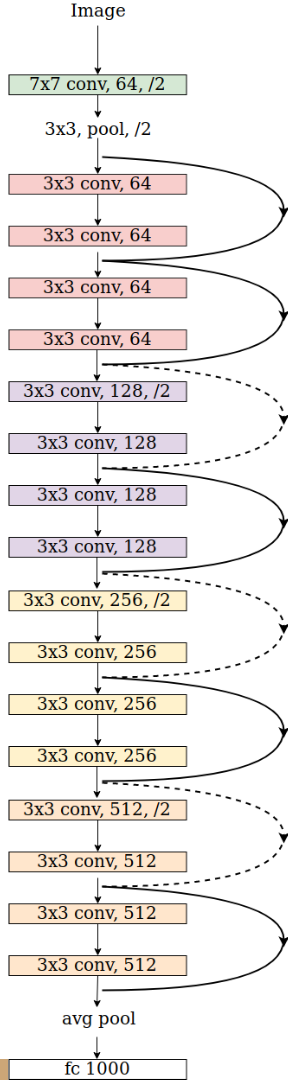
Performance of ResNet Architecture With Different SGD Optimizers

Tong Wu, tw2906
Yu Wu, yw3748

Residual neural network (ResNet)

- Deep neural network architecture
- Addresses the problem of vanishing gradients in very deep networks
- Introduces residual connections
- Allows the network to learn residual mappings
- Widely used in computer vision tasks
- Achive top performance in image classification, object detection and semantic segmentation



What did ResNet Optimize

- Vanishing gradients and exploding gradients problem
  - make network difficult to converge
- Degradation problem of deeper network
  - performance decline as number of layers increase, even if the network converge
- ResNet introduced residual connections
  - skip some intermediate layers
  - gradients can propagate through the residual connections

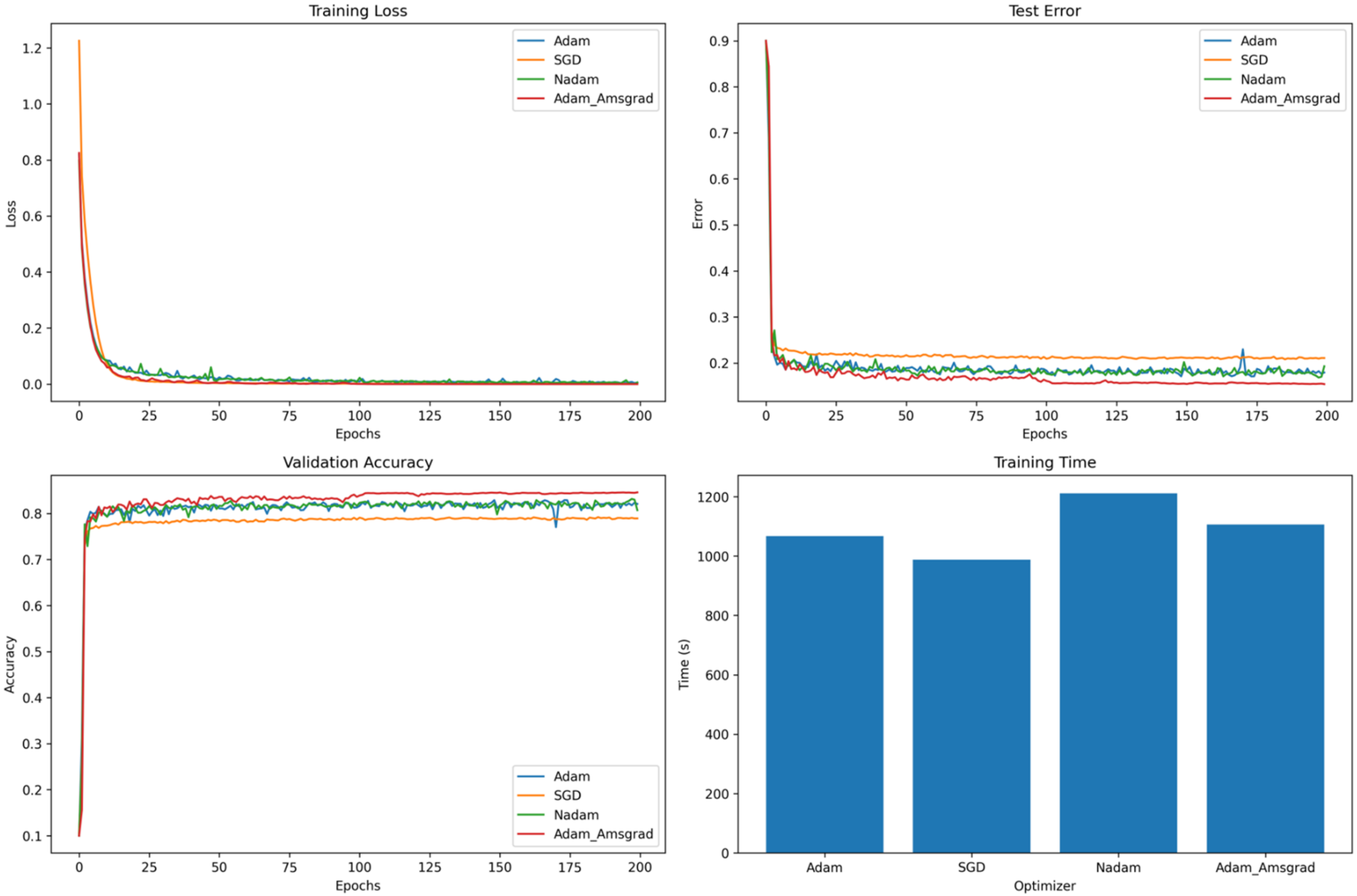
Optimizers

- Optimizer updates the network weight, minimize the loss function
- Helps network adapt, improve performance in tasks
- SGD, Adam, Nadam… May tunning learning rate manually
- Optimizer influence the converge speed, robustness and performance of the model
- Original ResNet paper use SGD
- Now more optimizers add to comparison

Paper: Analysis of Gradient Descent Optimization Algorithms on ResNet

- Discuss different optimizers performance mathematically
- Apply optimizers on ResNet architecture
- Train and test the model use CIFAR-10 dataset
- Compare the performance, including test error and train loss, of each optimizers.
- In this part, we will show the reproduce result of this paper and analysis it.

Reproduce results



Optimizers Analysis

- Adam: Compute adaptive learning rate for each parameter, has high performance.
- Nadam: Extension of Adam, introduce Nestrov acceleration speed up convergence and extra bias-correction step to reduce oscillation.
- AMSGard: Retains the maximum of all past second moment estimates. Correction factor to ensure the gradient average is not underestimated.
- SGD: only compute the loss function according to the gradient.

Adam:

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

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$$

Nadam:

$$g_t = \nabla_\theta J(\theta_{t-1}; x_t, y_t)$$

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

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$r_t = \frac{m_t}{\sqrt{v_t}+\epsilon}$$

$$\theta_t = \theta_{t-1} - \alpha \frac{r_t}{\sqrt{\sum_{i=1}^{t}(r_i^2)}}$$

AMSGard:

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

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$m_t^{\text{correction}} = \frac{m_t}{1-\beta_1^t}$$

$$v_t^{\text{correction}} = \frac{v_t}{1-\beta_2^t}$$

$$g_{t+\frac{1}{2}} = (1-\beta_1) g_t + \beta_1 m_t$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \left(g_{t+\frac{1}{2}} + \frac{\beta_1 g_t - \beta_1 m_t^{\text{correction}}}{1-\beta_1} \right)$$

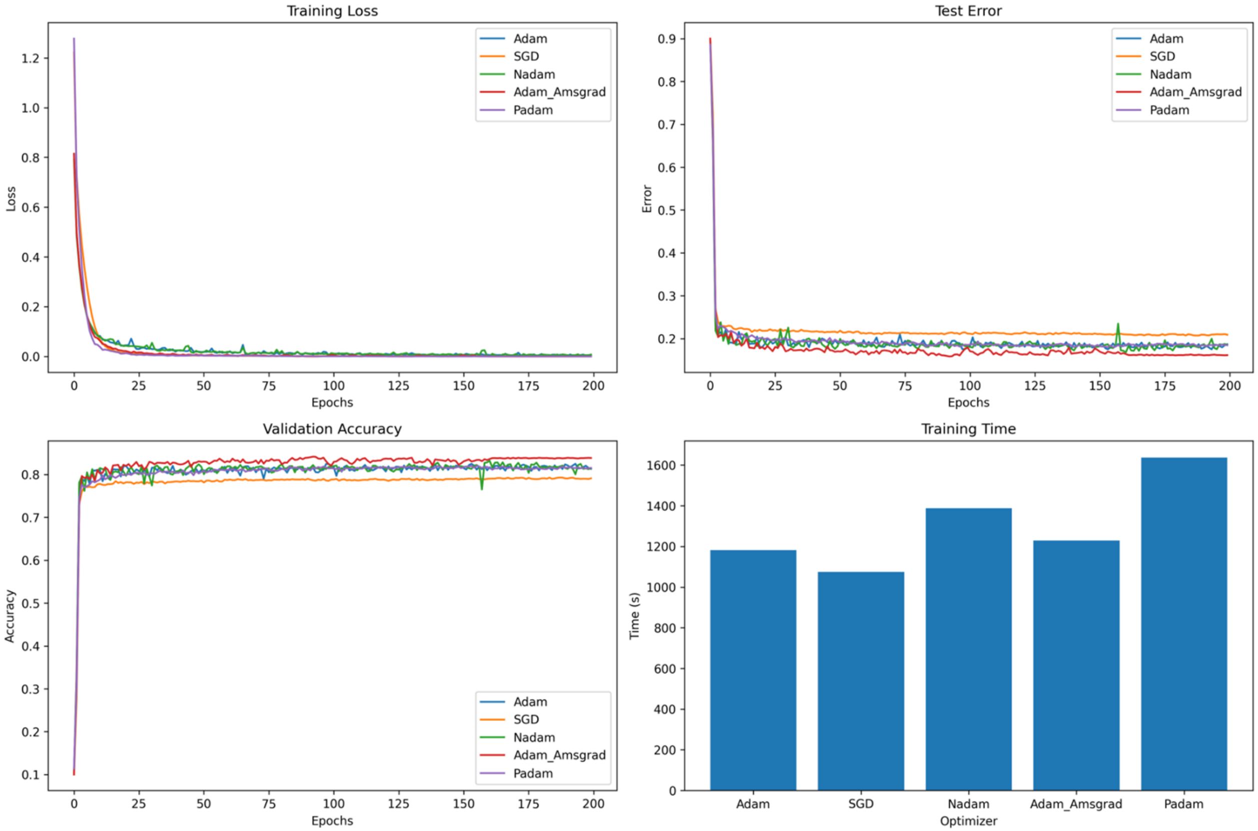
SGD:

$$\theta_t = \theta_{t-1} - \alpha \nabla_\theta J(\theta_{t-1}; x_t, y_t)$$

Padam

- Padam: new algorithm.
- Combines adaptive gradient methods & SGD.
- Introduces partial adaptive parameter.
- Addresses generalization gap.
- Fast convergence & good generalization.

Padam vs Other optimizers



Padam Analysis

- PADAM underperforms Adam_AMSGrad due to extra hyperparameter: momentum reservoir.
- Momentum reservoir affects convergence speed.
- Appropriate value crucial for convergence balance.
- Smoother test error line from adaptive momentum mechanism.
- Mechanism stabilizes optimization by updating based on gradient variance.

Conclusions

In this project, we:

- Discuss the performance of ResNet
- Reproduce a paper that compare performance of different optimizers on ResNet
- Analysis each optimizer’s performance with its update equation
- Discuss the improvement on Padam
- Implement Padam on ResNet and analysis its performance

Work Cited

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778).

[2] Grag, C., Grag, A., Raina, A. Analysis of Gradient Descent Optimization Algorithms on ResNet

[3] Chen, J., & Gu, Q. (2022). Padam: Closing the Generalization Gap of Adaptive Gradient Methods in Training Deep Neural Networks.