DenseNets & DelugeNets

Jiyuu Yi

Densely Connected Convolutional Networks

Gao Huang, Zhuang Liu, Kilian Q. Weinberger

Motivation1

- Problem of Deep Networks: vanishing gradient

- Solutions to improve the information propagation:
  - Companion loss (auxiliary classifier)
  - Skip connection
    - Highway networks
    - ResNet-v2 (pre-activation residual networks)
    - Wide Residual networks

Motivation2

- Problems of Residual networks
  - Simple addition (residual) operation of ResNet
    - May squash useful features of preceding layers ➔ hinders feature reuse
  - Cross-layer connection between preceding and succeeding layers are not selective
    - Hinders learning cross-layer interactions and correlations
  - Stochastic depth improved the performance of ResNet
    - Shows that not all layers may be needed => great amount of redundancy

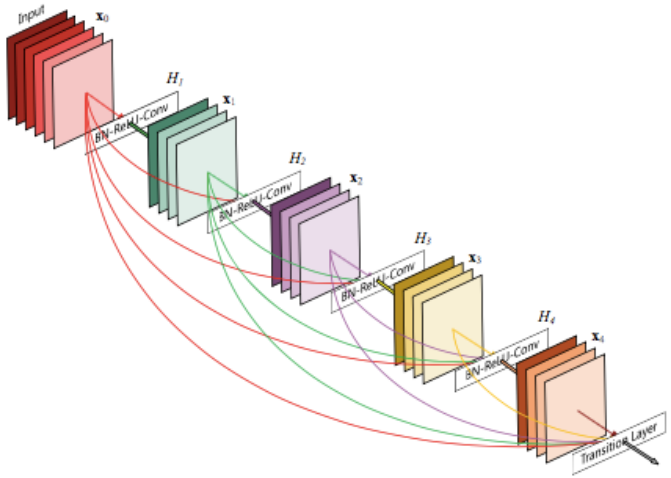
DenseNets: Dense connectivity

- Identity function of ResNets

$$\mathbf{x}_\ell = H_\ell(\mathbf{x}_{\ell-1}) + \mathbf{x}_{\ell-1}$$

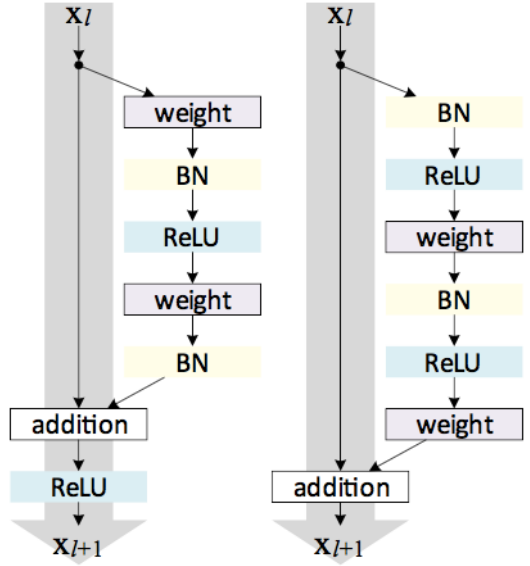
- Dense connectivity
  - Direct connections from any layer to all subsequent layers
  - Concatenation instead of addition operation

$$\mathbf{x}_\ell = H_\ell([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}])$$

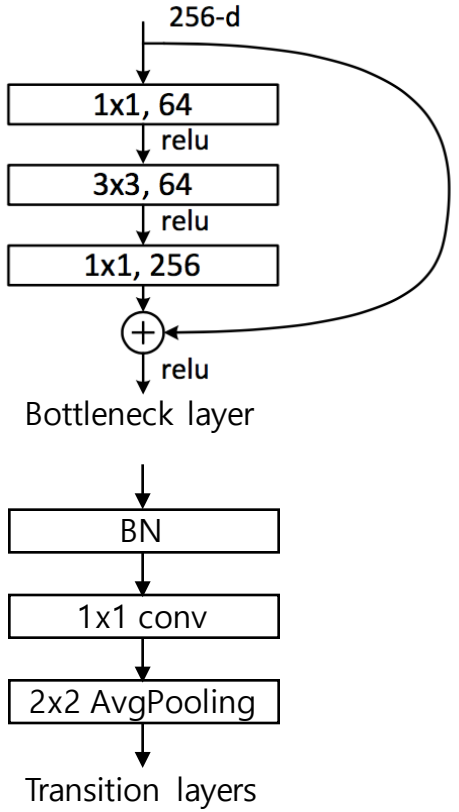


DenseNets: Architecture

- Composite layers
- Transition layers
- Bottleneck layers

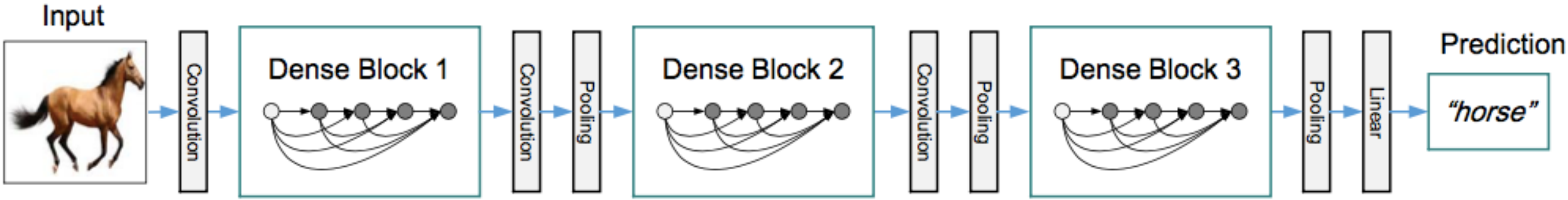


Composite layer



Bottleneck layer

Transition layers

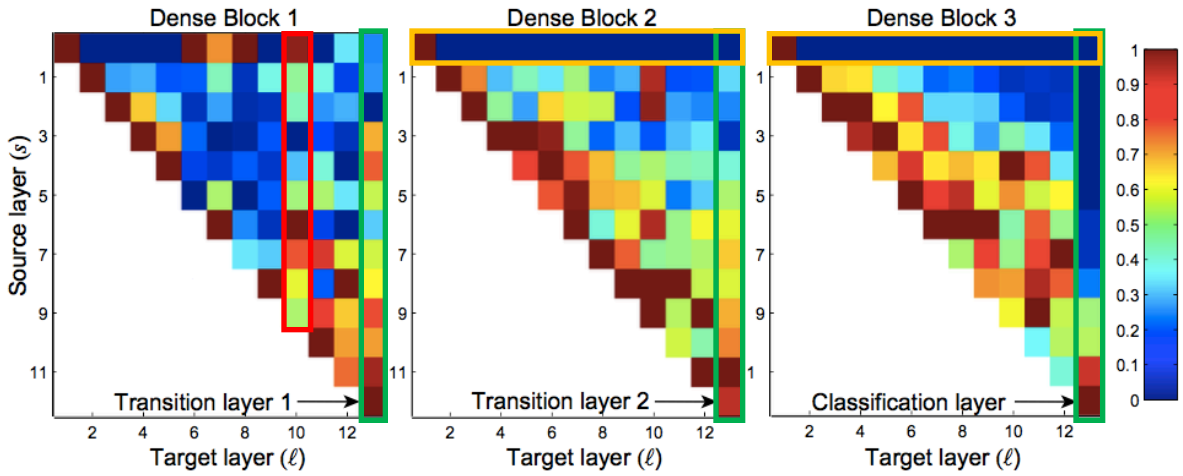


DenseNets: Experiments & Advantages

Method	Depth	Params	C10	C10+	C100	C100+	SVHN
Network in Network [22]	-	-	10.41	8.81	35.68	-	2.35
All-CNN [31]	-	-	9.08	7.25	-	33.71	-
Deeply Supervised Net [20]	-	-	9.69	7.97	-	34.57	1.92
Highway Network [33]	-	-	-	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 [41]	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**	-

- Model capacity
  - Bigger models have increased representational power
- Parameter efficiency
  - Less prone to overfitting

DenseNets: Experiments & Advantages



Average L1 norm of the weights connecting convolutional layers s to layers ℓ

- Feature reuse
  - All layers spread their weights
    - ➔ features extracted from early layers are used directly
  - The weights of transition layers also spread their weights across all layers
    - ➔ Information flow from the first to the last layers within dense block
  - The layers within the second and third dense block assign the least weight to the outputs of transition layers
    - ➔ Transition layers output many redundant features
    - ➔ Compressing output of transition layers help make better results

DelugeNets: Deep Networks with Massive and Flexible Cross-layer Information Inflows

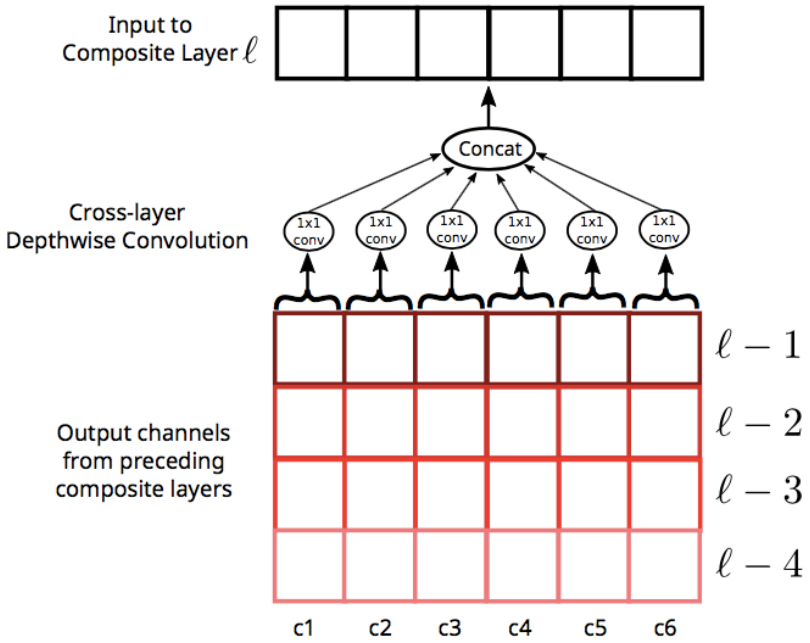
Jason Kuen, Xiangfei Kong, Gang Wang

Motivation

- It is crucial to consider width of network for its representational power
  - S. Zagoruyko and N. Komodakis, **Wide residual networks**, BMVC, 2016
- DenseNets have heavy amount of parameters
  - [filter_width * filter_height * #input_channel * #output_channel * #preceding_layers]
  - Fatally DenseNets are configured to have lower output width to deal with excessive parameter growth.

DelugeNets: Cross-layer Depthwise Convolutional Layers

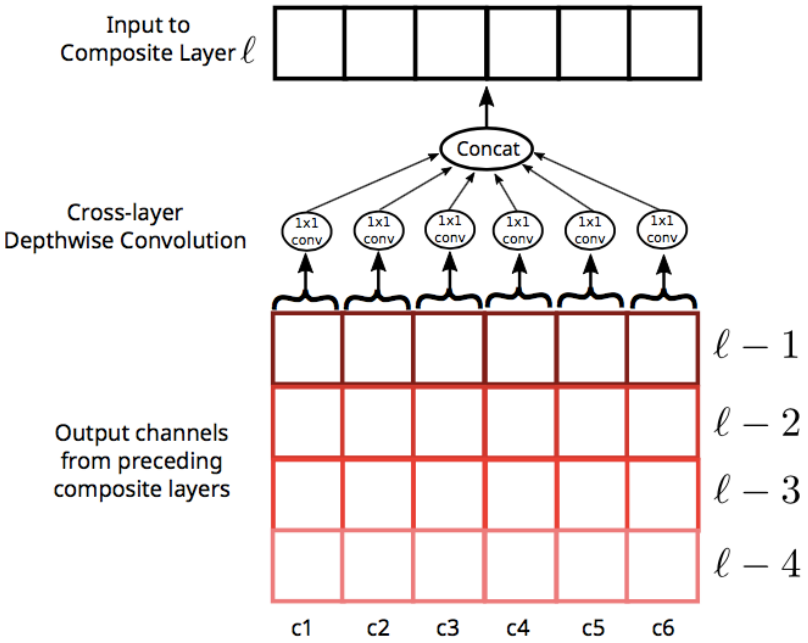
- Input: concatenated channels of feature map outputs of many layers
- Use (channel, spatial) independent filters
  - Can learn cross-layer interaction
- Facilitate the inflows of information from preceding composite layer to succeeding layer
- Additional # of parameters:
  - 1 * 1 * #_of_preceding_layers * #_of_channel
  - Much smaller # compared to DenseNets
    - width x height x (#_of_in_channel x (#_of_preceding_layers -1)) x #_of_out_channel



$$x_\ell^c = \sum_{i=1}^{N+1} w_{\ell-i}^c h_{\ell-i}^c + b_\ell^c$$

DelugeNets: Cross-layer Depthwise Convolutional Layers

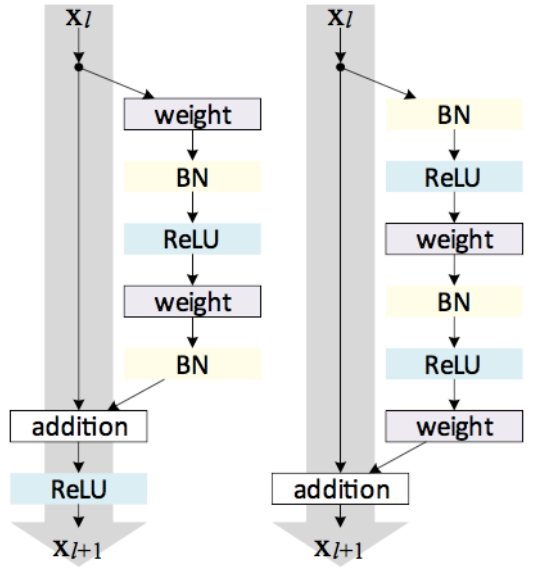
- Advantages:
  - Feature reuse: encourages features to be taken as input for many times
  - Parameter efficiency
    - No need to redundantly learn filters which generates same features in succeeding layers
  - Unique gradient signals
    - Cross-layer Depthwise Conv layers: multiplicative interaction with filter
    - ResNet: simple addition



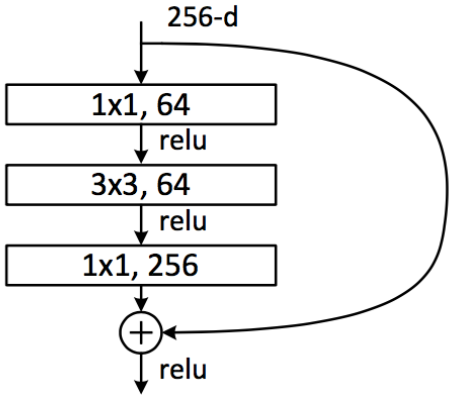
$$x_\ell^c = \sum_{i=1}^{N+1} w_{\ell-i}^c h_{\ell-i}^c + b_\ell^c$$

DelugeNets: Architecture

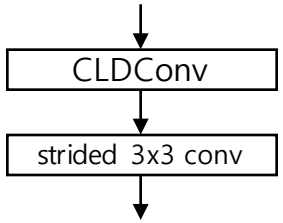
- Composite layers
- Bottleneck layers
- Transition layers
  - Transforms the summarized feature maps
  - 2 strided 3x3 conv for preserving information



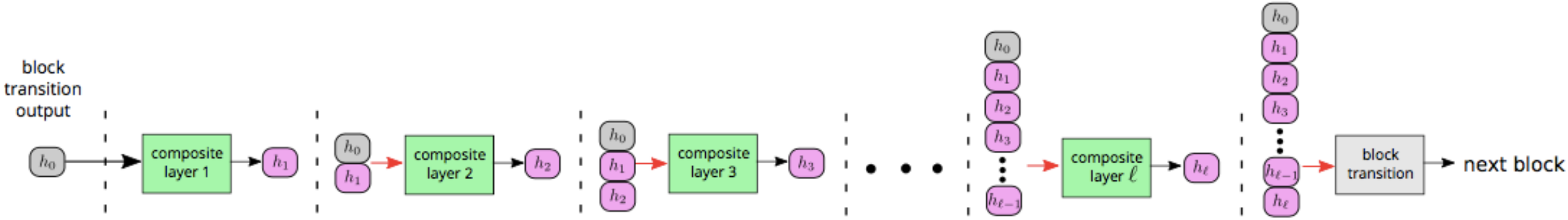
Composite layer



Bottleneck layer



Transition layers



DelugeNets: Experiment

Model	#Params	Depth	CIFAR-10	CIFAR-100
Maxout Network [7]	-	-	9.38	38.57
Network-in-Network [24]	-	-	8.81	35.68
Deeply Supervised Net [23]	-	-	7.97	34.57
All-CNN [30]	-	-	7.25	33.71
Fractional Max-Pooling [8]	-	-	4.50	27.62
ELU-Net [4]	-	-	6.55	24.28
Highway Network [31]	-	-	7.60	32.24
FractalNet [21]	38.6M	20	4.59	22.85
ResNet [12]	1.7M	164	5.93	25.16
ResNet [12]	10.2M	1001	7.61	27.82
ResNet with ELU [27]	-	110	5.62	26.55
ResNet with Identity Mappings [13]	1.7M	164	5.46	24.33
ResNet with Identity Mappings [13]	10.2M	1001	4.62	22.71
ResNet with Swapout [29]	7.4M	32	4.76	22.72
ResNet with Stochastic Depth [17]	1.7M	32	5.23	24.98
ResNet with Stochastic Depth [17]	10.2M	1202	4.91	-
Wide-ResNet (04×width) [39]	8.7M	40	4.97	22.89
Wide-ResNet (08×width) [39]	11.0M	16	4.81	22.07
Wide-ResNet (10×width) [39]	36.5M	28	4.17	20.50
DenseNet ($k = 12$) [16]	7.0M	100	4.10	20.20
DenseNet ($k = 24$) [16]	27.2M	100	**3.74**	**19.25**
DelugeNet-146	6.7M	146	3.98	19.72
DelugeNet-218	10.0M	218	3.88	19.31
Wide-DelugeNet-146	20.2M	146	**3.76**	**19.02**