# Intro

- In this video we will overview modern architectures of neural networks

# ImageNet classification dataset

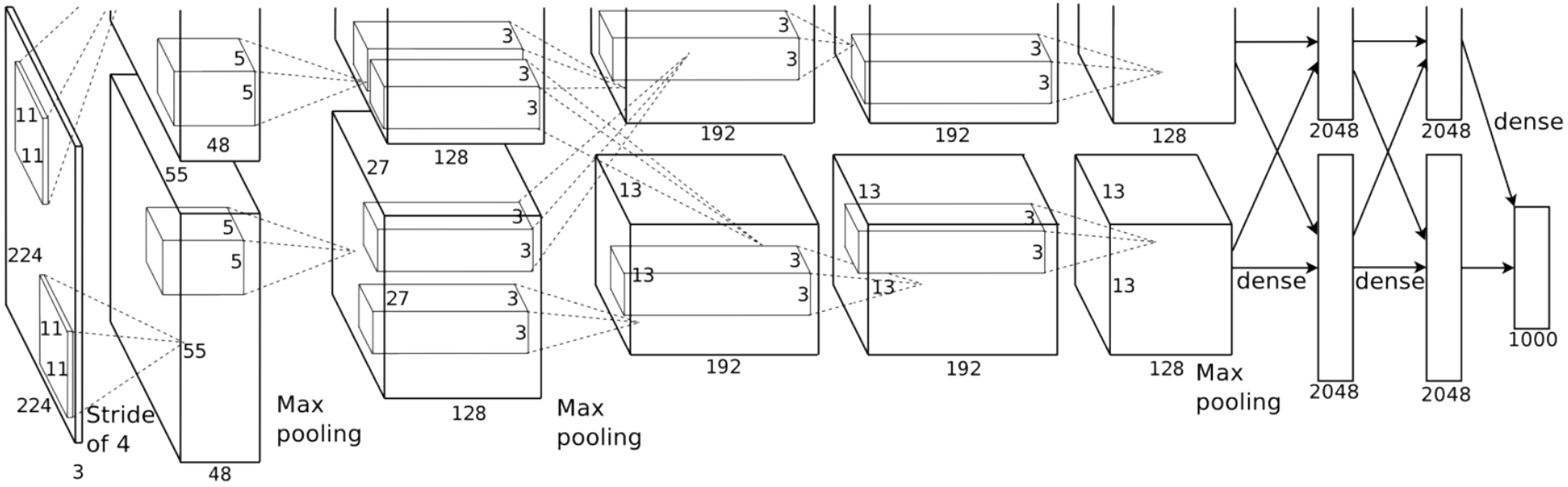
1000 classes, 1.2 million labeled photos

Human top 5 error: ~5%


Olga Russakovsky, https://arxiv.org/pdf/1409.0575.pdf

# AlexNet (2012)

- First deep convolutional neural net for ImageNet
- Significantly reduced top 5 error from 26% to 15%

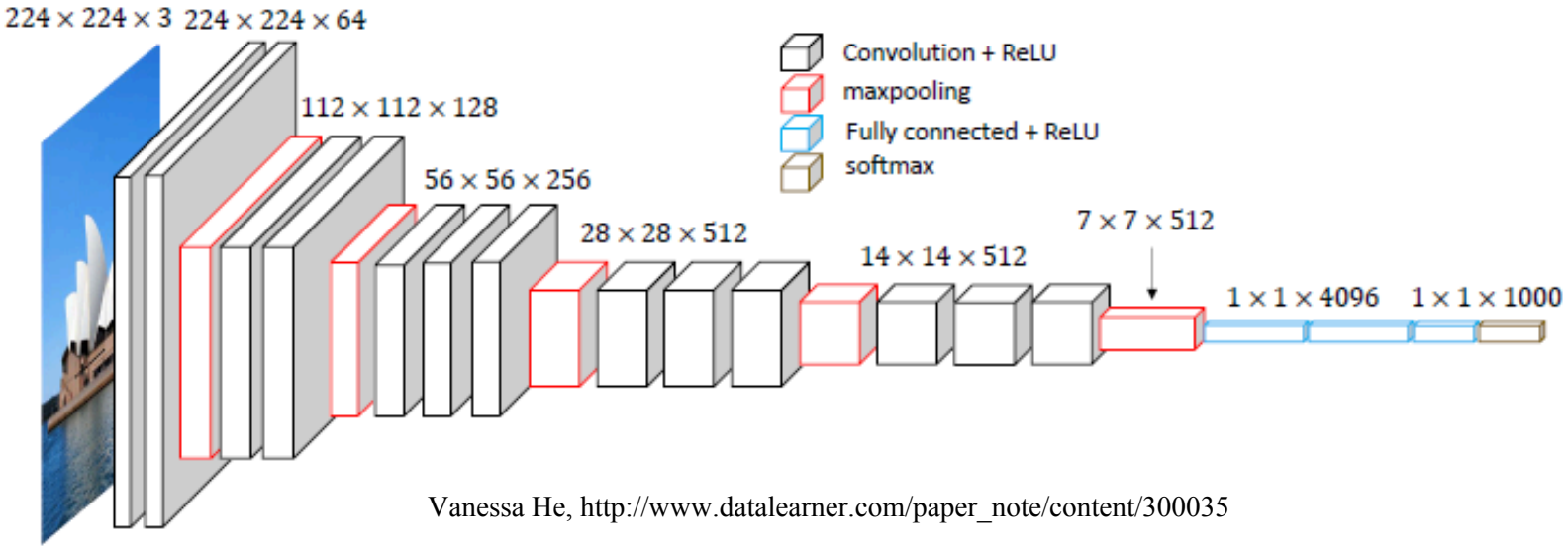



Alex Krizhevsky, https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

- 11x11, 5x5, 3x3 convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum
- 60 million parameters
- Trains on 2 GPUs for 6 days

# VGG (2015)

- Similar to AlexNet, only 3x3 convolutions, but lots of filters!
- ImageNet top 5 error: 8.0% (single model)

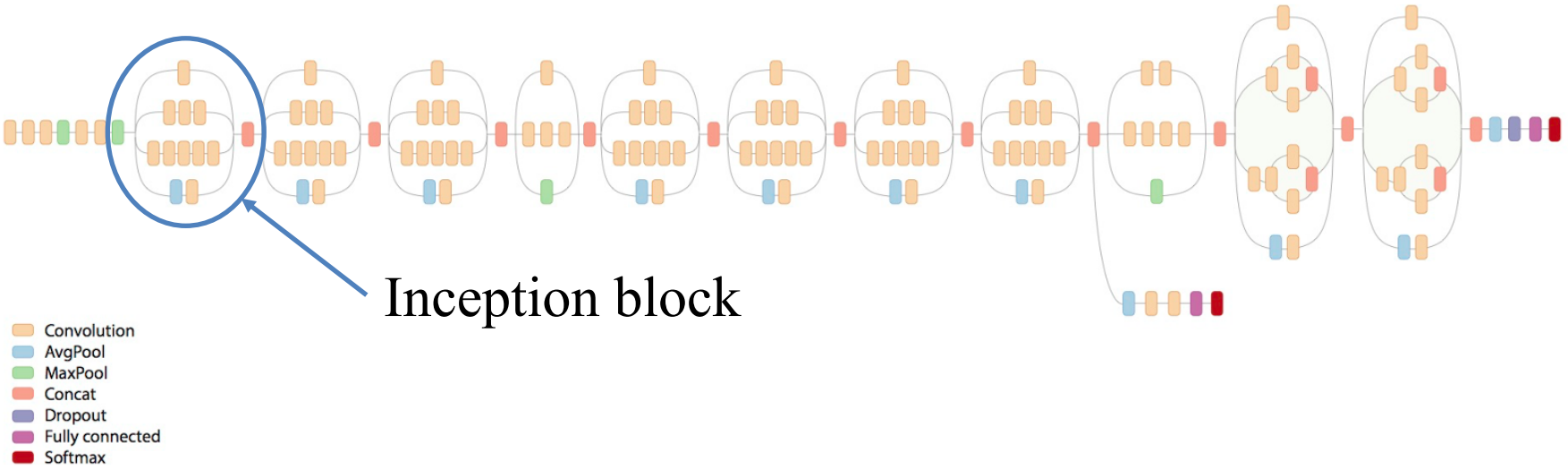



Vanessa He, http://www.datalearner.com/paper_note/content/300035

- Training similar to AlexNet with additional multi-scale cropping.
- 138 million parameters
- Trains on 4 GPUs for 2-3 weeks

# Inception V3 (2015)

- Similar to AlexNet? Not quite, uses Inception block introduced in GoogLeNet (a.k.a. Inception V1)
- ImageNet top 5 error: 5.6% (single model), 3.6% (ensemble)

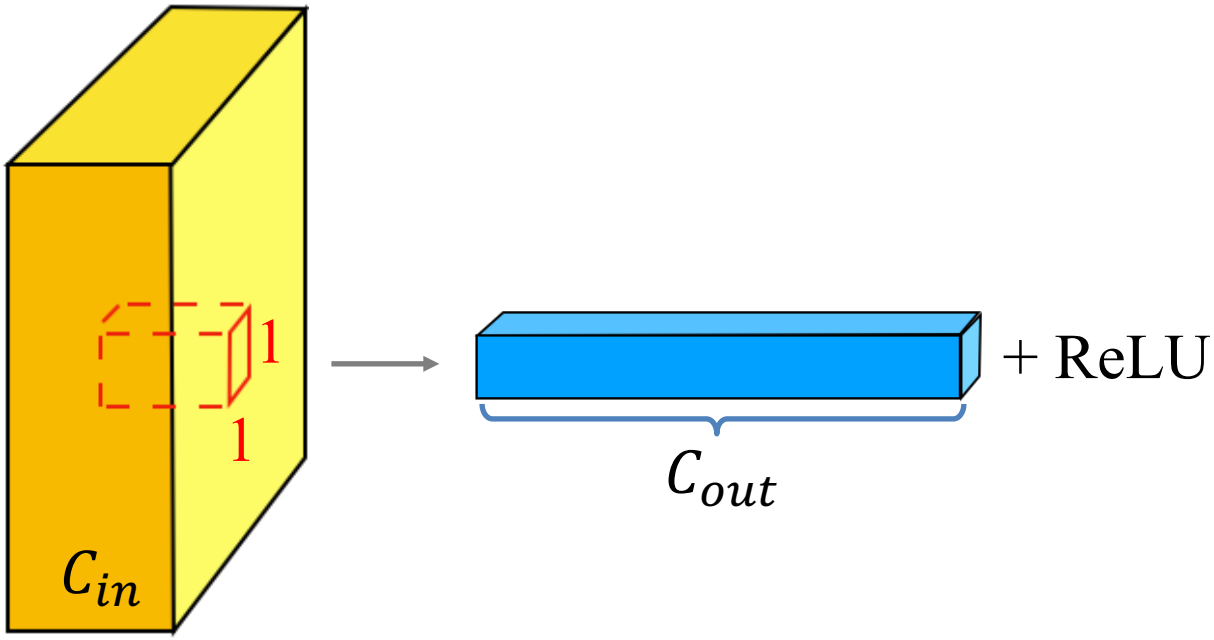



Jon Shlens, https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html

- Batch normalization, image distortions, RMSProp
- 25 million parameters!
- Trains on 8 GPUs for 2 weeks

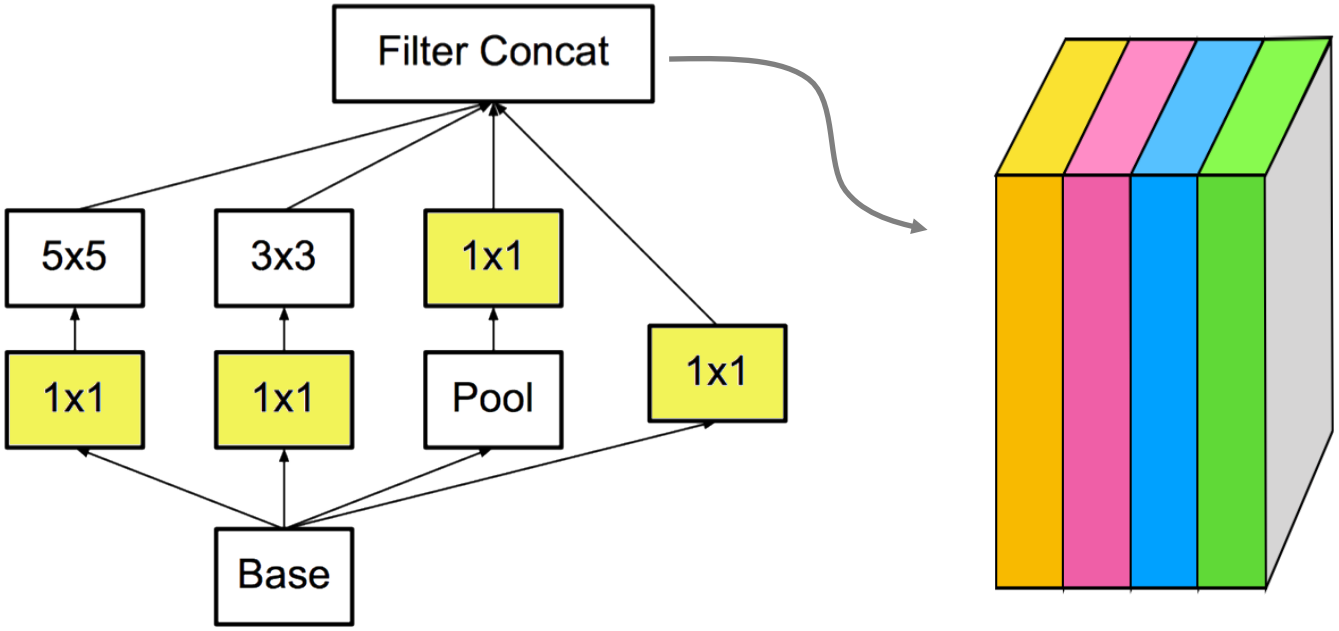
# 1x1 convolutions

- Such convolutions capture interactions of input channels in one "pixel" of feature map
- They can reduce the number of channels not hurting the quality of the model, because different channels can correlate
- Dimensionality reduction with added ReLU activation



# Basic Inception block

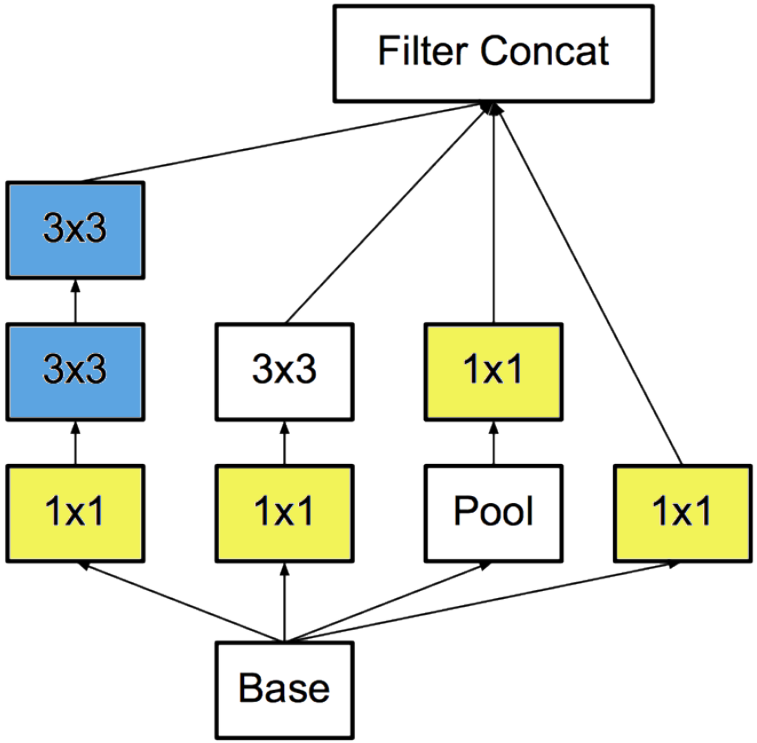
- All operations inside a block use stride 1 and enough padding to output the same spatial dimensions ($W \times H$) of feature map.
- 4 different feature maps are concatenated on depth at the end

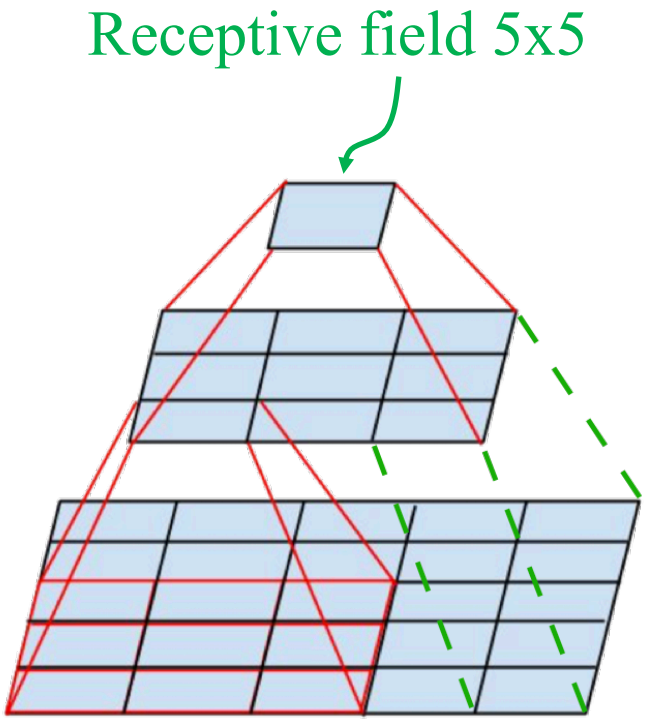



Christian Szegedy, https://arxiv.org/pdf/1512.00567.pdf

# Replace 5x5 convolutions

5x5 convolutions are expensive! Let's replace them with two layers of 3x3 convolutions which have an effective receptive field of 5x5.

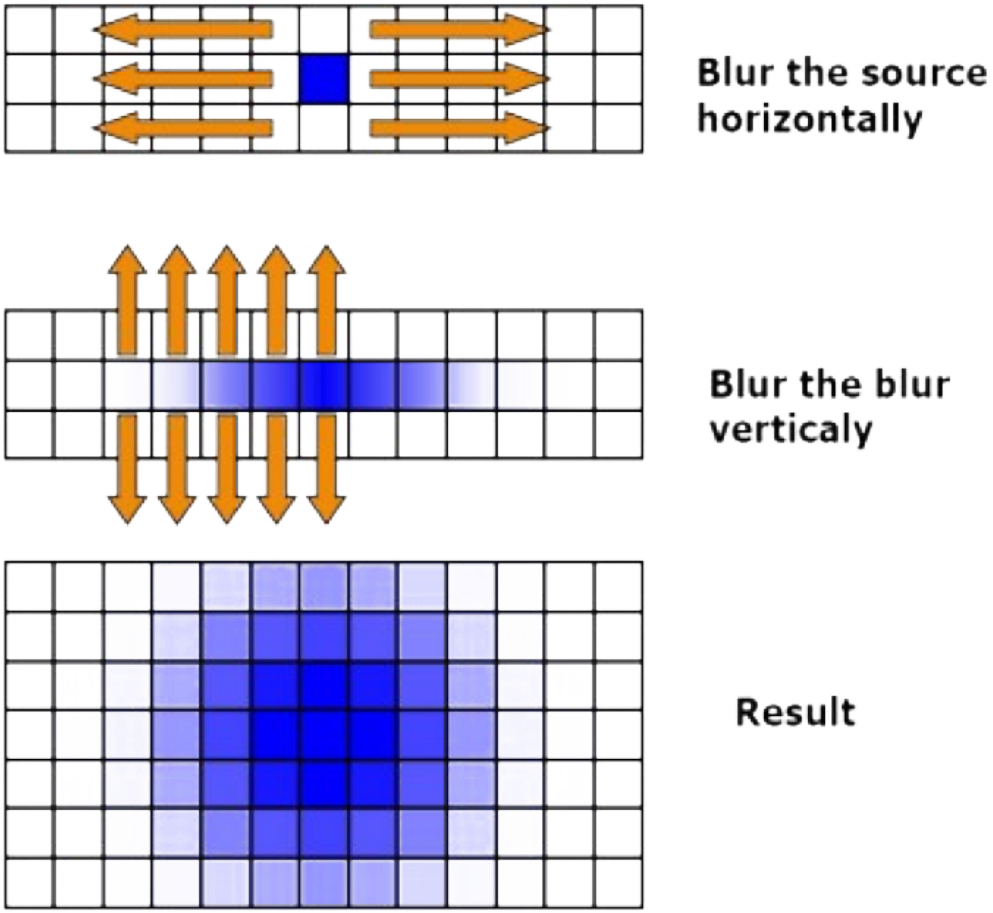



Christian Szegedy, https://arxiv.org/pdf/1512.00567.pdf

# Filter decomposition

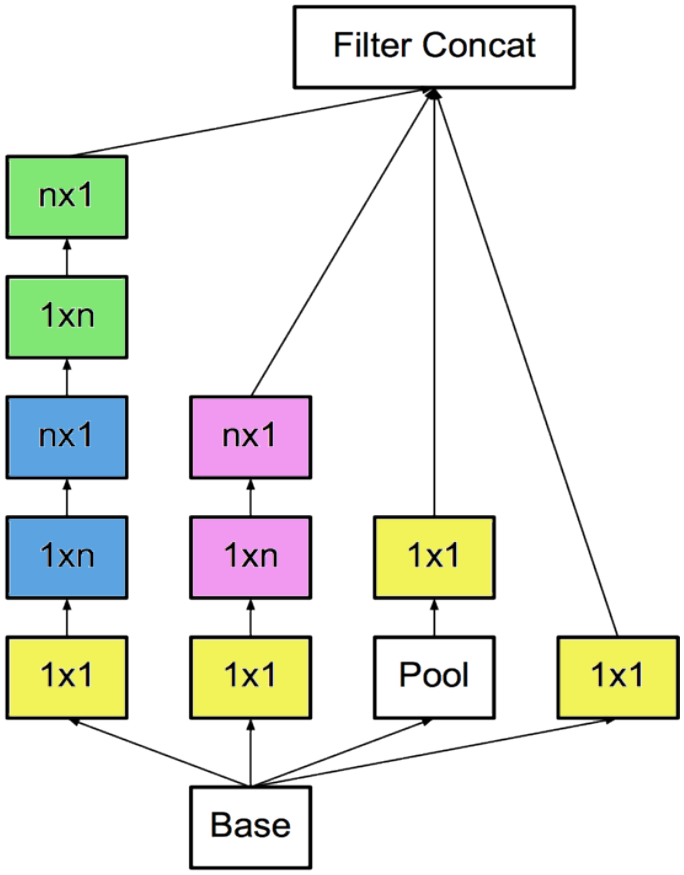
It’s known that a Gaussian blur filter can be decomposed in two 1 dimensional filters:




Ati’s presentation, http://www.florian-oeser.de/rtr/ue5/

# Filter decomposition in Inception block

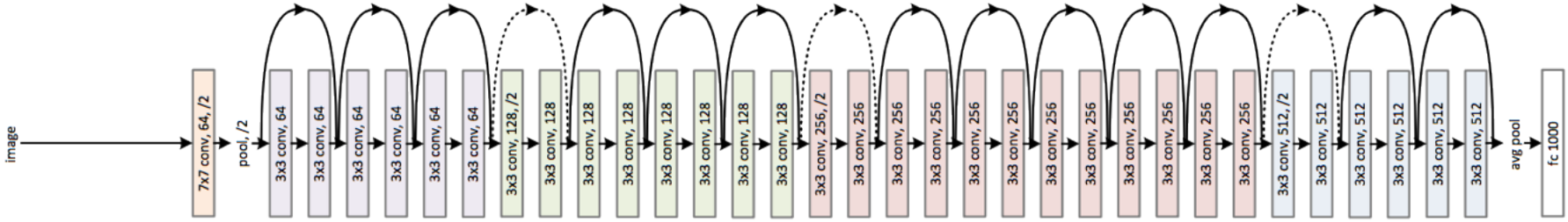
- 3x3 convolutions are currently the most expensive parts!
- Let's replace each 3x3 layer with 1x3 layer followed by 3x1 layer.




Christian Szegedy, https://arxiv.org/pdf/1512.00567.pdf

# ResNet (2015)

- Introduces residual connections
- ImageNet top 5 error: 4.5% (single model), 3.5% (ensemble)

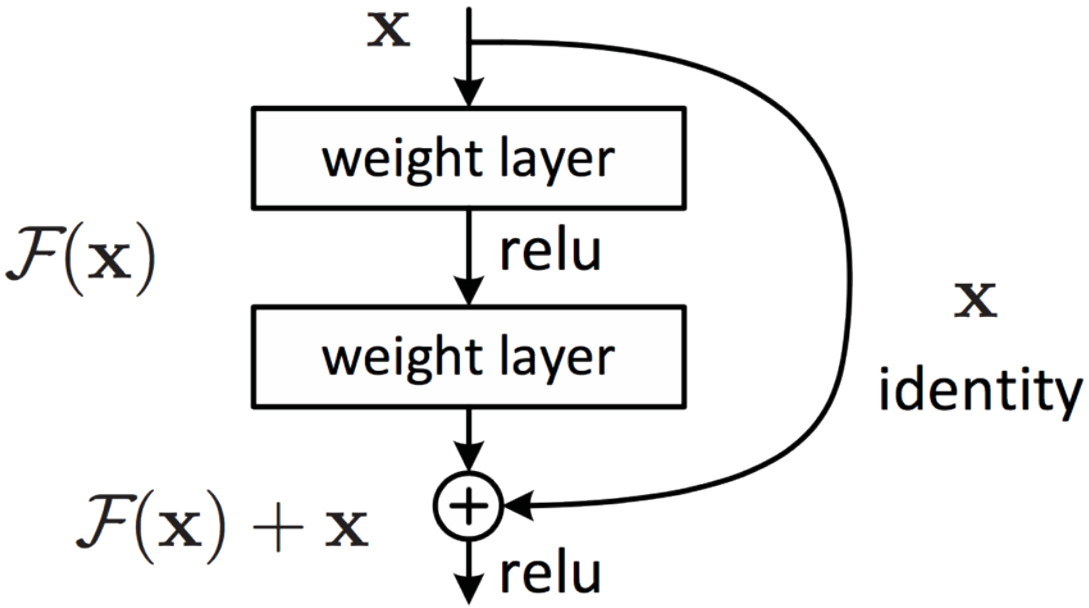



Kaiming He, https://arxiv.org/pdf/1512.03385.pdf

- 152 layers, few 7x7 convolutional layers, the rest are 3x3, batch normalization, max and average pooling.
- 60 million parameters
- Trains on 8 GPUs for 2-3 weeks.

# Residual connections

- We create output channels adding a small delta $F(x)$ to original input channels $x$:




Kaiming He, https://arxiv.org/pdf/1512.03385.pdf

- This way we can stack thousands of layers and gradients do not vanish thanks to residual connections

# Summary

- By stacking more convolution and pooling layers you can reduce the error! Like in AlexNet or VGG.
- But you cannot do that forever, you need to utilize new kind of layers like Inception block or residual connections.
- You've probably noticed that one needs a lot of time to train her neural network!
- In the following video we'll discuss the principle known as transfer learning that will help us to reduce the training time for a new task!