Wide Residual Networks

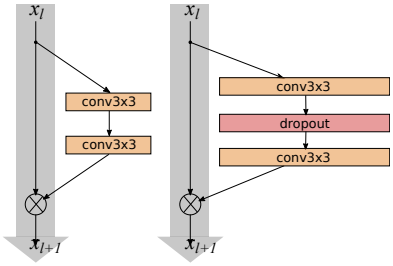
Sergey Zagoruyko
sergey.zagoruyko@enpc.fr
Nikos Komodakis
nikos.komodakis@enpc.fr

Université Paris-Est, École des Ponts ParisTech
Paris, France

Deep residual networks were shown to be able to scale up to thousands of layers and still have improving performance. However, each fraction of a percent of improved accuracy costs nearly doubling the number of layers, and so training very deep residual networks has a problem of diminishing feature reuse, which makes these networks very slow to train. To tackle these problems we conduct a detailed experimental study on the architecture of ResNet blocks, based on which we propose a novel architecture where we decrease depth and increase width of residual networks. In addition, we propose a new way of utilizing dropout within deep residual networks so as to properly regularize them and prevent overfitting during training. We call the resulting network structures wide residual networks (WRNs) and show that these are far superior over their commonly used thin and very deep counterparts. Our experiments show that:

- our widened architecture consistently improves performance across residual networks of different depth;
- increasing both depth and width helps until the number of parameters becomes too high and stronger regularization is needed;
- there doesn't seem to be a regularization effect from very high depth in residual networks as wide networks with the same number of parameters as thin ones can learn same or better representations. Furthermore, wide networks can successfully learn with a lot more parameters than thin ones, which would require doubling the depth of thin networks, making them infeasibly expensive to train.

Overall, we demonstrate that even a simple 16-layer-deep wide residual network outperforms in accuracy and efficiency all previous deep residual networks, including thousand-layer-deep networks, achieving new state-of-the-art results on CIFAR-10, CIFAR-100 and SVHN (table 1). Our code is available at https://github.com/szagoruyko/wide-residual-networks.



(a) basic (b) wide-dropout

Figure 1: Basic and wide-dropout residual blocks. Batch normalization and ReLU precede each convolution

	depth-k	CIFAR-10	CIFAR-100
NIN		8.81	35.67
FitNet		8.39	35.04
Highway [4]		7.72	32.39
ResNet[1]	110	6.43	25.16
	1202	7.93	27.82
stoc-depth[3]	110	5.23	24.58
	1202	4.91	-
pre-ResNet[2]	110	6.37	-
	164	5.46	24.33
	1001	4.64	22.71
WRN (ours)	40-4	4.97	22.89
	16-8	4.81	22.07
	28-10	**4.17**	**20.50**

Table 1: Test error on CIFAR-10 and CIFAR-100 with moderate data augmentation (flip/translation). k is a widening factor. We don't use dropout for these results.

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016.

[3] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. *CoRR*, abs/1603.09382, 2016.

[4] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.