AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients

임희주

Introduction

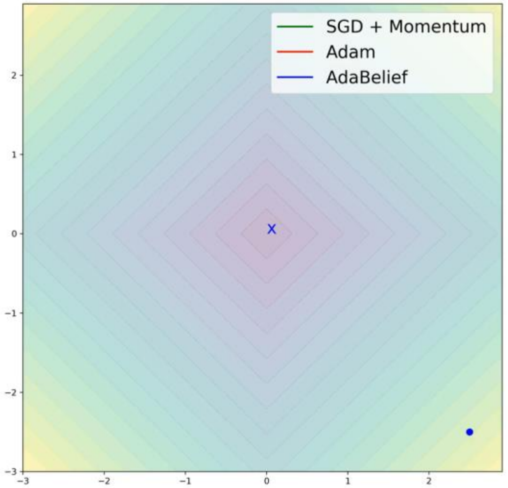
1. Introduction
AdaBelief

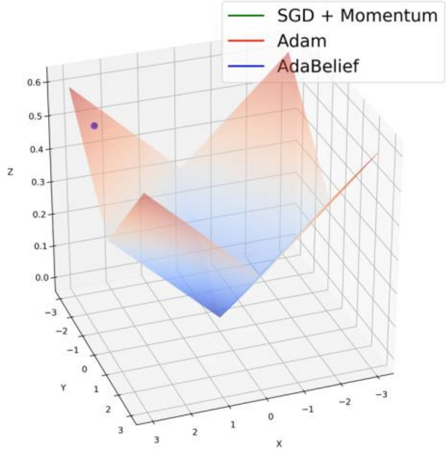
fast as Adam, generalizes as good
as SGD, and sufficiently stable to train GANs.

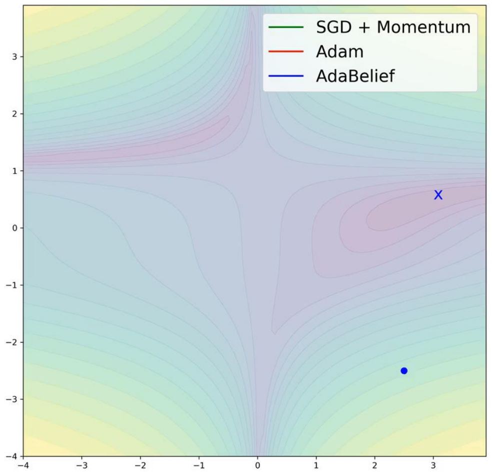
$\Rightarrow$ **AdaBelief optimizer**

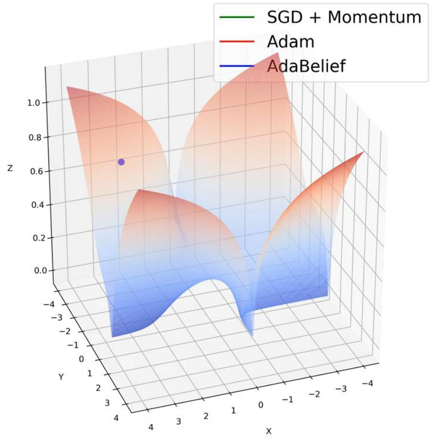
$\Rightarrow$ Adam 에서 한 줄만 변경했는데 성능향상

1. Introduction
AdaBelief

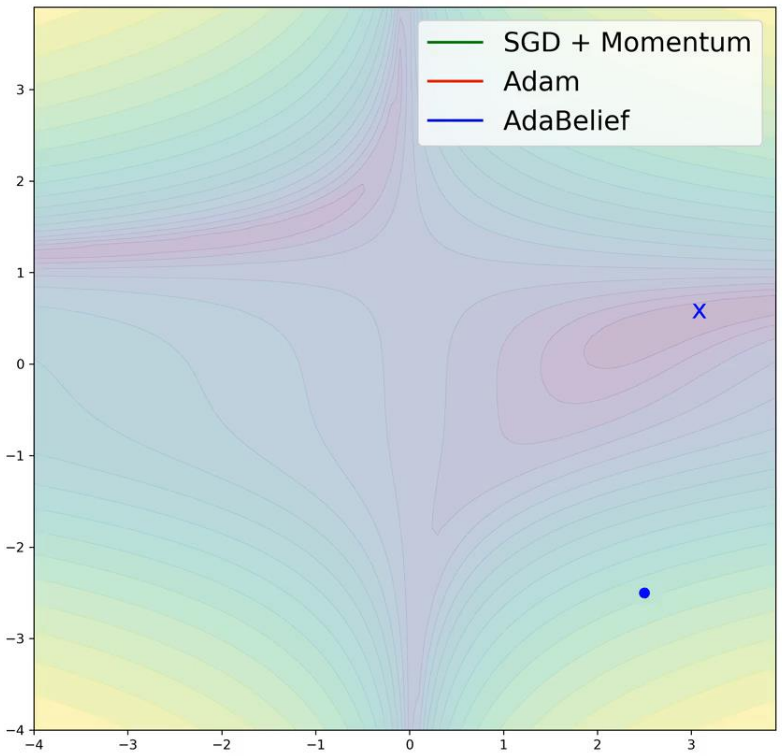


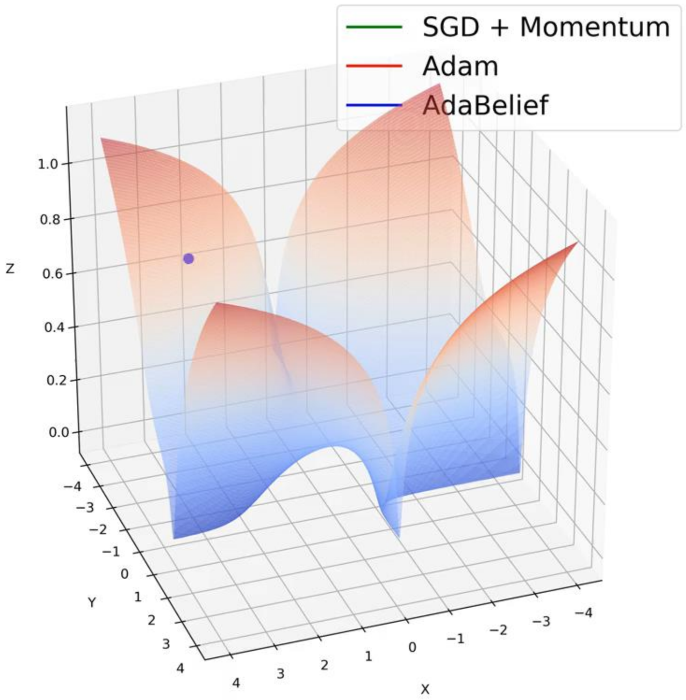






1. Introduction
AdaBelief





1. Introduction
Adam

Algorithm 1: Adam Optimizer

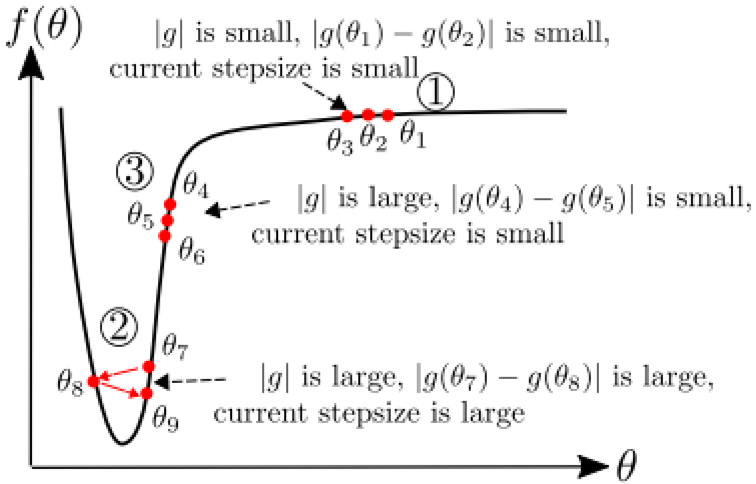
Initialize $\theta_0, m_0 \leftarrow 0$, $v_0 \leftarrow 0, t \leftarrow 0$
While θ_t not converged
$t \leftarrow t+1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
Bias Correction
$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}, \widehat{v_t} \leftarrow \frac{v_t}{1-\beta_2^t}$
Update
$\theta_t \leftarrow \prod_{\mathcal{F},\sqrt{\widehat{v_t}}} \left(\theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}\right)$

- g_t: the gradient and step t
- m_t: exponential moving average (EMA) of g_t
- v_t, s_t: v_t is the EMA of g_t^2, s_t is the EMA of $(g_t - m_t)^2$
- α, ϵ: α is the learning rate, default is 10^{-3}; ϵ is a small number, typically set as 10^{-8}
- β_1, β_2: smoothing parameters, typical values are $\beta_1 = 0.9, \beta_2 = 0.999$
- β_{1t}, β_{2t} are the momentum for m_t and v_t respectively at step t, and typically set as constant (e.g. $\beta_{1t} = \beta_1, \beta_{2t} = \beta_2, \forall t \in \{1, 2, ...T\}$

$Bias\ Correction$: 학습 초기 가중치들이 0으로 편향되는 것 방지.
m_t : 이전 $gradient$ 들의 1차 $moment$ 에 대한 추정
v_t : 이전 $gradient$ 들의 2차 $moment$ 에 대한 추정

Method

2. Method
Problem - Adam



In Case 3
Ideal optimizer는 큰 step size를 가져야 하지만
Adam 은 오히려 작은 step size를 가짐

$$\theta_t \leftarrow \prod_{\mathcal{F},\sqrt{\hat{v}_t}} \left(\theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\hat{v}_t}+\epsilon}\right) \qquad v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

Adam 의 step size 는 v_t 에 영향을 받음
이 v_t 는 g_t^2 에 따라 변하기 때문에 g_t가 커지는
Case 3 구간에서 step size가 작음

2. Method

Problem - solution

Algorithm 2: AdaBelief Optimizer

Initialize θ_0, $m_0 \leftarrow 0$, $s_0 \leftarrow 0$, $t \leftarrow 0$

While θ_t not converged

$t \leftarrow t+1$

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$

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

$s_t \leftarrow \beta_2 s_{t-1} + (1-\beta_2)(g_t - m_t)^2 + \epsilon$

Bias Correction

$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}$, $\widehat{s_t} \leftarrow \frac{s_t}{1-\beta_2^t}$

Update

$\theta_t \leftarrow \prod_{\mathcal{F},\sqrt{\widehat{s_t}}} \left(\theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\widehat{s_t}}+\epsilon}\right)$

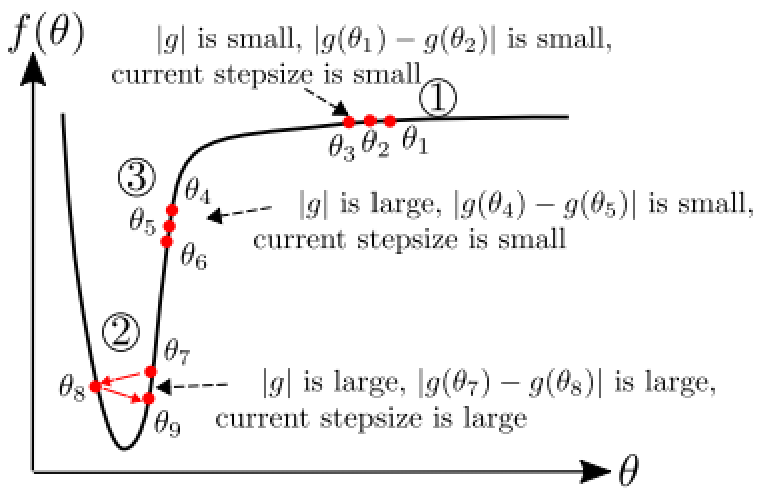
$g_t^2 \Rightarrow (g_t^2 - m_t^2)^2 + \varepsilon$

v_t를 s_t 로 바꿈으로써

Case 3 에서 Ideal 한 step size (large)

$$s_t = EMA\left((g_0 - m_0)^2, \ldots (g_t - m_t)^2\right) \approx \mathbb{E}\left[(g_t - \mathbb{E}g_t)^2\right] = \mathbf{Var} g_t$$

2. Method
Problem - solution



모든 구간에서 ideal 한 step size를 가짐

Case1 : g_t 의 Variance 가 작기 때문에 step size is **large**
Case2 : g_t 의 Variance 가 크기 때문에 step size is **small**
Case3 : g_t 의 Variance 가 작기 때문에 step size is **large**

	Case 1			Case 2			Case 3		
$\|g_t\|, v_t$	S			L			L		
$\|g_t - g_{t-1}\|, s_t$	S			L			S		
$\|\Delta\theta_t\|_{ideal}$	L			S			L		
$\|\Delta\theta_t\|$	SGD	Adam	AdaBelief	SGD	Adam	AdaBelief	SGD	Adam	AdaBelief
	S	L	L	L	S	S	L	S	L

2. Method

Convergence analysis in convex and non convex-optimization

Optimization problem For deterministic problems, the problem to be optimized is $\min_{\theta \in \mathcal{F}} f(\theta)$; for online optimization, the problem is $\min_{\theta \in \mathcal{F}} \sum_{t=1}^{T} f_t(\theta)$, where f_t can be interpreted as loss of the model with the chosen parameters in the t-th step.

Theorem 2.1. *(Convergence in convex optimization) Let $\{\theta_t\}$ and $\{s_t\}$ be the sequence obtained by AdaBelief, let $0 \leq \beta_2 < 1, \alpha_t = \frac{\alpha}{\sqrt{t}}$, $\beta_{11} = \beta_1$, $0 \leq \beta_{1t} \leq \beta_1 < 1, s_t \leq s_{t+1}, \forall t \in [T]$. Let $\theta \in \mathcal{F}$, where $\mathcal{F} \subset \mathbb{R}^d$ is a convex feasible set with bounded diameter D_∞. Assume $f(\theta)$ is a convex function optimal point as θ^*. For θ_t generated with AdaBelief, we have the following bound on the regret:*

$$\sum_{t=1}^{T}[f_t(\theta_t) - f_t(\theta^*)] \leq \frac{D_\infty^2 \sqrt{T}}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} s_{T,i}^{1/2} + \frac{(1+\beta_1)\alpha\sqrt{1+\log T}}{2\sqrt{c}(1-\beta_1)^3} \sum_{i=1}^{d} \left\| g_{1:T,i}^2 \right\|_2 + \frac{D_\infty^2}{2(1-\beta_1)} \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{\beta_{1t} s_{t,i}^{1/2}}{\alpha_t}$$

Corollary 2.1.1. *Suppose $\beta_{1,t} = \beta_1 \lambda^t$, $0 < \lambda < 1$ in Theorem (2.1), then we have:*

$$\sum_{t=1}^{T}[f_t(\theta_t) - f_t(\theta^*)] \leq \frac{D_\infty^2 \sqrt{T}}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} s_{T,i}^{1/2} + \frac{(1+\beta_1)\alpha\sqrt{1+\log T}}{2\sqrt{c}(1-\beta_1)^3} \sum_{i=1}^{d} \left\| g_{1:T,i}^2 \right\|_2 + \frac{D_\infty^2 \beta_1 G_\infty}{2(1-\beta_1)(1-\lambda)^2\alpha}$$

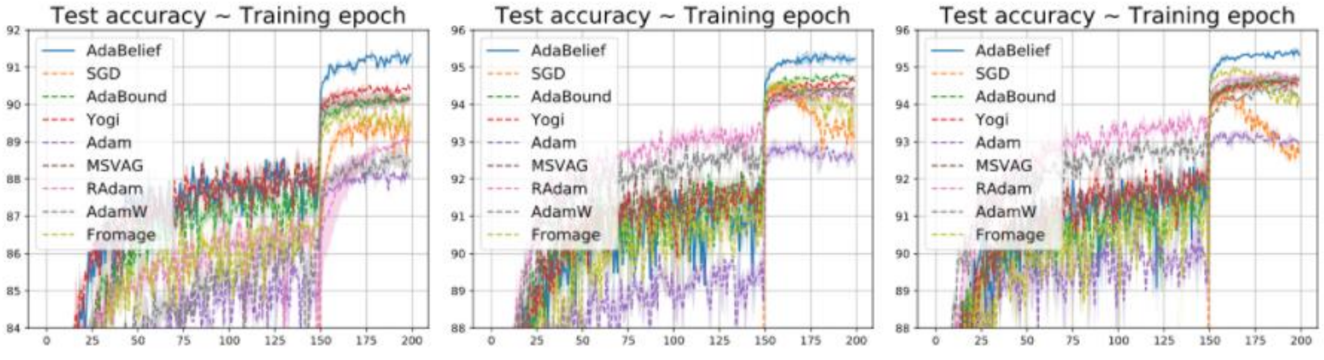
Proof in paper Appendix

Experiments

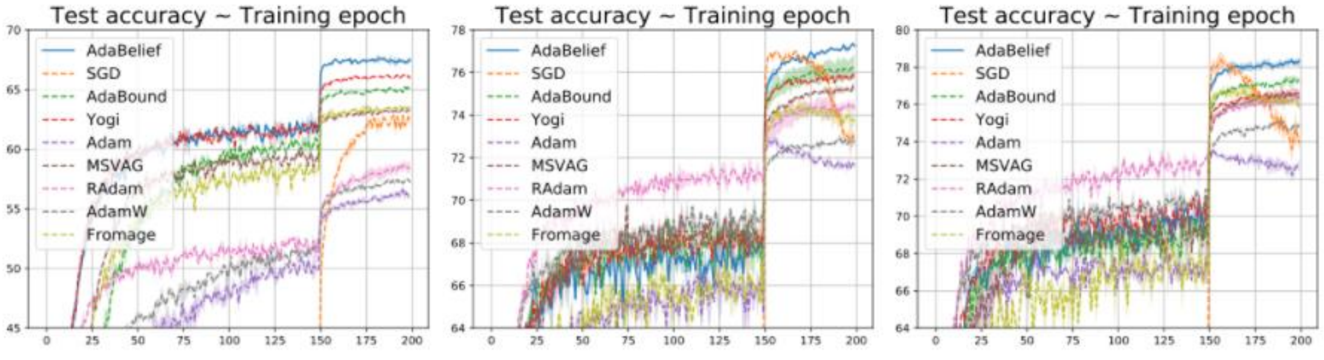
3.Experiments

Image Classification

Image Classification



(a) VGG11 on Cifar10 (b) ResNet34 on Cifar10 (c) DenseNet121 on Cifar10



(d) VGG11 on Cifar100 (e) ResNet34 on Cifar100 (f) DenseNet121 on Cifar100

Figure 4: Test accuracy ($[\mu \pm \sigma]$) on Cifar. Code modified from official implementation of AdaBound.

3.Experiments
Language Modeling

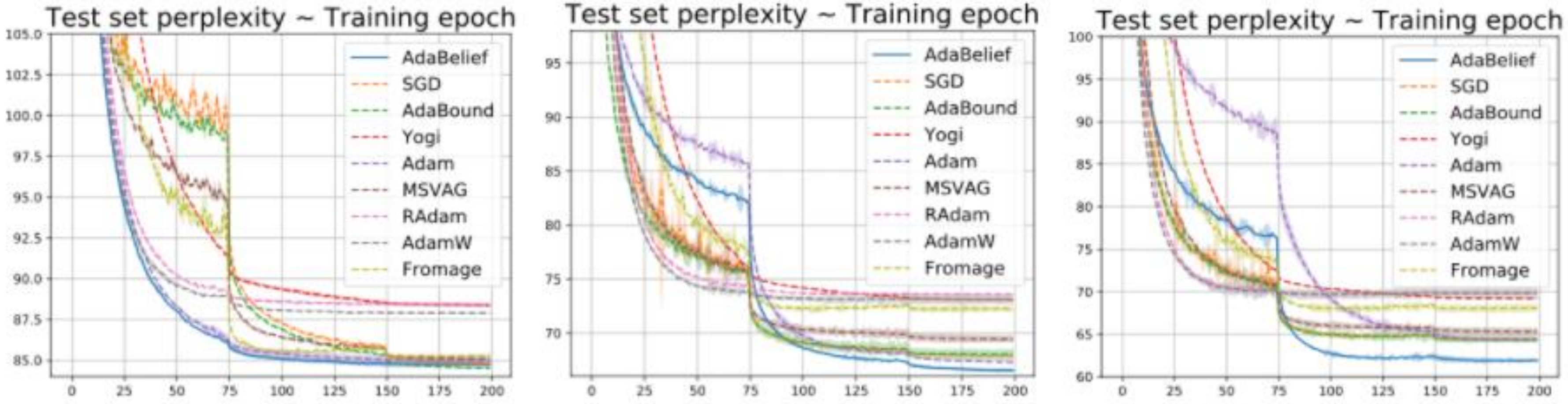
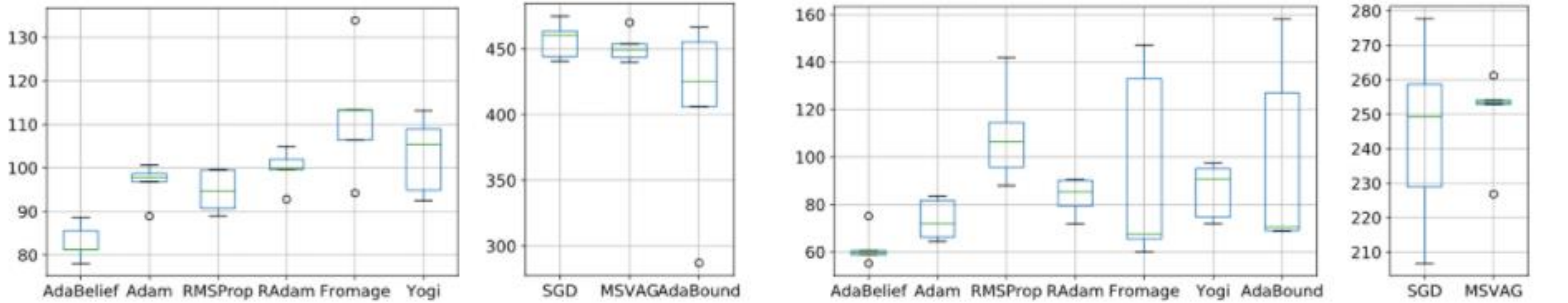


Figure 5: Left to right: perplexity ($[\mu \pm \sigma]$) on Penn Treebank for 1,2,3-layer LSTM. **Lower** is better.

3.Experiments
GAN

Generative Adversarial Network



(a) FID score of WGAN. (b) FID score of WGAN-GP.

Figure 6: FID score of WGAN and WGAN-GP on Cifar10. **Lower** is better. For each model, success and failure optimizers are shown in the left and right respectively, with different ranges in y value.

https://juntang-zhuang.github.io/adabelief/
https://arxiv.org/pdf/2010.07468v5.pdf