

Adam and Decoupled Weight Decay Regularization

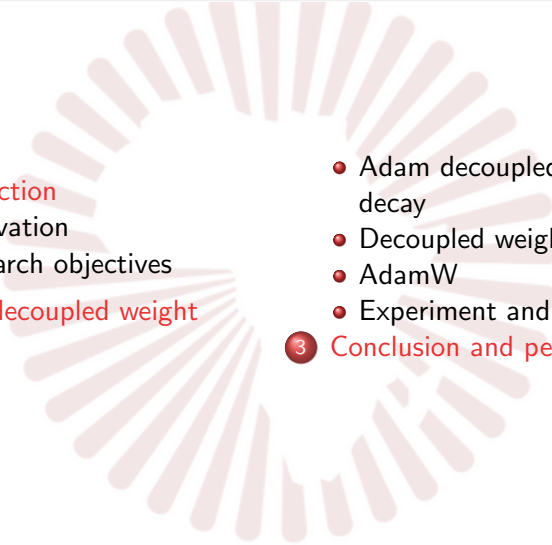
Benjamin Benteke Longau

African Institute for Mathematical Sciences, AMMI-Senegal

August 27, 2021



Overview

- 
- 1 Introduction
 - Motivation
 - Research objectives
 - 2 Adam decoupled weight decay
 - Adam decoupled weight decay
 - Decoupled weight decay
 - AdamW
 - Experiment and Results
 - 3 Conclusion and perspectives



Overview

- 
- 1 **Introduction**
 - Motivation
 - Research objectives
 - 2 Adam decoupled weight decay
 - Adam decoupled weight decay
 - Decoupled weight decay
 - AdamW
 - Experiment and Results
 - 3 Conclusion and perspectives



Introduction

Motivation

Problem of inference

After training a Neural network, and it got high test error. This could occur maybe when the Gradient is too noisy.

Solution 1

Gradient Descent (GD), Stochastic GD and Mini-batch GD. The choice of a proper learning rate and applying this one to parameters update

Solution 2

Adaptive Gradient algorithms as, Adam, RMSprop, etc.



Overview

- 
- 1 Introduction
 - Motivation
 - Research objectives
 - 2 Adam decoupled weight decay
 - Adam decoupled weight decay
 - Decoupled weight decay
 - AdamW
 - Experiment and Results
 - 3 Conclusion and perspectives



Introduction

Research objectives

As research objectives, we have:

- Adam and Adam weight decay;
- Experimentation on real dataset.



Overview

- 
- 1 Introduction
 - Motivation
 - Research objectives
 - 2 Adam decoupled weight decay
 - Adam decoupled weight decay
 - Decoupled weight decay
 - AdamW
 - Experiment and Results
 - 3 Conclusion and perspectives



SGD vs. Adam decoupled weight decay (AdamW)

Problem

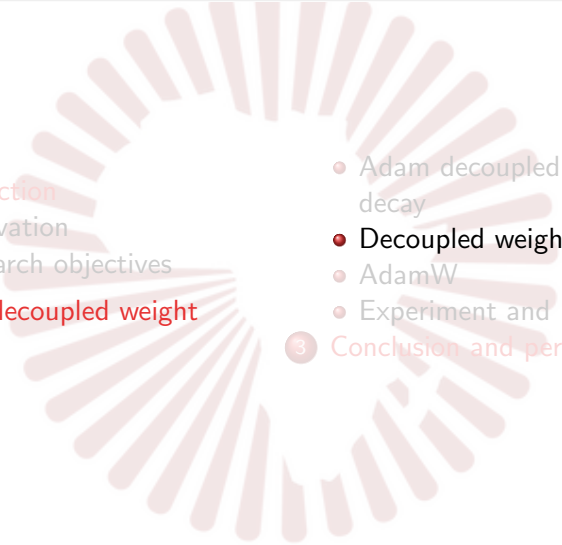
Question: Is Adam poor in generalization than SGD? if Yes/or, why?

Answer: Yes! because, $L2$ regularization that we apply to Adam is not effective as for SGD.

Solutions One of the solution, is to use *decoupled weight decay*.



Overview

- 
- 1 Introduction
 - Motivation
 - Research objectives
 - 2 Adam decoupled weight decay
 - Adam decoupled weight decay
 - Decoupled weight decay
 - AdamW
 - Experiment and Results
 - 3 Conclusion and perspectives



Decoupled weight decay

Decoupled weight decay

Weight decay

$$\theta_{t+1} = (1 - \lambda)\theta_t - \alpha \nabla f_t(\theta_t) \quad (2.1)$$

where λ defines the rate of the weight decay rate and α learning rate for batch t . For SGD, weight decay is similar to $L2$ regularization.


Proposition

$$f_t^{reg}(\theta_t) = f_t(\theta_t) + \frac{\lambda'}{2} \|\theta_t\|_2^2 \quad (2.2)$$

the $L2$ reg. term λ' , must be $\lambda' = \frac{\lambda}{\alpha}$ (couple between, weight decay and learning rate).



Overview

- 
- 1 Introduction
 - Motivation
 - Research objectives
 - 2 Adam decoupled weight decay
 - Adam decoupled weight decay
 - Decoupled weight decay
 - **AdamW**
 - Experiment and Results
 - 3 Conclusion and perspectives



AdamW

AdamW

Decoupled weight decay

It decouples weight decay from the gradient update by adding it after the parameter update as in the original definition.

Algorithm 2 Adam with L_2 regularization and Adam with decoupled weight decay (AdamW)

```

1: given  $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \lambda \in \mathbb{R}$ 
2: initialize time step  $t \leftarrow 0$ , parameter vector  $\theta_{t=0} \in \mathbb{R}^n$ , first moment vector  $m_{t=0} \leftarrow \mathbf{0}$ , second moment vector  $v_{t=0} \leftarrow \mathbf{0}$ , schedule multiplier  $\eta_{t=0} \in \mathbb{R}$ 
3: repeat
4:    $t \leftarrow t + 1$ 
5:    $\nabla f_t(\theta_{t-1}) \leftarrow \text{SelectBatch}(\theta_{t-1})$  ▷ select batch and return the corresponding gradient
6:    $g_t \leftarrow \nabla f_t(\theta_{t-1}) + \lambda \theta_{t-1}$ 
7:    $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$  ▷ here and below all operations are element-wise
8:    $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ 
9:    $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$  ▷  $\beta_1$  is taken to the power of  $t$ 
10:   $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$  ▷  $\beta_2$  is taken to the power of  $t$ 
11:   $\eta_t \leftarrow \text{SetScheduleMultiplier}(t)$  ▷ can be fixed, decay, or also be used for warm restarts
12:   $\theta_t \leftarrow \theta_{t-1} - \eta_t \left( \alpha \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) + \lambda \theta_{t-1} \right)$ 
13: until stopping criterion is met
14: return optimized parameters  $\theta_t$ 

```

Figure 1: AdamW optimizer



Overview

- 
- 1 Introduction
 - Motivation
 - Research objectives
 - 2 Adam decoupled weight decay
 - Adam decoupled weight decay
 - Decoupled weight decay
 - AdamW
 - Experiment and Results
 - 3 Conclusion and perspectives



Experiments and results

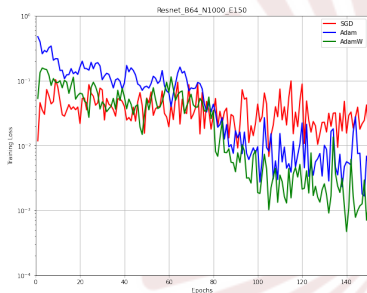
Setup

- Weight decay rate: 10^{-4} ;
- Dataset: CIFAR-10;
- Batch size: 64;
- Learning rate: 0.001;
- Epochs: 150;
- Training and test sample size: 1000 and 750;
- Model: resnet18;
- Train loss: cross entropy.

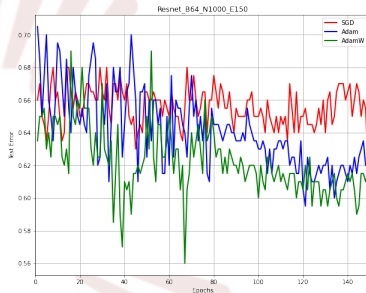


Experiments and results

Results and evaluation



(a) Train loss of SGD, Adam and AdamW with $lr=0.001$ and weight decay $=10^{-4}$



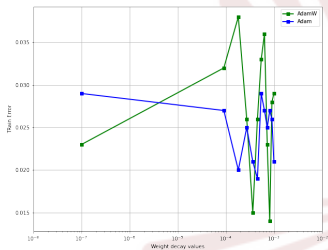
(b) Test loss of SGD, Adam and AdamW with $lr=0.001$ and weight decay $=10^{-4}$

Figure 2: Train loss vs. test loss

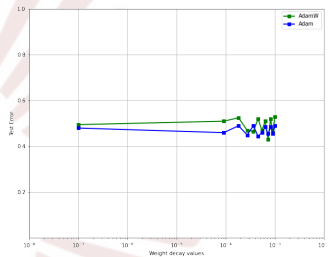


Experiments and results

Results and evaluation



(a) Train loss for Adam and AdamW for 12 different weight decay values.



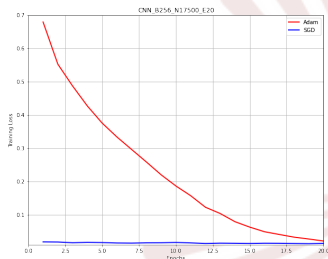
(b) Test loss for Adam and AdamW for 12 different weight decay values.

Figure 3: Train and test loss w.r.t to 12 weight decay

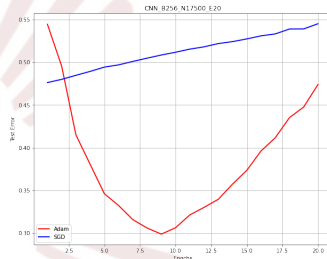


Experiments and results

Results and evaluation



(a) Train loss for Adam and SGD, on IMDB datasets



(b) Train loss for Adam and SGD, on IMDB datasets

Figure 4: Adam vs. SGD on IMDB dataset.



Conclusion and Perspectives

In conclusion, we have been able to:

- understanding AdamW.
- AdamW is the robust function of Adam.



Further Reading



Loshchilov, Ilya and Hutter, Frank

Decoupled weight decay regularization 2017.





Thank You for listening!

