_t$ corresponds to the direction of the true gradient

- The effective stepsize is also invariant to the scale of the gradients $(c \cdot \hat{m}_t) / \sqrt{c^2 \cdot \hat{v}_t} = \hat{m}_t / \sqrt{\hat{v}_t}$

Initialization Bias Correction

- Here derive the term for the second moment estimate; the derivation for the first moment estimate is completely analogous

$$v_t = \beta_2 v_{t-1} + \left(1-\beta_2\right) g_t^2,\ v_0 = 0$$

$$v_t = \left(1-\beta_2\right) \sum_{j=1}^{t} \beta_2^{t-j} g_j^2$$

$$E\left[v_t\right] = E\left[\left(1-\beta_2\right) \sum_{j=1}^{t} \beta_2^{t-j} \cdot g_j^2\right]$$

$$E\left[g_t^2\right] \cdot \left(1-\beta_2\right) \sum_{j=1}^{t} \beta_2^{t-j} + \xi$$

$$= E\left[g_t^2\right] \cdot \left(1-\beta_2^t\right) + \xi$$

- If the true second moment is stationary, $\xi = 0$

Initialization Bias Correction

- The term $(1 - \beta_2^t)$ is caused by initializing the running average with zeros
- We therefore divide by this term to correct the initialization bias

Convergence Analysis

- Analyze the convergence of Adam using the online learning framework proposed in (Zinkevich, 2003).
- Given an arbitrary, unknown sequence of convex cost functions $f_1(\theta), f_2(\theta), \ldots, f_T(\theta)$
- At each time t, our goal is to predict the parameter θ_t and evaluate it on a previously unknown cost

function f_t.

- Since the nature of the sequence is unknown in advance, we evaluate our algorithm

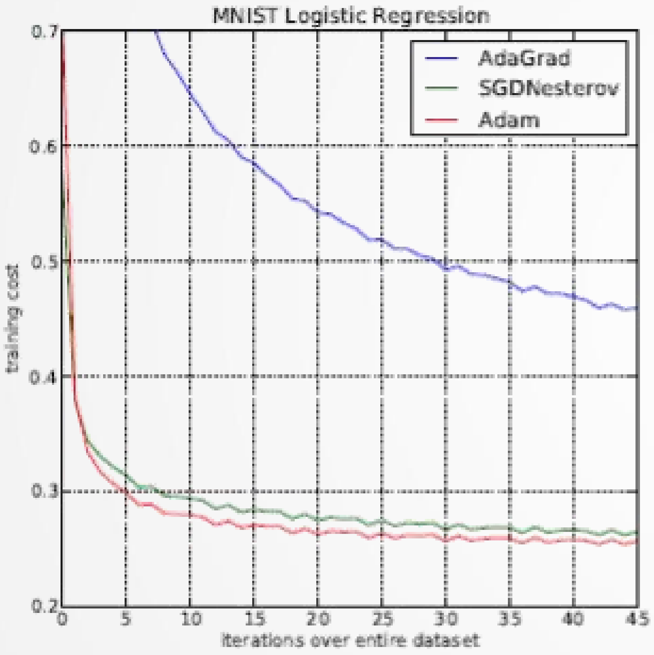
using the regret

$$R(T) = \sum_{t=1}^{T} \left[f_t(\theta_t) - f_t(\theta^*) \right]$$

$$\theta^* = \arg\min_{\theta \in X} \sum_{t=1}^{T} f_t(\theta)$$

- We show Adam has $O(\sqrt{T})$ regret bound

Experiment



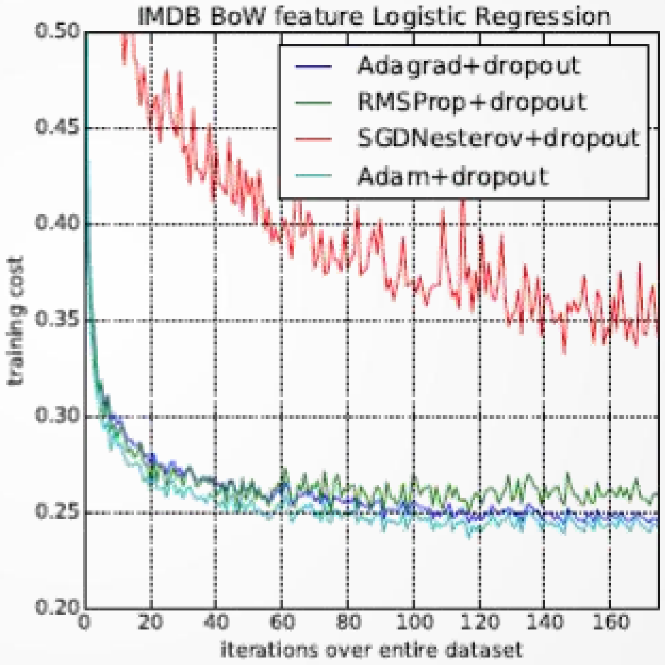
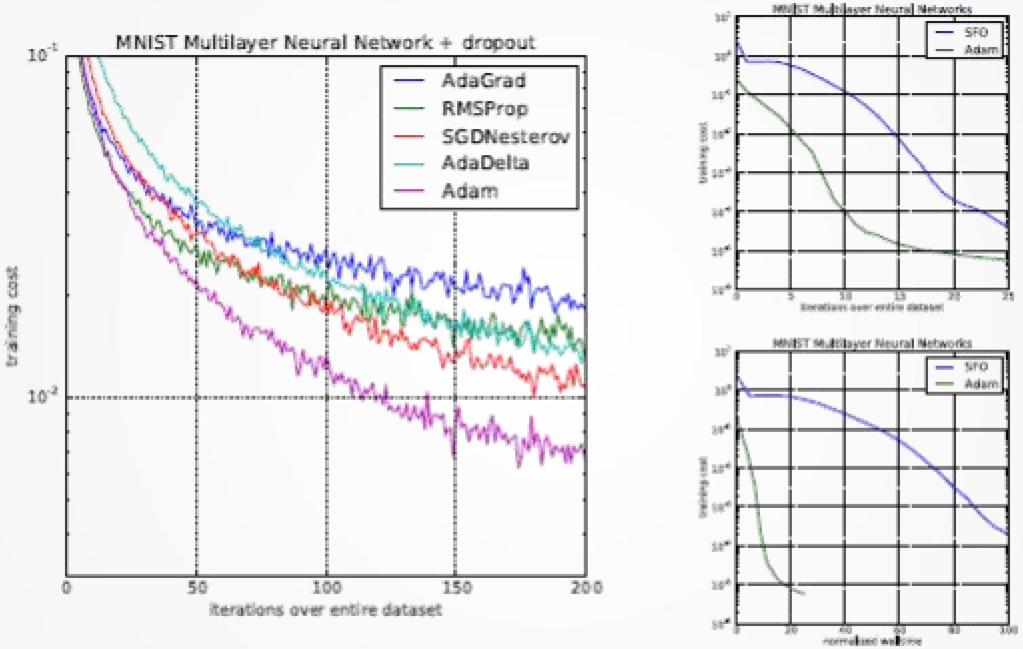


Figure 1: Logistic regression training negative log likelihood on MNIST images and IMDB movie reviews with 10,000 bag-of-words (BoW) feature vectors.

Experiment

- Non-convex objective functions.



(a)

(b)

Figure 2: Training of multilayer neural networks on MNIST images. (a) Neural networks using dropout stochastic regularization. (b) Neural networks with deterministic cost function. We compare with the sum-of-functions (SFO) optimizer (Sohl-Dickstein et al., 2014)

Experiment

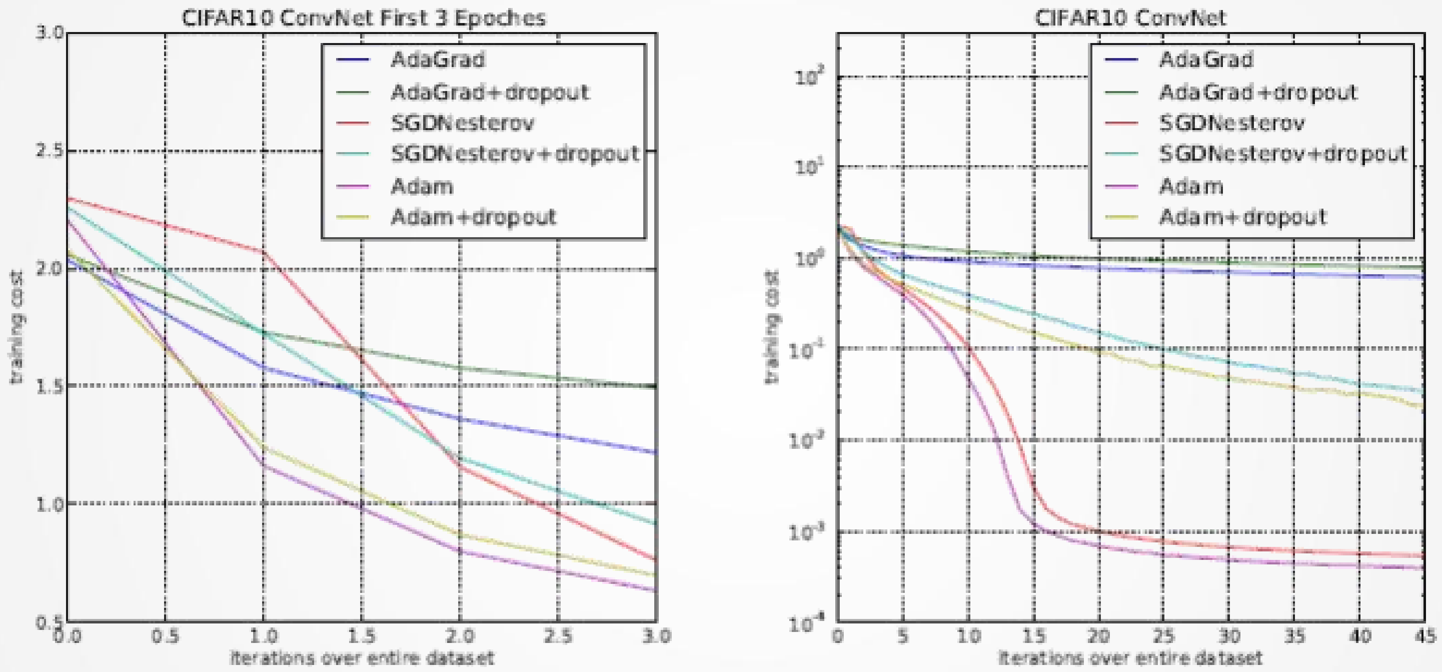


Figure 3: Convolutional neural networks training cost. (left) Training cost for the first three epochs. (right) Training cost over 45 epochs. CIFAR-10 with c64-c64-c128-1000 architecture.

Extensions

- AdaMax

- We can generalize the L2 norm based update rule to a Lp norm based update rule

$$\begin{aligned} v_t &= \beta_2^p v_{t-1} + (1-\beta_2^p)|g_t|^p \\ &= (1-\beta_2^p)\sum_{i=1}^{t} \beta_2^{p(t-i)} \cdot |g_i|^p \end{aligned}$$

$$\begin{aligned} u_t = \lim_{p\to\infty} (v_t)^{1/p} &= \lim_{p\to\infty} \left((1-\beta_2^p)\sum_{i=1}^{t} \beta_2^{p(t-i)} \cdot |g_i|^p \right)^{1/p} \\ &= \lim_{p\to\infty} (1-\beta_2^p)^{1/p} \left(\sum_{i=1}^{t} \beta_2^{p(t-i)} \cdot |g_i|^p \right)^{1/p} \\ &= \lim_{p\to\infty} \left(\sum_{i=1}^{t} \left(\beta_2^{(t-i)} \cdot |g_i| \right)^p \right)^{1/p} \\ &= \max\left(\beta_2^{t-1}|g_1|, \beta_2^{t-2}|g_2|, \ldots, \beta_2|g_{t-1}|, |g_t|\right) \end{aligned}$$

$$u_t = \max(\beta_2 \cdot u_{t-1}, |g_t|)$$

Extensions

$$m_t = \beta_1 m_{t-1} + \left(1 - \beta_1\right) g_t$$

$$u_t = \max\left(\beta_2 \cdot u_{t-1}, \left|g_t\right|\right)$$

$$\theta_t = \theta_{t-1} - \frac{a}{1 - \beta_1^t} \cdot \frac{m_t}{u_t}$$

- The magnitude of parameter updates has a simpler bound $\left|\Delta_t\right| \leq a$

Conclusion

- Adam is aimed towards machine learning problems with large datasets and/or high dimensional parameter spaces
- The method is straightforward and requires little memory.
- Adam is well-suited to a wide range of non-convex optimization problems.
- Easy to know the right scale of α in advance
- Provide bound for general convex online learning problem

復旦大學

FUDAN UNIVERSITY

Thanks！