Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

CNN Architectures and Transfer Learning

02.12.2022

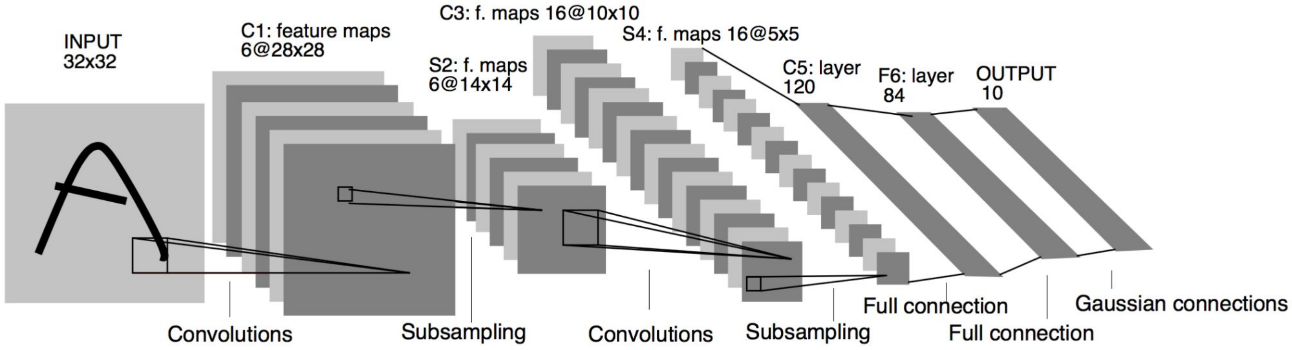
PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

Contact: villar@ais.uni-bonn.de

Early Architectures

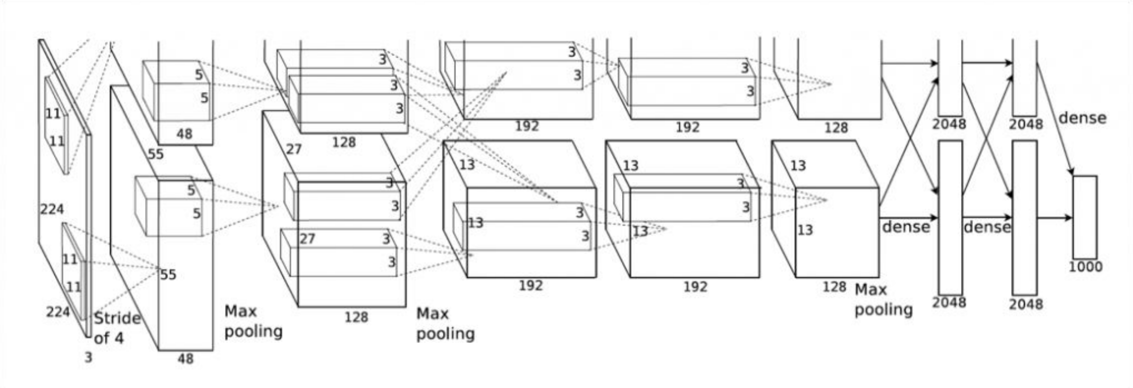
LeNet-5 (1998)

- Very first CNN, and inspiration for future architectures
- Key Features:
  - Conv. of spatial features
  - Subsampling through average pooling
  - Convolutional feature extractor
  - MLP classifier head



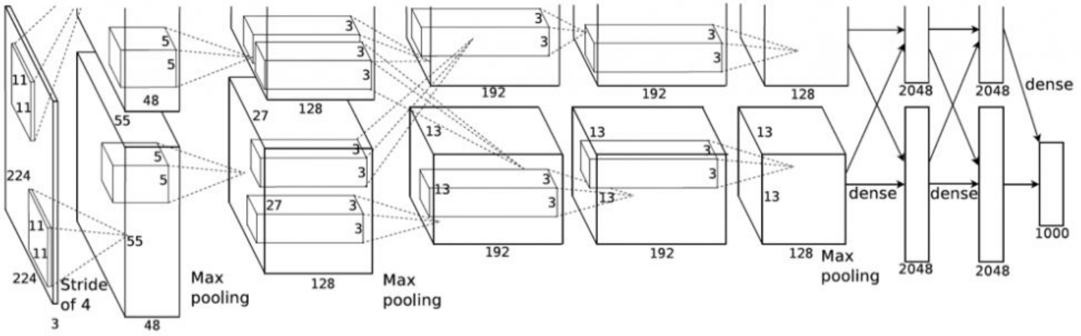
AlexNet (2012)

- Winner of 2012 Imagenet challenge ⇒ Breakthrough of CNNs
- Architectural Features:
  - 8 layers deep
  - Big convolutional kernels
  - Overlapping max-pooling
  - ➢ ReLU activation function



AlexNet (2012)

- Winner of 2012 Imagenet challenge ⇒ Breakthrough of CNNs
- Regularization Features:
  - ➢ Dropout regularization with p=0.5 in Fully-Connected layers
  - ➢ Data augmentation

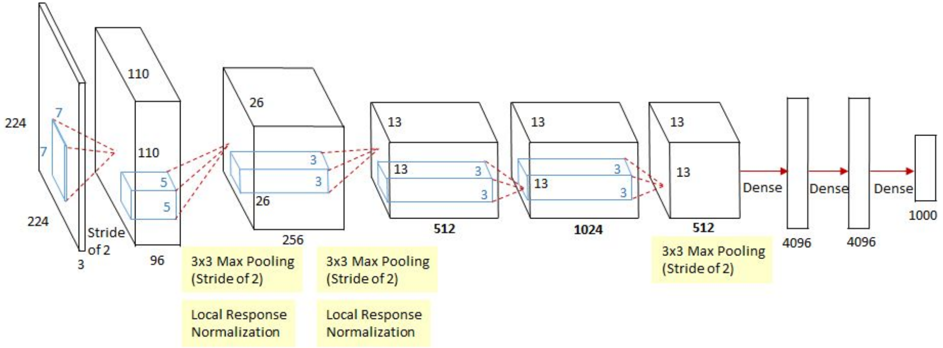


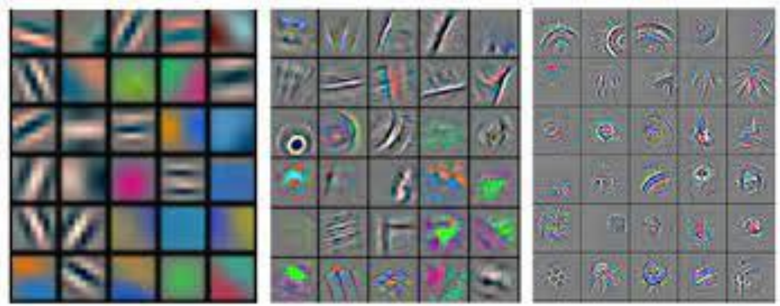
ZF-Net (2013)

- Mostly a fine-tuned version of AlexNet

➢ Use of smaller convolutional kernels
  - Initial 7x7 convolution with stride of 2

- Gave insights about what CNNs learn



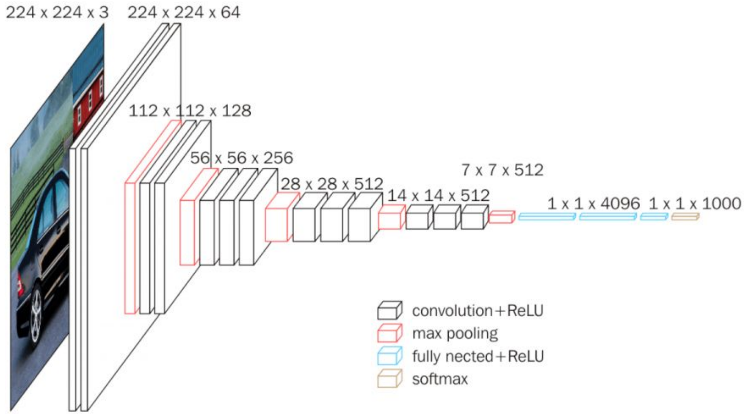


VGG (2014)

- Based on two pillars: simplicity and depth

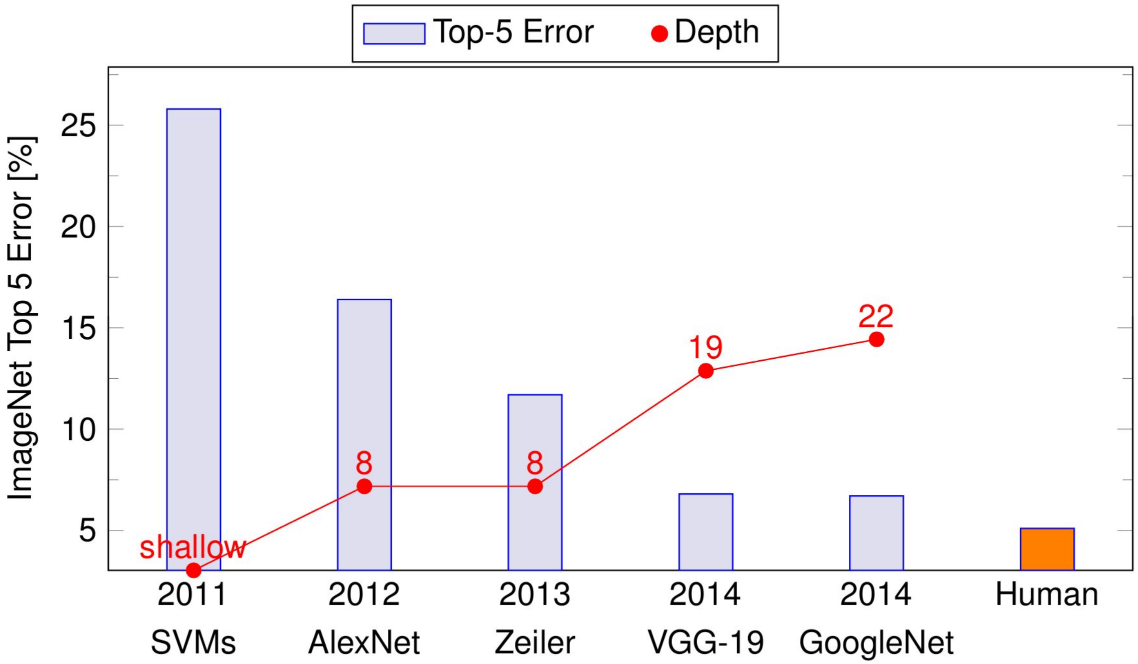
➢ Consolidated rules for modern convolutional layers
  - ➢ Small convolutional kernels
  - ➢ Many kernels per layer

➢ Exploiting hierarchy of features
  - ➢ Spatial size decreases
  - ➢ Depth increases



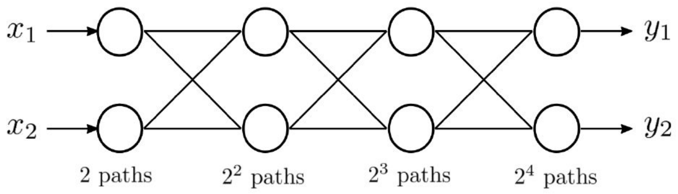
Deeper Models

Evolution of Depth

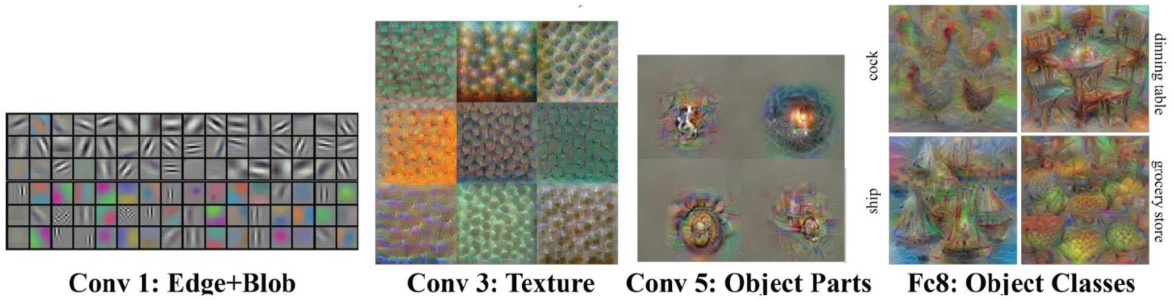


Advantages of Deeper Networks

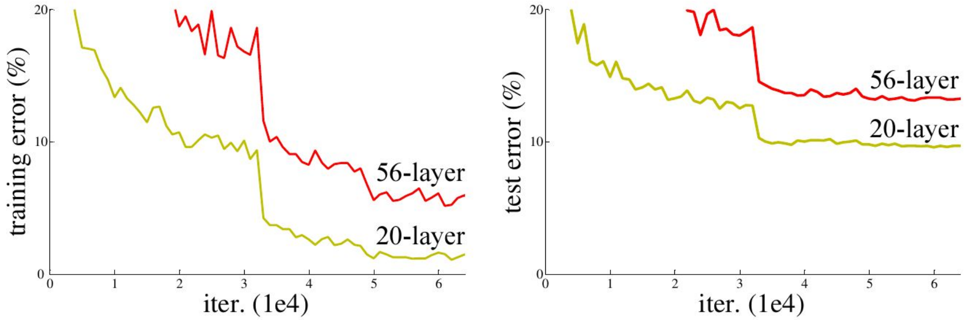
- Exponential feature reuse



- Hierarchical and increasingly abstract features



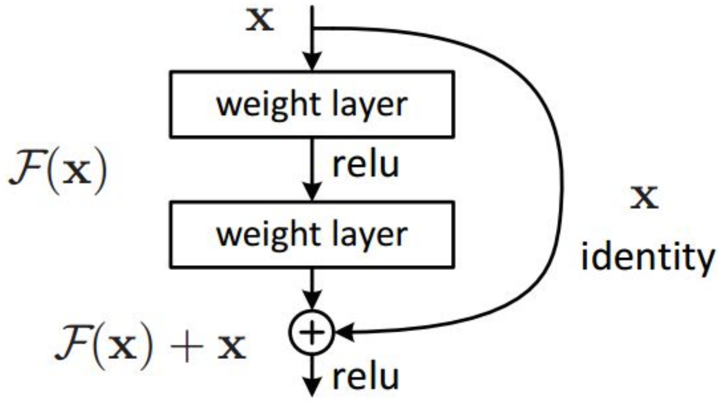
The Degradation Problem



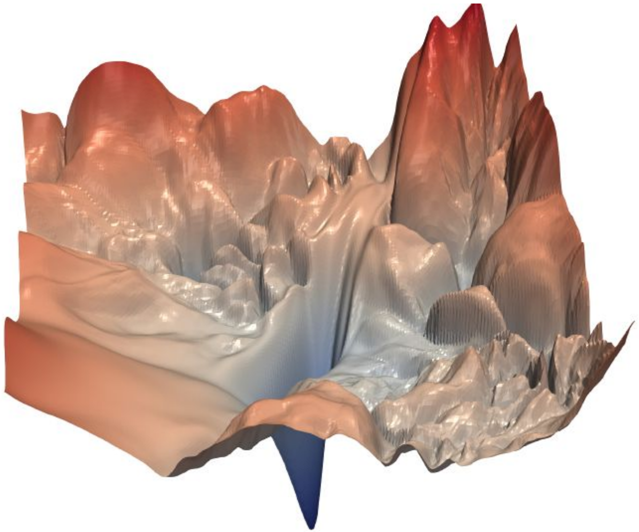
- Deeper models tend to have higher **training & test error** than shallow ones
  - Not just due to overfitting!
- Possible reasons:
  - **Vanishing gradients** due to activations
  - **Co-variate shifts** due to non-centered activations or normalizations
  - Poor **backpropagation** of activations and gradients

Residual Learning

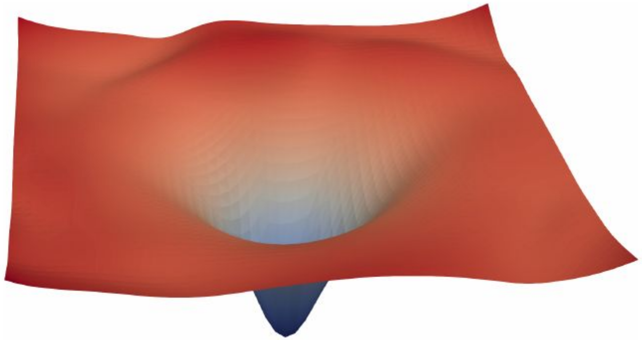
- Learning a residual mapping $\mathcal{F}(\mathbf{x}) + \mathbf{x}$ instead of a direct one $\mathcal{F}(\mathbf{x})$
- Reformulating learning as a refinement of the inputs
- Gradients backpropagated through identity do not vanish



Loss Landscapes



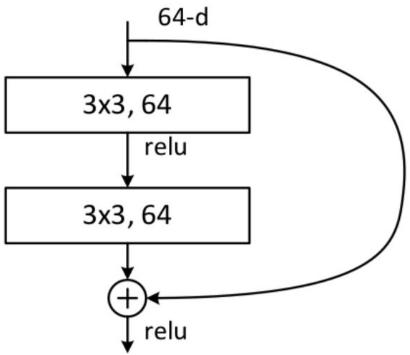
(a) without skip connections

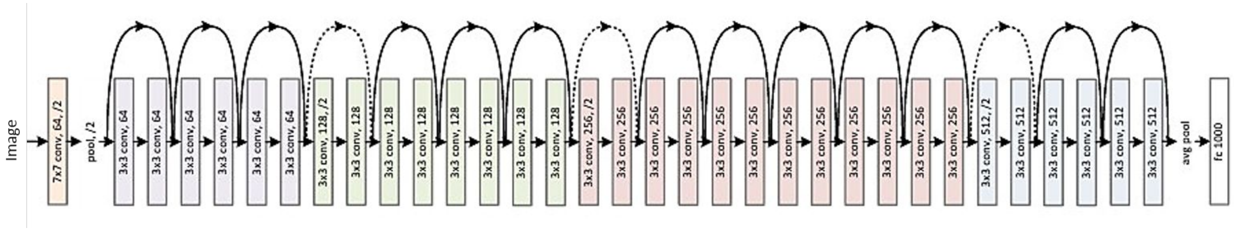


(b) with skip connections

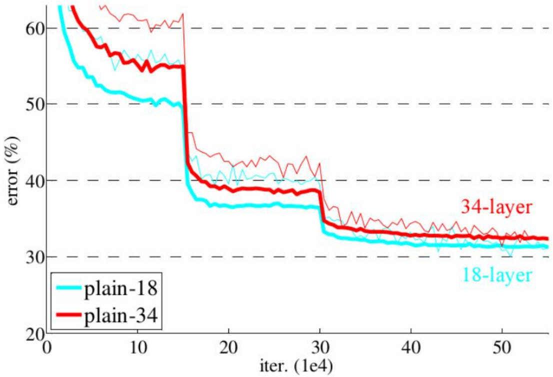
Residual Networks (2015)

- Winner of the 2015 ImageNet Challenge
- Deep cascade of residual blocks
- Super-human performance of several computer vision tasks

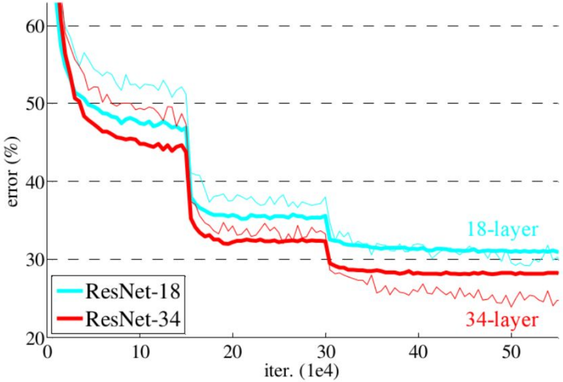




Residual Networks (2015)



Plain networks

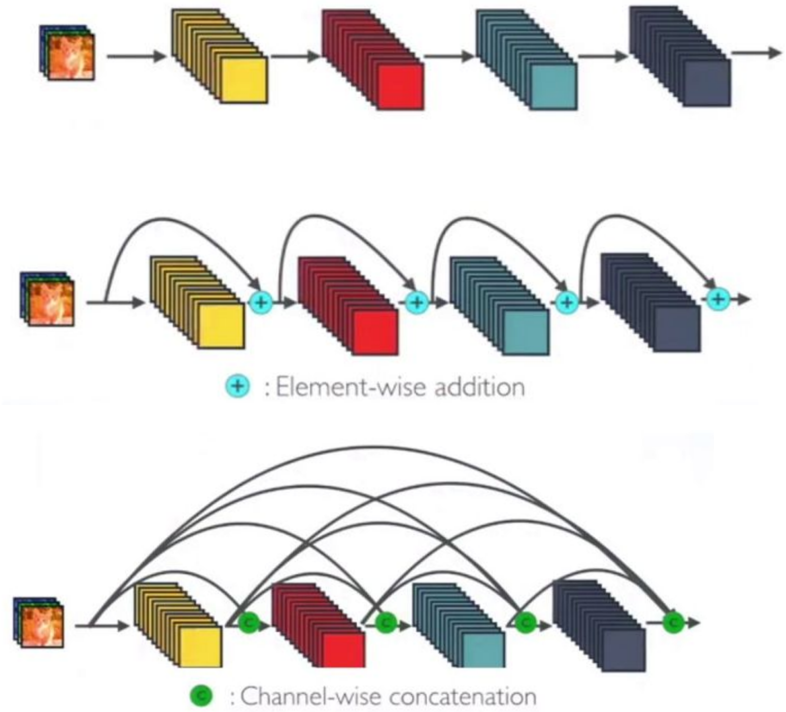


Residual networks

- Gradients backpropagated through residual connections do not vanish
- Deeper networks obtain better train & validation loss!

DenseNet (2016)

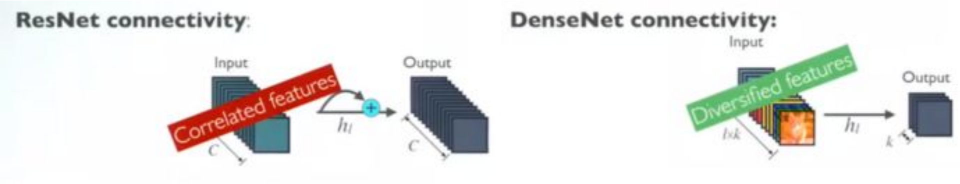
- **Standard CNNs**:
  - Cascade of convolutional layers
- **ResNets**:
  - Element-wise addition of residual and convolved features
- **DenseNets**:
  - Channel-wise concatenation



DenseNet (2016)

DenseNet (2016)

- Advantages:
  - Strong gradient flow
  - Diversified features
  - Classifier uses feature of all levels of complexity

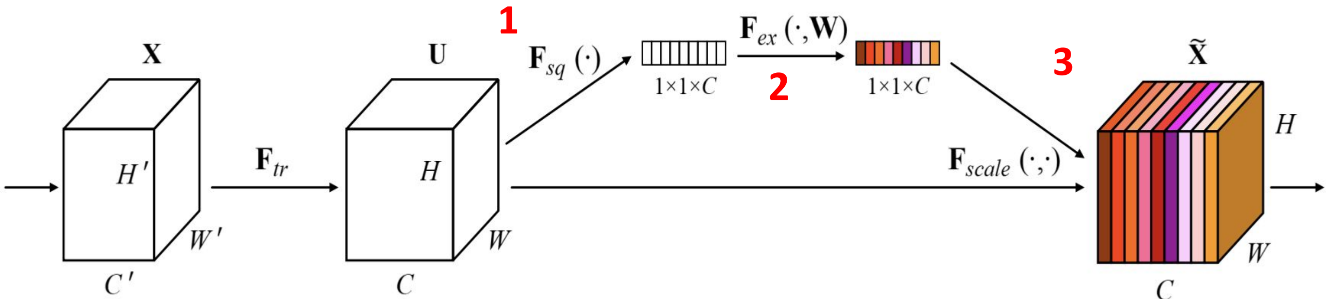


- Disadvantages:
  - Large number of parameters
  - Low parameter efficiency
  - Excessive computational power required

Squeeze and Excitation Net (2017)

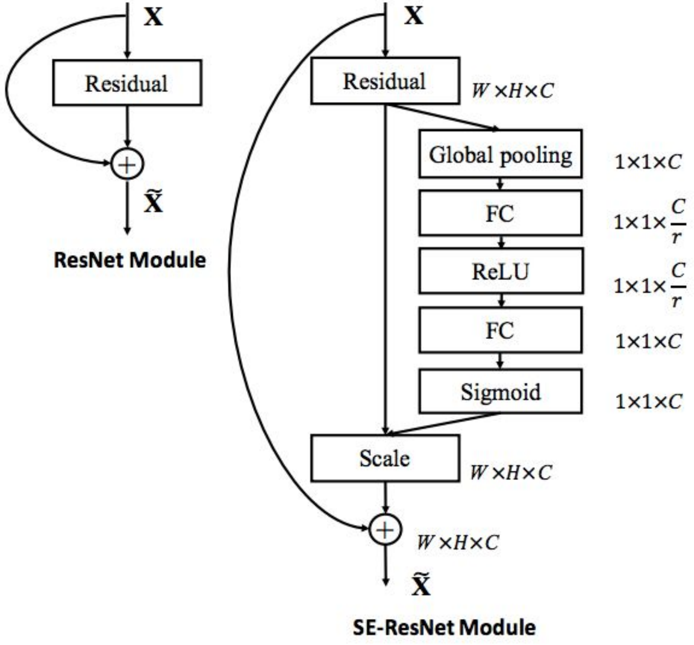
- Composed of squeeze and excitation blocks

1. Channels are squeezed into a single value using average pooling
2. Squeezed vector processed with fully-connected layers + sigmoid gating
3. Gated values are used to weight (excite) the conv. feature maps



Squeeze and Excitation Net (2017)

- Networks with high computational efficiency and representational power
- Perform dynamic channel-wise calibration
- Baseline model for channelwise attention mechanisms



More CNN Architectures...

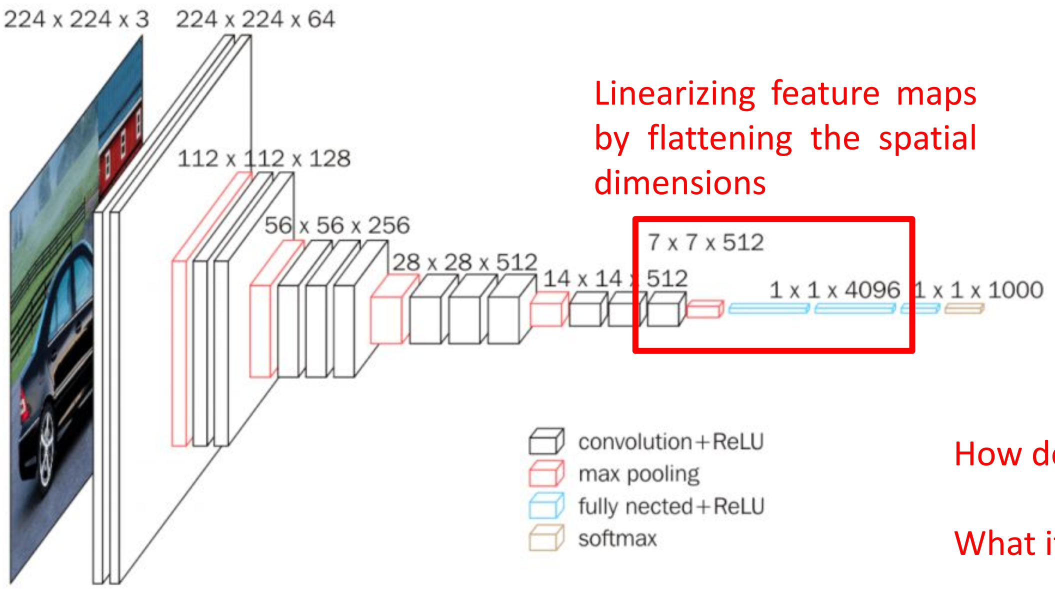
SparseNet ResNeXt EfficientNet ResNet-in-ResNet

InceptionNet Inception-Resnet NasNet ConvNexT Wide-Residual Networks

- Architectures for Object Detection
- Architectures for Semantic Segmentation
- Transformer-Bases Architectures

Global Average Pooling

From Convolutional to MLP



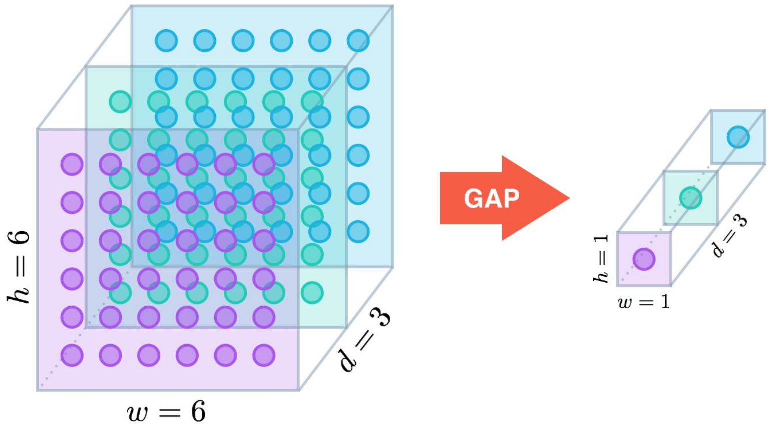
Linearizing feature maps by flattening the spatial dimensions

How do we know the exact dimensionality?

What if images have different sizes?

Global Average Pooling

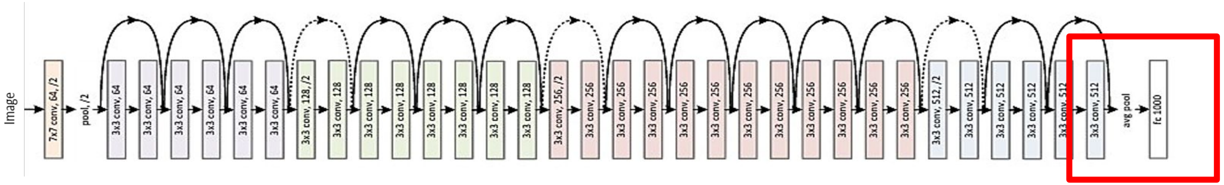
- Take the average activation over all spatial values
- C x H x W ⇒ C x 1 x 1
- Advantages:
  - Less constraints on input size
  - Reduce overfitting
  - Significantly less parameters



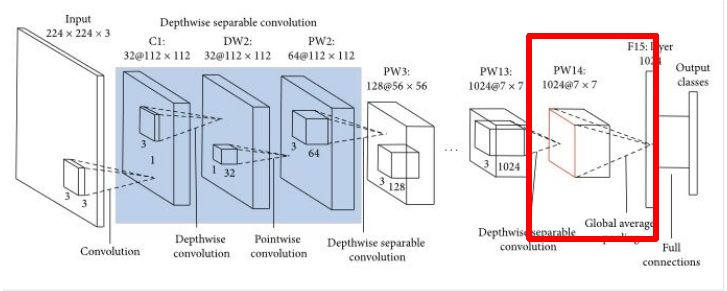

```
avgpool = nn.AdaptiveAvgPool2d((1, 1))
```

Architectures with GAP

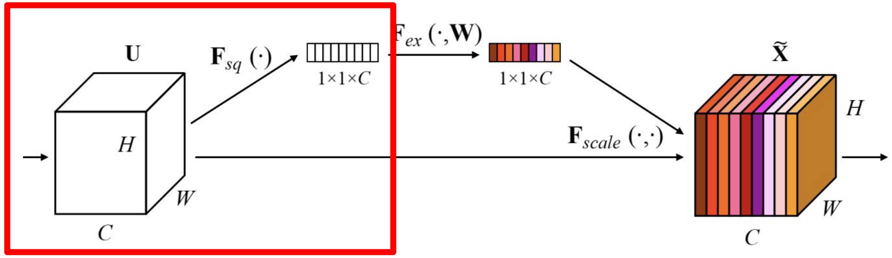
ResNet



MobileNet



SENet



and many more...

Fine-Tuning

Problem

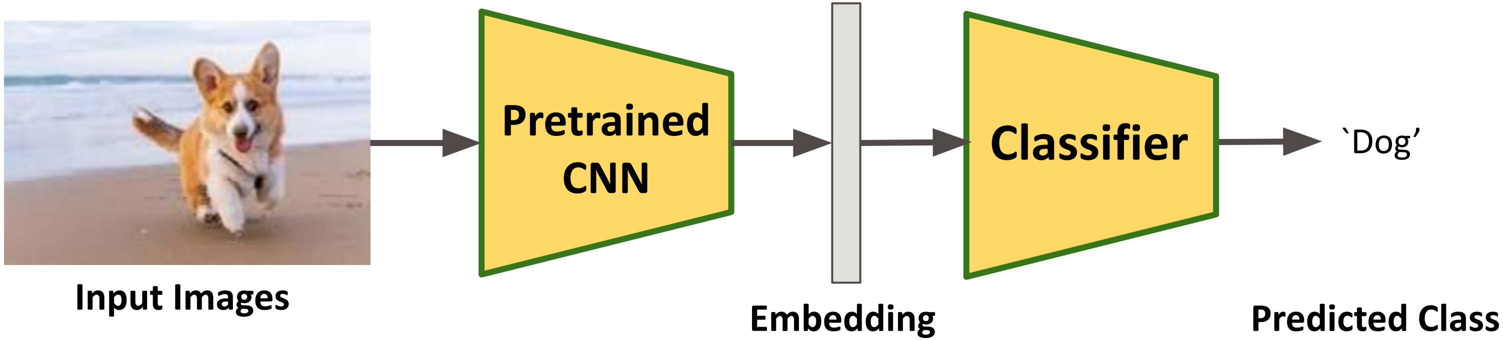
- Randomly initialized CNN's cannot learn from few labelled examples
- Most datasets are not large enough to train a deep CNN

Solutions

- **Fine-Tuning:** Training a CNN on a large dataset (e.g. ImageNet), and use these pretrained parameters as initialization when training on our dataset
- **Augmentation:** Increase the size of the dataset by applying data augmentation
- **Training Recipe:** Selecting a good learning rate and scheduler is especially important for small datasets

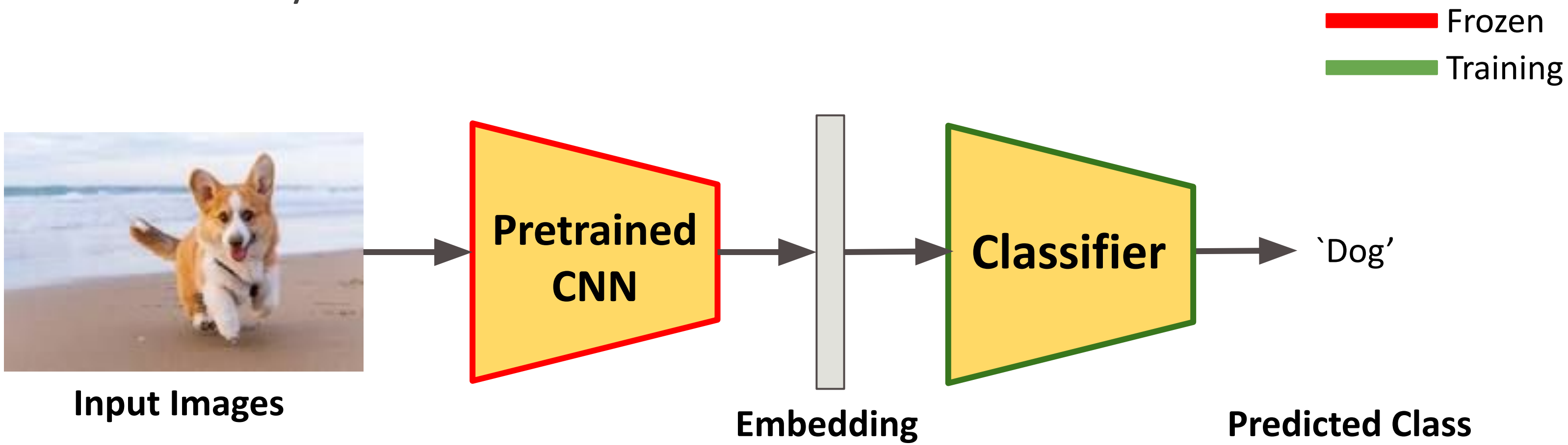
Fine-Tuning

- Directly training pretrained network



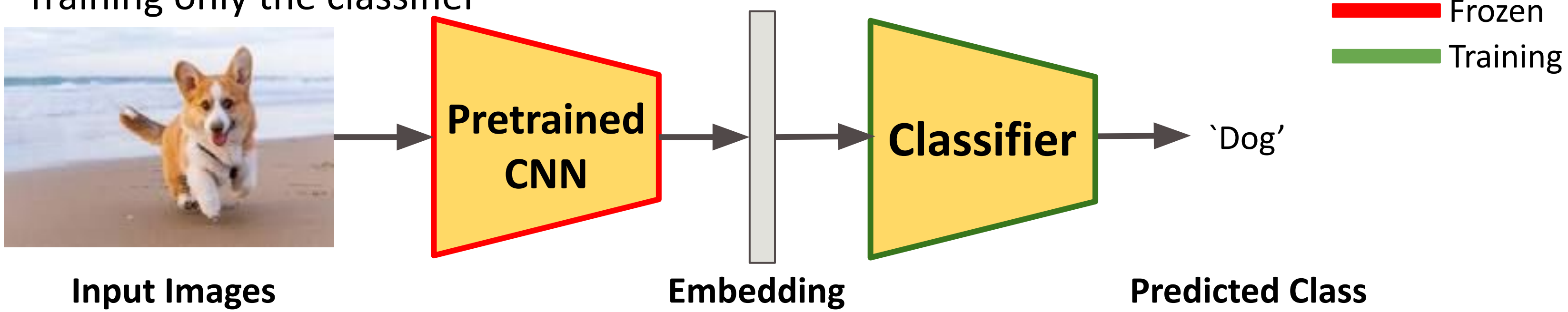
CNN as Feature Extractor

- Freeze pretrained CNN
- Train only the classifier

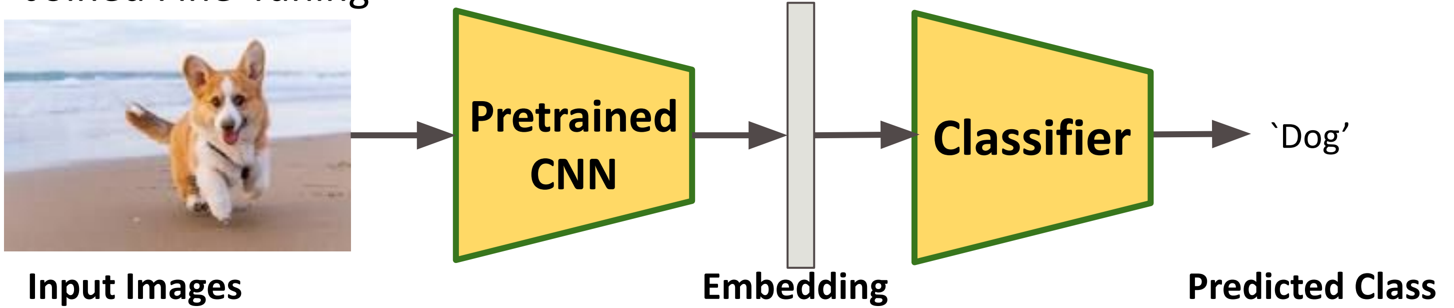


Combined Approach

1. Training only the classifier



2. Joined Fine-Tuning



References

1. Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115.3 (2015): 211-252.
2. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
3. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105.
4. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.
5. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
6. Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

References

7. Li, Hao, et al. "Visualizing the loss landscape of neural nets." arXiv preprint arXiv:1712.09913 (2017).
8. https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803
9. http://ethereon.github.io/netscope/#/gist/db945b393d40bfa26006
10. https://alexisbcook.github.io/2017/global-average-pooling-layers-for-object-localization/