University of Antwerp

Artificial Neural Networks

[2500WETANN]

José Oramas

Convolutional Neural Networks

[Part 2 – Relevant Architectures & Components]

José Oramas

University of Antwerp

Announcement

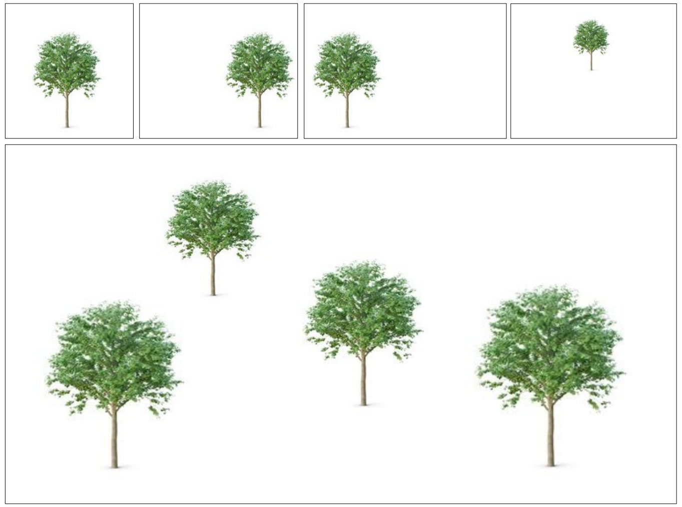
- **Research Paper Assignment**
  - Groups of two students
  - Submission
    - 26/03/2025
    - Send group information via email (add “[RPA]” in the subject of your email)
  - 27/03/2025: students without a group will be randomly assigned.

Recap: Convolutional Neural Networks [CNNs, ConvNets]

Some Characteristics of Visual Data

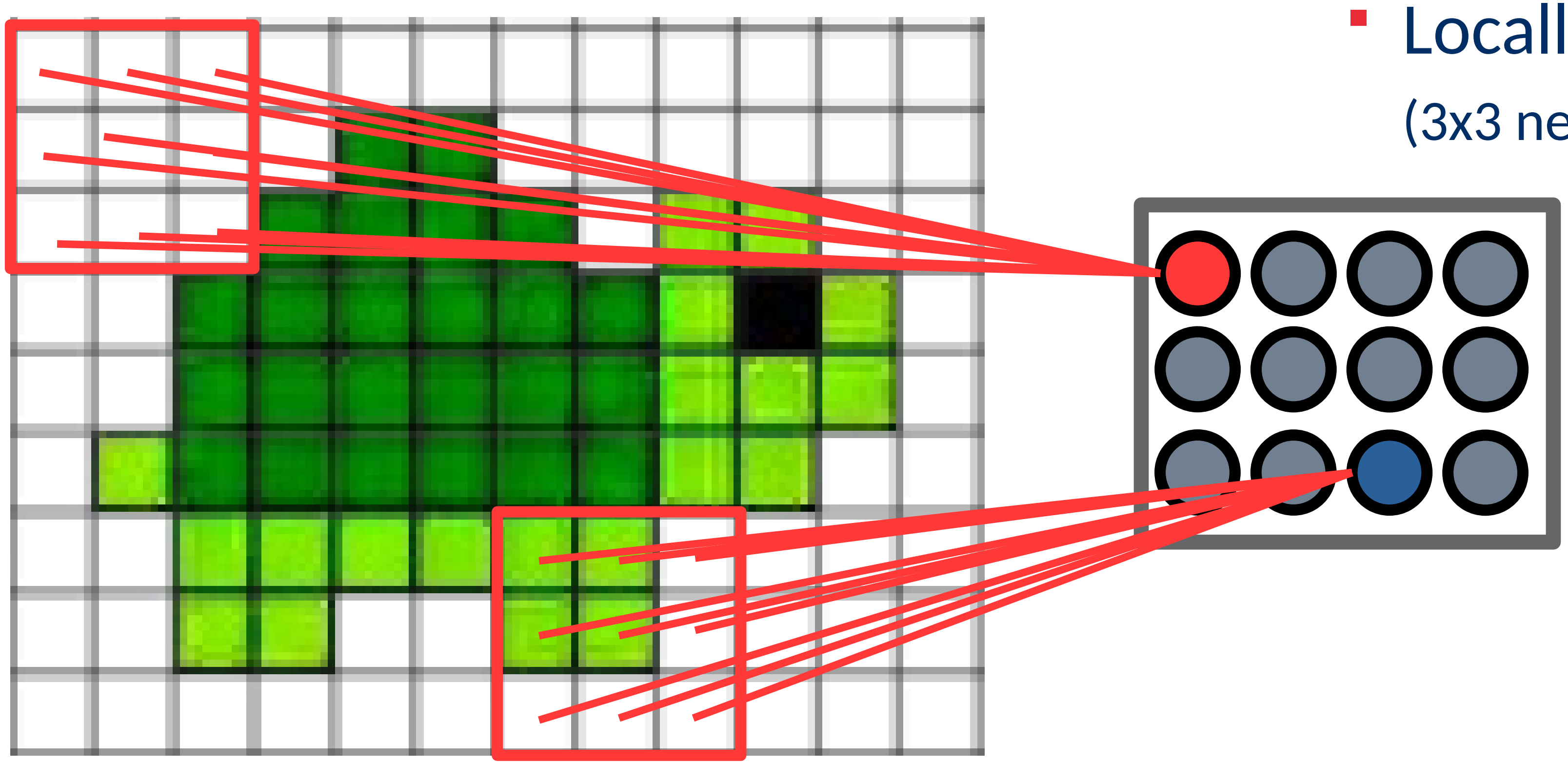
Locality

Translation Invariance



Compositionality

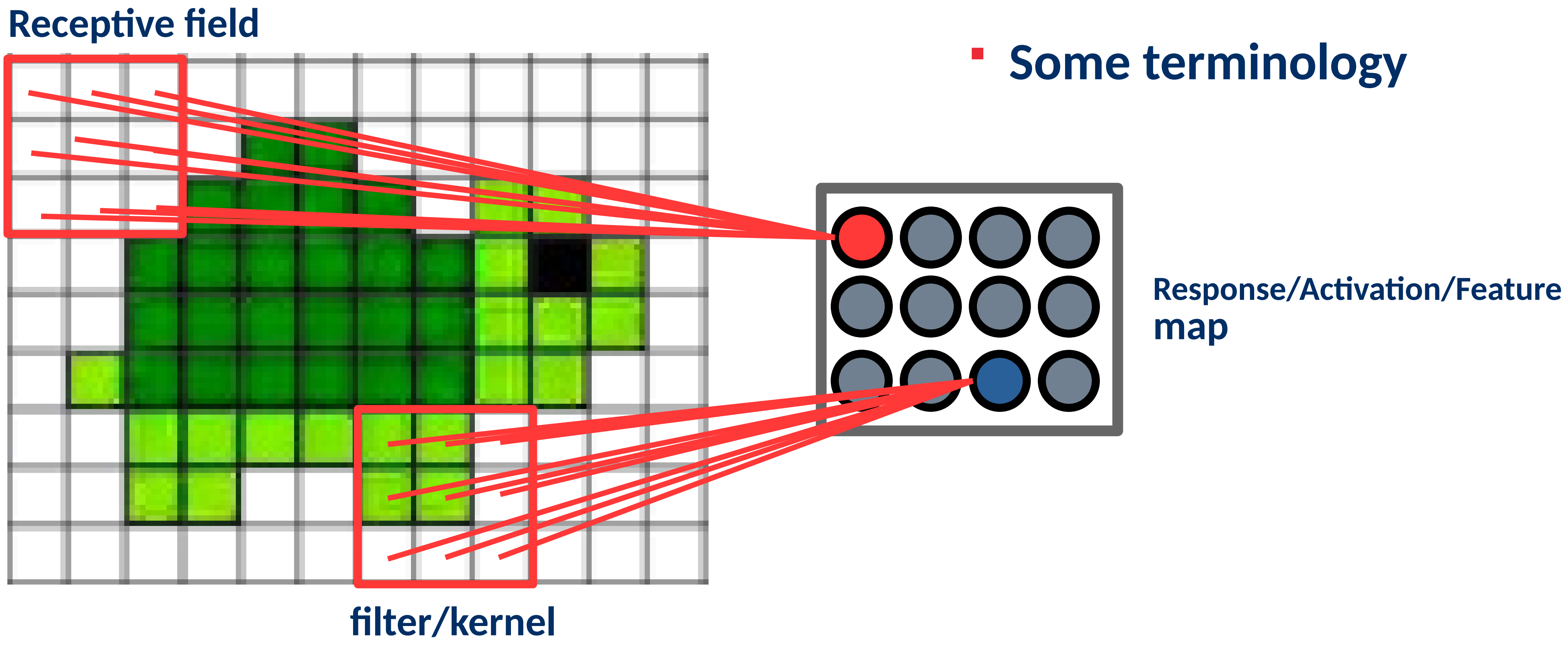
Recap: Convolutional Neural Networks [CNNs, ConvNets]



- Locally-connected Neurons
  (3x3 neighborhood)
- Weights are shared
  (over the space)

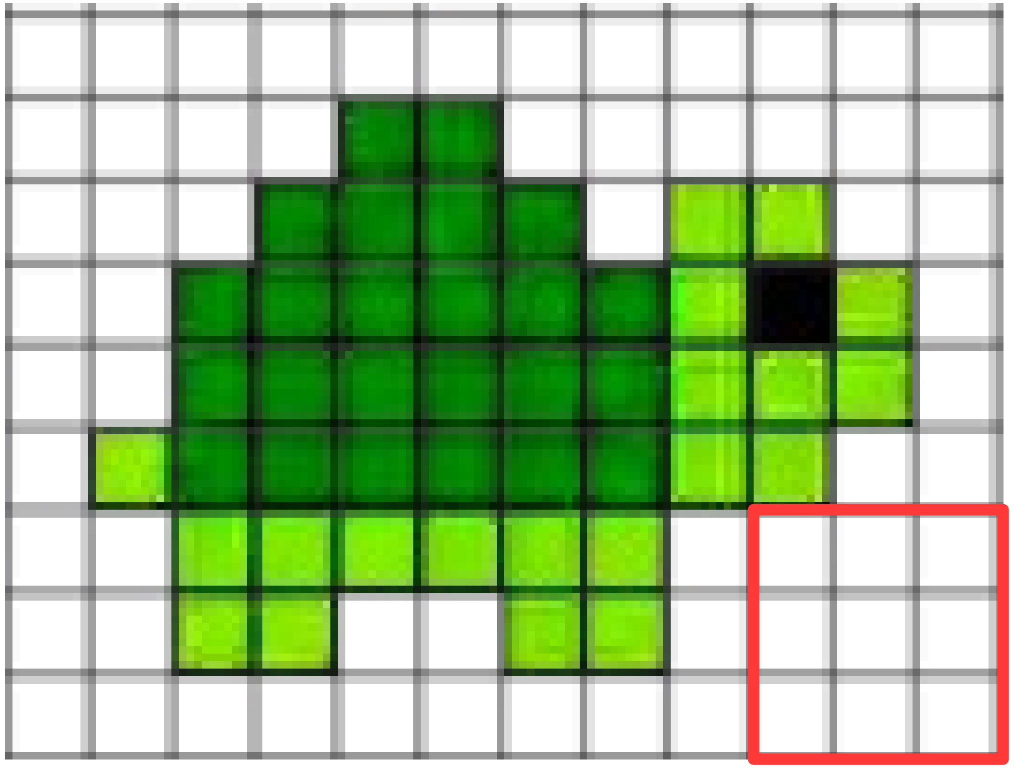
*** Translation Invariance***

Recap: Convolutional Neural Networks [CNNs, ConvNets]

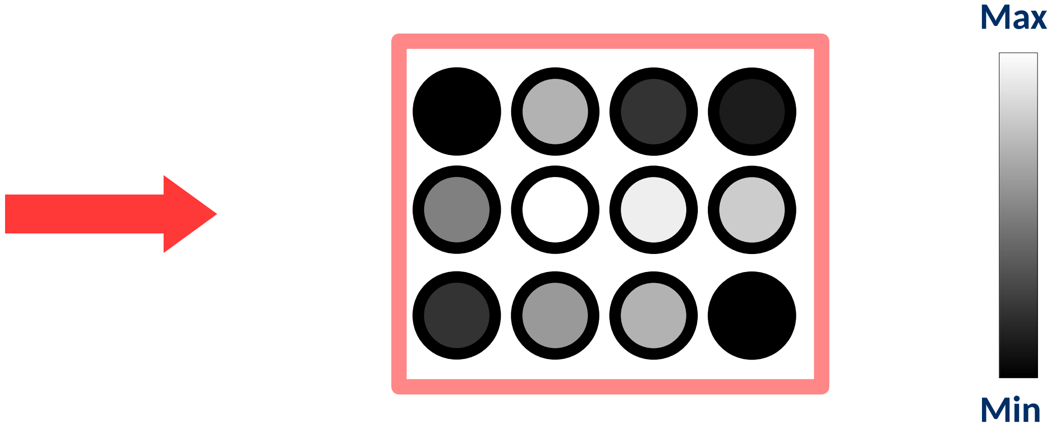


- Some terminology

Recap: Convolutional Neural Networks [CNNs, ConvNets]



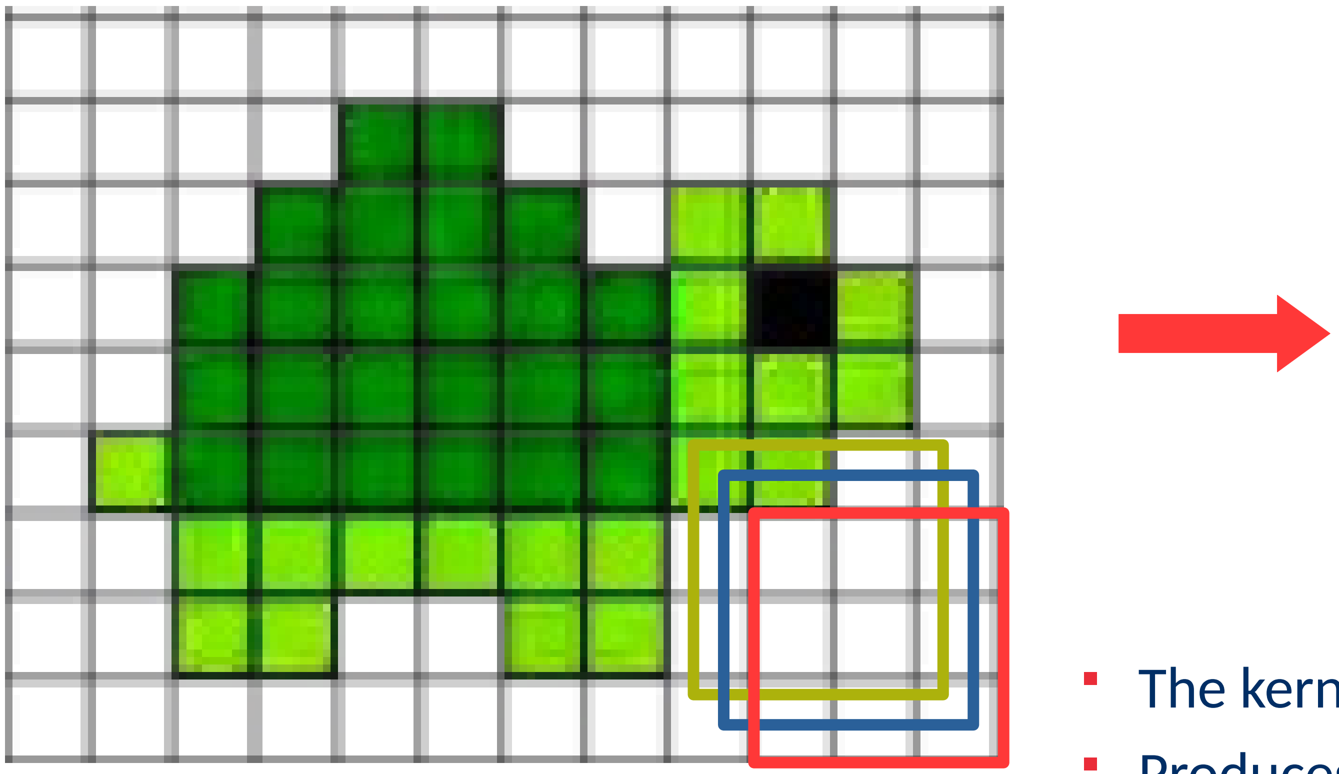
** Translation Equivariance*

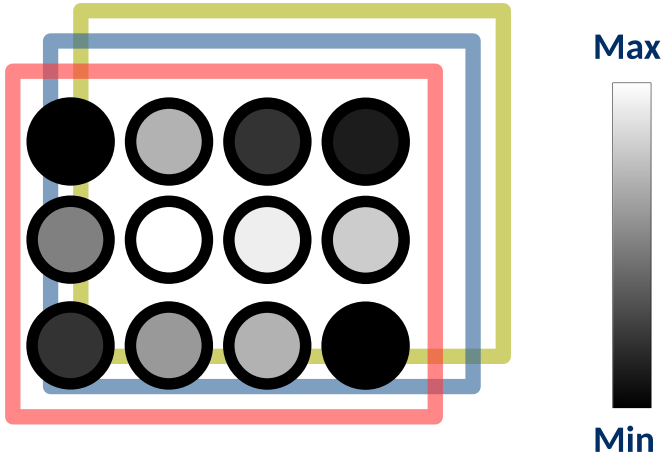


Response/feature map

- The kernel slides across the input
- Produces an output (or response) for every location where it is evaluated

Recap: Convolutional Neural Networks [CNNs, ConvNets]

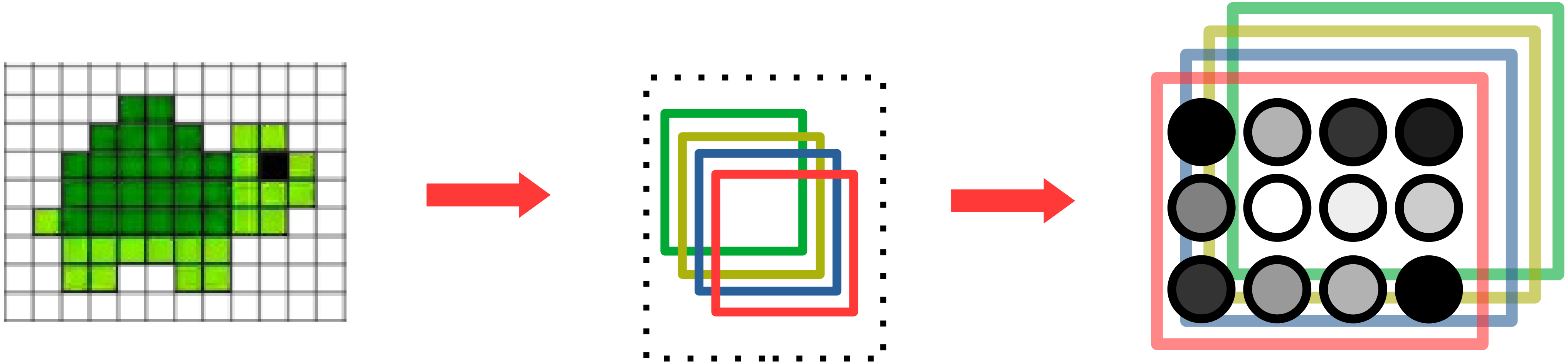




Response/feature map

- The kernel slides across the input
- Produces an output (or response) for every location where it is evaluated
- Repeating the process with *k* multiple kernels produces multiple features maps (channels)

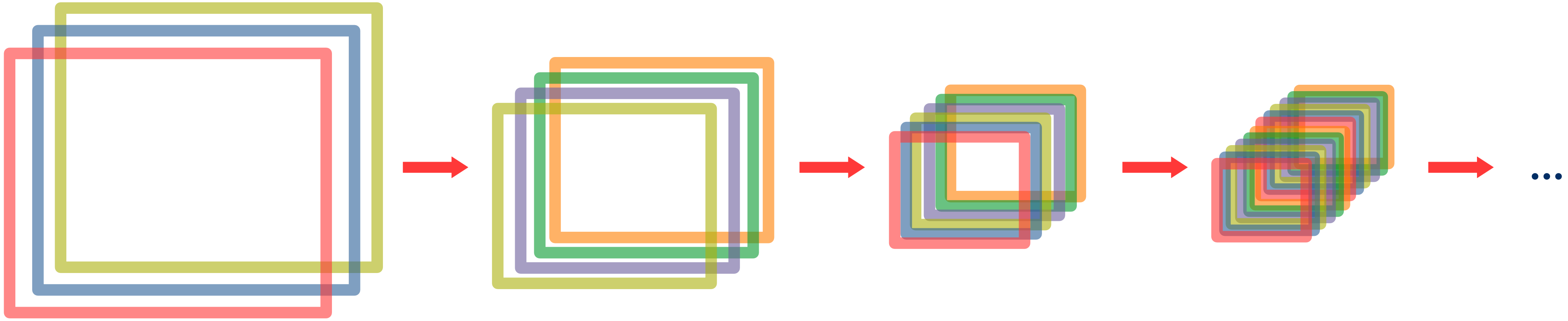
Recap: Convolutional Neural Networks [CNNs, ConvNets]



Input [h x w x 3] **Convolutional Layer** **Output** [h' x w' x k]

- Inputs an ouputs are usually "data cubes" [**Tensors**]
- Filter reponses across inputs are aggregated

Recap: Convolutional Neural Networks [CNNs, ConvNets]



Input [h x w x 3]

Convolutional Neural Network

*Promotes Compositionality

Useful Techniques

[Data Augmentation & Dropout]

Data Augmentation

What?

- Apply a set of operations on a given data sample to produce additional samples

Original Image

Data Augmentation

What?

- Apply a set of operations on a given data sample to produce additional samples

Original Image

Cropped samples

Data Augmentation

What?

- Apply a set of operations on a given data sample to produce additional samples

Original Image

Cropped samples

Mirrored samples

Data Augmentation

What?

- Apply a set of operations on a given data sample to produce additional samples

Benefits

- Increase training data
- Introduce variability

Original Image

Cropped samples

Mirrored samples

OK, but...

Can I apply any random operation?

Data Augmentation

Original Image

Applying any random operation for augmentation

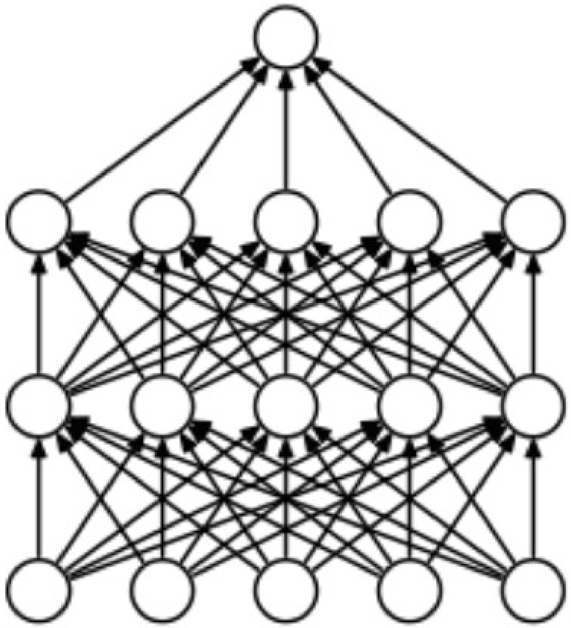
Data Augmentation

Original Image

Applying any random operation for augmentation

Randomly Cropped Samples

Dropout [Krizhevsky et al., 2012]



Standard Neural Net

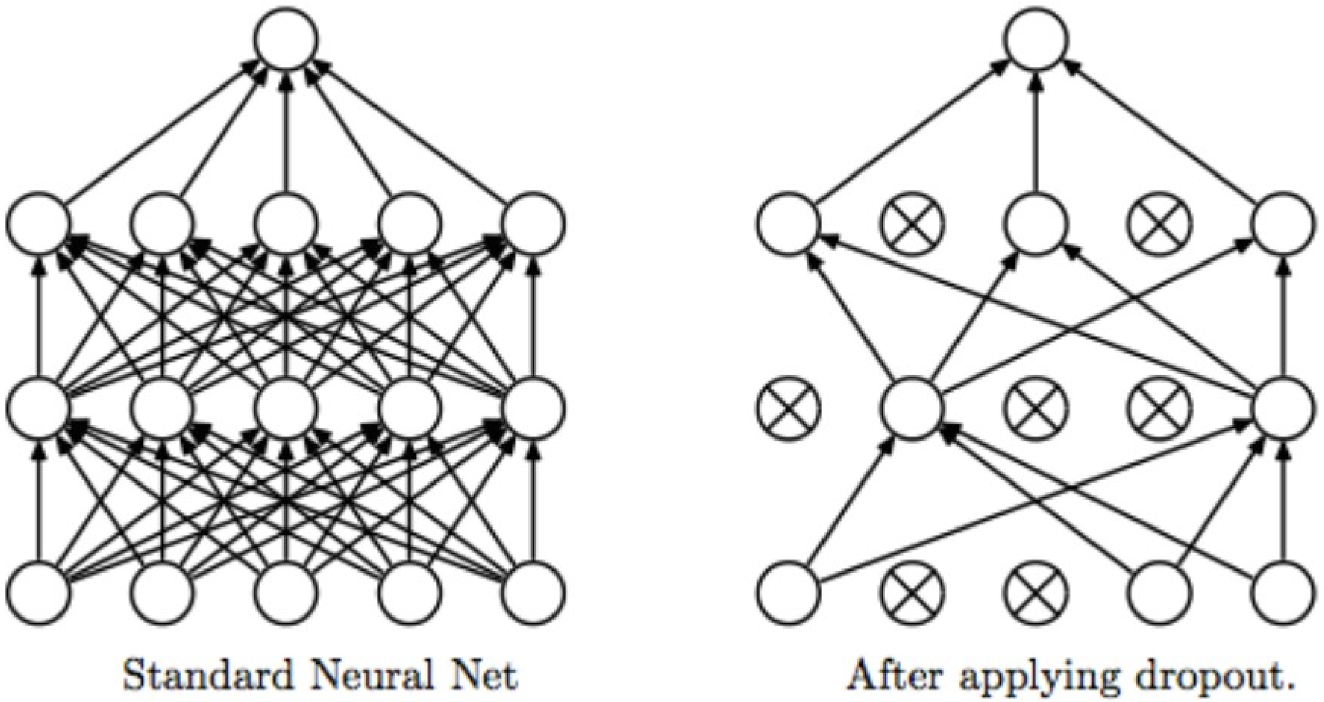
How?

- Deactivate a neuron with a given probability.

Benefits

- Avoid overfitting
- Promote ensemble learning

Dropout [Krizhevsky et al., 2012]



How?

- Deactivate a neuron with a given probability.

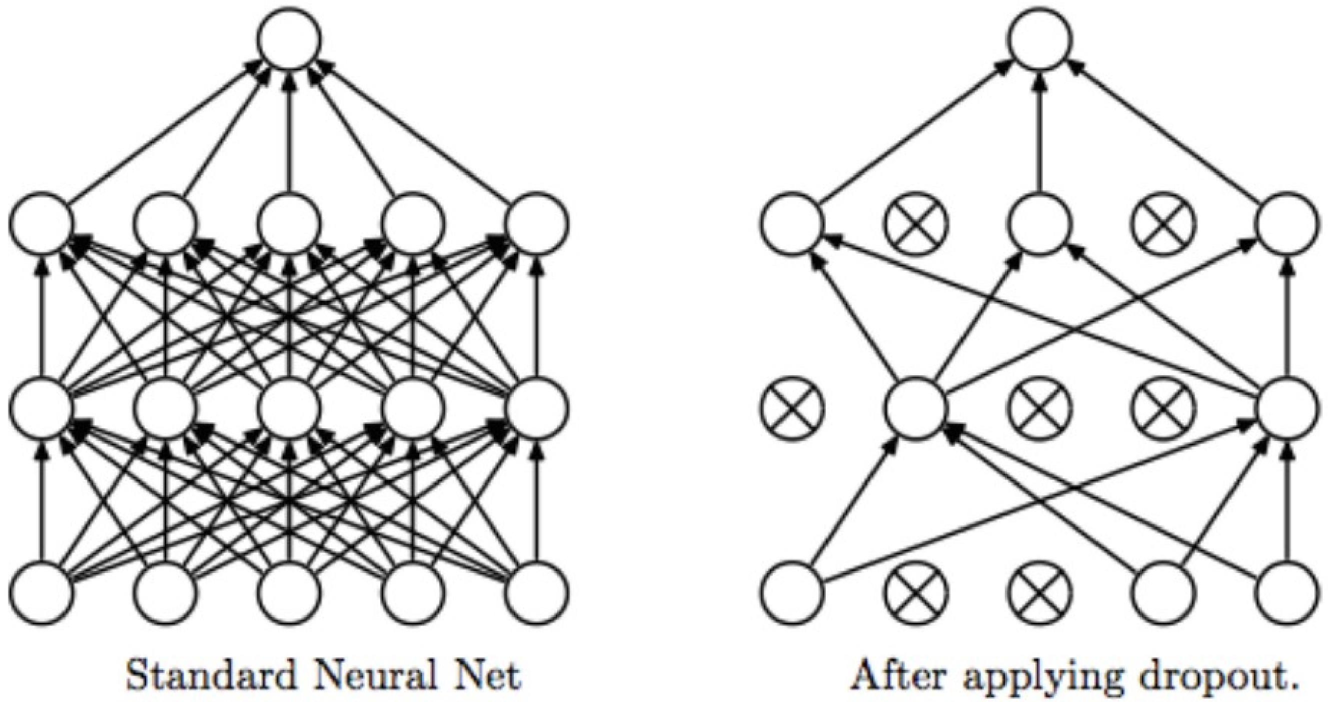
Benefits

- Avoid overfitting
- Promote ensemble learning

OK, but…

How this helps in practice?

Dropout [Krizhevsky et al., 2012]



How?

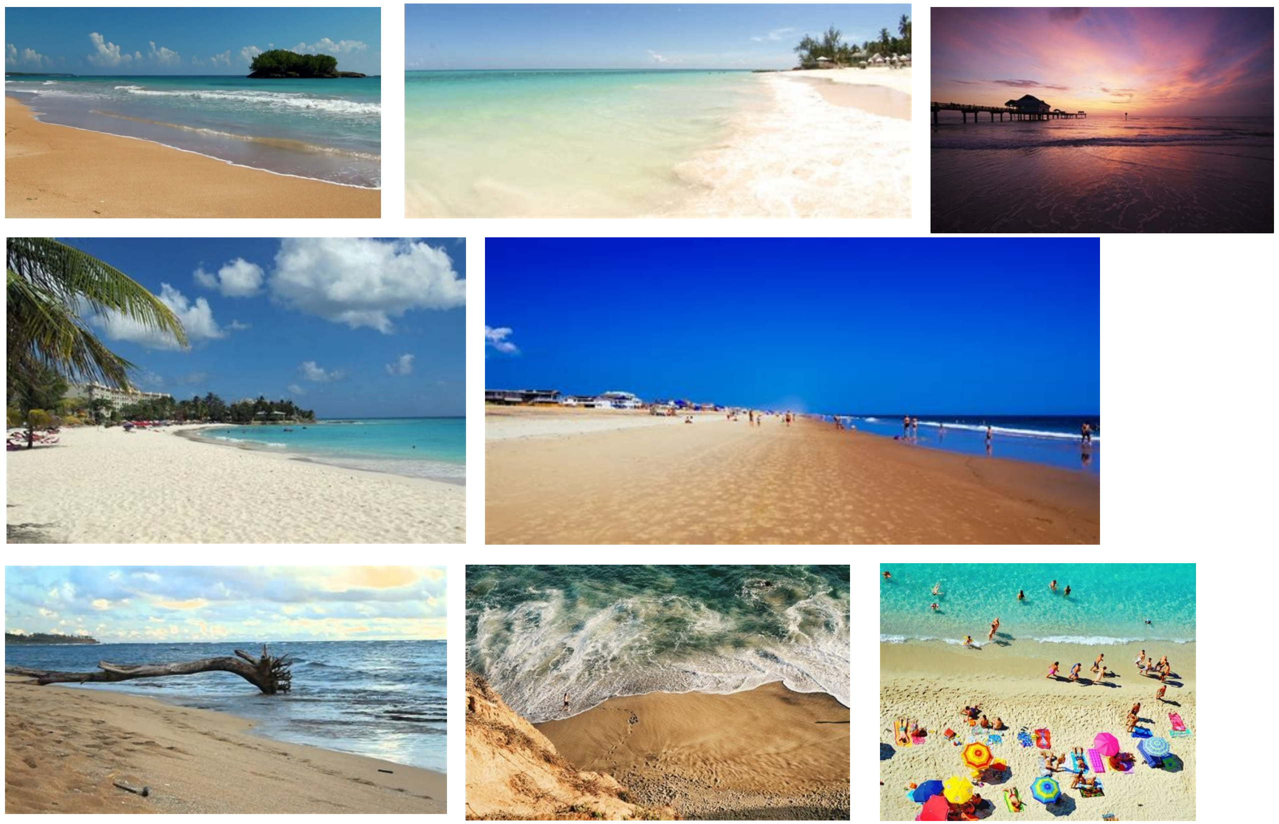
- Deactivate a neuron with a given probability.

Benefits

- Avoid overfitting
- Promote ensemble learning

Dropout [Krizhevsky et al., 2012]

Scene Recognition



How?

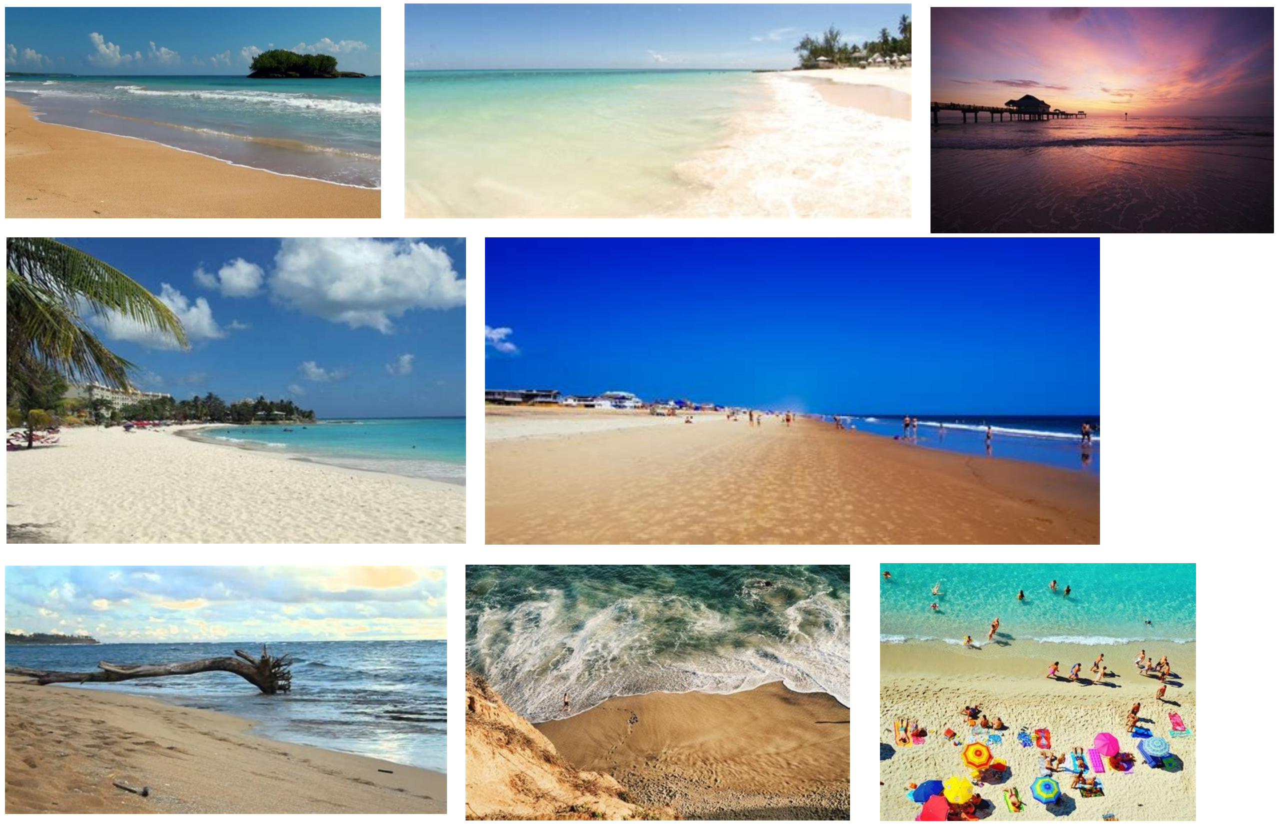
- Deactivate a neuron with a given probability.

Benefits

- Avoid overfitting
- Promote ensemble learning

Dropout [Krizhevsky et al., 2012]

Scene Recognition



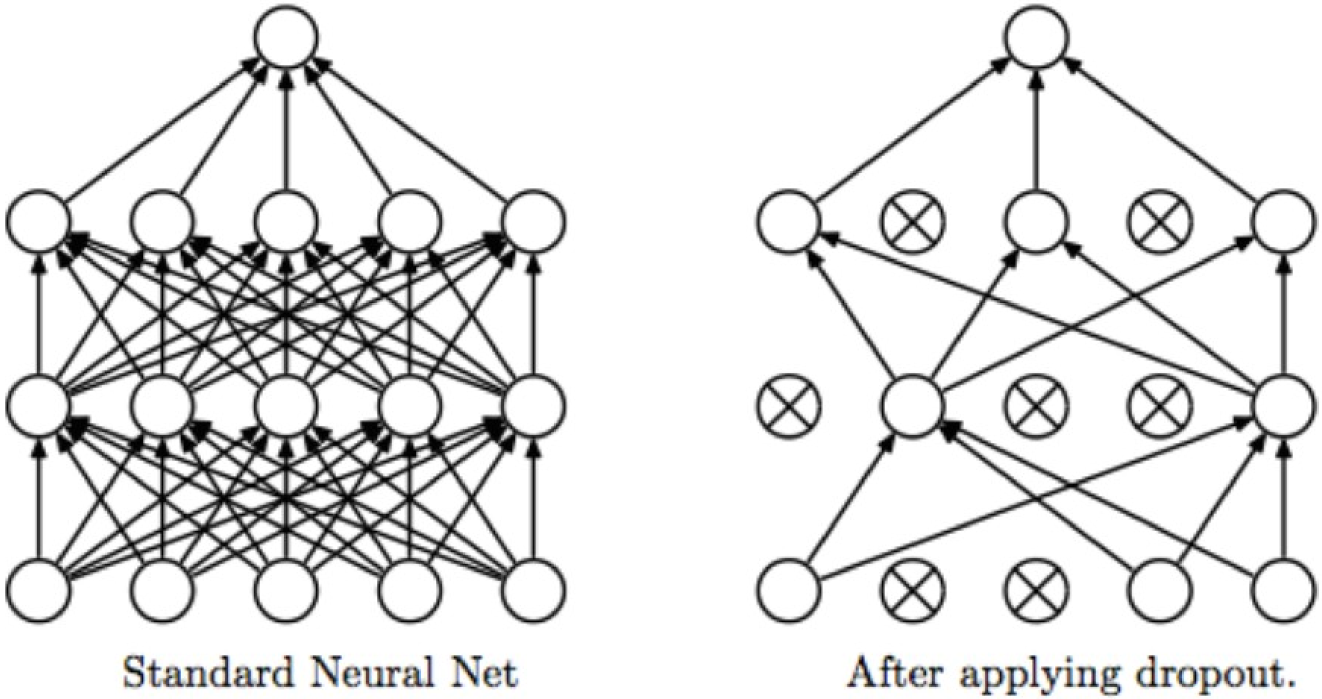
How?

- Deactivate a neuron with a given probability.

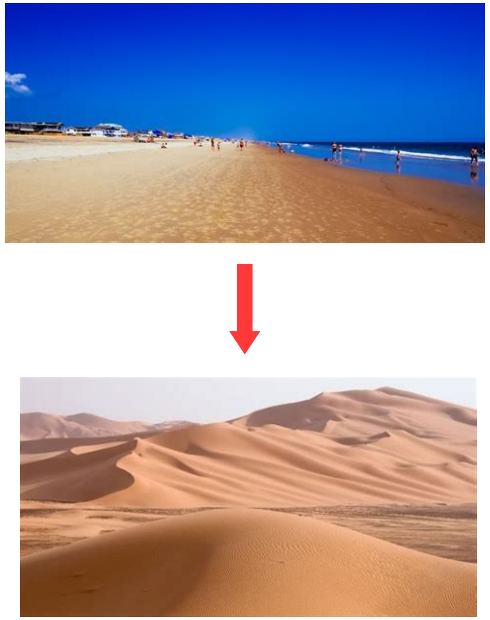
Benefits

- Avoid overfitting
- Promote ensemble learning

Dropout [Krizhevsky et al., 2012]



How would it help?



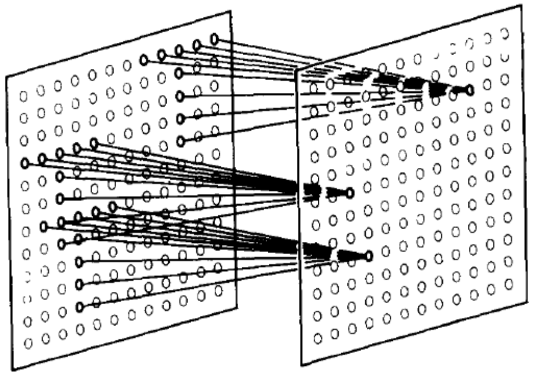
Nice, but...

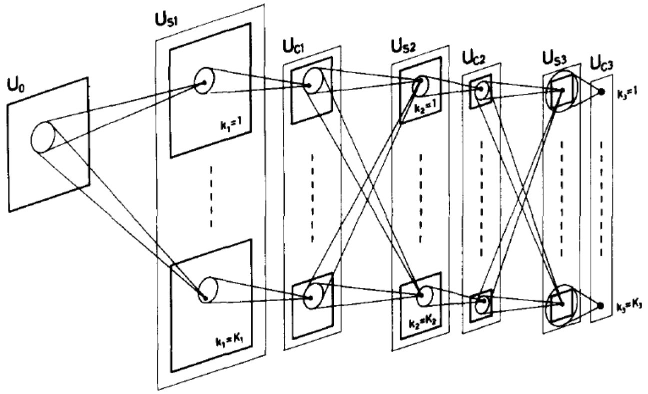
How did we get there?

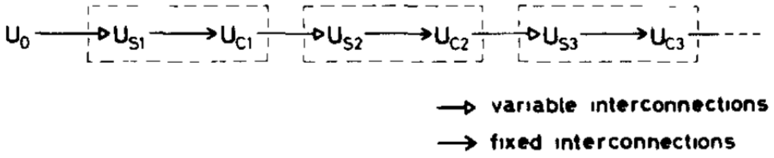
Relevant Architectures

[AlexNet, VGG-Net, GoogLeNet, ResNet, *Net]

1982: Neocognitron [Fukushima & Miyake., 1982]

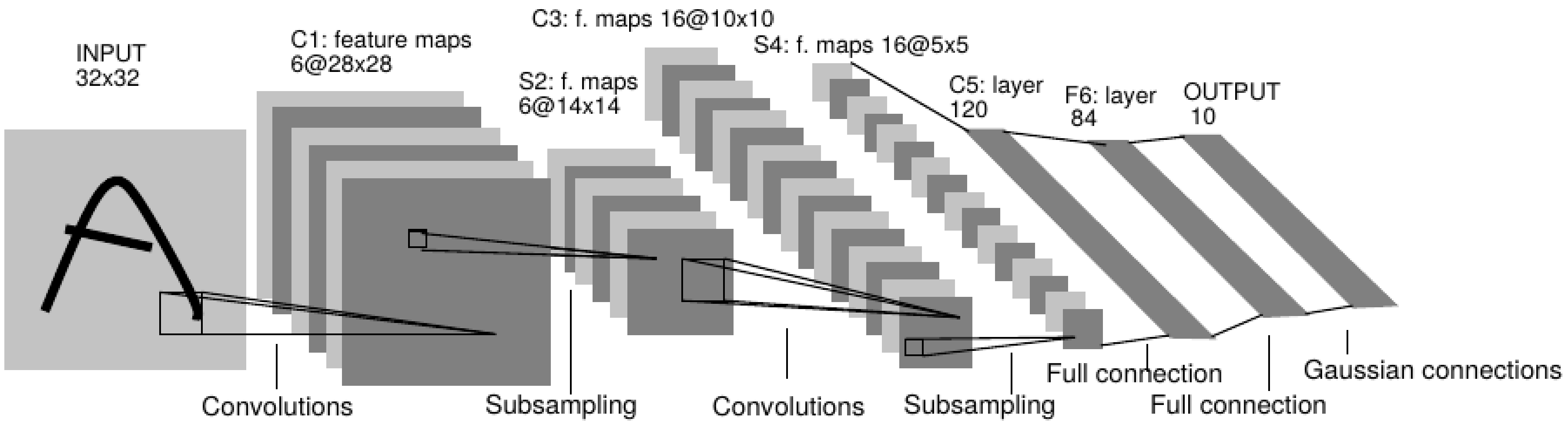






- Goal: Recognition of position-shifted / shape-distorted patterns
- Proposed the cell-plane arrangement (convolution)
- Hierarchical structure
- Convolution/sub-sampling combination

1998: LeNet-5 [Lecun et al., 1998]

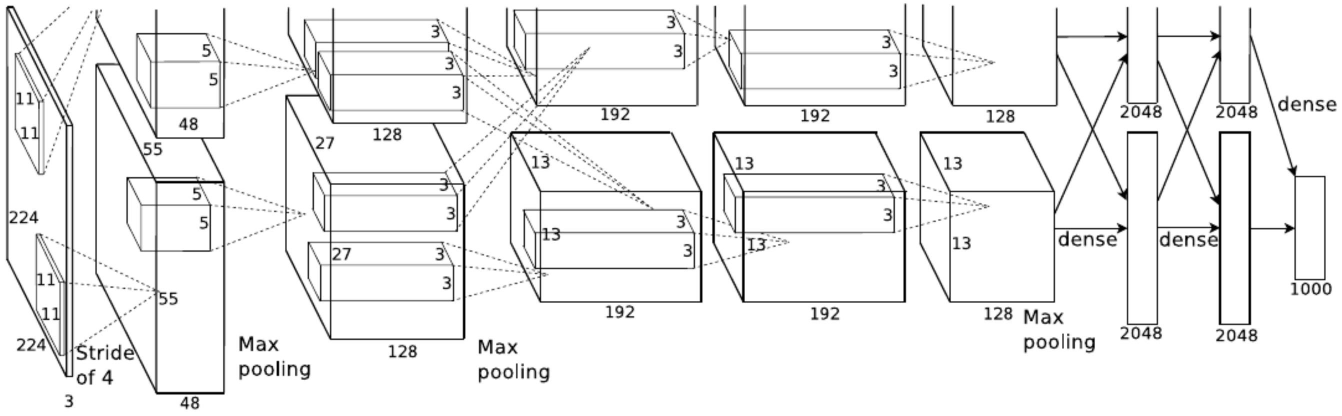


7 layers

- 3 conv. layers
- 2 subsampling layers
- 2 FC layers

- **Addressed handwritten digit recognition task**
- **Modified NIST (MNIST) dataset was proposed**
- **One of the first use of ConvNets + Backprop**

2012: AlexNet [Krizhevsky et al., 2012]



- 5 conv. layers + 3 FC layers
- 60M param. , 650K neurons
- Trained across 2 GPUs
  (Model Parallelism)

- **No need to pair convolutional with pooling layers**
- **ReLU for Convolutional Layers**
- **Data Augmentation and Dropout**

2012: AlexNet [Krizhevsky et al., 2012]

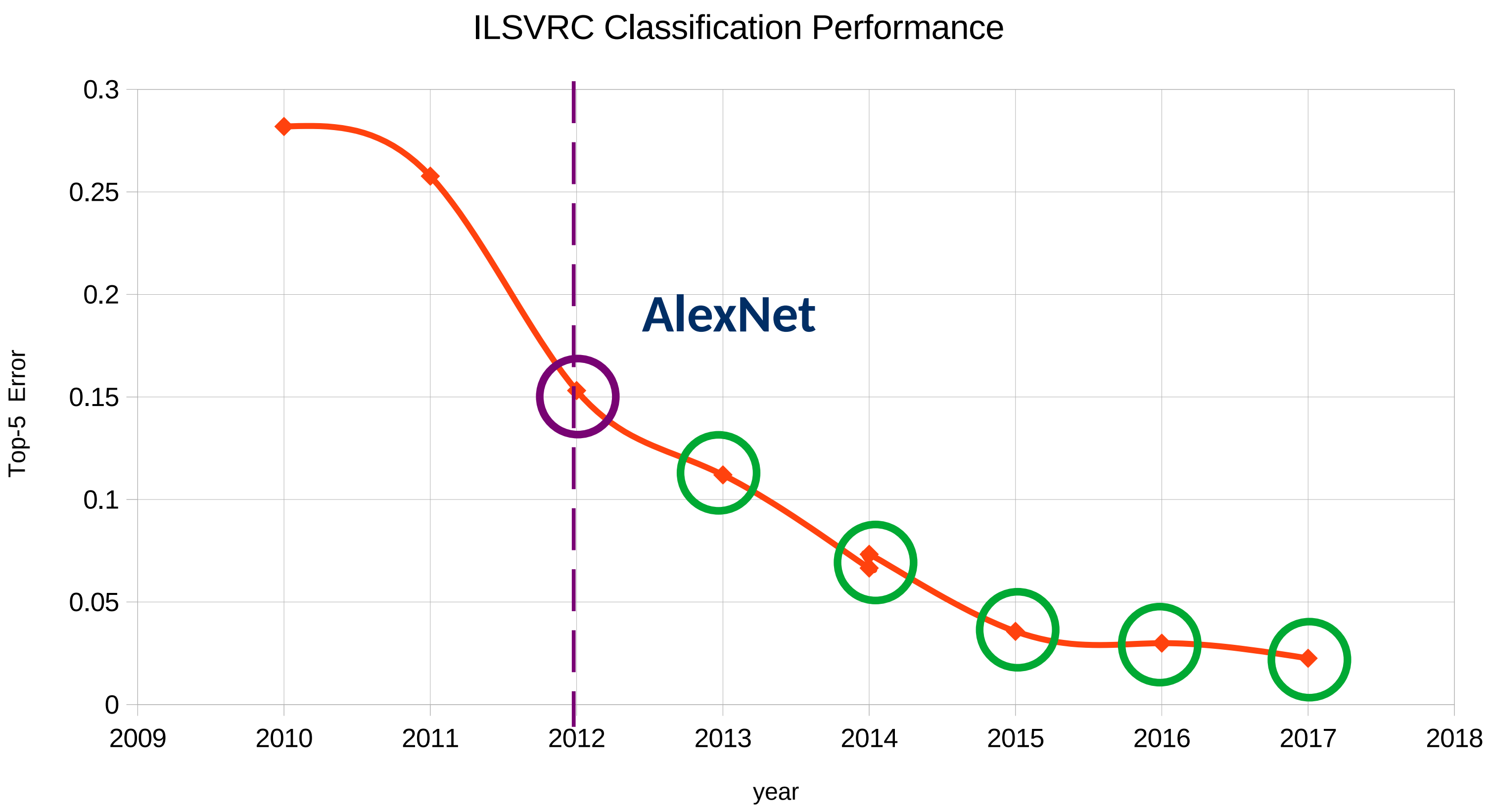
Relevance → Winner: ILSVRC 2012 (1K categories, 1.2M images)

- Challenges?
- Performance Metrics?

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- Winner: ILSVRC 2012 (1K categories, 1.2M images)



2012: AlexNet [Krizhevsky et al., 2012]

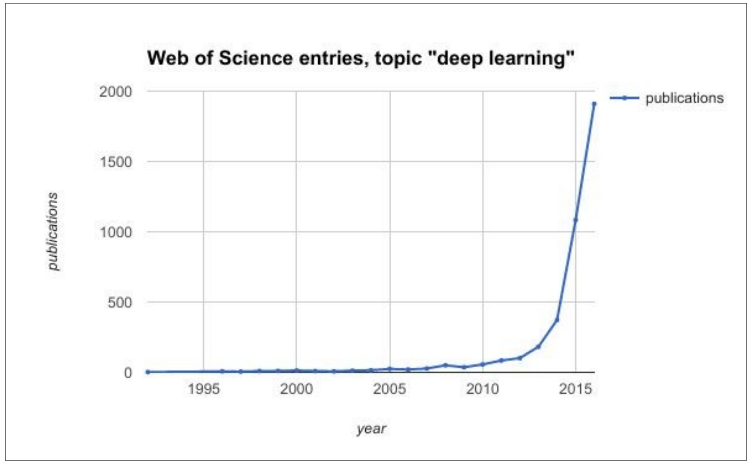
Relevance

- “Deep Learning” goes mainstream

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- “Deep Learning” goes mainstream



2012: AlexNet [Krizhevsky et al., 2012]

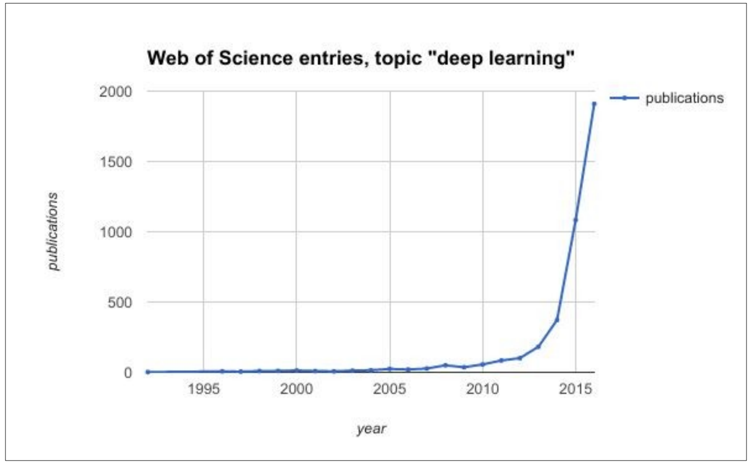
Relevance

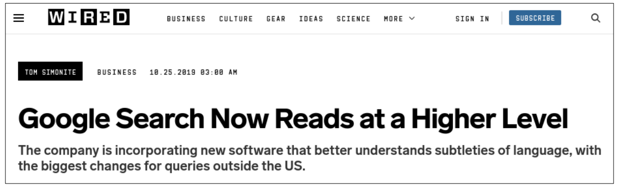
- "Deep Learning" goes mainstream


Microsoft's speech recognition engine listens as well as a human

"This is an historic achievement" - Xuedong Huang

Andrew Tarantola, @terrortola
10.18.16 in Personal Computing




WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

TOM SIMONITE BUSINESS 10.25.2019 03:00 AM

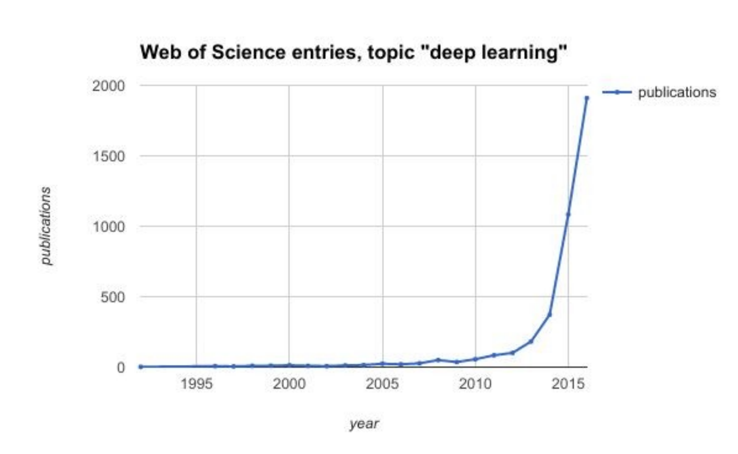
Google Search Now Reads at a Higher Level

The company is incorporating new software that better understands subtleties of language, with the biggest changes for queries outside the US.

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- "Deep Learning" goes mainstream



Microsoft's speech recognition engine listens as well as a human

"This is an historic achievement" - Xuedong Huang

Andrew Tarantola, @terrortola
10.18.16 in Personal Computing

WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

TOM SIMONITE BUSINESS 10.25.2019 03:00 AM

Google Search Now Reads at a Higher Level

The company is incorporating new software that better understands subtleties of language, with the biggest changes for queries outside the US.

Intelligent Machines

Deep-Learning Machine Listens to Bach, Then Writes Its Own Music in the Same Style

Can you tell the difference between music composed by Bach and by a neural network?

by Emerging Technology from the arXiv December 14, 2016

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- "Deep Learning" goes mainstream

Microsoft's speech recognition engine listens as well as a human

"This is an historic achievement" - Xuedong Huang

Andrew Tarantola, @terrortola
10.18.16 in Personal Computing

The Big Read Driverless vehicles + Add to myFT

Driverless cars inspire a new gold rush in California

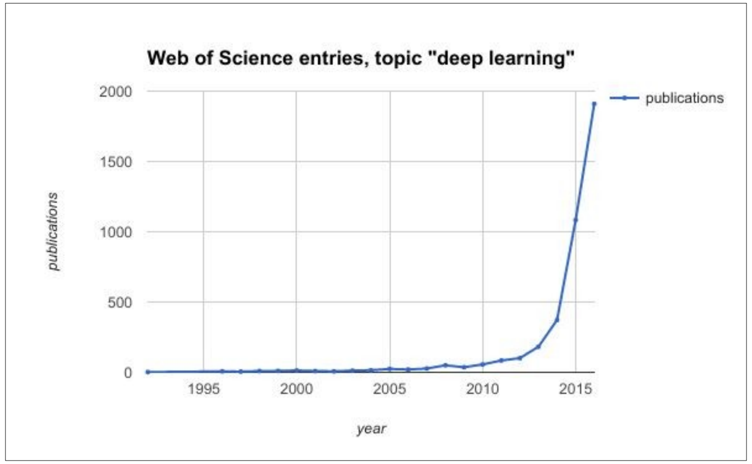
MAY 23, 2017 by: Leslie Hook and Tim Bradshaw

Intelligent Machines

Deep-Learning Machine Listens to Bach, Then Writes Its Own Music in the Same Style

Can you tell the difference between music composed by Bach and by a neural network?

by Emerging Technology from the arXiv December 14, 2016



WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

TOM SIMONITE BUSINESS 10.25.2019 03:00 AM

Google Search Now Reads at a Higher Level

The company is incorporating new software that better understands subtleties of language, with the biggest changes for queries outside the US.

Article | Open Access | Published: 29 August 2019

Deep Learning to Improve Breast Cancer Detection on Screening Mammography

Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh

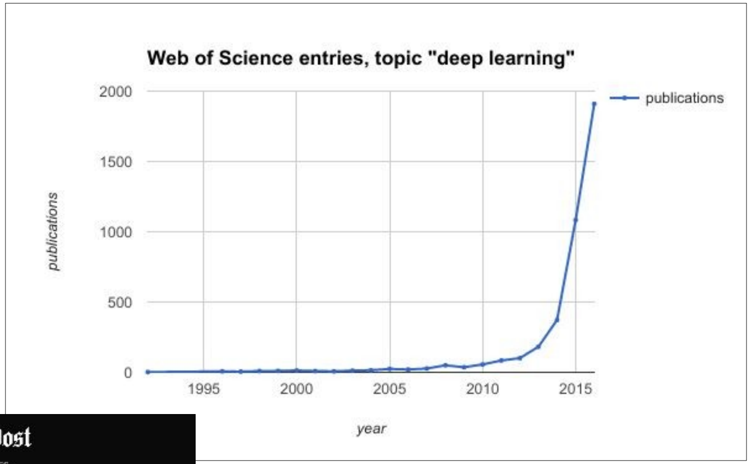
Scientific Reports 9, Article number: 12495 (2019) | Cite this article

9229 Accesses | 2 Citations | 27 Altmetric | Metrics

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- "Deep Learning" goes mainstream



Microsoft's speech recognition engine listens as well as a human

"This is an historic achievement" - Xuedong Huang

Andrew Tarantola, @terrortola
10.18.16 in Personal Computing

The Big Read Driverless vehicles + Add to myFT

Driverless cars inspire a new gold rush in California

MAY 23, 2017 by: Leslie Hook and Tim Bradshaw

The Washington Post
Democracy Dies in Darkness

Innovations

Google's AlphaGo beats the world's best Go player — again

By Hamza Shaban May 26

WIRED

TOM SIMONITE BUSINESS 10.25.2019 03:00 AM

Google Search Now Reads at a Higher Level

The company is incorporating new software that better understands subtleties of language, with the biggest changes for queries outside the US.

Intelligent Machines

Deep-Learning Machine Listens to Bach, Then Writes Its Own Music in the Same Style

Can you tell the difference between music composed by Bach and by a neural network?

by Emerging Technology from the arXiv December 14, 2016

Article | Open Access | Published: 29 August 2019

Deep Learning to Improve Breast Cancer Detection on Screening Mammography

Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh

Scientific Reports 9, Article number: 12495 (2019) | Cite this article

9229 Accesses | 2 Citations | 27 Altmetric | Metrics

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- "Deep Learning" goes mainstream

deepBlue - Chess

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- "Deep Learning" goes mainstream

deepBlue - Chess

Watson - Jeopardy

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- "Deep Learning" goes mainstream

deepBlue - Chess

Watson - Jeopardy

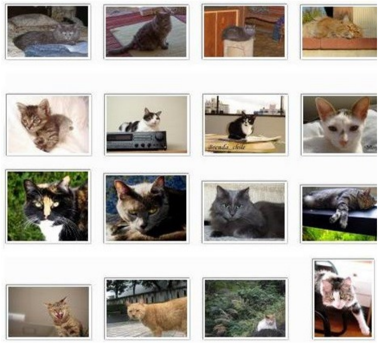
AlphaGo - Go

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- From Engineered Features to Learning-based Representations

dataset



Feature extraction

Interest Points
(SIFT, SURF, kAS, etc.)

Machine Learning

ML Method
(SVM, RF, Boosting, etc.)

Output

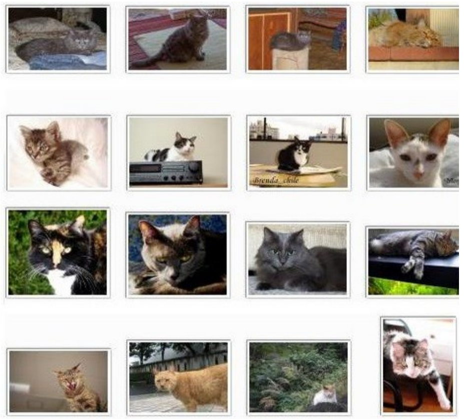
- **Idea:** Engineer informative features + Use ML to discriminate between those features

2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- From Engineered Features to Learning-based Representations

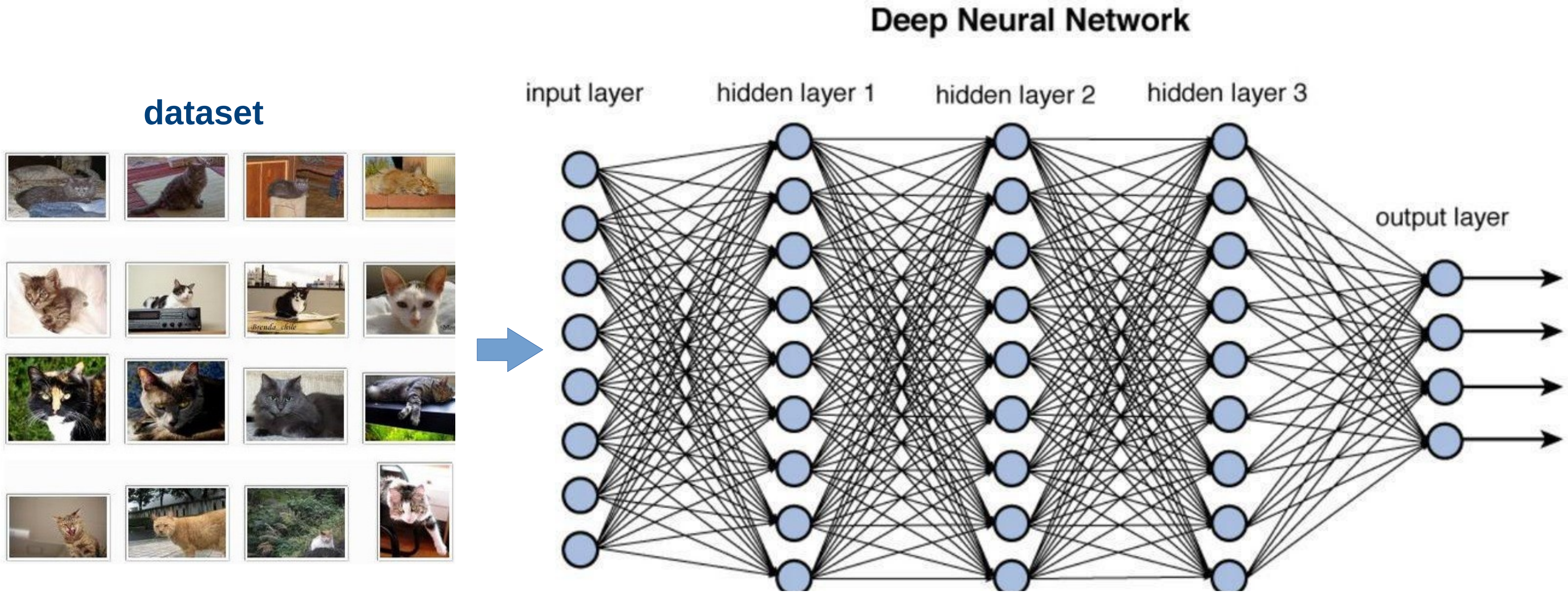
dataset



2012: AlexNet [Krizhevsky et al., 2012]

Relevance

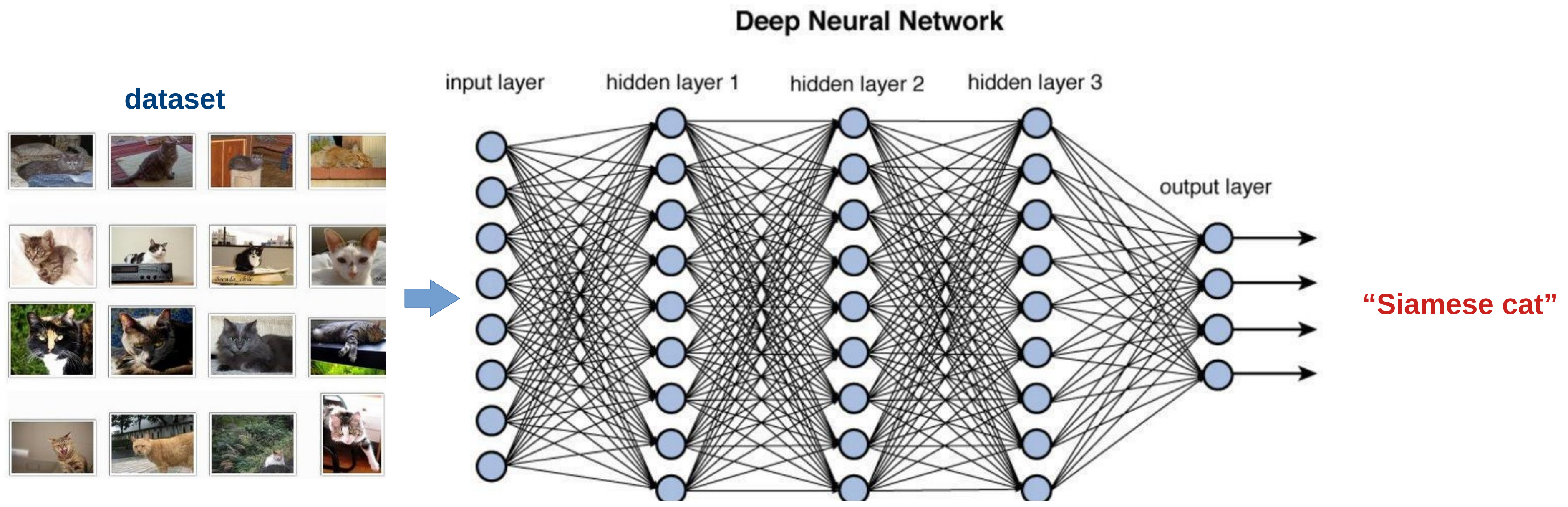
- From Engineered Features to Learning-based Representations



2012: AlexNet [Krizhevsky et al., 2012]

Relevance

- From Engineered Features to Learning-based Representations

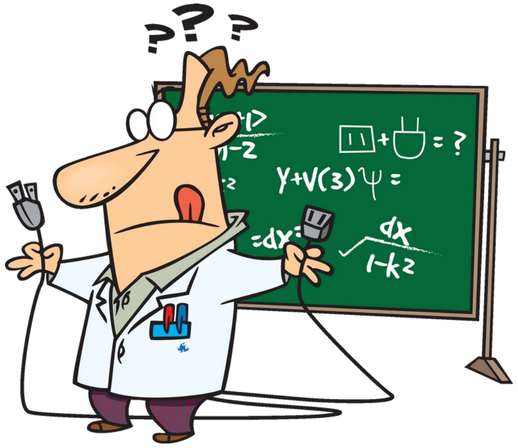


Learning-based Representations

Enablers

2012: AlexNet [Krizhevsky et al., 2012]

Enablers

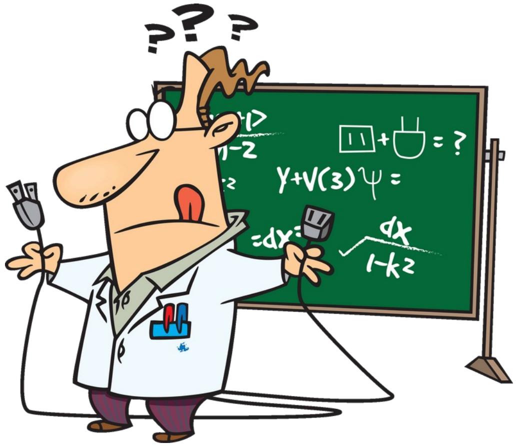


Scientific Community

2012: AlexNet [Krizhevsky et al., 2012]

Enablers

Open-Access Datasets

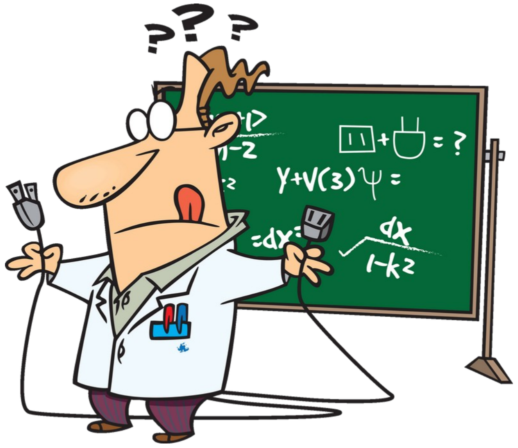


Scientific Community

2012: AlexNet [Krizhevsky et al., 2012]

Enablers

Open-Access Datasets



Scientific Community

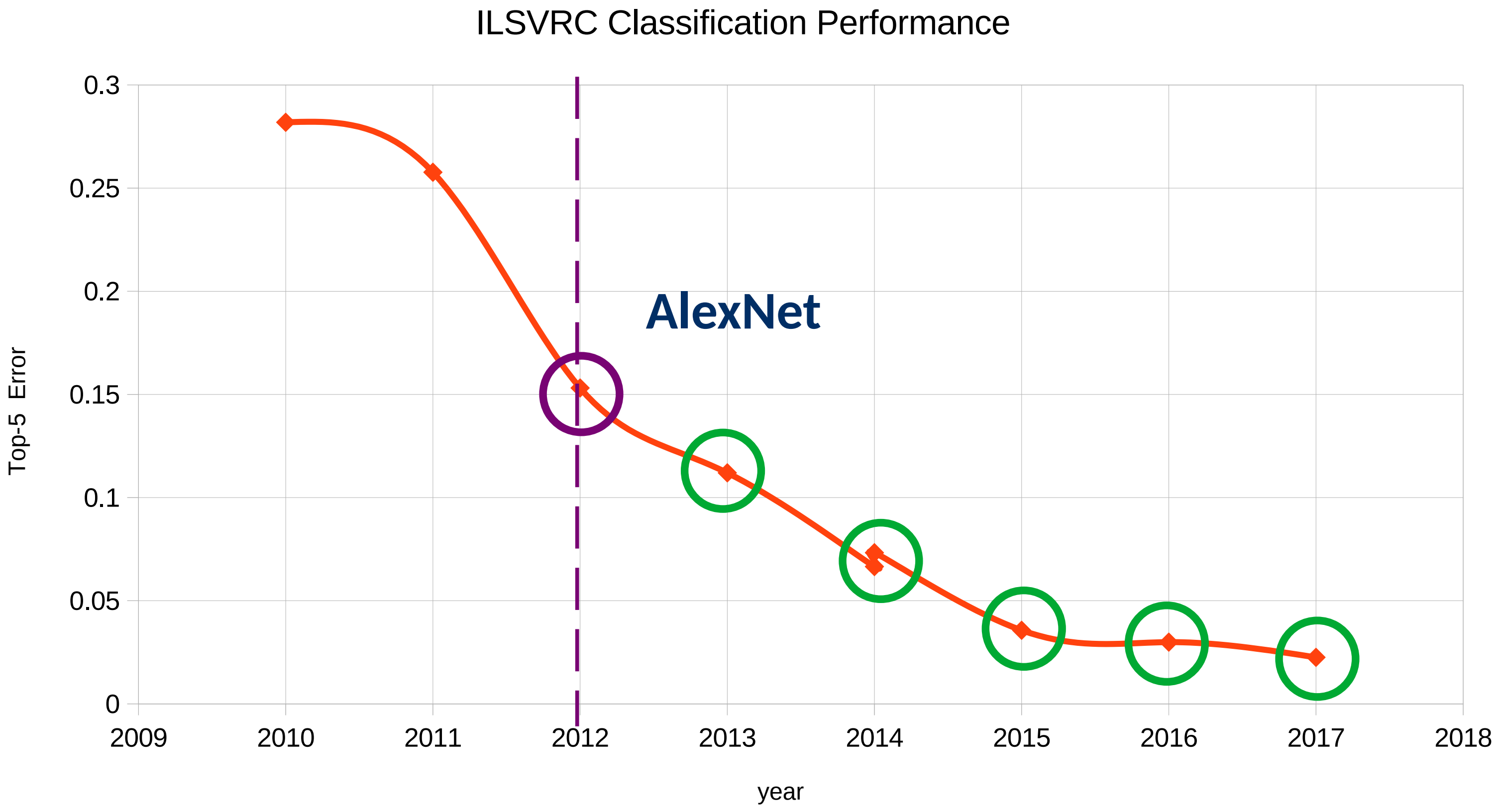
Hardware
Developments

The Post-AlexNet Era

[The Birth of “Deep Learning”]

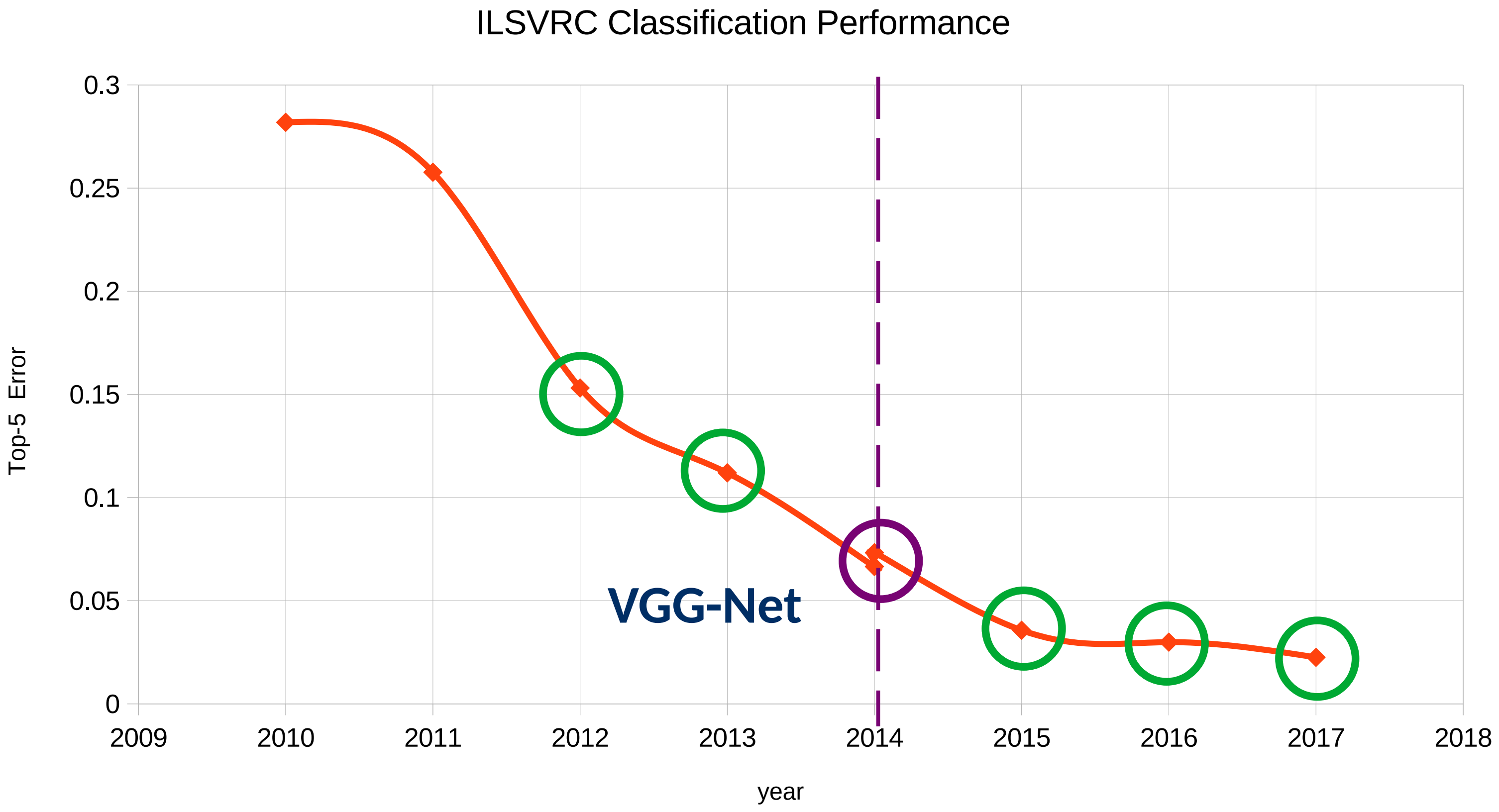
The Post-AlexNet Era

- Everything was built on top of deep models



2014: VGG-Net [Simoyan & Zisserman, 2015]

Going Very Deep



2014: VGG-Net [Simoyan & Zisserman, 2015]

Going Very Deep

- Fixed-size 3x3 kernels
- Use *same* conv. to preserve resolution
- Trained by splitting data across 4 copies of the same model → *data parallelism*

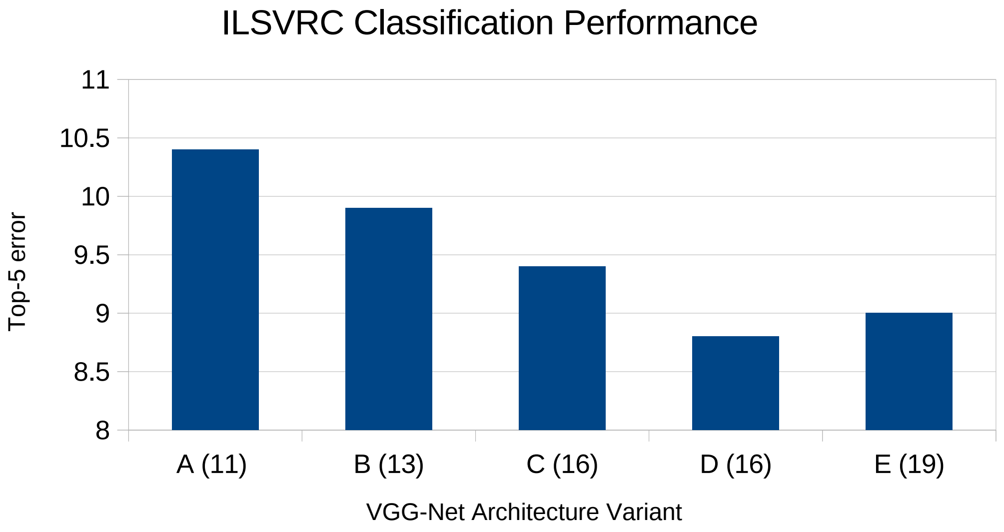
[Simoyan & Zisserman., 2015]

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

2014: VGG-Net [Simoyan & Zisserman, 2015]

Going Very Deep

- Fixed-size 3x3 kernels
- Use *same* conv. to preserve resolution
- Trained by splitting data across 4 copies of the same model → *data parallelism*

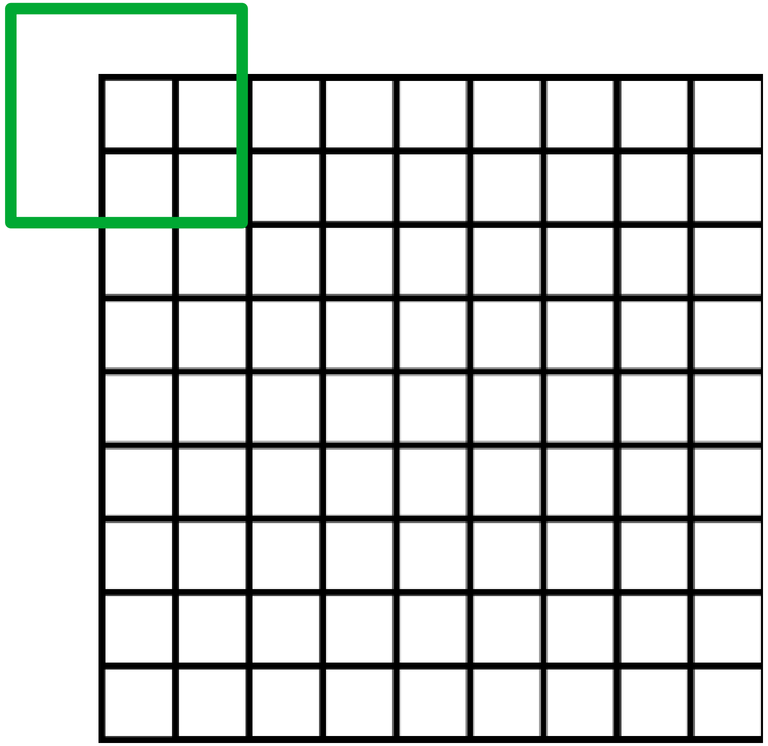


[Simoyan & Zisserman., 2015]

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

2014: VGG-Net [Simoyan & Zisserman, 2015]

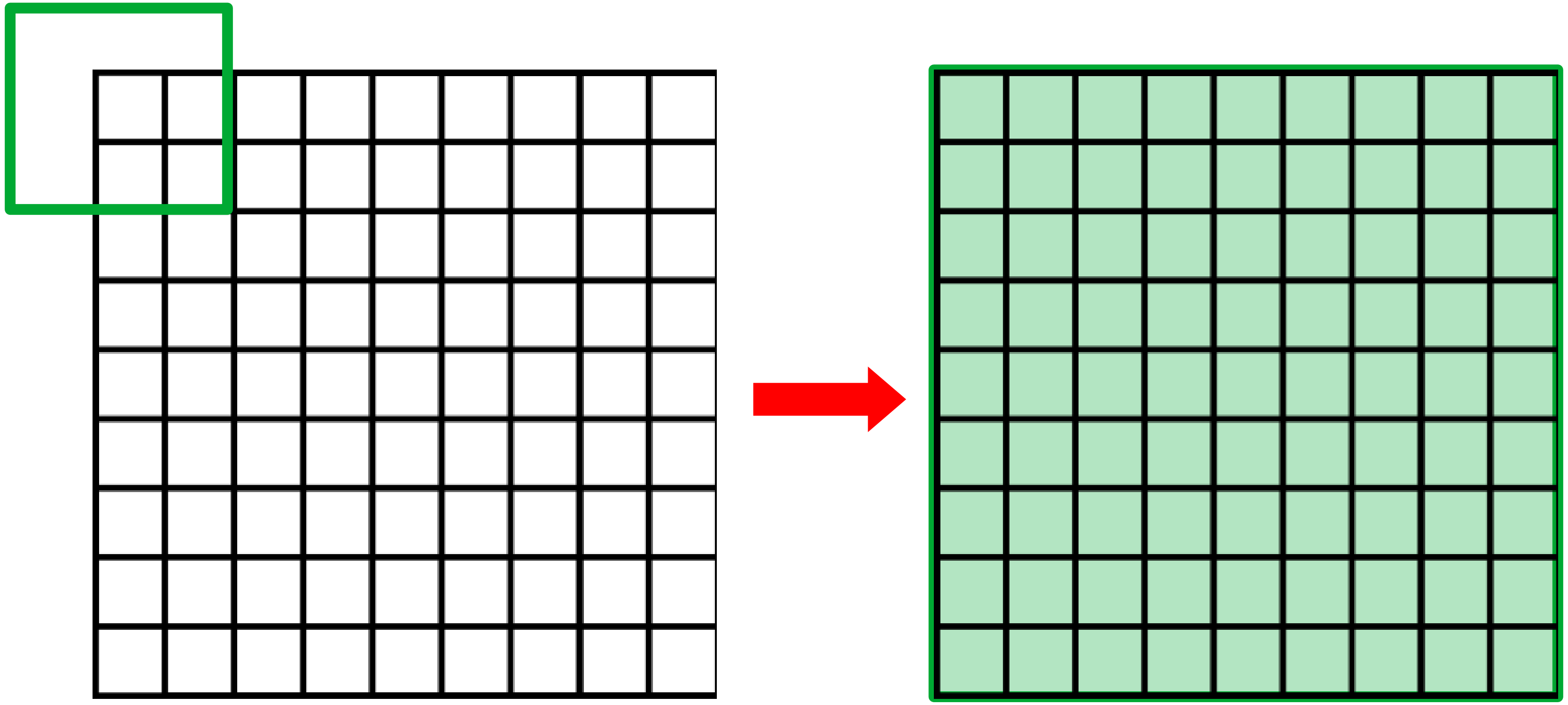
Going Very Deep via Stacked kernels and Same Convolutions



1st Conv. Layer (3x3)

2014: VGG-Net [Simoyan & Zisserman, 2015]

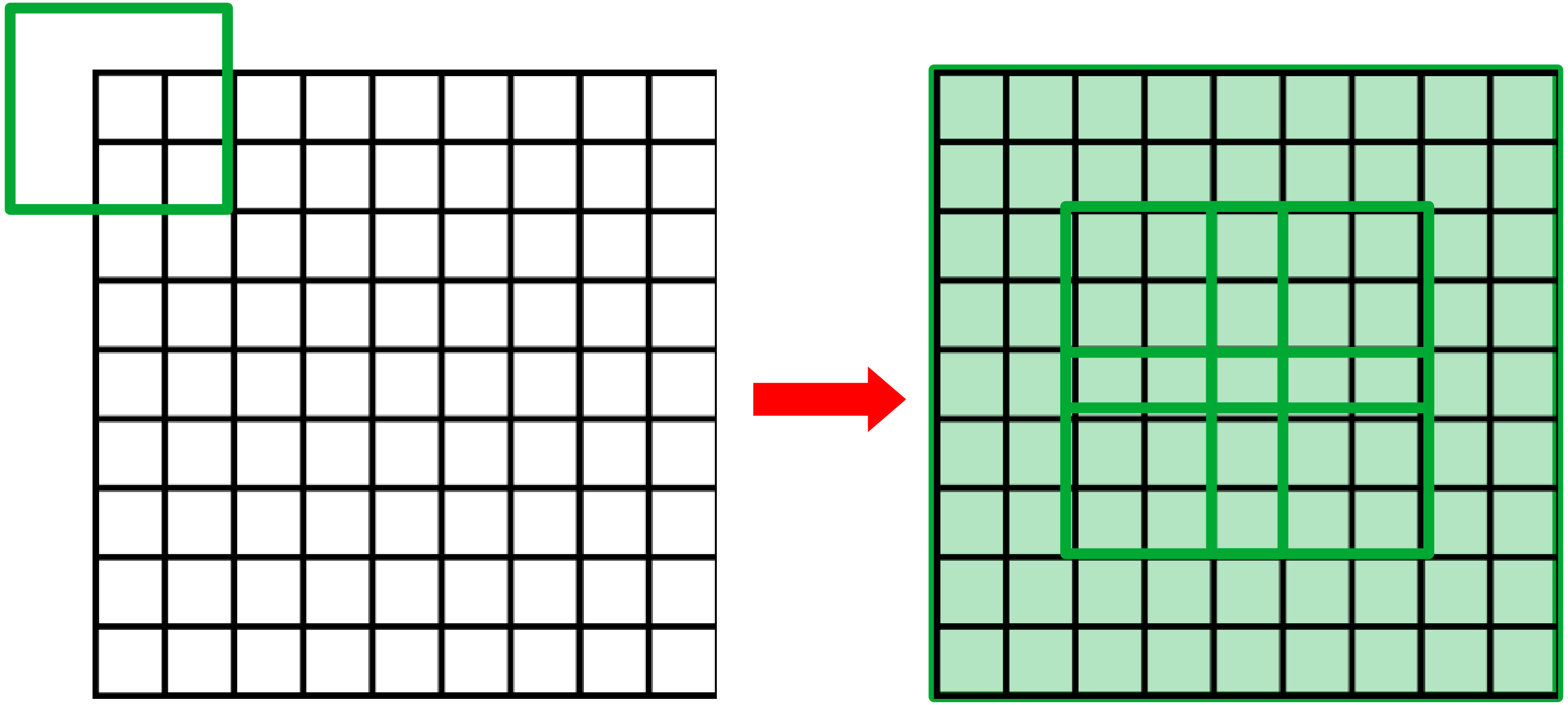
Going Very Deep via Stacked kernels and Same Convolutions



1^{st} Conv. Layer (3x3)

2014: VGG-Net [Simoyan & Zisserman, 2015]

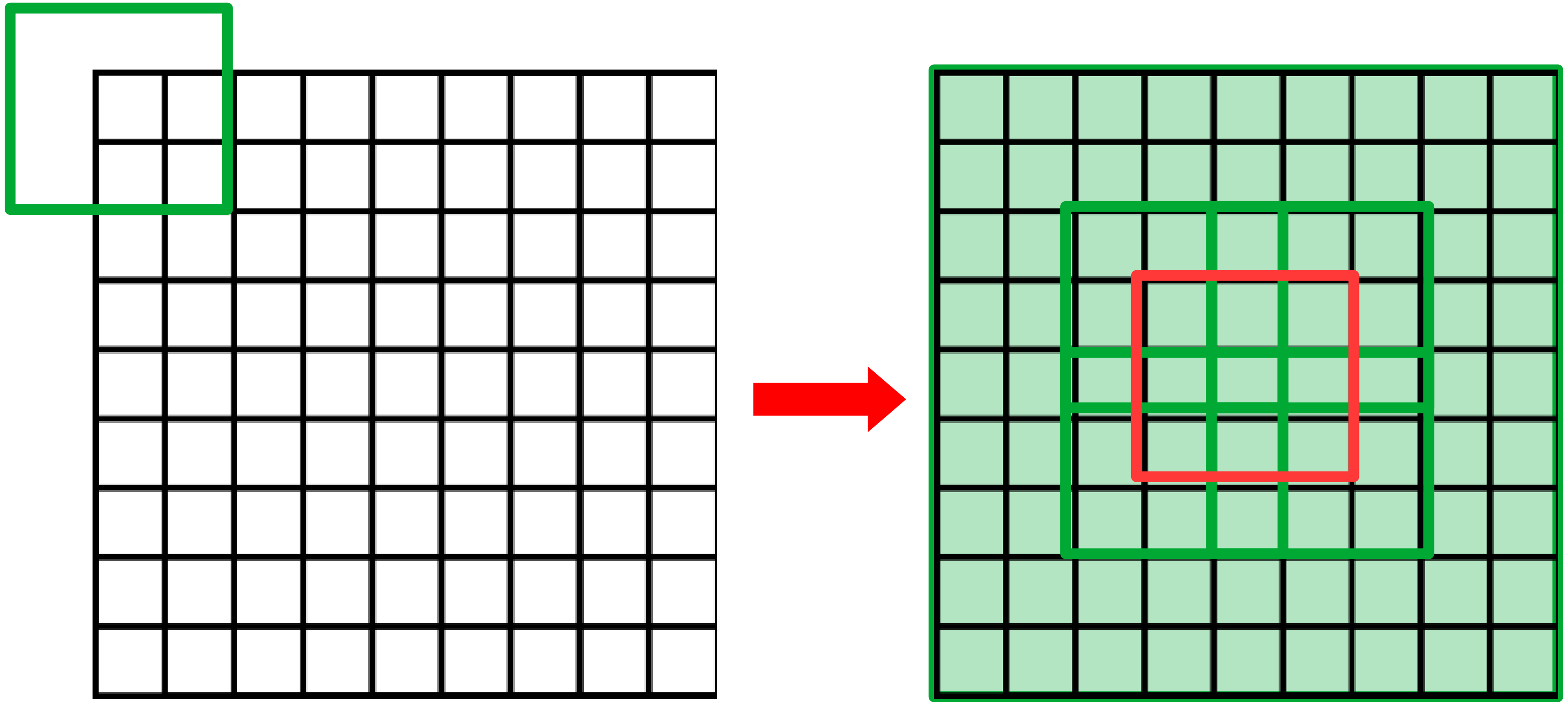
Going Very Deep via Stacked kernels and Same Convolutions



1st Conv. Layer (3x3)

2014: VGG-Net [Simoyan & Zisserman, 2015]

Going Very Deep via Stacked kernels and Same Convolutions

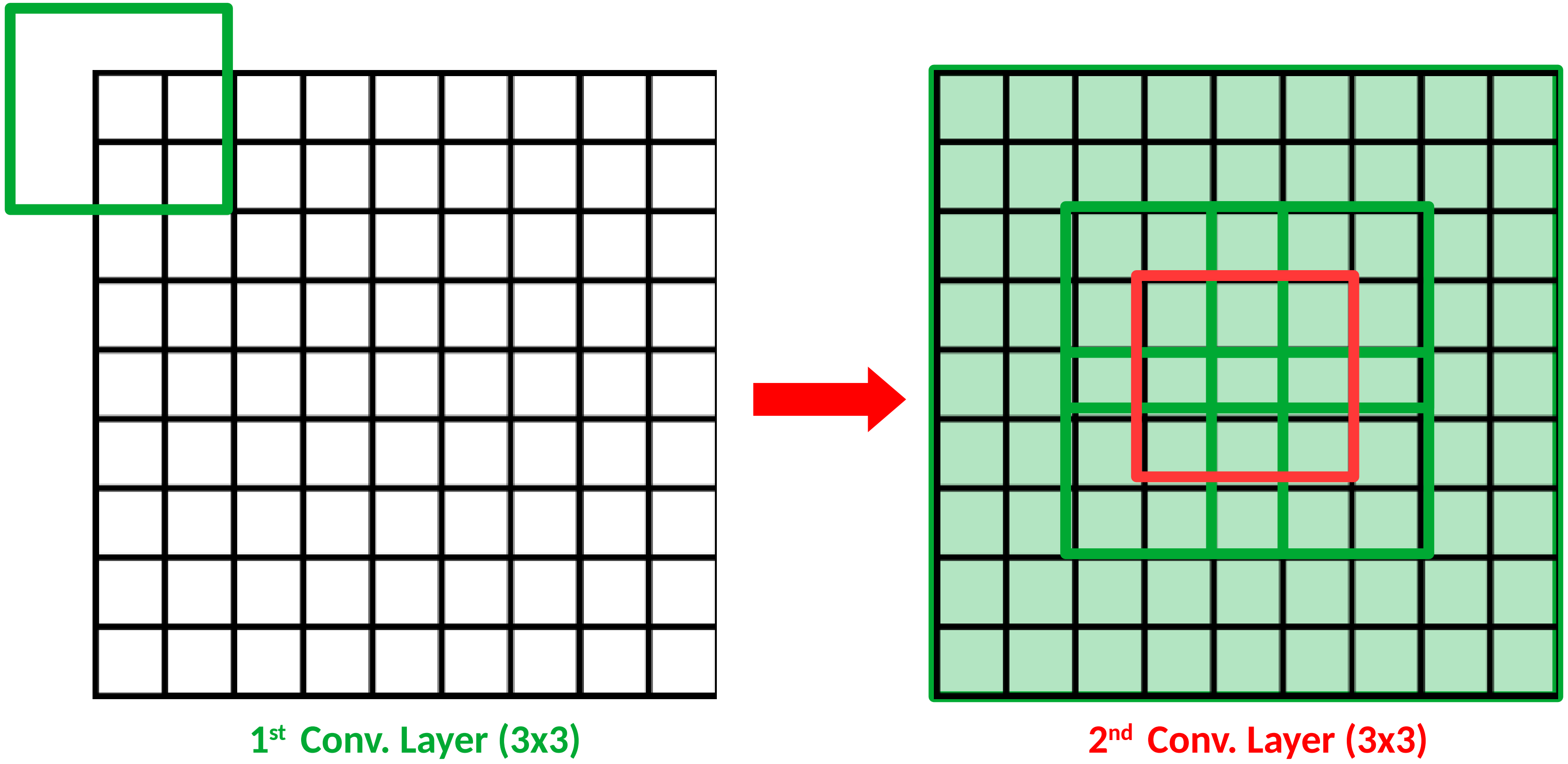


1st Conv. Layer (3x3)

2nd Conv. Layer (3x3)

2014: VGG-Net [Simoyan & Zisserman, 2015]

Going Very Deep via Stacked kernels and Same Convolutions

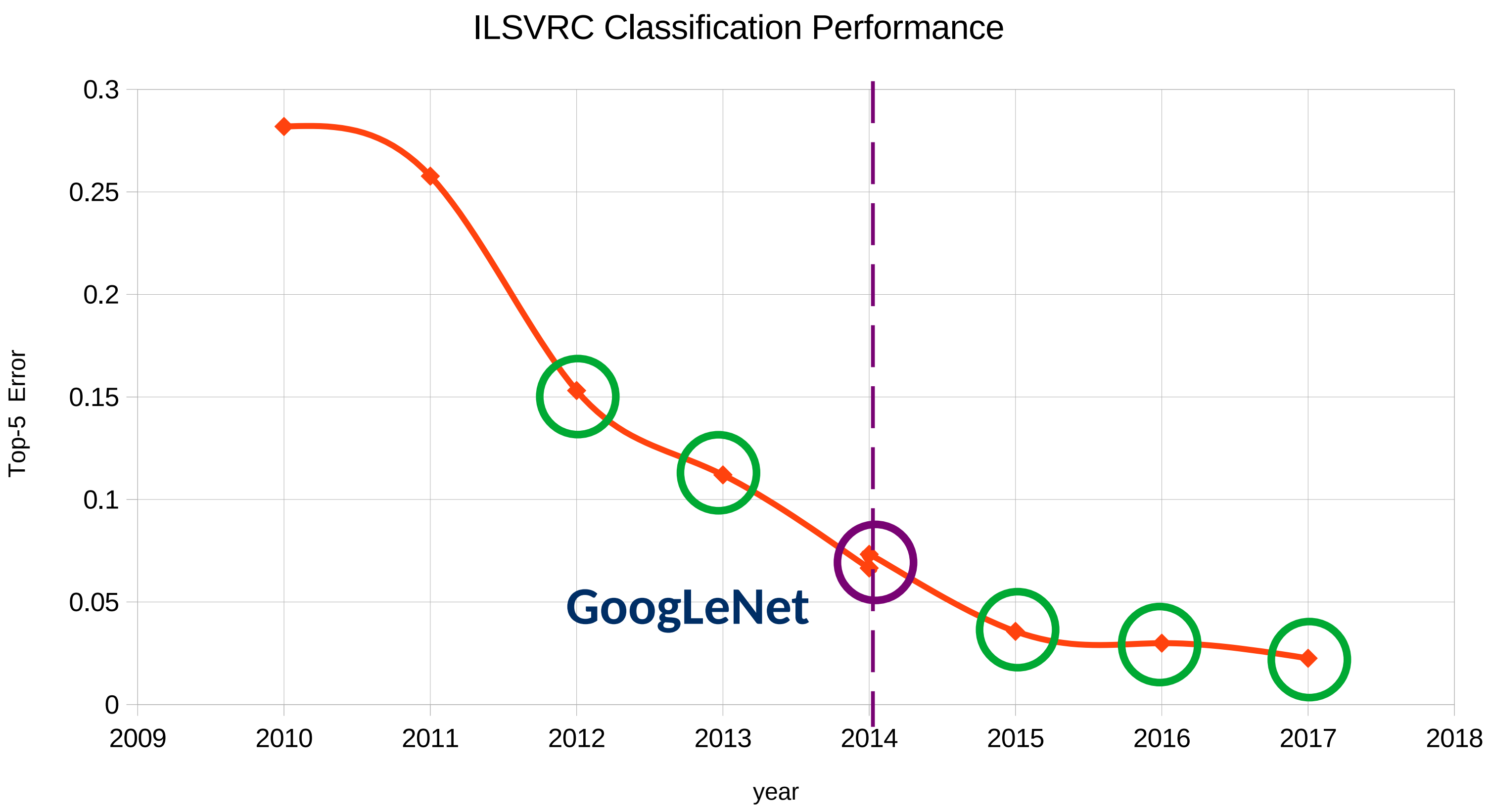


Some Benefits

- Smaller kernels → less parameters to estimate.
- Larger receptive field with less parameters.

2014: GoogLeNet [Simoyan & Zisserman, 2015]

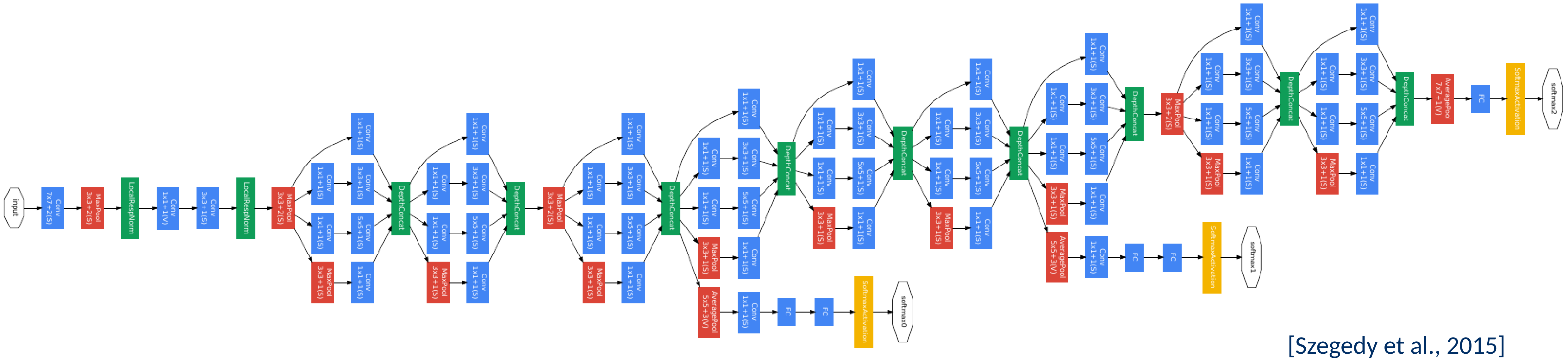
Going Deeper



2014: GoogLeNet [Simoyan & Zisserman, 2015]

Going Deeper

- Branching Architecture
- Aggregate the output of different branches.

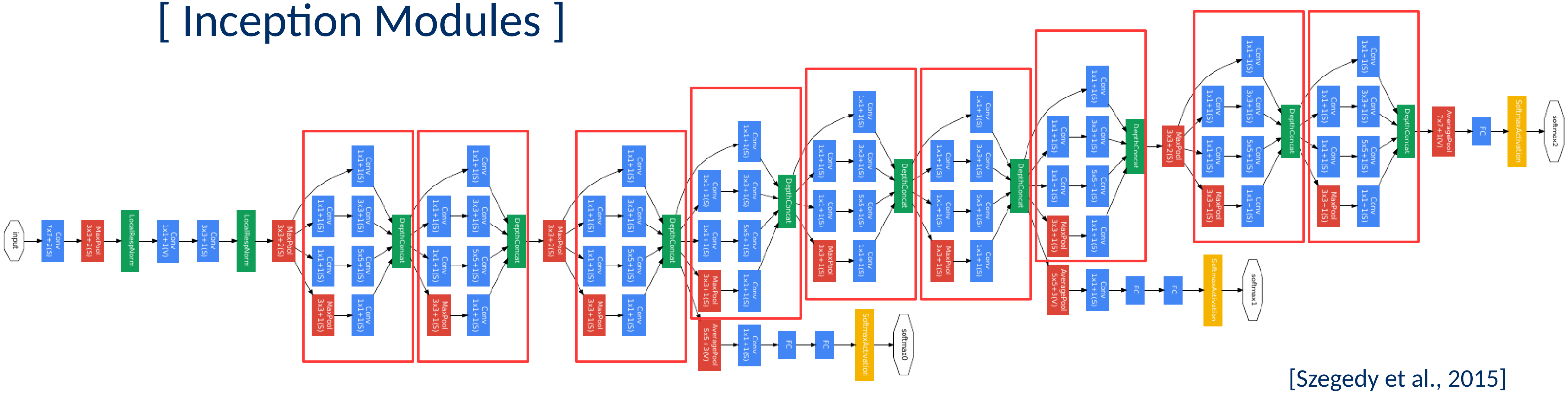


[Szegedy et al., 2015]

2014: GoogLeNet [Simoyan & Zisserman, 2015]

Going Deeper

- Branching Architecture
- Aggregate the output of different branches.
  [Inception Modules]



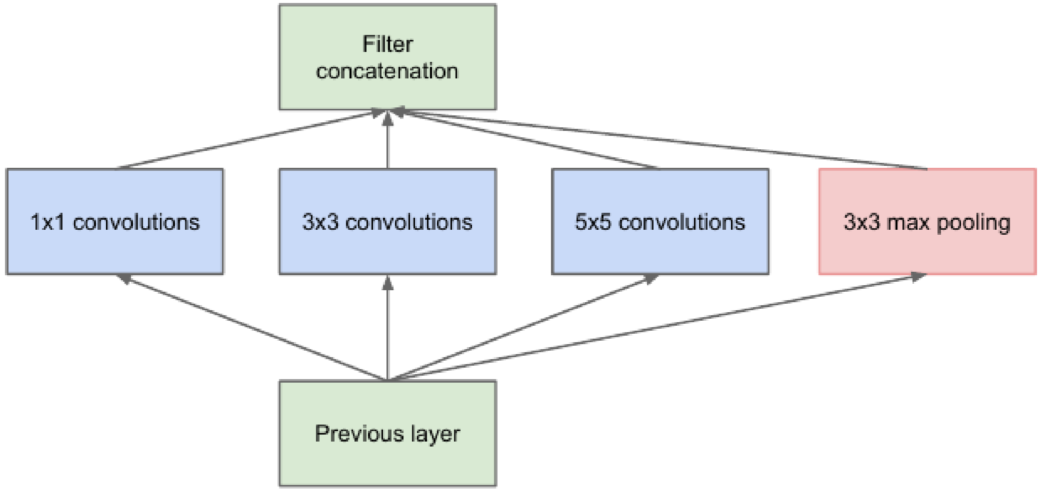
[Szegedy et al., 2015]

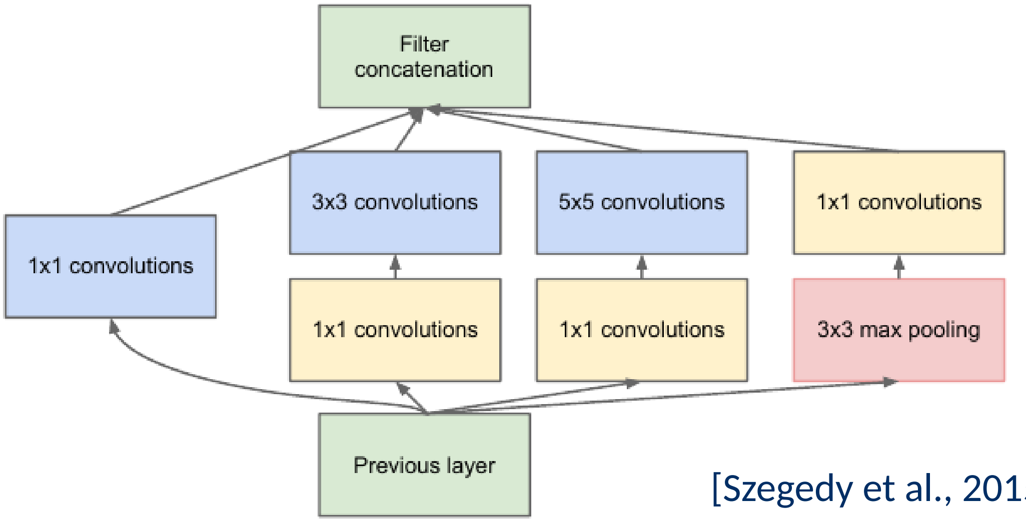
2014: GoogLeNet [Simoyan & Zisserman, 2015]

Going Deeper

Inception Module

- Aggregate the output of different branches.

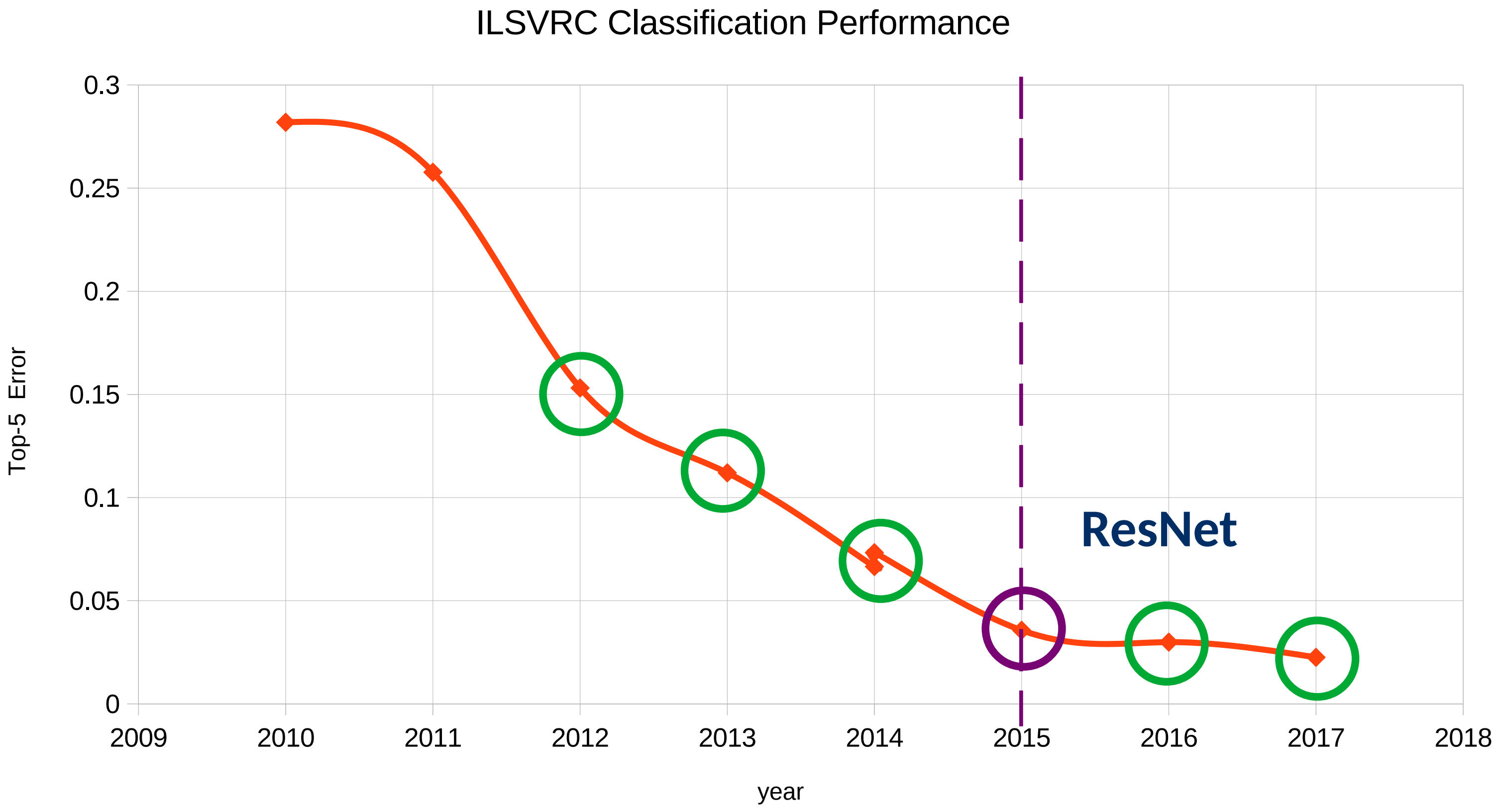




[Szegedy et al., 2015]

2015: ResNet [He et al., 2016]

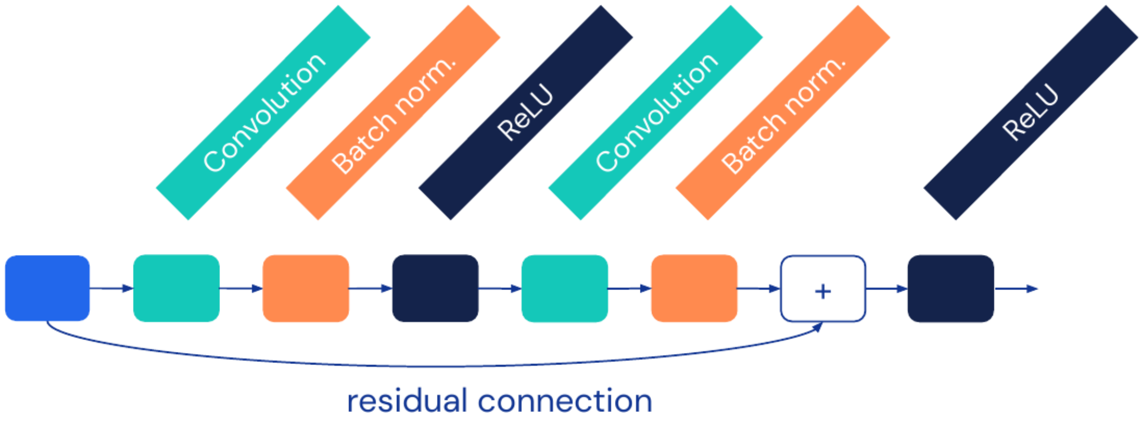
Going even Deeper



2015: ResNet [He et al., 2016]

Going even Deeper

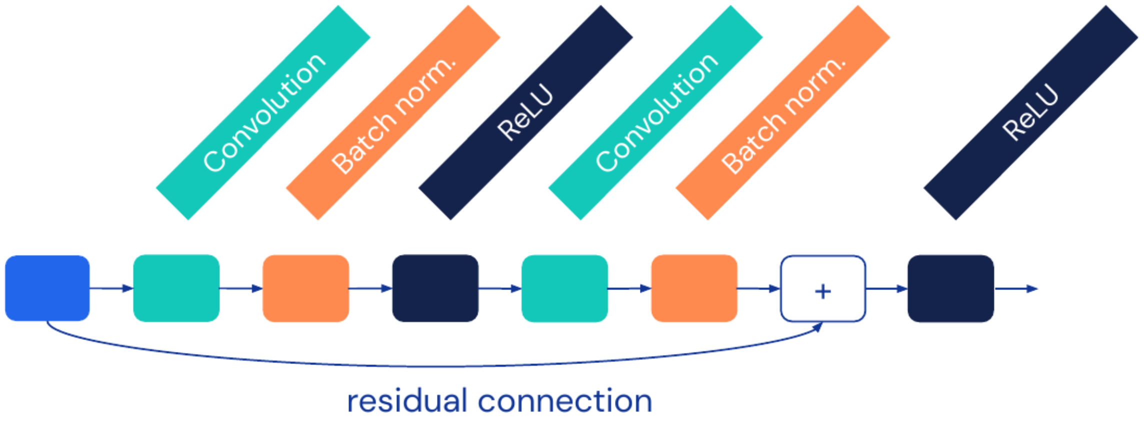
- Provide a skip mechanism to assist the backpropagation of gradients.
- Enable going deeper (18, 34, … , 152 layers!)



2015: ResNet [He et al., 2016]

Going even Deeper

- Provide a skip mechanism to assist the backpropagation of gradients.
- Enable going deeper (18, 34, … , 152 layers!)



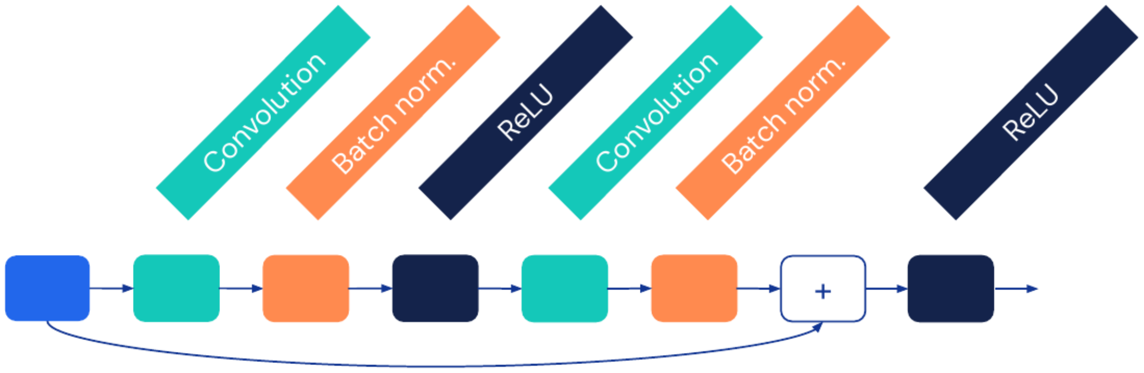
$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}.$$

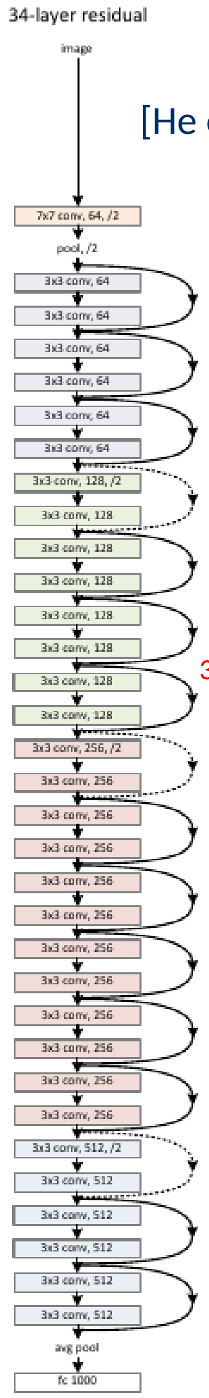
$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s\mathbf{x}.$$

2015: ResNet [He et al., 2016]

Going even Deeper

- Provide a skip mechanism to assist the backpropagation of gradients.
- Enable going deeper (18, 34, … , 152 layers!)





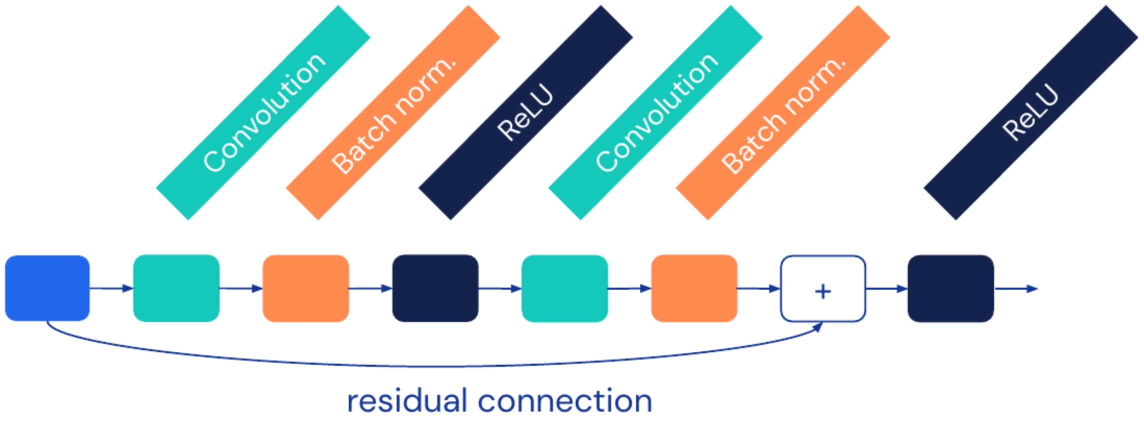
[He et al., 2016]

3.6 billion FLOPS

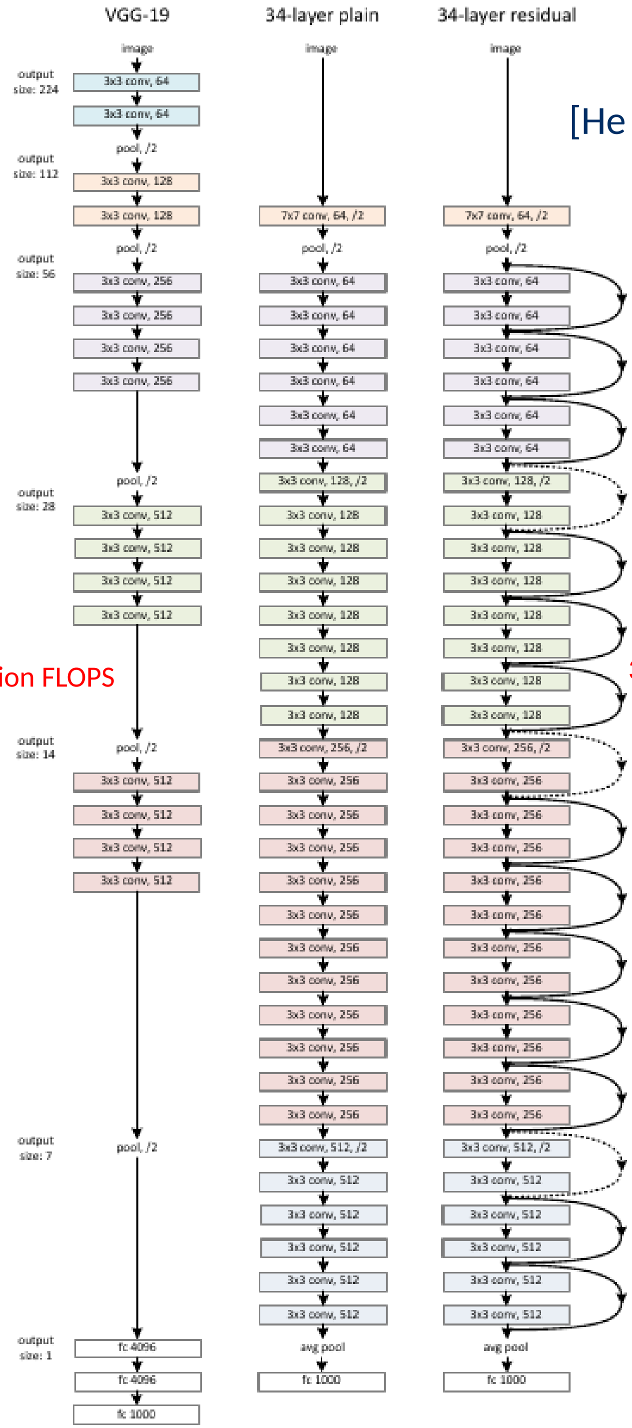
2015: ResNet [He et al., 2016]

Going even Deeper

- Provide a skip mechanism to assist the backpropagation of gradients.
- Enable going deeper (18, 34, … , 152 layers!)



19.6 billion FLOPS

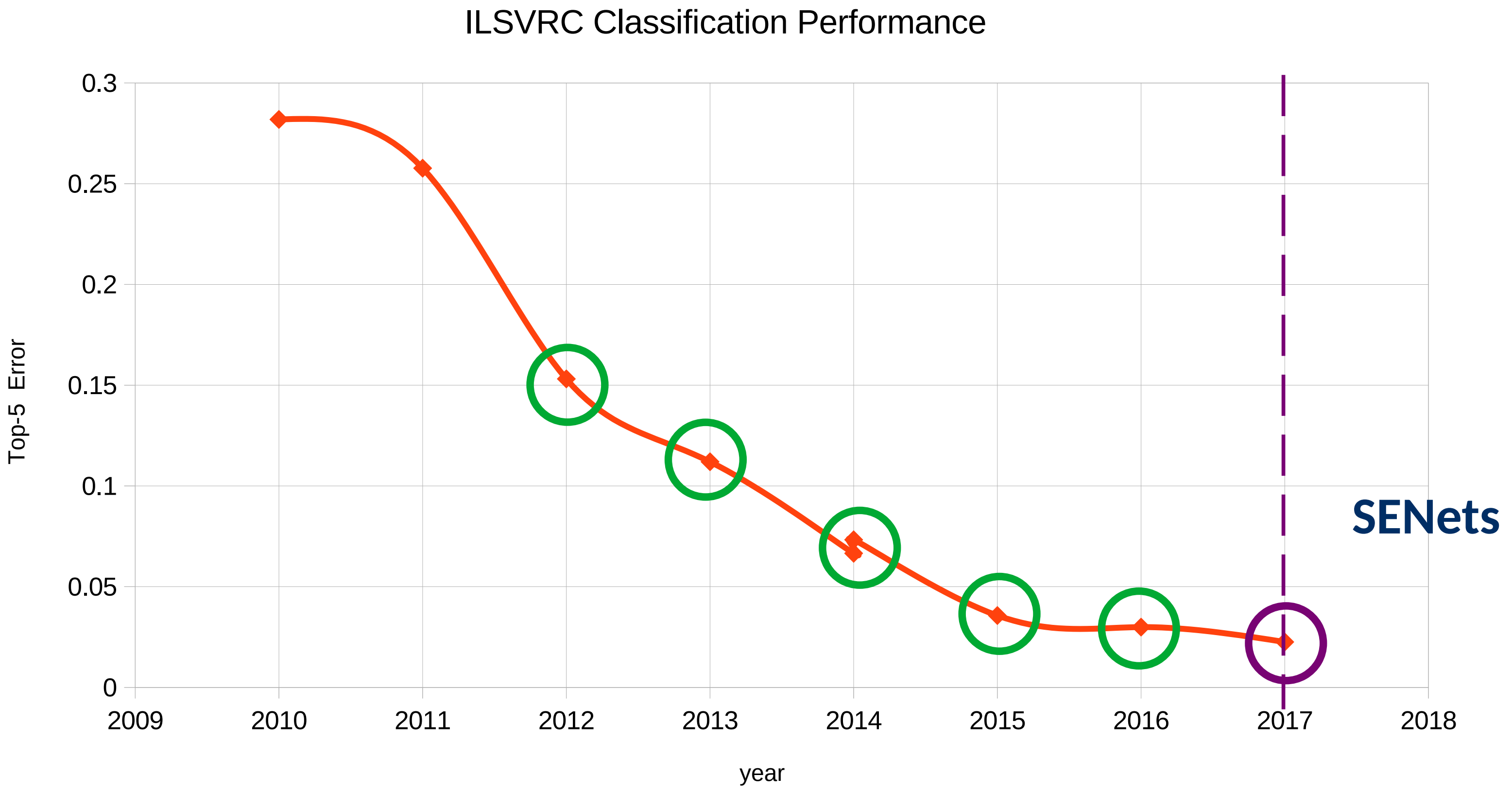


[He et al., 2016]

3.6 billion FLOPS

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

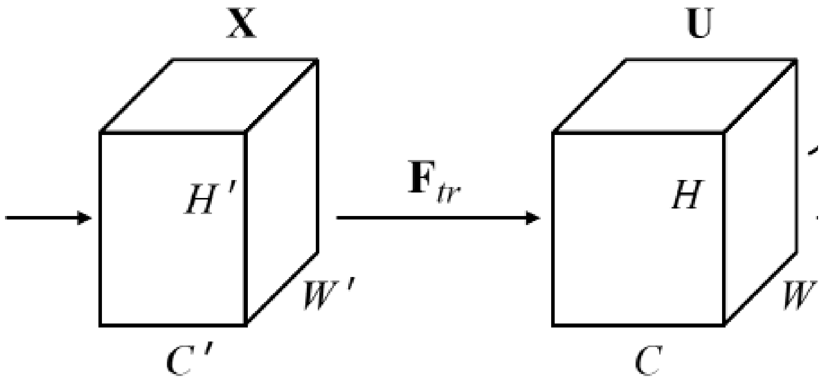


University of Antwerp

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations

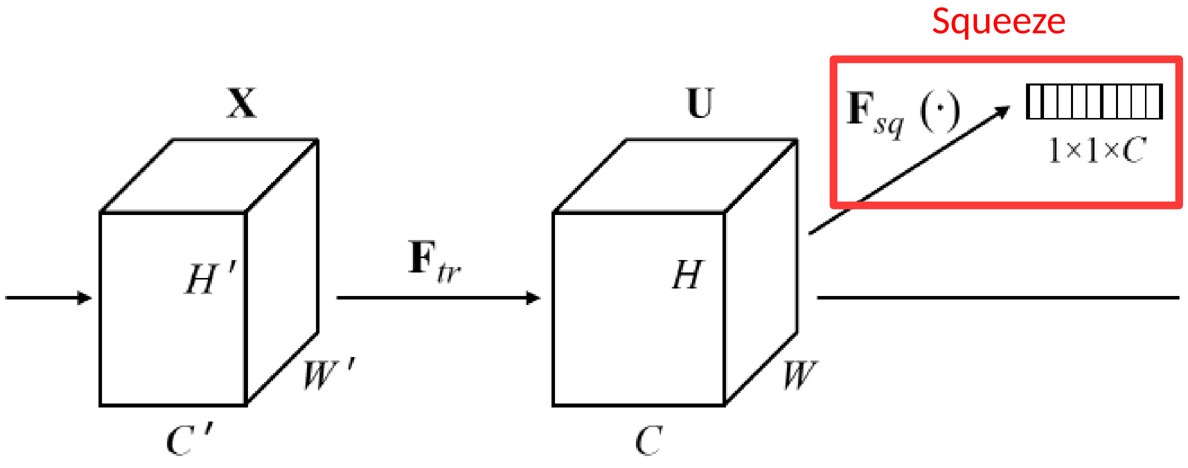


[Hu et al., 2017]

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations
  - **Squeeze:** produce a channel-wise descriptor



[Hu et al., 2017]

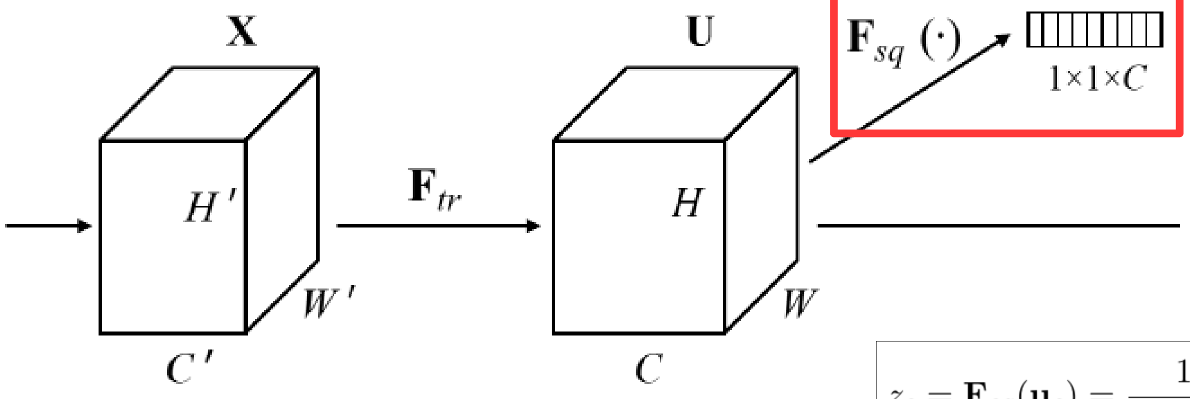
University of Antwerp

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations
  - **Squeeze:** produce a channel-wise descriptor

Squeeze



$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j)$$

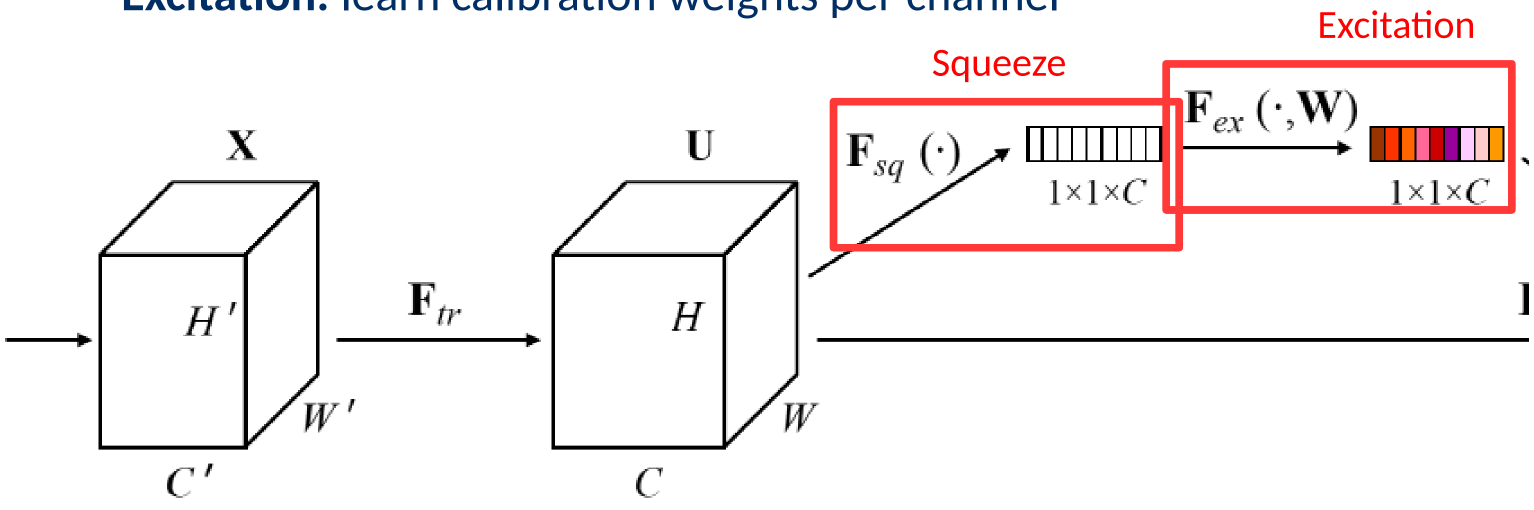
Global Average Pooling (GAP)

[Hu et al., 2017]

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations
  - **Squeeze:** produce a channel-wise descriptor
  - **Excitation:** learn calibration weights per channel

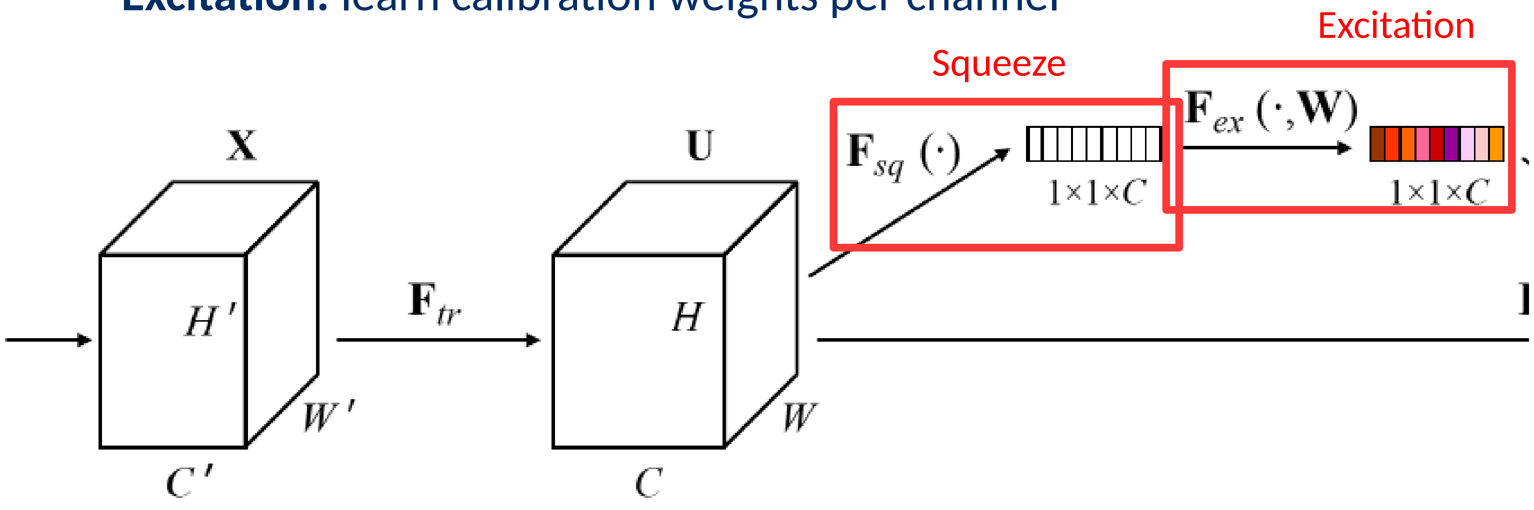


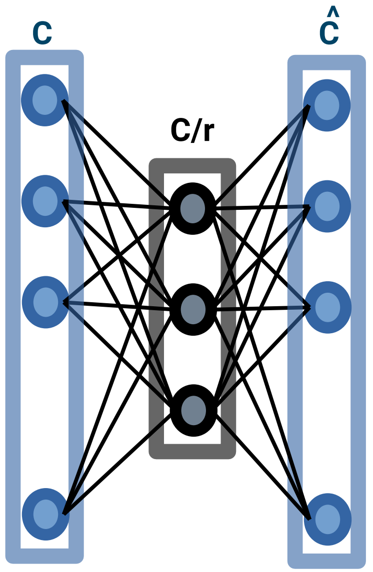
[Hu et al., 2017]

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations
  - **Squeeze:** produce a channel-wise descriptor
  - **Excitation:** learn calibration weights per channel



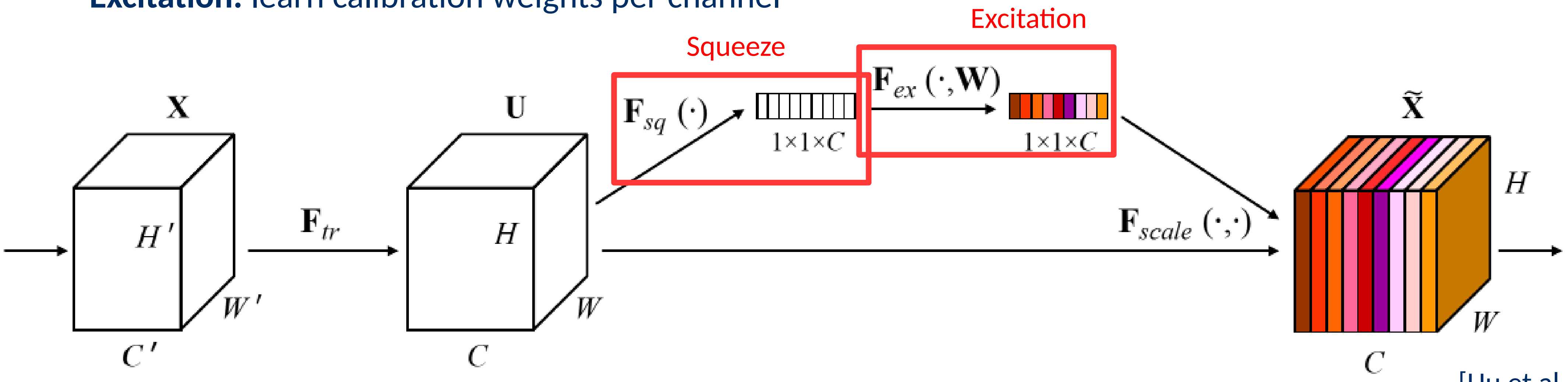


[Hu et al., 2017]

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations
  - **Squeeze:** produce a channel-wise descriptor
  - **Excitation:** learn calibration weights per channel

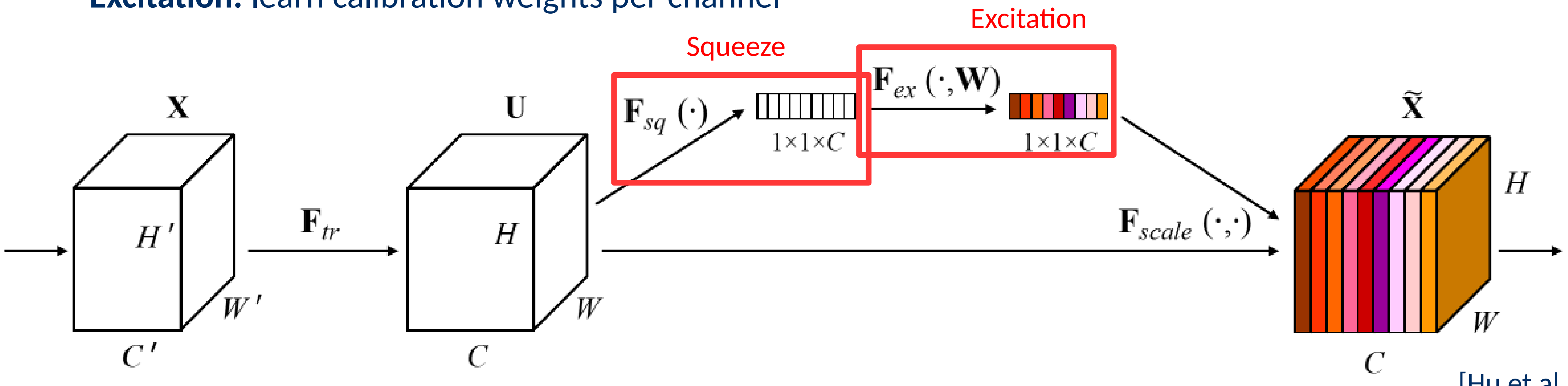


[Hu et al., 2017]

2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations
  - **Squeeze:** produce a channel-wise descriptor
  - **Excitation:** learn calibration weights per channel



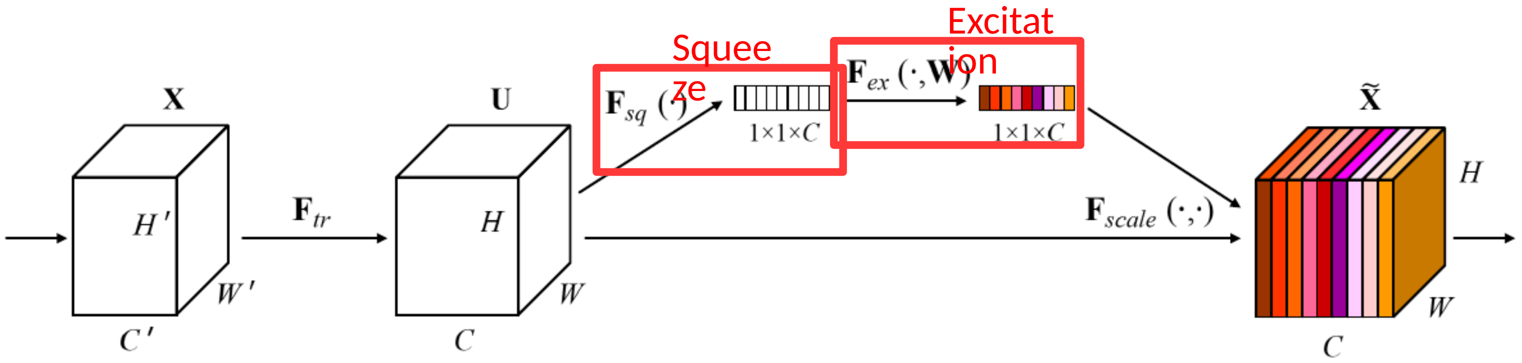
[Hu et al., 2017]

A.k.a. *Channel Attention* or *Self-Attention*

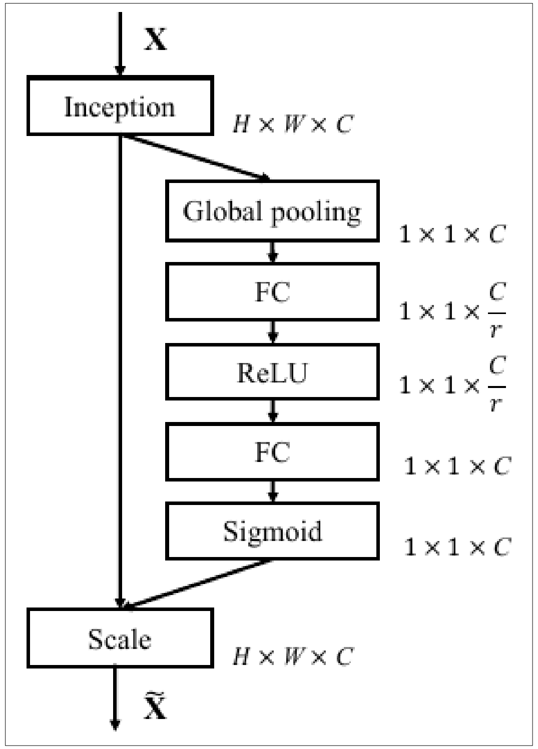
2017: SENet [Hu et al., 2017]

Squeeze and Excitation Networks

- **Problem:** Convolution is a very local operation
- **Do:** Propagate channel information at different spatial locations
  - **Squeeze:** produce a channel-wise descriptor
  - **Excitation:** learn calibration weights per channel



[Hu et al., 2017]



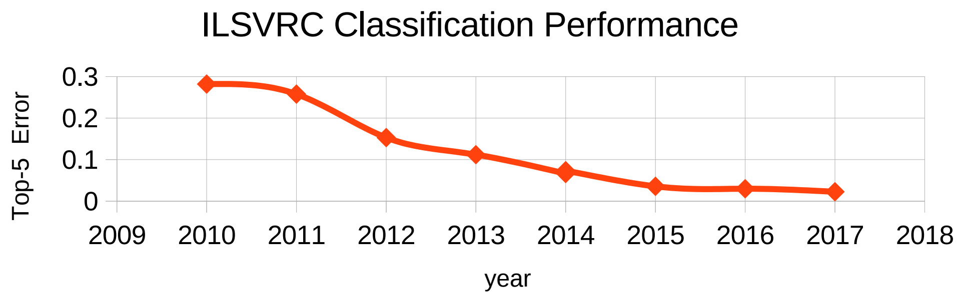
Summarizing

[Finally :D]

Summarizing

- **ConvNets are not new**

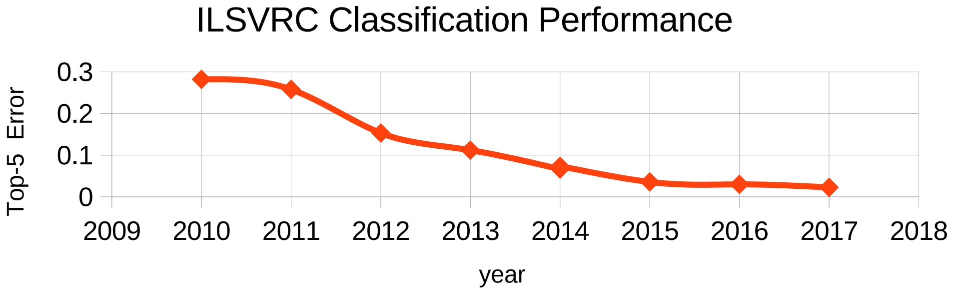
  lots of progress in the last decade



Summarizing

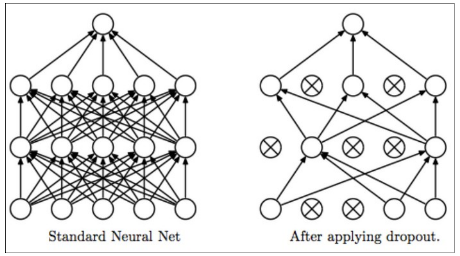
- **ConvNets are not new**

  lots of progress in the last decade



- **Serveral techniques to assist training**

  Data augmentation | Dropout

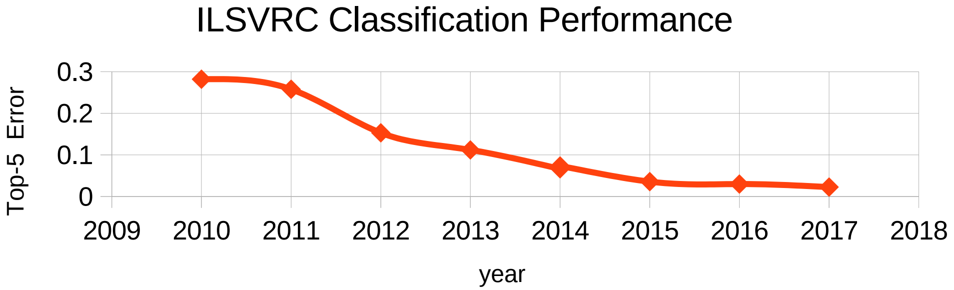


University of Antwerp

Summarizing

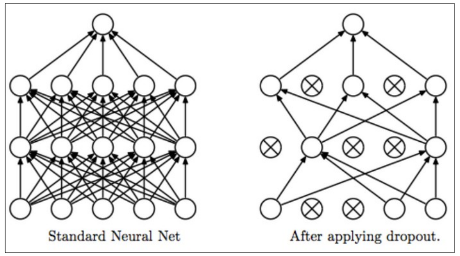
- **ConvNets are not new**

  lots of progress in the last decade



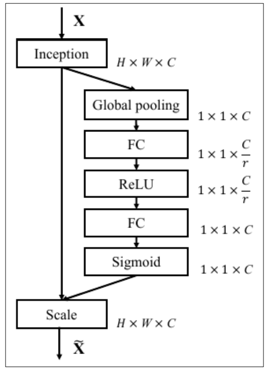
- **Serveral techniques to assist training**

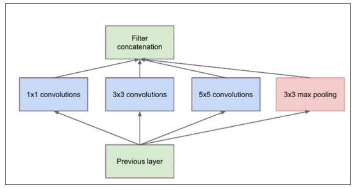
  Data augmentation | Dropout

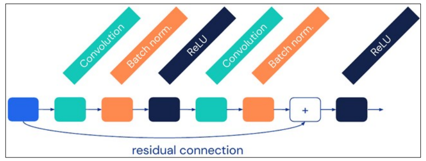


- **Relevant new components**

  inception | Residual | Squeeze-Excitation blocks







References

- Kunihiko Fukushima, Sei Miyake, ***Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position***, Pattern Recognition, Volume 15, Issue 6. 1982.
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, ***Handwritten digit recognition with a back-propagation network***. NeurIPS 1989
- Y. Lecun, L. Bottou, Y. Bengio and P. Haffner. ***Gradient-based Learning Applied to Document Recognition***. Proceedings of IEEE, 1998
- A. Krizhevsky, I. Sutskever, G. E. Hinton. ***ImageNet Classification with Deep Convolutional Neural Networks***. NeurIPS 2012
- K. Simonyan & A. Zisserman, ***Very Deep Convolutional Networks for large-scale Image Recognition***, ICLR 2015
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, ***Going deeper with convolutions***, CVPR 2015.
- K. He, X. Zhang, S. Ren, J. Sun, ***Deep Residual Learning for Image Recognition***, CVPR 2016.
- J. Long, E. Shelhamer, T. Darrell, ***Fully Convolutional Networks for Semantic Segmentation***, CVPR 2015.
- J. Hu1, L. Shen, G. Sun, ***Squeeze-and-Excitation Networks***, CVPR 2017
- D. E. Rumelhart, G. E. Hinton & R. J. Williams. ***Learning representations by back-propagating errors.*** 1986
- L. Antanas, M. van Otterlo, J. Oramas, T. Tuytelaars and L. De Raedt. ***There are Plenty of Places like Home: Using Hierarchies and Relational Representations for Distance-based Image Understanding***. Neurocomputing 2014.

Convolutional Neural Networks

[ConvNets, CNNs]