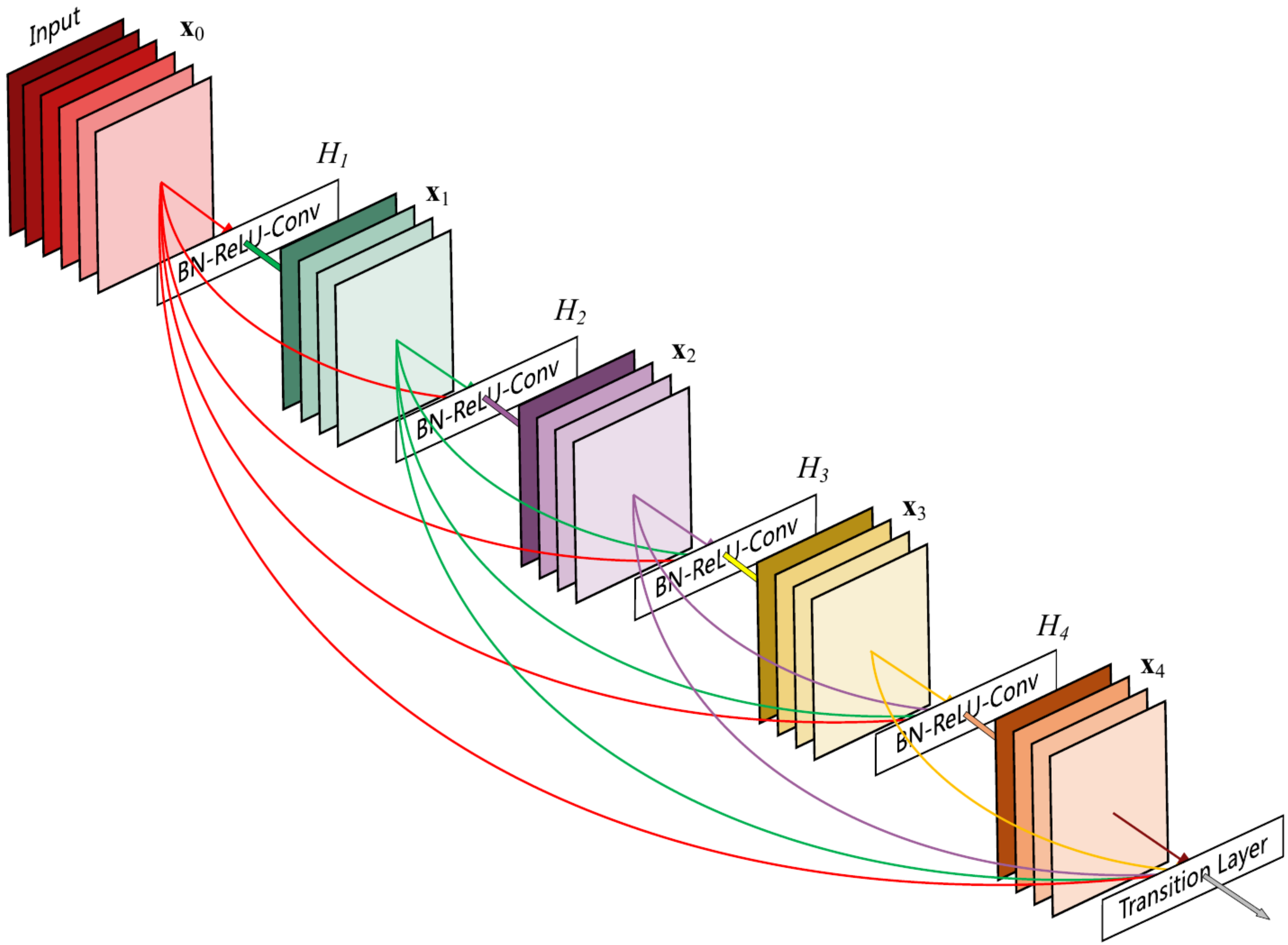
Densely Connected Convolutional Networks

presented by Elmar Stellnberger

a 5-layer dense block, k=4



Densely Connected CNNs

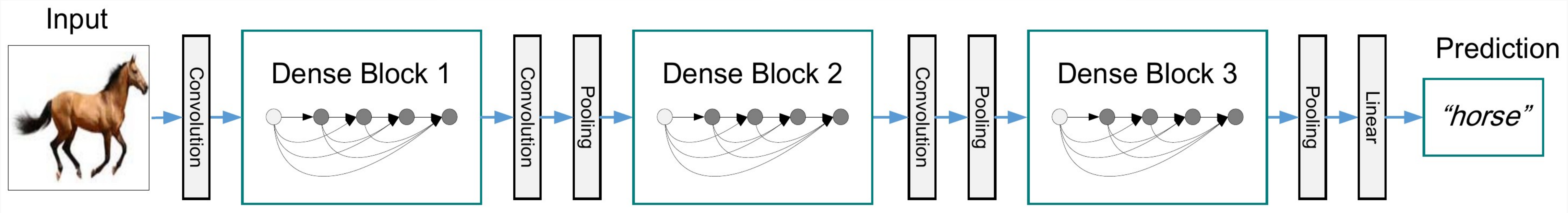
- better feature propagation & feature reuse
- alleviate the vanishing gradient problem
- parameter-efficient
- less prone to overfitting even without data augmentation
- naturally scale to hundreds of layers yielding a consistent improvement in accuracy

DenseNet Architecture

- Traditional CNNs: $x_l = H_l(x_{l-1})$
- ResNets: $x_l = H_l(x_{l-1}) + x_{l-1}$
- DenseNets: $x_l = H_l([x_0, x_1, .., ..., x_{l-2}, x_{l-1}])$
- $H_l(x)$ in DenseNets ~ Batch Normalization (BN), rectified linear units (ReLU), 3x3 Convolution
- $k_0 + k \cdot (l-1)$ input activation maps for layer l

  but: data reduction required, f.i. by max-pooling with stride $\geqslant 2$

DenseNet Architecture



- only dense blocks are fully connected
- between dense blocks: convolution & 2x2 average pooling

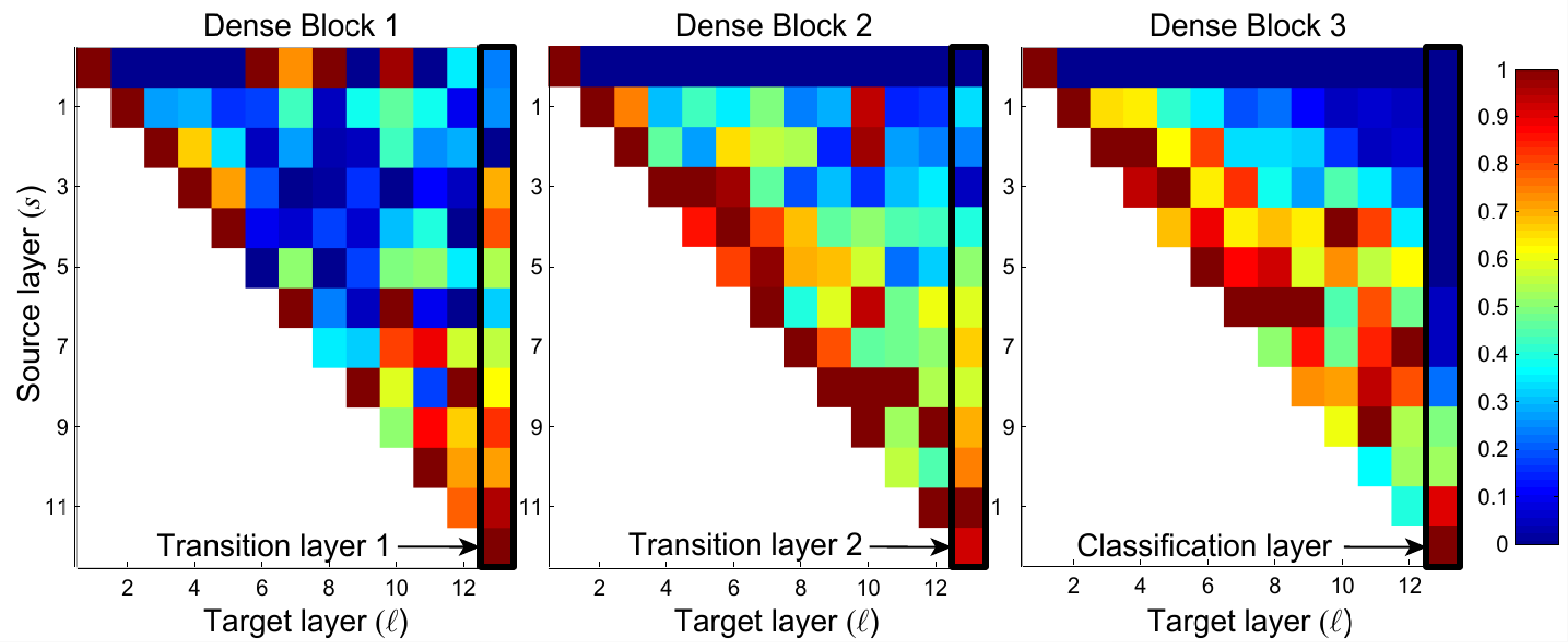
  → transition layers

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112×112	7×7 conv, stride 2			
Pooling	56×56	3×3 max pool, stride 2			
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56×56	1×1 conv			
	28×28	2×2 average pool, stride 2			
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28×28	1×1 conv			
	14×14	2×2 average pool, stride 2			
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 64$
Transition Layer (3)	14×14	1×1 conv			
	7×7	2×2 average pool, stride 2			
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$
Classification Layer	1×1	7×7 global average pool			
		1000D fully-connected, softmax			

DenseNet Variants

- DenseNet-B: 1x1 convolution bottleneck layer (including BN & ReLU activation function), reduces the number of input feature maps, more computationally efficient
- DenseNet-C: compression at transition layers, here: $\theta = 0.5$, only ½ of the activation maps are forwarded
- DenseNet-BC

average abs. filter weights



Comparable Architectures

- Identity connections: Highway Networks: gating units, ResNets: $x_l = H_l(x_{l-1}) + x_{l-1}$
- +width & + depth: GoogleNets: 5x5, 3x3, 1x1 convolution and 3x3 pooling in parallel
- Deeply-Supervised Nets: classifiers at every layer
- Stochastic depth: drop layers randomly

→ shorter paths from the beginning to the end which do not pass through all layers

Experiments & Evaluation

- CIFAR data set (C10, C100), +data augemntation C10+, C100+ (mirroring, shifting), training/test/validation = 50,000/10,000/5,000
- SVHN: Street View House Numbers, training/test/validation = 73,000/26,000/6,000, relatively easy task
- ImageNet: 1,2 million images for training, 50,000 for validation

ImageNet results

- 4 dense blocks instead of three
- no comparison with performance of other arches
- bottom: Deeply-Supervised Nets

Model	top-1	top-5
DenseNet-121	25.02 / 23.61	7.71 / 6.66
DenseNet-169	23.80 / 22.08	6.85 / 5.92
DenseNet-201	22.58 / 21.46	6.34 / 5.54
DenseNet-264	22.15 / 20.80	6.12 / 5.29

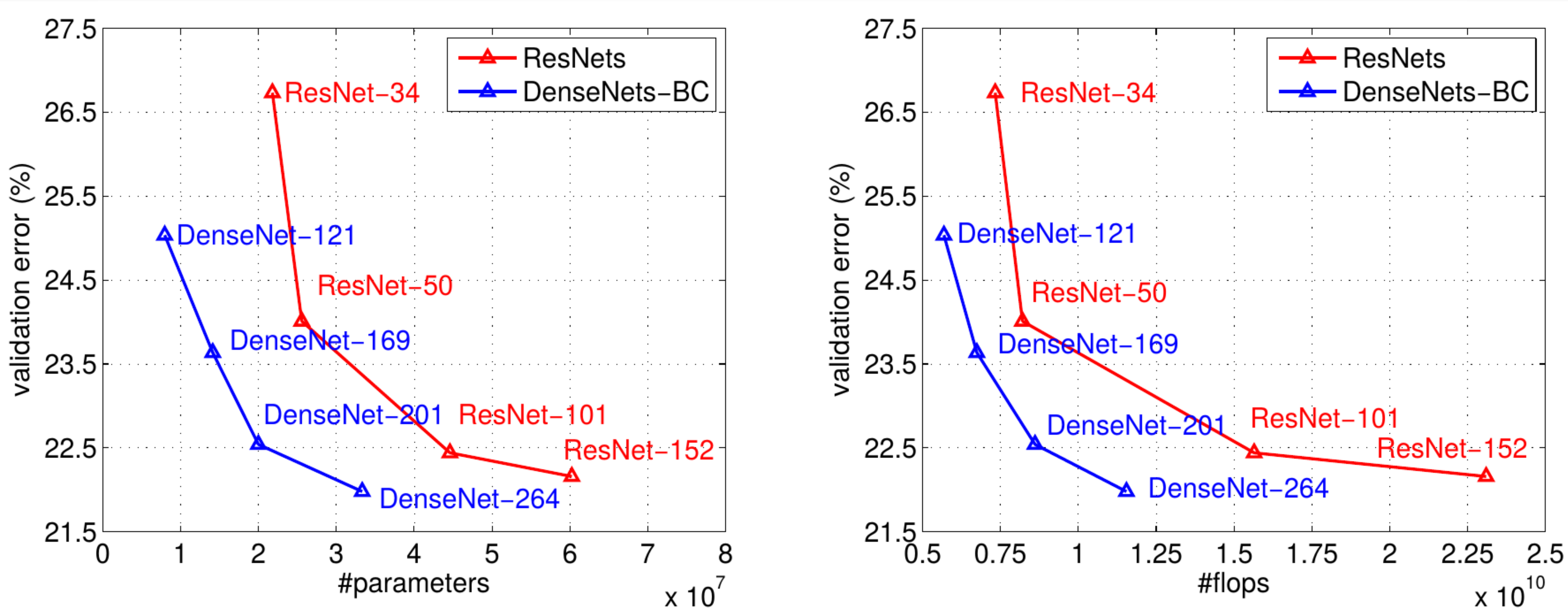
Table 5: ImageNet 2012 classification error.

Method	top-1 val. error(%)	top-5 val. error(%)
CNN 8-layer [14]	40.7	18.2
DSN 8-layer (ours)	39.6	17.8
CNN 11-layer	34.5	13.9
DSN 11-layer (ours)	33.7	13.1

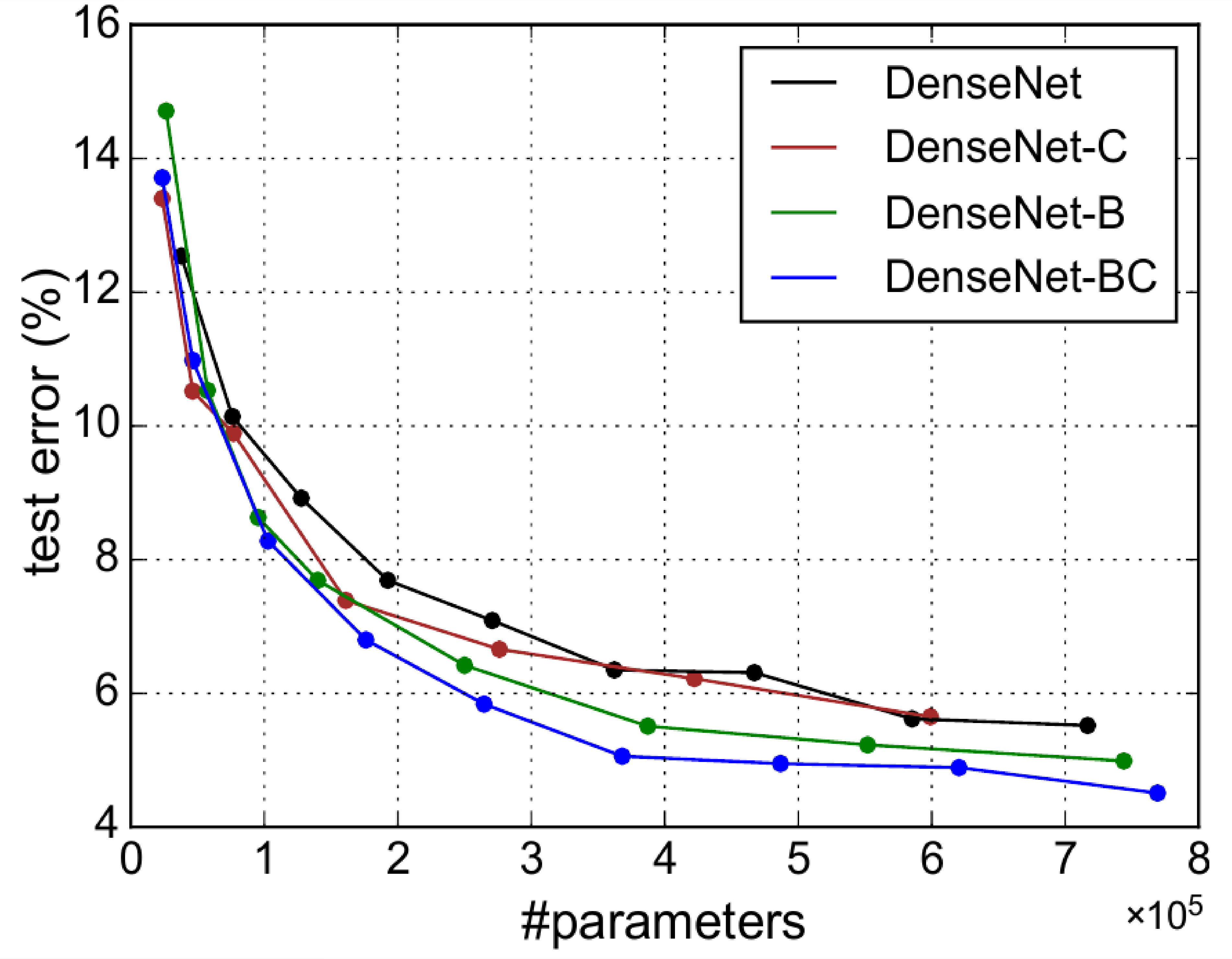
Method	Depth	Params	C10	C10+	C100	C100+	SVHN
Network in Network [22]	-	-	10.41	8.81	35.68	-	2.35
All-CNN [32]	-	-	9.08	7.25	-	33.71	-
Deeply Supervised Net [20]	-	-	9.69	7.97	-	34.57	1.92
Highway Network [34]	-	-	-	7.72	-	32.39	-
FractalNet [17]	21	38.6M	10.18	5.22	35.34	23.30	2.01
with Dropout/Drop-path	21	38.6M	7.33	4.60	28.20	23.73	1.87
ResNet [11]	110	1.7M	-	6.61	-	-	-
ResNet (reported by [13])	110	1.7M	13.63	6.41	44.74	27.22	2.01
ResNet with Stochastic Depth [13]	110	1.7M	11.66	5.23	37.80	24.58	1.75
	1202	10.2M	-	4.91	-	-	-
Wide ResNet [42]	16	11.0M	-	4.81	-	22.07	-
	28	36.5M	-	4.17	-	20.50	-
with Dropout	16	2.7M	-	-	-	-	1.64
ResNet (pre-activation) [12]	164	1.7M	11.26*	5.46	35.58*	24.33	-
	1001	10.2M	10.56*	4.62	33.47*	22.71	-
DenseNet ($k = 12$)	40	1.0M	**7.00**	5.24	**27.55**	24.42	1.79
DenseNet ($k = 12$)	100	7.0M	**5.77**	**4.10**	**23.79**	**20.20**	1.67
DenseNet ($k = 24$)	100	27.2M	**5.83**	**3.74**	**23.42**	**19.25**	**1.59**
DenseNet-BC ($k = 12$)	100	0.8M	**5.92**	4.51	**24.15**	22.27	1.76
DenseNet-BC ($k = 24$)	250	15.3M	**5.19**	**3.62**	**19.64**	**17.60**	1.74
DenseNet-BC ($k = 40$)	190	25.6M	-	**3.46**	-	**17.18**	-

Evaluation Results

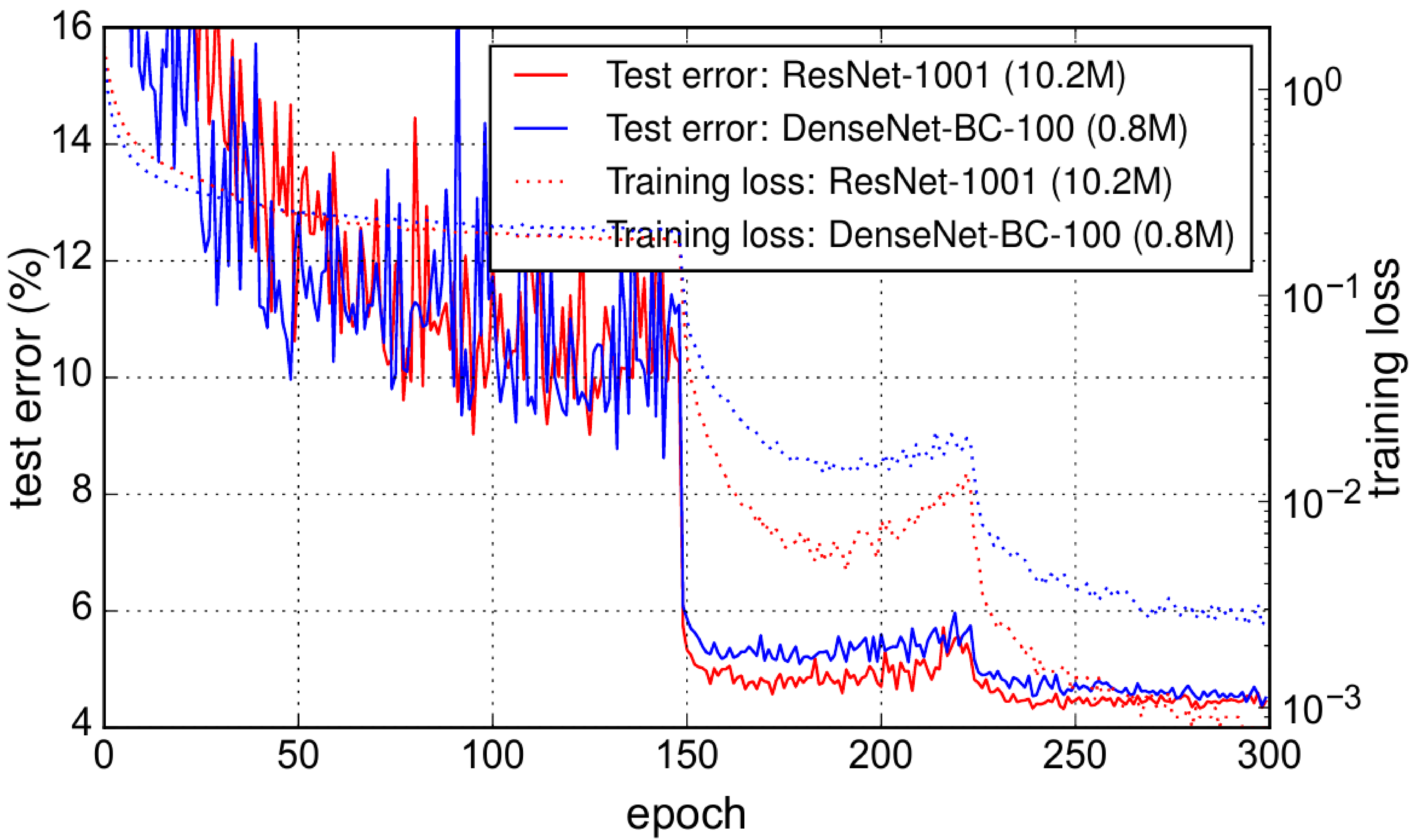
- CIFAR: DenseNet-BC better, SVHN: DenseNet
- better performance as L (deepness) & k (growth factor) increase
- more efficient usage of parameters: better performance with same number of parameters
- less prone to overfitting: differences are particularely pronounced for the data sets without data augmentation



more parameter efficient,
less computationally itensive



C10+ data set: comparison of DenseNet variants


Test error: ResNet-1001 (10.2M)
Test error: DenseNet-BC-100 (0.8M)
Training loss: ResNet-1001 (10.2M)
Training loss: DenseNet-BC-100 (0.8M)
test error (%)
training loss
epoch
16
14
12
10
8
6
4
0
50
100
150
200
250
300
10^0
10^{-1}
10^{-2}
10^{-3}

G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, “Densely Connected Convolutional Networks”, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700-4708.

C-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, “Deeply-Supervised Nets”, in AISTATS 2015.