Survey: Learning Strategies for Self-Attention

Jie Hao

March 20, 2019

Outline

1. Overview for Optimization Methods
   - $SGD \rightarrow SGDM \rightarrow NAG \rightarrow AdaGrad \rightarrow AdaDelta \rightarrow Adam$
2. Learning Rate Decay
3. Exploratory Experiments

Framework

Objective function: $f(\theta)$, **Parameters**: θ, **Initial learning rate**: α

1. Compute the gradient for current parameters: $g_t = \nabla f(\theta_t)$.

2. Based on history gradients to compute the first-order momentum and second-order momentum: $m_t = \phi(g_1, ..., g_t)$, $V_t = \psi(g_1, ..., g_t)$.

3. Compute current descent gradient: $\eta_t = \alpha \cdot m_t / V_t$

4. Update parameters: $\theta_{t+1} = \theta_t - \eta_t$.

SGD: $m_t = g_t, V_t = I^2 \to \eta_t = \alpha \cdot g_t$

SGD with Momentum(SGDM): $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$

SGD with Nesterov Acceleration(NAG):
$g_t = \nabla f(\theta_t - \alpha \cdot m_{t-1}/\sqrt{V_{t-1}})$

AdaGrad: $V_t = \Sigma_{\tau=1}^{t} g_\tau^2$

AdaDelta: $V_t = \beta_2 \cdot V_{t-1} + (1 - \beta_2) g_t^2$.

Adam: $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t, V_t = \beta_2 \cdot V_{t-1} + (1 - \beta_2) g_t^2$.

Overview

Methods	First-order Momentum	Second-order Momentum
SGD	$\times$	$\times$
SGDM	$\surd$	$\times$
NAG	$\surd$	$\times$
AdaGrad	$\times$	$\surd$
AdaDelta	$\times$	$\surd$
Adam	$\surd$	$\surd$

Learning Rate Decay

Why decay the learning rate?

- To reach the minimum of strongly convex function, we should decay the learning rate to slow down the learning process (Robbins et al., 1951; Smith et al., 2018).
- The scale of random fluctuations in the SGD dynamics: $g = \epsilon(N/B - 1)$, where ϵ is learning rate, N is training set size, B is batch size. (Smith&Le et al., 2017)

Learning Rate Decay Strategies

Methods	**Fomula**
Discrete staircase	e.g. halve the lr after fixed steps
Exponential	e.g. $lr = \alpha_0 * 0.95^{epochs}$
Natural exponential	e.g. $lr = \alpha_0 * e^{epochs}$
Inverse square root	e.g. $lr = \alpha_0 * k/\sqrt{epochs}$
Cosine	e.g. $lr = 0.5 * \alpha_0 * (1 + cos(\pi * epochs))$

Learning Strategies for Self-Attention on Different Tasks

SAN on NMT (Vaswani et al., 2017)

Adam. $lr = d_{model}^{-0.5} * min(step_num^{-0.5}, step_num * warmup\_steps^{-1.5})$

SAN on SRL (Tan et al., 2018)

Adadelta. $lr = 1.0$, halve the learning rate every 100K steps.

SAN on SNLI (Shen et al., 2018)

Adadelta. $lr = 0.5$, exponential decay.

BERT (Devlin et al., 2019)

Adam. $lr = 0.0001$, Warm-up 10000 steps, linear decay.

Exploratory Experiments

Transfomer on WMT14 EN-De, training step=10000.

Learning Strategies	Dev Bleu	Decay	Warm-up
Baseline	22.38	Y	4000
Baseline	0.35	Y	N
lr=0.001	19.50	N	N
lr=0.0001	18.38	N	N
lr=0.0005	22.69	N	N

References

[1] Herbert Robbins, Sutton Monro. 1951. A Stochastic Approximation Method. The Annals of Mathematical Statistics.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, Quoc V. Le. 2018. Don't Decay the Learning Rate, Increase the Batch Size. In ICLR.

[3] Samuel L. Smith, Quoc V. Le. 2018. A Bayesian Perspective on Generalization and Stochastic Gradient Descent. In ICLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N Gomez, ukaszKaiser, and Illia Polosukhin. 2017. Attention is allyou need. In NIPS.

[5] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang,Shirui Pan, and Chengqi Zhang. 2018a. DiSAN: di-rectional self-attention network for RNN/CNN-freelanguage understanding. In AAAI.

[6] Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, Xiaodong Shi. 2018. Deep Semantic Role Labeling with Self-Attention. In AAAI.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL.

Thank you!