

# Weight Prediction Boosts the Convergence of AdamW

Lei Guan

Department of Mathematics, National University of Defense Technology  
guanleimath@163.com

**Abstract.** In this paper, we introduce weight prediction into the AdamW optimizer to boost its convergence when training the deep neural network (DNN) models. In particular, ahead of each mini-batch training, we predict the future weights according to the update rule of AdamW and then apply the predicted future weights to do both forward pass and backward propagation. In this way, the AdamW optimizer always utilizes the gradients w.r.t. the future weights instead of current weights to update the DNN parameters, making the AdamW optimizer achieve better convergence. Our proposal is simple and straightforward to implement but effective in boosting the convergence of DNN training. We performed extensive experimental evaluations on image classification and language modeling tasks to verify the effectiveness of our proposal. The experimental results validate that our proposal can boost the convergence of AdamW and achieve better accuracy than AdamW when training the DNN models.

**Keywords:** Deep learning · Weight prediction · Convergence · AdamW.

## 1 Introduction

The optimization of deep neural network (DNN) models is to find the optimal parameters using an optimizer which has a decisive influence on the convergence of the models and thus directly affects the total training time. Adaptive gradient methods, such as RMSprop [20], AdaGrad [3], Adam [7] and AdamW [11], are currently of core practical importance in deep learning training as they are able to attain rapid training of modern deep neural network (DNN) models. Particularly, AdamW [11], also known as Adam with decoupled weight decay, has been used as a default optimizer for training various DNN models [1, 10, 11, 20, 21]. The major advantage of AdamW lies in that it improves the generalization performance of Adam [7] and thus works as effectively as SGD with momentum [18] on image classification tasks.

As with other popular gradient-based optimization methods, when using AdamW as an optimizer, each iteration of DNN training, *i.e.*, a mini-batch training, generally consists of one forward pass and one backward propagation, where the gradients w.r.t. all the parameters (also known as weights) are computed during the backward propagation. The generated gradients are then used

by the AdamW optimizer to calculate the update values for all parameters, which are finally applied to updating the weights. The remarkable features of using AdamW to update parameters include: 1) the updates of weights are continuous; 2) each mini-batch uses the currently available weights to do both forward pass and backward propagation.

Motivated by the fact that DNN weights are updated in a continuous manner and the update values calculated by the AdamW should reflect the "correct" direction for updating the weights, we introduce weight prediction [2,4] into the DNN training to further boost the convergence of AdamW. Concretely, ahead of each mini-batch training, we first perform weight prediction according to the currently available weights and the update rule of AdamW. Following that, we use the predicted future weights instead of current weights to perform both forward pass and backward propagation. Finally, the AdamW optimizer utilizes the gradients w.r.t. the predicted weights to update the DNN parameters. We experiment with two typical machine learning tasks, including image classification and language modeling. The experimental results demonstrate that our proposal outperforms AdamW in terms of convergence and accuracy. For instance, when training four convolution neural network (CNN) models on CIFAR-10 dataset, our proposal yields an average accuracy improvement of 0.47% (up to 0.74%) over AdamW. When training LSTMs on Penn TreeBank dataset, our proposal achieves 5.52 less perplexity than AdamW on average.

The contributions of this paper can be summarized as follows:

- (1) We, for the first time, construct the mathematical relationship between currently available weights and future weights after several continuous updates when using AdamW as an optimizer.
- (2) We devise an effective way to incorporate weight prediction into AdamW. To the best of our knowledge, this is the first time that uses weight prediction strategy to boost the convergence of AdamW. The proposed weight prediction strategy is believed to be well suited for other popular optimization methods such as RMSprop [20], AdaGrad [3], Adam [7], et al.
- (3) We conducted extensive experimental evaluations to validate the effectiveness of our proposal, which demonstrates that our proposal is able to boost the convergence of AdamW when training the DNN models.

## 2 Related Work

When using the gradient-based optimization methods to train DNN models, the differences in optimization methods lie in that the ways using gradients to update model parameters are different. Generally, the commonly used first-order gradient methods can be categorized into two groups: the accelerated stochastic gradient descent (SGD) family [15, 16, 18] and adaptive gradient methods [7, 23, 24].

Adaptive gradient methods, also known as adaptive learning methods, have been heavily studied in prior research and widely used in deep learning training. Very different from the SGD methods (e.g., Momentum SGD [18]), which use a

unified learning rate for all parameters, adaptive gradient methods compute a specific learning rate for each individual parameter [24]. In 2011, Duchi et al. [3] proposed the AdaGrad, which can dynamically adjust the learning rate according to the history gradients from previous iterations and utilize the quadratic sum of all previous gradients to update the model parameters. Zeiler [23] proposed AdaDelta, seeking to alleviate the continual decay of the learning rate of AdaGrad. AdaDelta does not require manual tuning of a learning rate and is robust to noisy gradient information. Tieleman and Hinton [20] refined AdaGrad and proposed RMSprop. The same as AdaGrad, RMSprop adjusts the learning rate via element-wise computation and then updates the variables. One remarkable feature of RMSprop is that it can avoid decaying the learning rate too quickly. In order to combine the advantages of both AdaGrad and RMSprop, Kingma and Ba [7] proposed another famous adaptive gradient method, Adam, which has become an extremely important choice for deep learning training. Loshchilov and Hutter [11] found that the major factor of the poor generalization of Adam is due to that  $L_2$  regularization for it is not as effective as for its competitor, the Momentum SGD. They thus proposed decoupled weight decay regularization for Adam, which is also known as AdamW. The experimental results demonstrate that AdamW substantially improves the generalization performance of Adam and illustrates competitive performance as Momentum SGD [18] when tackling image classification tasks. To simultaneously achieve fast convergence and good generalization, Zhuang *et al.* [24] proposed another adaptive gradient method called AdaBelief, which adapts the stepsize according to the “belief” in the current gradient direction. Other adaptive gradient methods include AdaBound [12], RAdam [9], Yogi [22], et al. It is worth noting that all these adaptive gradient methods share a common feature: weight updates are continuous and each mini-batch training always uses currently available weights to perform both forward pass and backward propagation.

Weight prediction was previously used to overcome the weight inconsistency issue in the asynchronous pipeline parallelism. Chen et al. [2] used the smoothed gradient to replace the true gradient in order to predict future weights when using Momentum SGD [18] as the optimizer. Guan et al. [4] proposed using the update values of Adam [7] to make weight predictions. Yet, both approaches use weight prediction to ensure the weight consistency of pipeline training rather than considering the impact of weight prediction on the optimizers themselves.

### 3 Methods

Ahead of any  $t$ -th ( $t \geq 1$ ) iteration, we assume that the current available DNN weights are  $\theta_{t-1}$ . Given the initial learning rate  $\gamma \in \mathbb{R}$ , momentum factor  $\beta_1 \in \mathbb{R}$  and  $\beta_2 \in \mathbb{R}$ , and weight decay value  $\lambda \in \mathbb{R}$ , we reformulate the update of

AdamW [11] as

$$\begin{aligned} \boldsymbol{\theta}_t &= (1 - \gamma\lambda)\boldsymbol{\theta}_{t-1} - \frac{\gamma\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t + \epsilon}}, \\ \text{s.t. } \begin{cases} \mathbf{g}_t = \nabla_t f(\boldsymbol{\theta}_{t-1}), \\ \mathbf{m}_t = \beta_1 \cdot \mathbf{m}_{t-1} + (1 - \beta_1) \cdot \mathbf{g}_t, \\ \mathbf{v}_t = \beta_2 \cdot \mathbf{v}_{t-1} + (1 - \beta_2) \cdot \mathbf{g}_t^2, \\ \hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t}, \\ \hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t}. \end{cases} \end{aligned} \quad (1)$$

In (1),  $\mathbf{m}_t$  and  $\mathbf{v}_t$  refer to the first and second moment vector respectively,  $\epsilon$  is the smoothing term which can prevent division by zero.

Letting  $\boldsymbol{\theta}_0$  denote the initial weights of a DNN model, then in the following  $s$  times of continuous mini-batch training, the DNN weights are updated via

$$\begin{aligned} \boldsymbol{\theta}_1 &= (1 - \gamma\lambda)\boldsymbol{\theta}_0 - \frac{\gamma\hat{\mathbf{m}}_1}{\sqrt{\hat{\mathbf{v}}_1 + \epsilon}}, \\ \boldsymbol{\theta}_2 &= (1 - \gamma\lambda)\boldsymbol{\theta}_1 - \frac{\gamma\hat{\mathbf{m}}_2}{\sqrt{\hat{\mathbf{v}}_2 + \epsilon}}, \\ &\dots \\ \boldsymbol{\theta}_s &= (1 - \gamma\lambda)\boldsymbol{\theta}_{s-1} - \frac{\gamma\hat{\mathbf{m}}_s}{\sqrt{\hat{\mathbf{v}}_s + \epsilon}}, \end{aligned} \quad (2)$$

where for any  $i \in \{1, 2, \dots, s\}$ , we have

$$\begin{cases} \mathbf{g}_i = \nabla_i f(\boldsymbol{\theta}_{i-1}), \\ \mathbf{m}_i = \beta_1 \cdot \mathbf{m}_{i-1} + (1 - \beta_1) \cdot \mathbf{g}_i, \\ \mathbf{v}_i = \beta_2 \cdot \mathbf{v}_{i-1} + (1 - \beta_2) \cdot \mathbf{g}_i^2, \\ \hat{\mathbf{m}}_i = \frac{\mathbf{m}_i}{1 - \beta_1^i}, \\ \hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{1 - \beta_2^i}. \end{cases} \quad (3)$$

When summing up all weight update equations in (2), we have

$$\begin{aligned} \boldsymbol{\theta}_s &= \boldsymbol{\theta}_0 - \gamma\lambda \sum_{i=0}^{t-1} (\boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 + \dots + \boldsymbol{\theta}_{t-1}) - \sum_{i=1}^t \frac{\gamma\hat{\mathbf{m}}_i}{\sqrt{\hat{\mathbf{v}}_i + \epsilon}}, \\ \text{s.t. } \begin{cases} \mathbf{g}_i = \nabla_i f(\boldsymbol{\theta}_{i-1}), \\ \mathbf{m}_i = \beta_1 \cdot \mathbf{m}_{i-1} + (1 - \beta_1) \cdot \mathbf{g}_i, \\ \mathbf{v}_i = \beta_2 \cdot \mathbf{v}_{i-1} + (1 - \beta_2) \cdot \mathbf{g}_i^2, \\ \hat{\mathbf{m}}_i = \frac{\mathbf{m}_i}{1 - \beta_1^i}, \\ \hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{1 - \beta_2^i}. \end{cases} \end{aligned} \quad (4)$$

It is well known that the weight decay value  $\lambda$  is generally set to an extremely small value (e.g.,  $5e^{-4}$ ), and the learning rate  $\gamma$  is commonly set to a value smaller than 1 (e.g., 0.01). Consequently,  $\gamma\lambda$  is pretty close to zero, and thus, the second term of the right hand of (4) can be neglected. This, therefore, generates the

following equation:

$$\boldsymbol{\theta}_s \approx \boldsymbol{\theta}_0 - \sum_{i=1}^s \frac{\gamma \hat{\mathbf{m}}_i}{\sqrt{\hat{\mathbf{v}}_i + \epsilon}}. \quad (5)$$

(5) illustrates that given the initial weights  $\boldsymbol{\theta}_0$ , the weights after  $s$  times of continuous updates can be approximately calculated. Correspondingly, given  $\boldsymbol{\theta}_t$ , the weights after  $s$  times of continuous updates can be approximately calculated via

$$\boldsymbol{\theta}_{t+s} \approx \boldsymbol{\theta}_t - \sum_{i=t+1}^{t+s} \frac{\gamma \hat{\mathbf{m}}_i}{\sqrt{\hat{\mathbf{v}}_i + \epsilon}}. \quad (6)$$

From (6), we see that given the initial weights  $\boldsymbol{\theta}_t$ ,  $\boldsymbol{\theta}_{t+s}$  can be approximately calculated by letting  $\boldsymbol{\theta}_t$  subtract the sum of  $s$  continuous relative variation of the weights. Note that the relative increments of the weights in each iteration should reflect the trend of the weight updates in each iteration. In (6),  $\frac{\gamma \hat{\mathbf{m}}_i}{\sqrt{\hat{\mathbf{v}}_i + \epsilon}}$  should reflect the ‘‘correct’’ direction for updating the weights  $\boldsymbol{\theta}_t$  as it is calculated by the AdamW, and the weights are updated in a continuous manner and along the way of inertia directions.

We can therefore replace  $\sum_{i=t+1}^{t+s} \frac{\gamma \hat{\mathbf{m}}_i}{\sqrt{\hat{\mathbf{v}}_i + \epsilon}}$  in (6) with  $s \frac{\gamma \hat{\mathbf{m}}_{t+1}}{\sqrt{\hat{\mathbf{v}}_{t+1} + \epsilon}}$  in an effort to approximately predict  $\boldsymbol{\theta}_{t+s}$  for the case when only  $\boldsymbol{\theta}_t$ ,  $\mathbf{g}_t$  and the weight prediction steps  $s$  are available. Note that at any  $t$ -th iteration, the gradients of stochastic objective, *i.e.*,  $\mathbf{g}_t = \nabla_t(\boldsymbol{\theta}_{t-1})$ , can be calculated when the backward propagation is completed. Letting  $\hat{\boldsymbol{\theta}}_{t+s}$  denote the approximately predicted weights for  $\boldsymbol{\theta}_{t+s}$ , we can construct the mathematical relationship between  $\boldsymbol{\theta}_t$  and  $\hat{\boldsymbol{\theta}}_{t+s}$  as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{t+s} &= \boldsymbol{\theta}_t - s \frac{\gamma \hat{\mathbf{m}}_{t+1}}{\sqrt{\hat{\mathbf{v}}_{t+1} + \epsilon}}, \\ \text{s.t. } &\begin{cases} \mathbf{g}_t = \nabla_t f(\boldsymbol{\theta}_{t-1}), \\ \mathbf{m}_t = \beta_1 \cdot \mathbf{m}_{t-1} + (1 - \beta_1) \cdot \mathbf{g}_t, \\ \mathbf{v}_t = \beta_2 \cdot \mathbf{v}_{t-1} + (1 - \beta_2) \cdot \mathbf{g}_t^2, \\ \hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t}, \\ \hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t}. \end{cases} \end{aligned} \quad (7)$$

In the following, we showcase how to incorporate weight prediction into the DNN training when using AdamW [11] as an optimizer. Algorithm 1 illustrates the detailed information. The weight prediction step  $s$  and other hyperparameters are required ahead of the DNN training. At each iteration, the current available weights  $\boldsymbol{\theta}_t$  should be cached before the forward pass starts (Line 4). Then weight prediction are performed using (7) and the predicted weights  $\hat{\boldsymbol{\theta}}_{t+s}$  is generated (Line 5). Following that, the predicted weights  $\hat{\boldsymbol{\theta}}_{t+s}$  are used to do both forward pass and backward propagation (Lines 6 and 7). Finally, the cached weights  $\boldsymbol{\theta}_t$  is recovered and updated using the AdamW optimizer (Lines 8 and 9).

---

**Algorithm 1** Weight prediction for AdamW

---

**Require:** Weight prediction step  $s$ , other hyper-parameters such as  $\gamma, \beta_1, \beta_2, \gamma, \epsilon$ .

- 1: Initialize or load DNN weights  $\theta_0$ .
  - 2:  $t \leftarrow 1$ .
  - 3: **while** stopping criterion is not met **do**
  - 4:   Cache the current weights  $\theta_t$ .
  - 5:   Calculate  $\hat{\theta}_{t+s}$  using (7).
  - 6:   Do forward pass with  $\theta_{t+s}$ .
  - 7:   Do backward propagation with  $\hat{\theta}_{t+s}$ .
  - 8:   Recover the cached weights  $\theta_t$ .
  - 9:   Update the weights  $\theta_t$  using the AdamW optimizer.
  - 10:   $t \leftarrow t + 1$ .
  - 11: **end while**
- 

## 4 Experiments

### 4.1 Experiment Settings

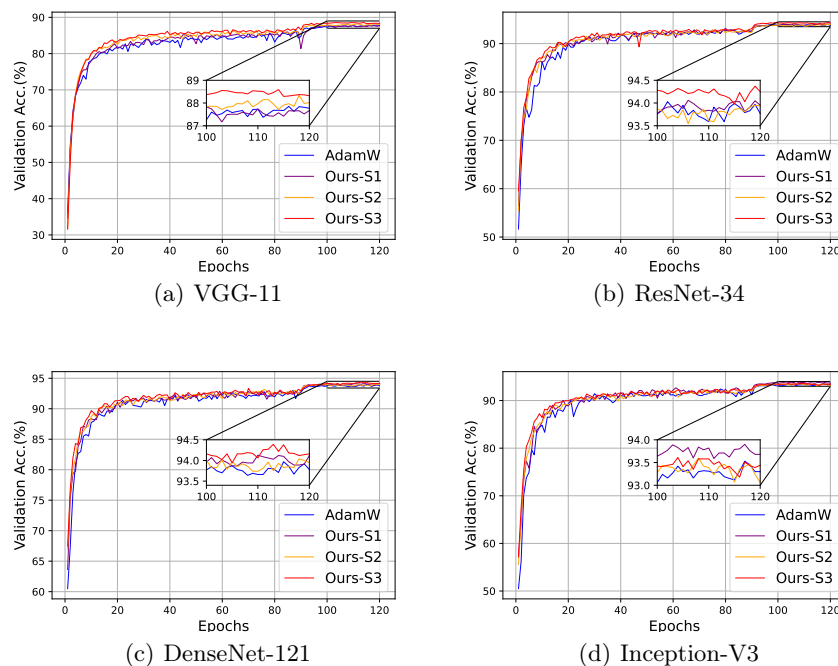
In this section, we mainly compare our proposal with AdamW [11]. We evaluated our proposal with three different weight prediction steps (*i.e.*,  $s = 1$ ,  $s = 2$ , and  $s = 3$ ), which were respectively denoted as Ours-S1, Ours-S2, and Ours-S3 for convenience purposes. We conducted experimental evaluations on two different machine learning tasks: image classification on the CIFAR-10 [8] dataset with four CNN models and language modeling on Penn TreeBank [14] dataset with two LSTM [13] models. All the experiments were conducted on a multi-GPU platform which is equipped with four NVIDIA Tesla P100 GPUs, each with 16GB of memory size. The CPU on the platform is Intel Xeon E5-2680 with 128GB DDR4-2400 off-chip main memory.

The CIFAR-10 dataset totally includes 60k  $32 \times 32$  images, 50k images for training, and 10k images for validation. The Penn TreeBank dataset consists of 929k training words, 73k validation words as well as 82k test words. For image classification, the used CNN models are VGG-11 [17], ResNet-34 [5], DenseNet-121 [6], and Inception-V3 [19]. For language modeling, we trained the LSTM models with two sizes: 1-layer LSTM and 2-layer LSTM. Each layer was configured with 650 units and was applied 50% dropout on the non-recurrent connections.

We trained all CNN models for 120 epochs with a mini-batch size of 128. The learning rate was initialized as  $1e^{-4}$ , and divided by ten at the 90th epoch. For training 1-layer and 2-layer LSTM models, we set the size of each mini-batch to 20. We trained both LSTM models for 100 epochs with an initial learning rate of 0.01 and decreased the learning rate by a factor of 10 at the 60th and 80th epochs. For AdamW [11] and our proposal, we always evaluated them with the default parameters, *i.e.*,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$ . The weight decay for both approaches was set to  $\lambda = 5e^{-4}$ .

## 4.2 CNNs on CIFAR-10

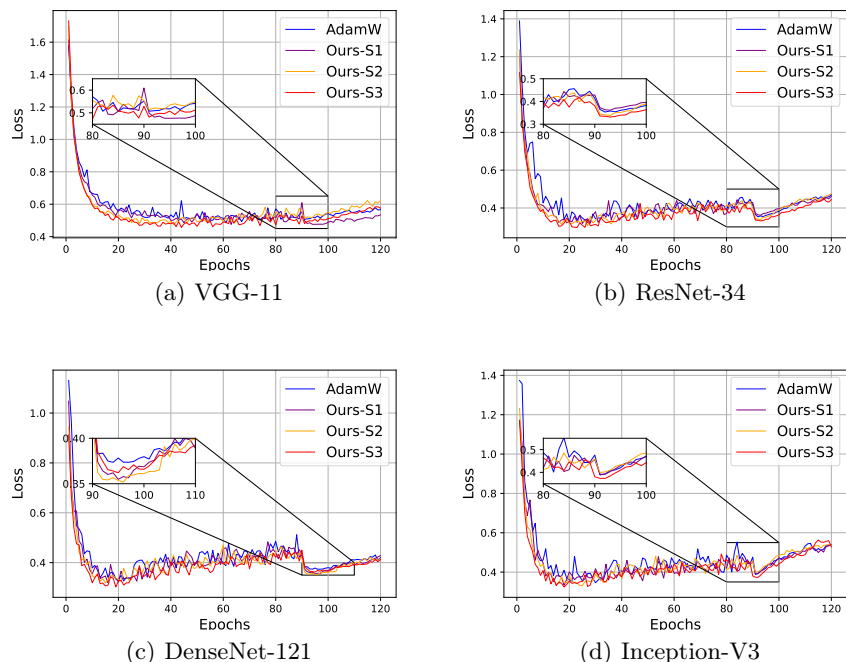
In this section, we report the experimental results when training four CNN models on the CIFAR-10 dataset. Table 1 summarizes the maximum validation top-1 accuracy and Table 2 presents the minimum validation loss. Figure 1 depicts the learning curves of validation accuracy vs. epochs for training CNNs using AdamW, Ours-S1, Ours-S2, and Ours-S3, respectively. The learning curves about validation loss vs. epochs are shown in Figure 2.



**Fig. 1.** Validation accuracy vs. epochs of training VGG-11, ResNet-34, DenseNet-121 and Inception-V3 on CIFAR-10.

**Table 1.** Maximum validation top-1 accuracy on CIFAR-10. **Higher** is better.

Models	AdamW	Ours-S1	Ours-S2	Ours-S3
VGG-11	87.85%	87.83%	88.34%	<b>88.59%</b>
ResNet-34	94.03%	94.06%	93.95%	<b>94.37%</b>
DenseNet-121	93.97%	94.13%	94.04%	<b>94.39%</b>
Inception-V3	93.53%	<b>93.90%</b>	93.60%	93.61%



**Fig. 2.** Validation loss vs. epochs of training VGG-11, ResNet-34, DenseNet-121 and Inception-V3 on CIFAR-10.

**Table 2.** Minimum validation loss on CIFAR-10. **Lower** is better.

Models	AdamW	Ours-S1	Ours-S2	Ours-S3
VGG-11	0.485	0.475	0.474	<b>0.456</b>
ResNet-34	0.323	0.318	0.305	<b>0.297</b>
DenseNet-121	0.327	0.319	0.319	<b>0.302</b>
Inception-V3	0.346	0.333	0.331	<b>0.324</b>

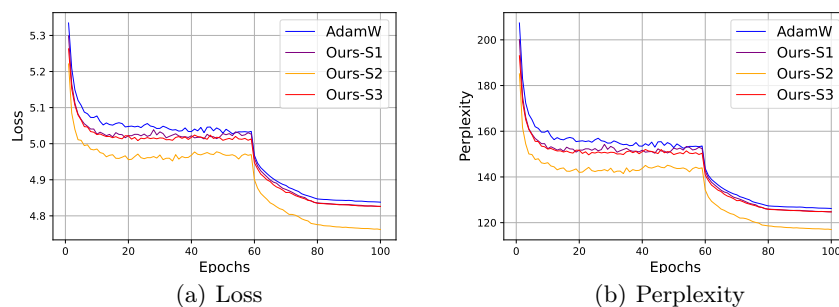
Based on the observation of Table 1 and Figure 1, we can immediately reach the following conclusions. First, Figure 1 shows that the learning curves of our proposal with different weight prediction steps match well with that of AdamW but converge faster than that of AdamW, especially at the beginning of training epochs. The learning curves in Figures 1(a), 1(b), 1(c), and 1(d) also illustrate that our proposal generally attains higher validation accuracy than AdamW at the end of the training. Second, Table 1 shows that our proposal outperforms AdamW on all evaluated CNN models in terms of the obtained maximum vali-



dation accuracy. In particular, our proposal achieves consistently higher validation top-1 accuracy than AdamW. Compared to AdamW, our proposal achieves 0.74%, 0.34%, 0.42%, and 0.37% for training VGG-11, ResNet-34, DenseNet-121, and Inception-V3, respectively. On average, our proposal yields 0.47% (up to 0.74%) top-1 accuracy improvement over AdamW. Third, comparing the experimental results of Ours-S1, Ours-S2, and Ours-S3, we can see that our proposal with different weight prediction steps consistently gets good results, which demonstrates that the performance of our proposal is independent of the settings of the weight prediction step. Particularly, the experimental results show that Ours-S3 works the best for VGG-11, ResNet-34, and DenseNet-121, while Ours-S1 works the best for Inception-V3. Similar conclusions can be drawn from the observation of Table 2 and Figure 2. Our proposal consistently obtains less validation loss than AdamW which again verifies that weight prediction can boost the convergence of AdamW when training DNN models.

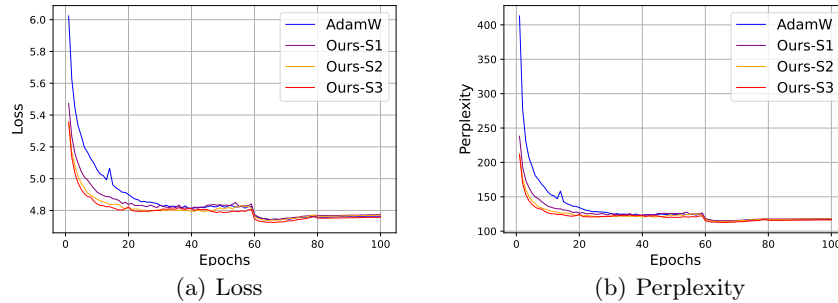
### 4.3 LSTMs on Penn TreeBank

In this section, we report the experimental results when training 1-layer and 2-layer LSTM models on the Penn TreeBank dataset [14]. Figures 3 and 4 depict the learning curves. Table 3 presents the obtained minimum perplexity (lower is better), and Table 4 summarizes the obtained minimum validation loss (lower is better).



**Fig. 3.** Training 1-layer LSTM on Penn TreeBank. Left: Loss vs. epochs; Right: Perplexity vs. epochs.

We can draw the following conclusions from the experiment results. First, as shown in Table 3, for both 1-layer and 2-layer LSTM models, our proposal achieves lower perplexity and validation loss than AdamW, validating the fast convergence and good accuracy of our proposal. Second, for 1-layer LSTM, our proposal with  $s = 2$  yields 9.22 less perplexity than AdamW. For 2-layer LSTM, our proposal with  $s = 3$  yields 2.02 less perplexity than AdamW. On average, our



**Fig. 4.** Training 2-layer LSTM on Penn TreeBank. Left: Loss vs. epochs; Right: Perplexity vs. epochs.

**Table 3.** Minimum perplexity on Penn TreeBank. **Lower** is better.

Models	AdamW	Ours-S1	Ours-S2	Ours-S3
1-Layer LSTM	126.21	124.75	<b>116.99</b>	124.71
2-Layer LSTM	114.68	114.80	113.99	<b>112.64</b>

**Table 4.** Minimum validation loss on Penn TreeBank. **Lower** is better.

Models	AdamW	Ours-S1	Ours-S2	Ours-S3
1-Layer LSTM	4.838	4.826	<b>4.762</b>	4.826
2-Layer LSTM	4.742	4.743	4.736	<b>4.724</b>

proposal achieves 5.52 less perplexity than AdamW. Second, similar conclusions can be drawn based on the observation of the loss vs. epochs learning curves in Figures 3(a) and 4(a) and Table 4. Our proposal consistently achieves less validation loss than AdamW, again validating that weight prediction can boost the convergence of AdamW.

## 5 Conclusions

To further boost the convergence of AdamW, in this paper, we introduce weight prediction into the DNN training. The remarkable feature of our proposal is that we perform both forward pass and backward propagation using the future weights which are predicted according to the update of AdamW. In particular, we construct the mathematical relationship between current weights and future weights and devise an effective way to incorporate weight prediction into DNN training. Our proposal is easy to implement and works well in boosting the convergence of DNN training. The experimental results on image classification and language modeling tasks verify the effectiveness of our proposal.

The weight prediction should also work well for other adaptive optimization methods such as RMSprop [20], AdaGrad [3], and Adam [7] et al. when training the DNN models. For future work, we would like to apply weight prediction to those popular optimization methods.

## References

1. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than cnns? *Advances in Neural Information Processing Systems* **34**, 26831–26843 (2021)
2. Chen, C.C., Yang, C.L., Cheng, H.Y.: Efficient and robust parallel dnn training through model parallelism on multi-gpu platform. *arXiv preprint arXiv:1809.02839* (2018)
3. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* **12**(7) (2011)
4. Guan, L., Yin, W., Li, D., Lu, X.: Xpipe: Efficient pipeline model parallelism for multi-gpu dnn training. *arXiv preprint arXiv:1911.04610* (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4700–4708 (2017)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
8. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
9. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265* (2019)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10012–10022 (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
12. Luo, L., Xiong, Y., Liu, Y., Sun, X.: Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843* (2019)
13. Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y.: Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies* **54**, 187–197 (2015)
14. Marcinkiewicz, M.A.: Building a large annotated corpus of english: The penn tree-bank. *Using Large Corpora* **273** (1994)
15. Nesterov, Y.: A method of solving a convex programming problem with convergence rate  $\mathcal{O}(1/k^2)$ . In: *Sov. Math. Dokl.* vol. 27
16. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics* **4**(5), 1–17 (1964)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
18. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: *International conference on machine learning*. pp. 1139–1147. PMLR (2013)

19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
20. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. University of Toronto, Technical Report **6** (2012)
21. Wang, P., Wang, X., Luo, H., Zhou, J., Zhou, Z., Wang, F., Li, H., Jin, R.: Scaled relu matters for training vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2495–2503 (2022)
22. Zaheer, M., Reddi, S., Sachan, D., Kale, S., Kumar, S.: Adaptive methods for nonconvex optimization. *Advances in neural information processing systems* **31** (2018)
23. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
24. Zhuang, J., Tang, T., Ding, Y., Tatikonda, S.C., Dvornek, N., Papademetris, X., Duncan, J.: Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems* **33**, 18795–18806 (2020)