A Study On Adam And AdamW

D11949006 Wei-Lun Chen
R11921118 Yong-Tai Qiu

1 Introduction

The field of machine learning has witnessed the advent of numerous optimization algorithms, among which the Adam and AdamW optimizers stand out due to their unique characteristics and widespread use. Despite their extensive application, certain features within these optimizers still call for comprehensive exploration and validation. Therefore, this research sets out to achieve three primary objectives centered around these optimizers.

Firstly,we strive to gain a comprehensive understanding of the Adam and AdamW optimizers. We seek to delve into their operational principles, understand the theoretical underpinnings that govern their functioning, and explore their common usage in machine learning tasks. This foundational understanding will set the stage for subsequent investigations and pave the way for our next objectives.

Our second goal is to evaluate the significance of the bias correction feature intrinsic to the Adam optimizer. This bias correction, integral to Adam's functionality, is designed to offset any underestimation of the gradient moments during the initial stages of training. However, the practical implications of this feature on model performance are not fully understood. We propose to probe into the role of bias correction through extensive experimentation and detailed analysis. Our study design includes varying learning rates and beta values while observing the impacts with and without bias correction.

Finally, our third aim is to draw comparisons between L2 regularization in Adam and weight decay in AdamW. AdamW was conceived as a solution to the issues found with weight decay in Adam, yet the distinction between the two and their consequent impact on model performance remain somewhat hazy. Our research intends to illuminate these subtle differences, offering strong empirical evidence to validate the distinction. This objective includes experiments with varying lambda values to understand the effects on L2 regularization and AdamW.

Through this study, we aspire to en-

rich our understanding of these optimizers and offer significant insights that may inform optimization decisions in machine learning projects. Our experimental outcomes, documented through loss and accuracy measures, will provide robust evidence to support our conclusions.

2 Methods

In our research, we conducted a comparative evaluation using a Convolutional Neural Network (CNN) model performing classification tasks on the MNIST dataset. To maintain a balanced comparison, we kept consistent the model architecture, initial weight settings, training epochs, batch size, and the loss function across all experimental setups. We scrutinized four unique settings of the optimizer: the conventional Adam optimizer, a modified Adam version without bias correction, Adam incorporating L2 regularization, and AdamW. These configurations were meticulously chosen to explore the impacts of bias correction and different weight decay methods on the performance of the model.

For every optimizer setup, we trained our CNN model across 100 epochs with a batch size of 64. To ensure consistency in the starting point, we initiated the model with identical random seeds across all conditions, guaranteeing uniform initial weights. The loss function employed was Cross Entropy Loss.

We undertook the training with three distinct learning rates (γ) – 1e-3, 1e-4, and 1e-5 – applied to each of the four optimizers. This procedure enabled us to study the variation in the model's performance under different learning rates spanning the optimizers. In all scenarios, the beta parameters (β_1 and β_2) were set at 0.9 and 0.999, respectively, while the epsilon (ϵ) was established at 1e-8. These configurations were selected to be in sync with commonly used values in existing literature.

Finally, we integrated weight decay in the Adam with L2 regularization and the AdamW optimizers, standardizing the decay factor (λ) to 0.01 in both instances. This facilitated us to draw direct comparisons between the effects of L2 regularization and weight decay decoupling on the model's performance.

Through this experimental design, we were able to systematically investigate the comparative impacts of bias correction and different weight decay methods on the performance of a CNN model in image classification tasks.

3 Results

Table 1 shows our experimental settings. In Experiment 1-1 and Experiment 1-2, we want to know whether or not bias correction will affect the accuracy under different learning rates and different betas?

In Experiment 2, I want to know whether it will affect the accuracy of Adam and AdamW under different lambda (weight decay)?

Table 1: Experimental Settings

Hyperparameter	**Exp1**	**Exp2**
Optimizer	Adam Adam w/o Bias Correction	Adam (L2) AdamW
Learning Rate (γ)	1e-3, 1e-4, 1e-5	1e-4
Beta Parameters (β_1, β_2)	(0.9, 0.999), (0.99, 0.999)	(0.9, 0.999)
Epsilon (ϵ)	1e-8	1e-8
Decay Factor (λ)	-	1e-2, 1e-3, 1e-4

Figure 1 shows the results of Experiment 1-1, represents the verification accuracy of Adam with or without bias correction under the same learning rate and different betas.

It can be found that under same learning rate, different betas, the with or without of bias correction has little effect on the accuracy rate.

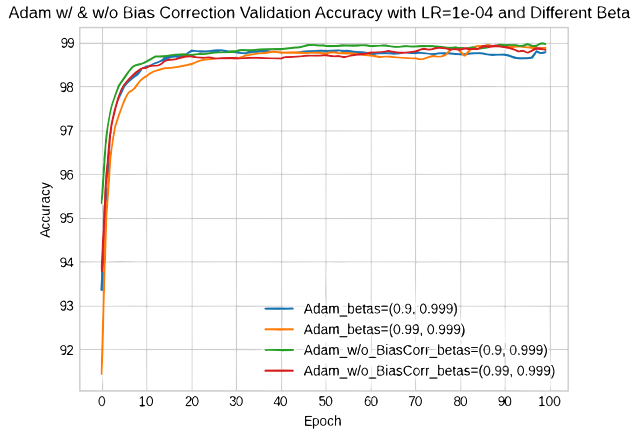

Figure 1: Adam with and without bias correction with learning rate 1e-04 and different betas validation accuracy

We only show results for learning rate 1e-4 here, but in fact learning rates 1e-3, 1e-5 have similar results.

Figure 2 shows the results of Experiment 1-2, represents the verification accuracy of Adam with or without bias correction under the same betas and different learning rate.

It can be found that under same betas, different learning rate, the with or without of bias correction has little effect on the accuracy rate.

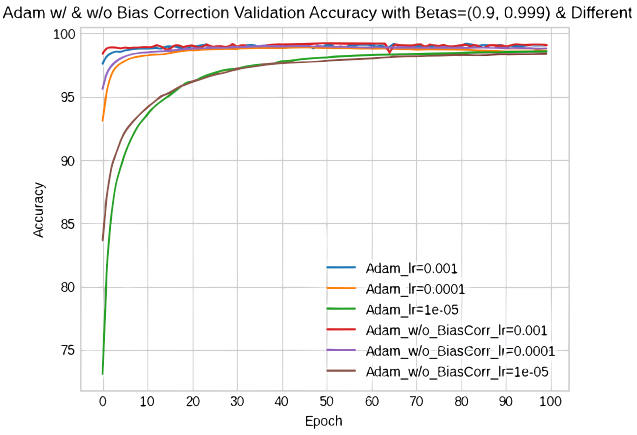

Figure 2: Adam with and without bias correction with betas (0.9, 0.999) and different learning rate validation accuracy

We only show results for betas (0.9, 0.999) here, but in fact betas (0.99, 0.999) have similar results.

Figure 3 shows the results of Experiment 2, under different lambdas(weight decay), whether Adam with L2 and adamW will affect the accuracy?

We can found in Experiment 2 that if Adam's L2 is larger, the accuracy rate will be smaller. In addition, the Adam+L2 scheme will lead to Suboptimal optimization process, and the L2 size of AdamW seems to have little effect on the accuracy.

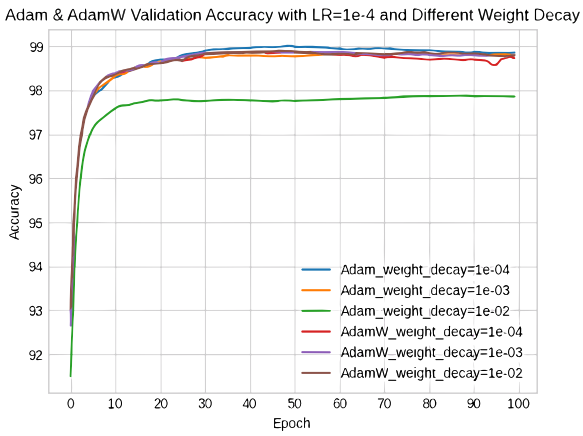

Figure 3: Adam and AdamW with different weight decay validation accuracy

4 Discussion and Conclusions

Let's break down and discuss each experiment individually:

Experiment 1-1 and Experiment 1-2: The results indicate that the presence or absence of bias correction does not significantly impact the accuracy rate, regardless of the learning rate and betas used. This suggests that, in these experiments, bias correction does not play a crucial role in improving the model's accuracy. It could imply that bias correction may not be essential in certain contexts or that other factors overshadow its influence on model performance. Further research is necessary to explore the underlying reasons for this observation.

Experiment 2: The findings from Experiment 2 reveal that higher L2 regularization parameter in Adam leads to lower accuracy rates. This could be attributed to the stronger constraint imposed on the model's weights by larger L2 regularization, limiting the model's expressive power and consequently affecting its performance. The weight decay and L2 regularization may reach equivalence under some conditions still are slightly different concepts and should be treated differently otherwise can lead to unexplained performance degradation or other practical problems.

However, the optimal value of the regularization parameter depends on the specific dataset and problem at hand. Therefore, in practical applications, fine-tuning the regularization parameter may be necessary to achieve optimal performance.

In summary, these conclusions highlight the effects of bias correction and L2 regularization in different scenarios. Further investigation is needed to gain a deeper understanding of these observations and to determine the generalizability of the findings across various datasets and models.