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 aim to undertake an in-depth understanding of the Adam and AdamW optimizers. We endeavor to dissect their mechanics, the theoretical basis for their construction, and their typical applications in machine learning. This groundwork will lay the foundation for further investigations, setting the stage for the subsequent objectives.

Secondly, we intend to validate the importance of the bias correction feature implemented in the Adam optimizer. Bias correction is an inherent element of Adam, designed to prevent underestimation of the gradient moments during initial training stages. However, its practical implications on model performance are yet to be exhaustively investigated. We aim to substantiate the role of bias correction through rigorous experimentation and analysis.

Lastly, our third objective revolves around a comparison between L2 regularization in Adam and weight decay in AdamW. AdamW was developed as a response to the issues concerning weight decay in Adam, but the distinction and the resulting impact on model performance remain somewhat ambiguous. Our study seeks to elucidate these nuances, providing robust evidence to validate the difference between the two.

Through this research, we hope to deepen our understanding of these optimizers and offer valuable insights that could guide the optimization choices in machine learning tasks.

2 Methods

In this study, we performed a comparative analysis using a Convolutional Neu-

ral Network (CNN) model for the classification task on the MNIST dataset. To ensure a fair comparison, we held constant the model architecture, initial weights, number of training epochs, batch size, and loss function across all conditions.

We evaluated four different settings of the optimizer: the standard Adam optimizer, a custom version of Adam without bias correction, Adam with L2 regularization, and AdamW. These settings were selected to test the effects of bias correction and weight decay method on model performance.

For each optimizer, we trained the CNN model for 100 epochs with a batch size of 64. The model was initialized with the same random seed in each condition to ensure consistent initial weights. The loss function used was Cross Entropy Loss.

We conducted the training with three different learning rates (γ) 1e-3, 1e-4, and 1e-5, for all the four optimizers. This allowed us to examine how the model's performance varied with different learning rates across the optimizers. In all cases, the beta parameters (β_1 and β_2) were set to 0.9 and 0.999, respectively, and the epsilon (ϵ) was set to 1e-8. These settings were chosen to align with the typical values used in the literature.

Lastly, we incorporated weight decay in the Adam with L2 regularization and the AdamW optimizers, setting the decay factor (λ) to 0.01 in both cases. This enabled us to directly compare the impact of L2 regularization and weight decay decoupling on model performance.

This experimental design allowed us to systematically examine the relative impact of bias correction and weight decay method on the performance of a CNN model for image classification.

3 Results

4 Discussion and Conclusions