Deep Networks with Stochastic Depth

Gao Huang*[1], Yu Sun*[1], Zhuang Liu[2], Daniel Sedra[1] , Kilian Weinberger[1]

[1]Cornell University, [2]Tsinghua University

*Equal contribution

Motivation

Training very deep networks is difficult:

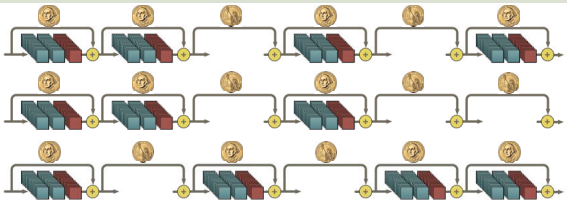
- Gradients vanish and forward signals diminishes
- Long training time
- Overfitting

Question: Can we use short networks during *training*, but use deep networks during *testing*?

Idea: For each mini-batch, randomly **drop** a subset of layers and bypass them with the identity function!

Method

Stochastic depth network at training time



A subset of layers are dropped at each mini-batch

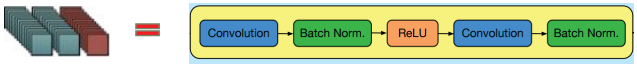
$$H_\ell = \texttt{ReLU}(b_\ell f_\ell(H_{\ell-1}) + \texttt{id}(H_{\ell-1}))$$

b_ℓ: Bernoulli random variable

- Linear decay rule for survival probabilities

$$b_\ell \sim \texttt{Bernoulli}(p_\ell) \quad \text{with} \quad p_\ell = (1 - \frac{\ell}{L}) \times 1 + \frac{\ell}{L} \times p_L$$

- Basic block (Similar to ResNets, He et al, CVPR'16)



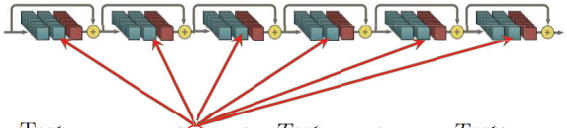
Expected network depth

$$E(\tilde{L}) = \sum_{\ell=1}^{L} p_\ell = (3L-1)/4 \approx 3L/4$$

~25% shorter

Stochastic depth network at test time

At test time



$$H_\ell^{\text{Test}} = \texttt{ReLU}(p_\ell f_\ell(H_{\ell-1}^{\text{Test}}; W_\ell) + H_{\ell-1}^{\text{Test}})$$

All layers are on, but outputs of f_ℓ are down weighted by their corresponding survival probabilities.

Advantages of stochastic depth

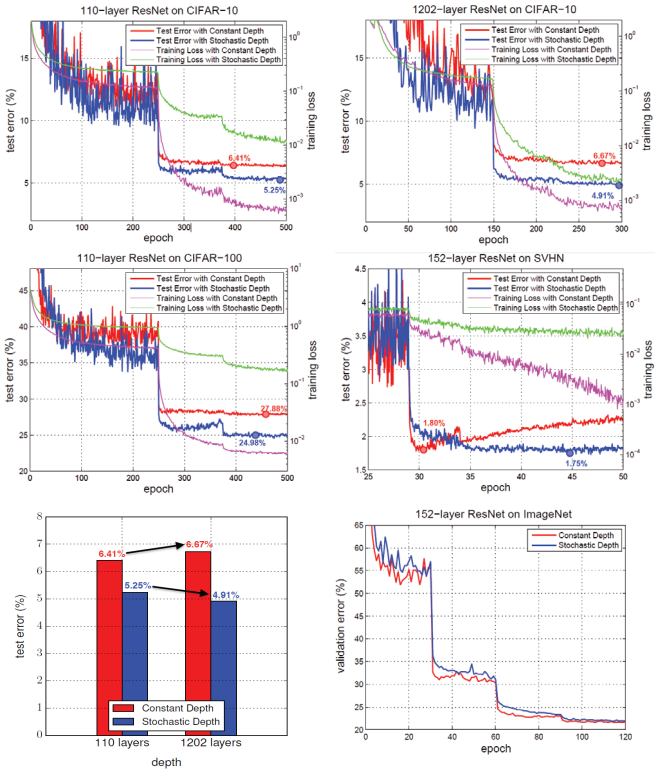
- Alleviates the gradient and signal vanishing problem
- Speeds up the training process
- Performs regularization and improves generalization (implicit ensemble of 2^L models)

Code

https://github.com/yueatsprograms/Stochastic_Depth

Results

Classification



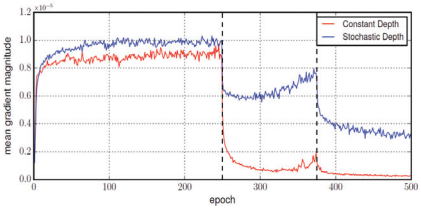
Training time

	CIFAR10+	CIFAR100+	SVHN
Constant Depth	20h 42m	20h 51m	33h 43m
Stochastic Depth	15h 7m	15h 20m	25h 33m

~25% faster

Analysis

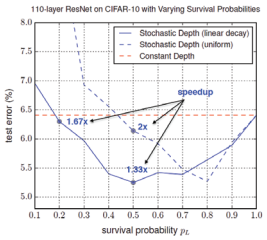
Gradient strength



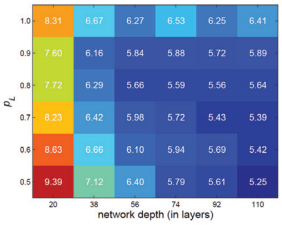
The gradient strength at the input layer

Hyper-parameter p_L

Varying p_L with fixed depth



Varying p_L with different depth



Extension (DenseNets)

Densely Connected Convolutional Networks (https://arxiv.org/abs/1608.06993)

-From implicit long-range connections to explicit long-range connections

-Learn more *compact* models!

-And more *accurate*!