Deep Networks with Stochastic Depth

Authors: Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, Kilian Q. Weinberger
Group: Jose Vizueth, Prakriti Tandon, Rachael Close, Tanay Punjabi, Tasmin Sangha

Introduction

Training deep neural networks allows models to capture increasingly complex features, which improves the model's ability to generalize from raw input to high-level concepts. However, these models often face challenges such as v**anishing gradients, diminished feature reuse, and extensive training times.** The goal of our project is to reproduce key results from the paper "Deep Networks with Stochastic Depth" by Huang et al., more specifically Figure 3, which illustrates the improvements in test error and training efficiency on CIFAR-10 dataset using stochastic depth in residual networks (ResNets).

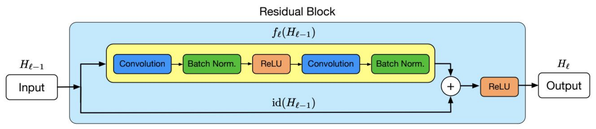


Fig. 1. A close look at the ℓ^{th} ResBlock in a ResNet.

We aim to reproduce the performance shown in Figure 3, where **stochastic depth lowers the test error compared to the traditional constant-depth ResNets**. By recreating these results, we will validate the authors' claims that stochastic depth effectively mitigates the vanishing gradient problem.

The paper introduces a novel training technique where ResNet layers are randomly dropped out during training according to a survival probability schedule. The main contribution of this paper include:

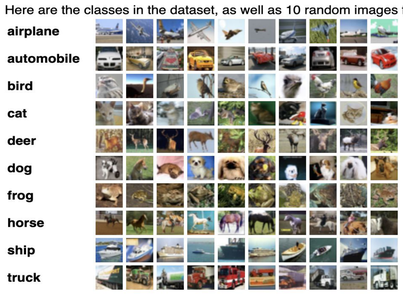
- Proposing this novel stochastic depth training approach for deep neural networks
- Demonstrating significant improvements in test error across different datasets such as CIFAR-10, CIFAR-1000, SVHN, and ImageNet
- Establishing stochastic depth as an effective method to mitigate the vanishing gradient problem.

Overall, we want deepen our understanding of this technique for training deep networks.

Dataset

CIFAR-10:

Here are the classes in the dataset, as well as 10 random images from each:



[2]

Methodology

- trained a **110 layer ResNet** with constant depth and another with stochastic depth
- trained with SGD, 500 epochs, batchsize = 128, initial learning rate 0.1 (divide by 10 at epochs 250 & 375)

Base Architecture (Constant Depth - Baseline ResNet-110)

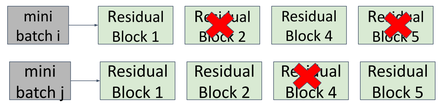
Always active. No skipping during training or testing

- 1 Convolution Layer
- 3 groups of 18 residual blocks (= 54 residual blocks)
  - each ResBlock has 2 Conv-BN layers
- final fully connected layer

= 110 layers

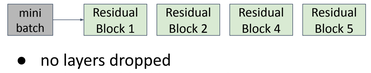
Model with Stochastic Depth (ResNetDrop)

ResNetDrop Training



- for a mini-batch, res block b has survival probability p_b
- if block b not dropped: $H_b = ReLU(f(H_{b-1}) + H_{b-1})$
- if dropped: $H_b = H_{b-1}$

ResNetDrop Inference



- no layers dropped
- scale each block's output by p_b

$H_b = ReLU(p_b f(H_{b-1}) + H_{b-1})$

- each block is active only with probability p_b

In expectation, 40 (out 54) Residual Blocks are active during training

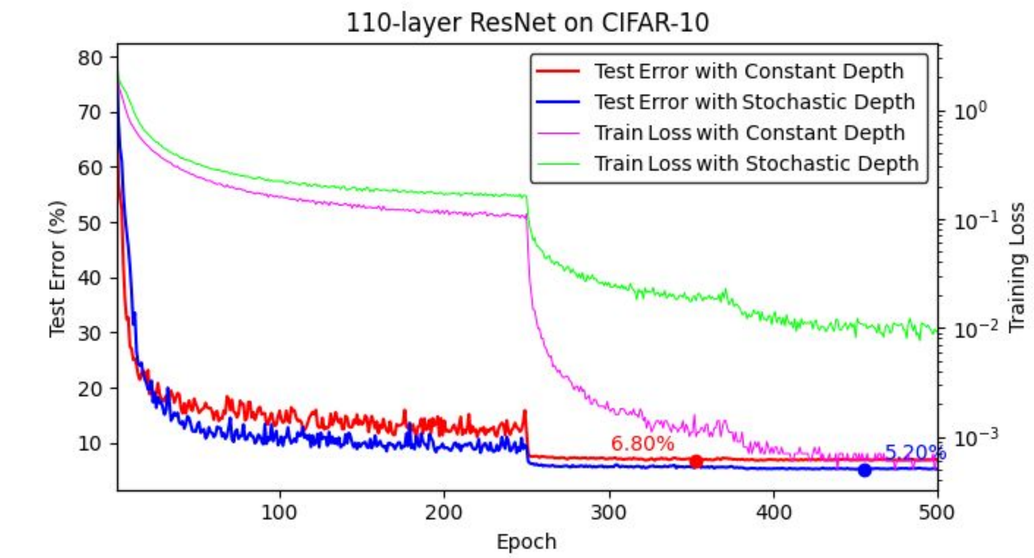
Survival Probability

- survival probability p_b linearly decays across network
- earlier block → higher chance to survive
- set $p_L = 0.5$ (L = last block)

for any residual block b

$$p_b = 1 - \frac{b}{L}(1 - p_L)$$

Results



Test error on CIFAR-10 using 110-layer ResNets. The points of lowest validation errors are highlighted for both models.

Performance:

- Lower test error with Stochastic Depth
- Train error higher with Stochastic Depth
- Slightly higher fluctuations with Stochastic Depth
- Errors matched those reported in the paper
  - Constant Depth: 6.41%; Stochastic Depth 5.25%

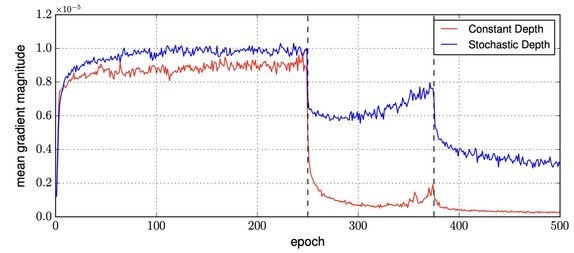
Timing (50 Epochs):

- Constant Depth: 57m 48s
- Stochastic Depth: 43m 26s

Discussion

Some things we observed but haven't plotted in our results:

- Stochastic depth reduces training time compared to constant-depth ResNets.
- Stochastic depth helps prevent vanishing gradients, resulting in stronger gradients during training. (Fig. 7)



mean gradient magnitude during constant and stochastic depth training (Fig 7 taken from [1])

Conclusion

The implementation of stochastic depth in deep neural networks demonstrates significant improvements in both performance and training efficiency:

- Stochastic depth significantly lowers test error compared to traditional constant-depth ResNets.
- Stochastic depth reduces training time.
- The approach successfully mitigates the vanishing gradient problem, showing consistently larger gradient magnitudes during training.

Future Work

Building on these findings, we could try exploring:

- Apply stochastic depth to even deeper networks (beyond 1202 layers)
- Test on larger and more complex datasets like ImageNet
- Explore different layer survival probability functions beyond linear decay
- Combine with other regularization techniques for further improvements

References

[1] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. 2016. Deep Networks with Stochastic Depth. arXiv preprint arXiv:1603.09382. https://arxiv.org/abs/1603.09382
[2] CIFAR-10 and CIFAR-100 Datasets, www.cs.toronto.edu/~kriz/cifar.html.