People with no idea about AI saying it will take over the world:

My Neural Network:

CIS 522: Lecture 7

CNN Architectures & Applications
02/6/20

Penn Engineering

Feedback (PollEV)

What you will learn today

Prehistoric convnets

How Alexnet works - i.e. how really good engineering looks like

How deeper networks win

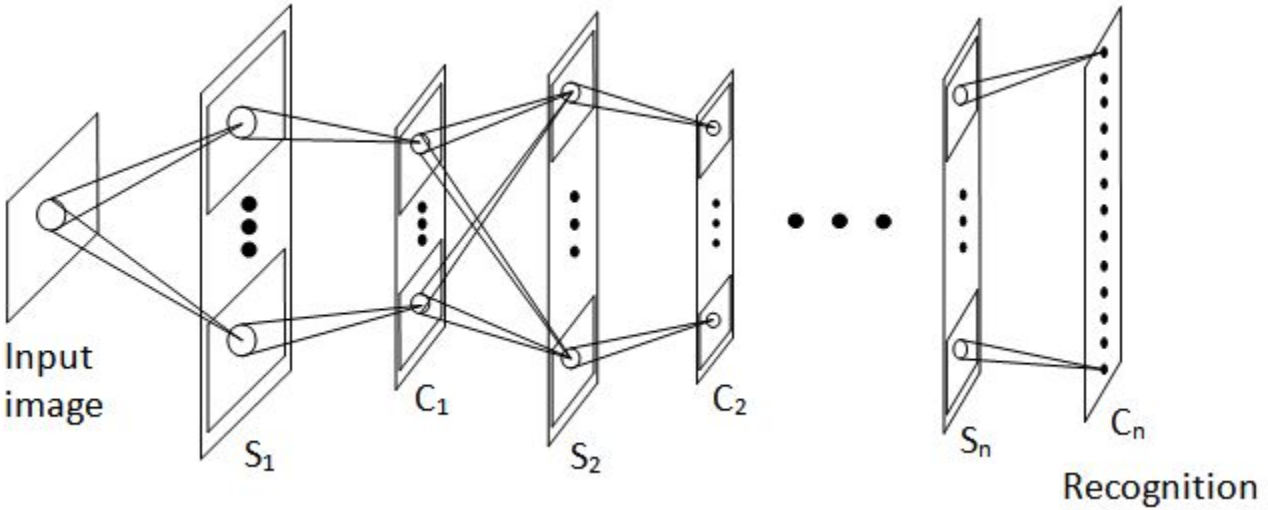
How skip connections help

How CNNs are broadly used

Some classic CNNs

Neocognitron -- Fukushima 1980

Alternating S (simple) and C (complex) layers, mimics visual cortex



Aki et al 2017 IEEE ICECS

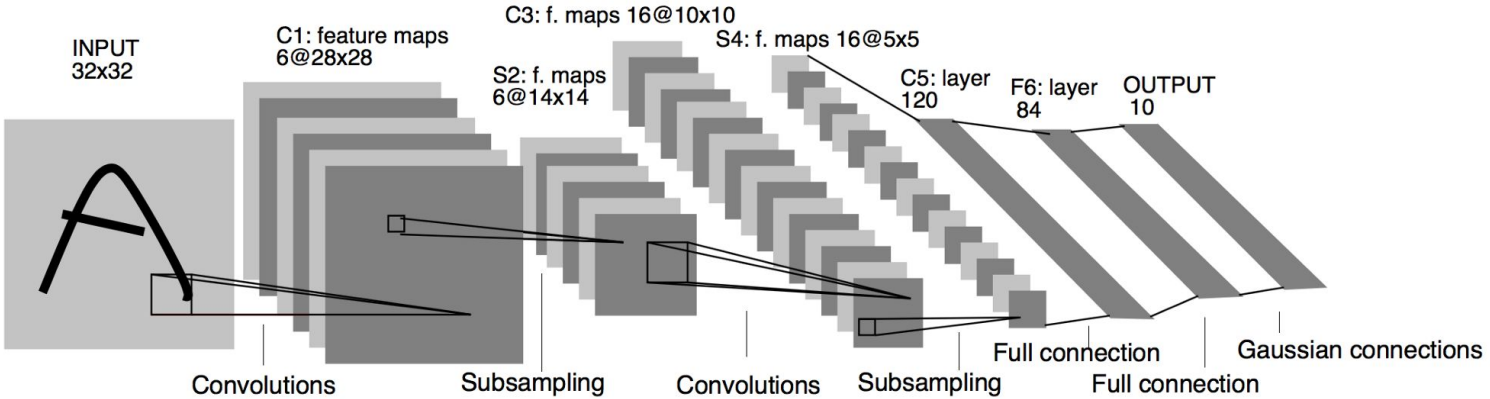
LeNet (1998) -- Background

- Developed by Yann LeCun
  - Worked as a postdoc at Geoffrey Hinton's lab
  - Chief AI scientist at Facebook AI Research
  - Wrote a whitepaper discovering backprop (although Werbos).
  - Co-founded ICLR
- Problem: classify 7x12 bit images of 80 classes of handwritten characters.

Fig. 3. Initial parameters of the output RBFs for recognizing the full ASCII set.

LeNet (1998) -- Architecture

- Convolution filter size: 5x5.
- Subsampling (pooling) kernel size: 2x2.
- Semi-sparse connections.



LeNet (1998) -- Results

- Successfully trained a 60K parameter neural network without GPU acceleration!
- Solved handwriting for banks -- pioneered automated check-reading.
- 0.8% error on MNIST; near state-of-the-art at the time.
    - Virtual SVM, kernelized by degree 9 polynomials, also achieves 0.8% error.

LeNet: http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

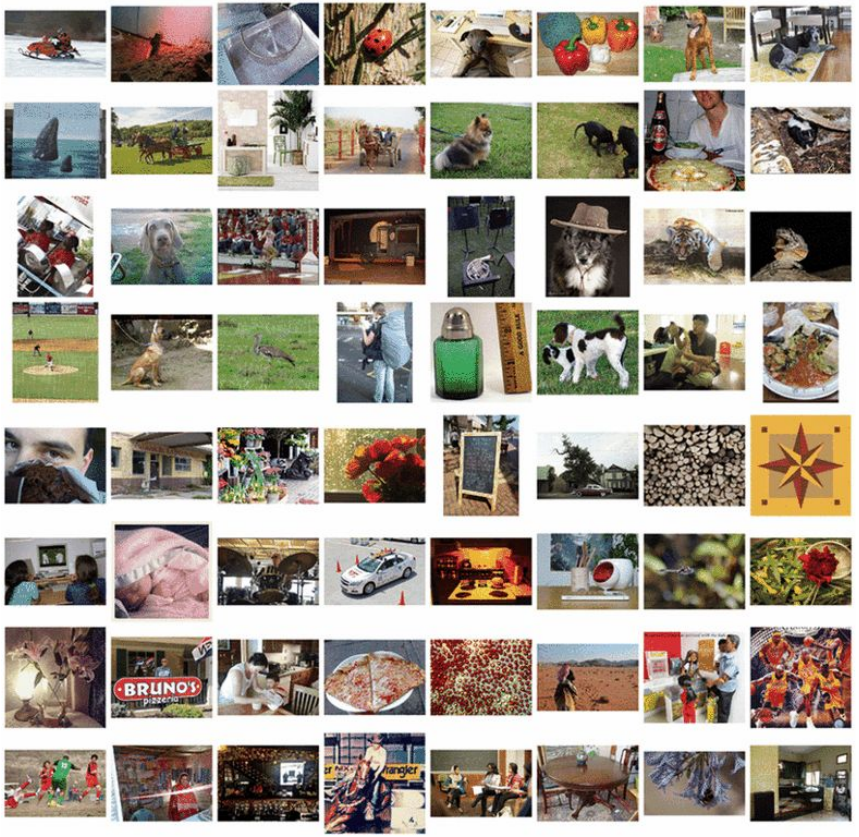
2012: Imagenet

>1M images

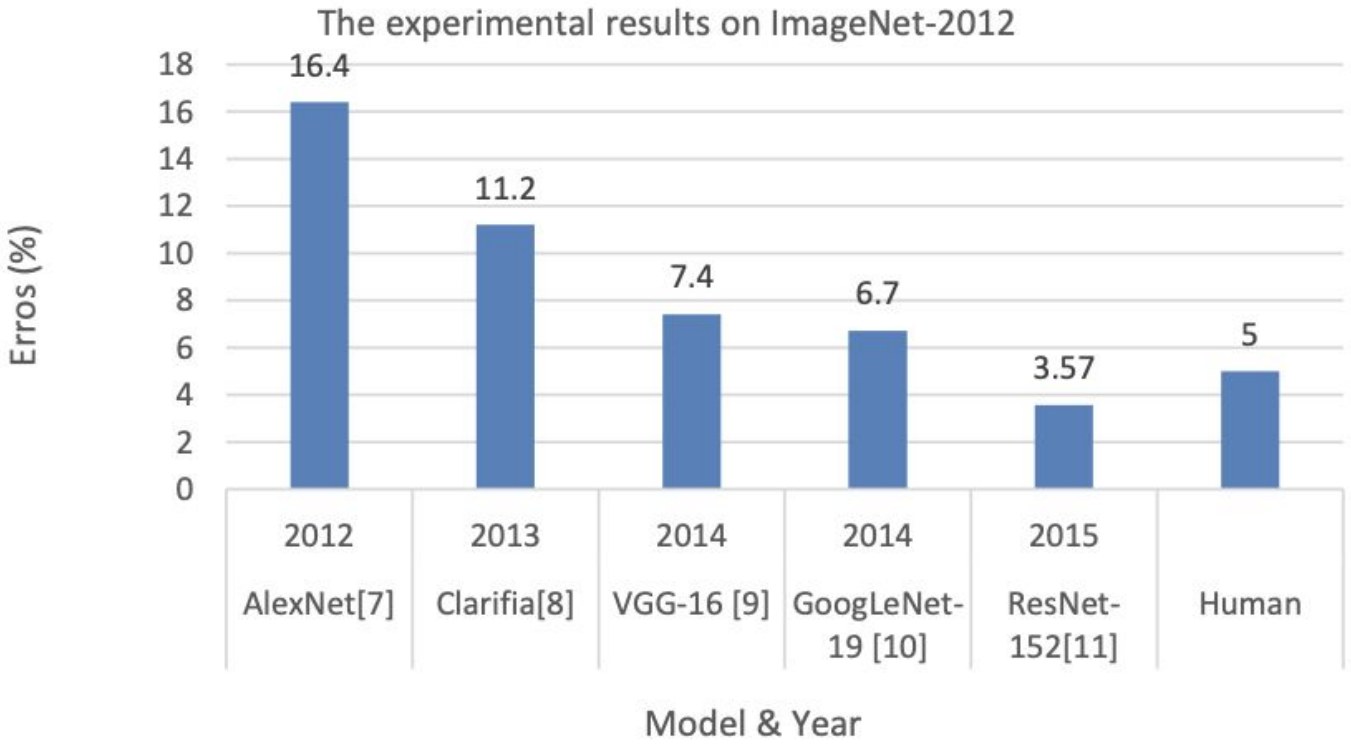
1000 categories

And lots of sift filter approaches

Pre-alexnet: SOTA ~30%



Model and year



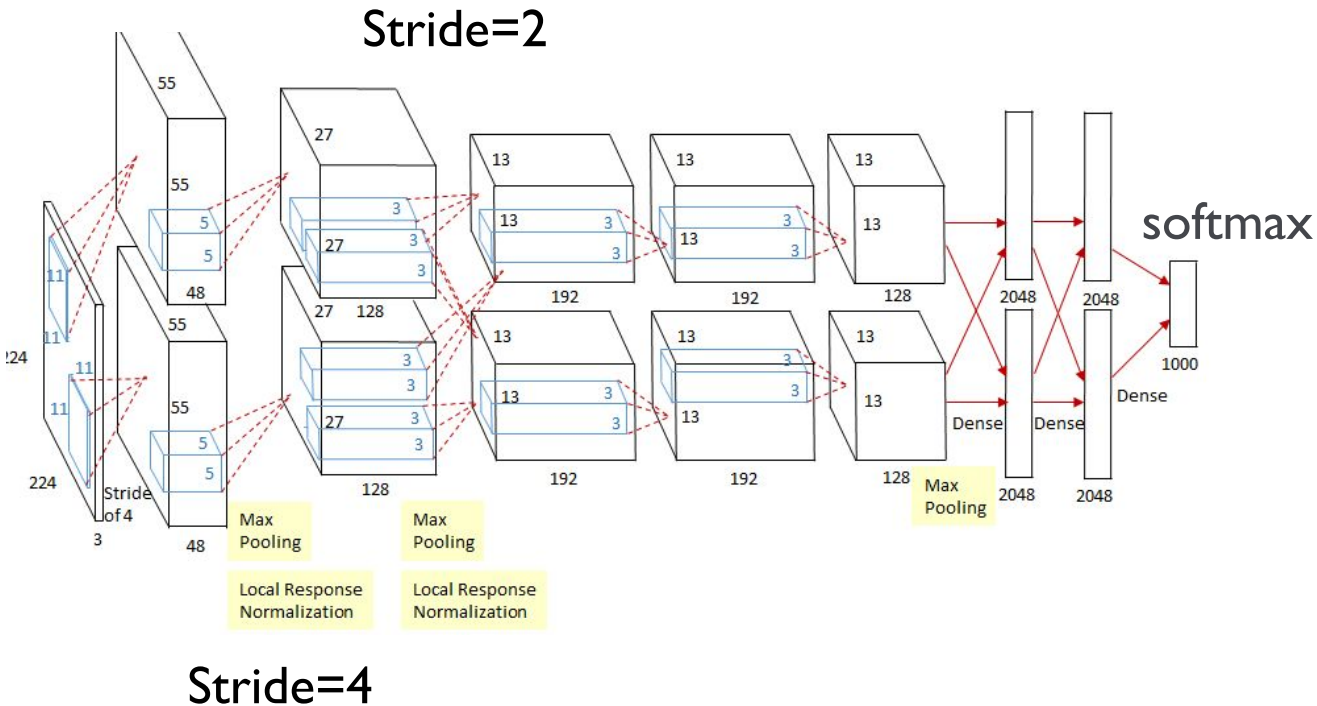
Alom, Taha, et al

AlexNet (2012) -- Background

- Developed by
  - Alex Krizhevsky
  - Ilya Sutskever
    - Chief scientist, OpenAI
    - Inventor of seq2seq learning.
  - Geoffrey Hinton, Alex Krizhevsky's PhD adviser
    - Co-invented Boltzmann machines
  - Hinton was against it. Believed that unsupervised was the future
- Problem: compete on ImageNet, Fei-Fei Li's dataset of 14 million images with more than 20,000 categories (e.g. strawberry, balloon).

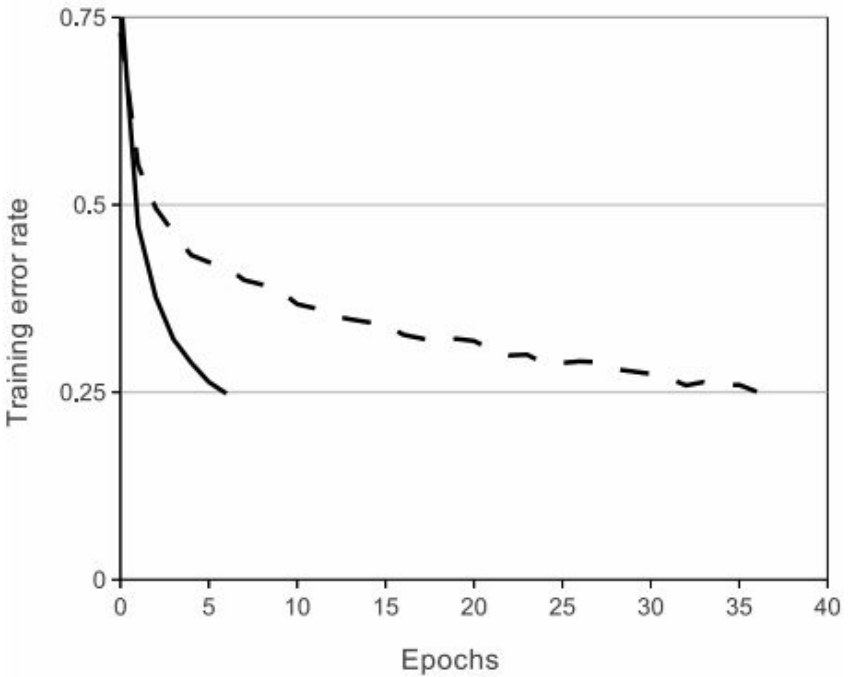
Uses two GPUs, basis for architecture

Because 3GB each is too little!

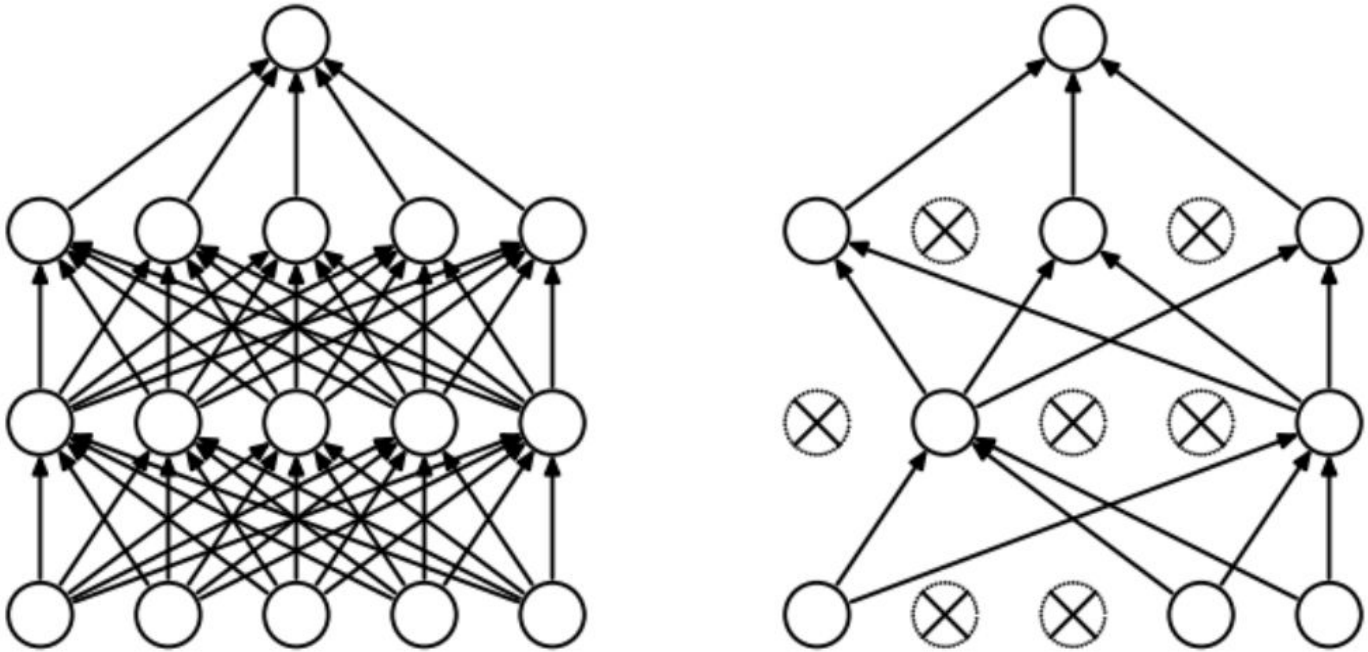


Poll: what does stacking conv layers do?

Relu is much better than tanh on alexnet



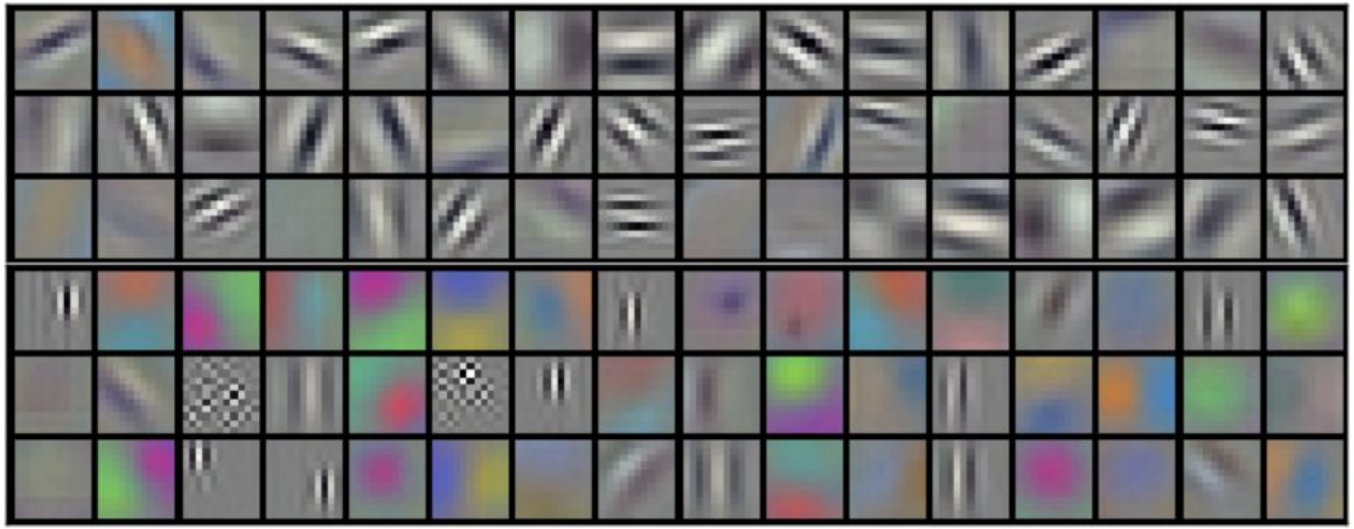
Use dropout



(a) Standard Neural Net

(b) After applying dropout.

Level 1 filters



Was there a disappointed supervisor?

To simplify our experiments, we did not use any unsupervised pre-training even though we expect that it will help, especially if we obtain enough computational power to significantly increase the size of the network without obtaining a corresponding increase in the amount of labeled data. Thus

Also, local normalization adds memory and is largely useless.

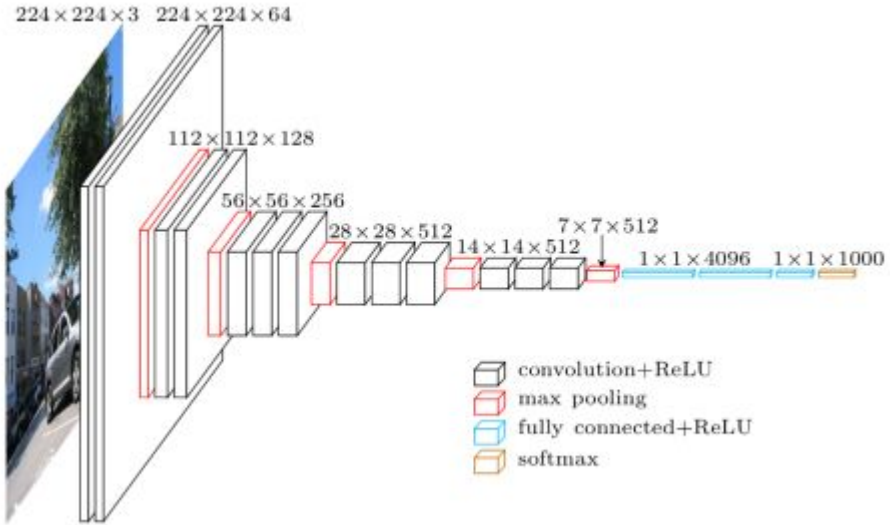
AlexNet (2012) -- Results

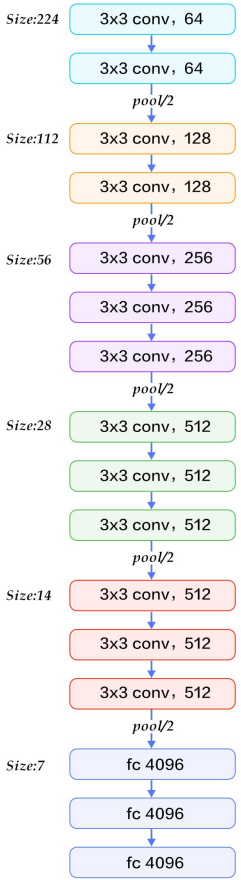
- Smoked the competition with 15.3% top-5 error (runner-up had 26.2%)
- One of the first neural nets trained on a GPU with CUDA.
  - (There had been 4 previous contest-winning CNNs)
  - Trained 60 million parameters
- Cited over 50,000 times: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

VGG (2014) -- Background

- Developed by the Visual Geometry Group (Oxford)
- Problem: beat AlexNet on ImageNet
- Uses smaller filters and deeper networks

VGG (2014) -- Architecture





Why does it help to have all the fully connected layers? (pollEV)

VGG (2014) -- Architecture

- Jumps from 8 layers in AlexNet to 16-19 layers
- Uses only 3*3 CONV stride 1, pad 1 and 2*2 MAX POOL stride 2
  - Stack of three 3*3 conv (stride 1) layers has the same receptive field as one 7*7 conv layer and the resultant network is deeper with more non-linearities.
  - It also has fewer parameters: 3*(3*3*C*C) vs. 7*7*C*C for C channels per layer.

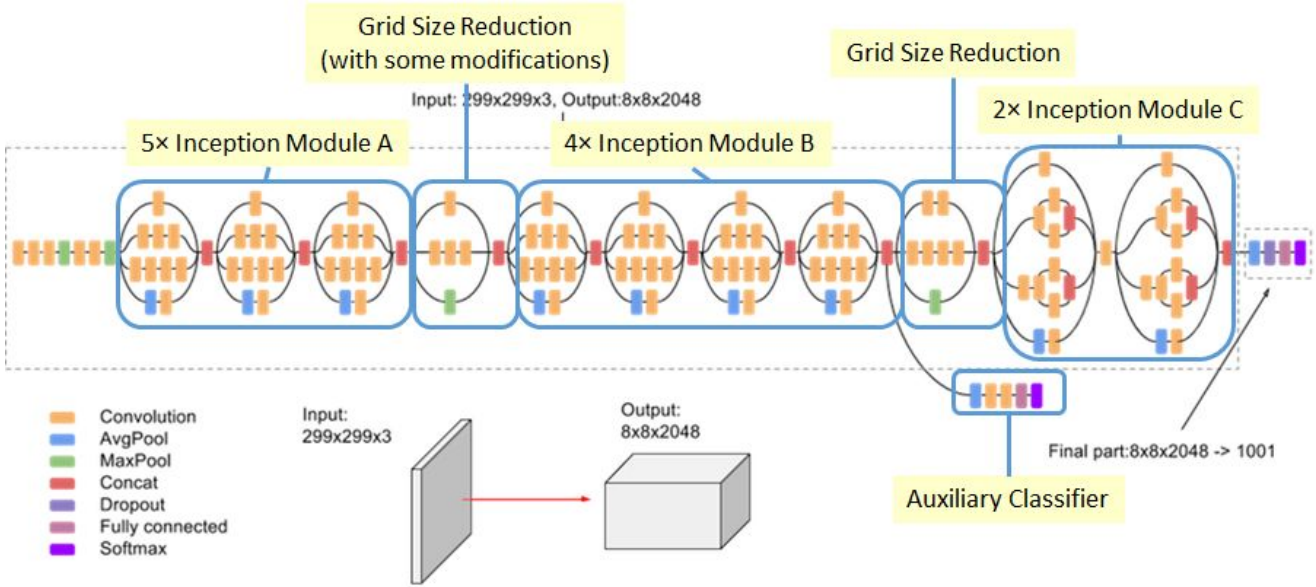
VGG (2014) -- Results

- Configuration E (19 weighted layers) achieved 8.0% top-5 error.
- https://arxiv.org/pdf/1409.1556.pdf

VGG (2014) -- Drawbacks

- Slow to train.
- The weights themselves are quite large (in terms of disk/bandwidth).
- The size of VGG is over 530 MB which makes deploying it a tiresome task.

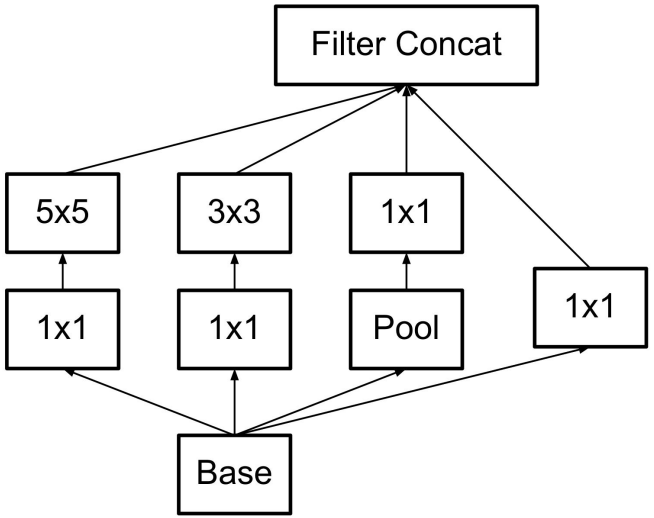
Inception / GoogLeNet (2015)



Inception v1: tuning kernel size is hard.

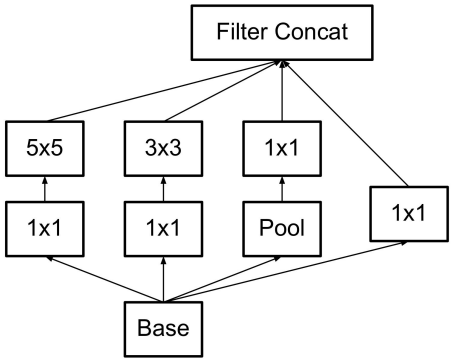
- A computer scientist's solution: toss 'em all together.

A single "Inception Module"

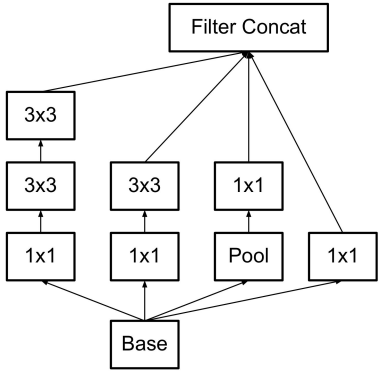


Inception v3: compute, representational bottlenecks

Original



Later version

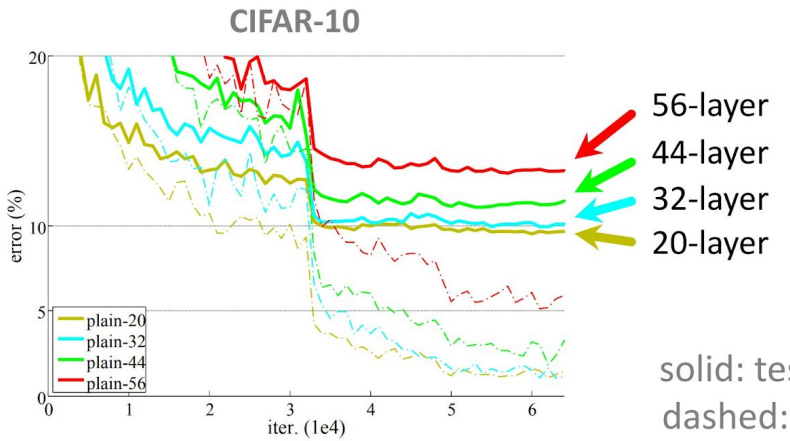


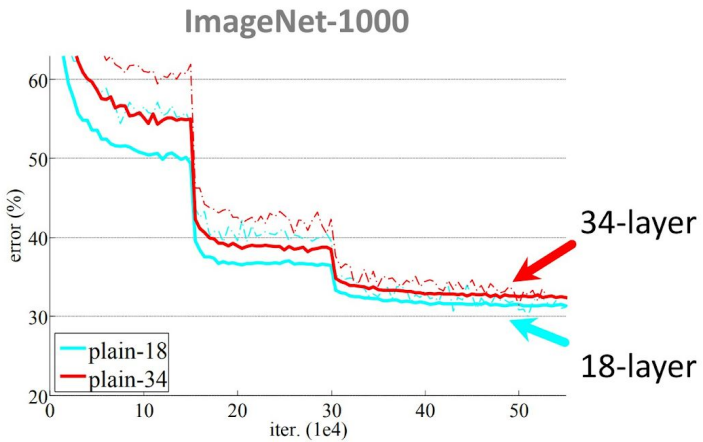
- Representational bottleneck: worse learning properties when the dimensions of the data are drastically changed all at once (more on that later in course).

Inception (2014) -- Results

- 6.67% top-5 error rate!
- Andrej Karpathy achieved 5.1% top-5 error rate at Human labeling.
- Humans are still better

ResNet

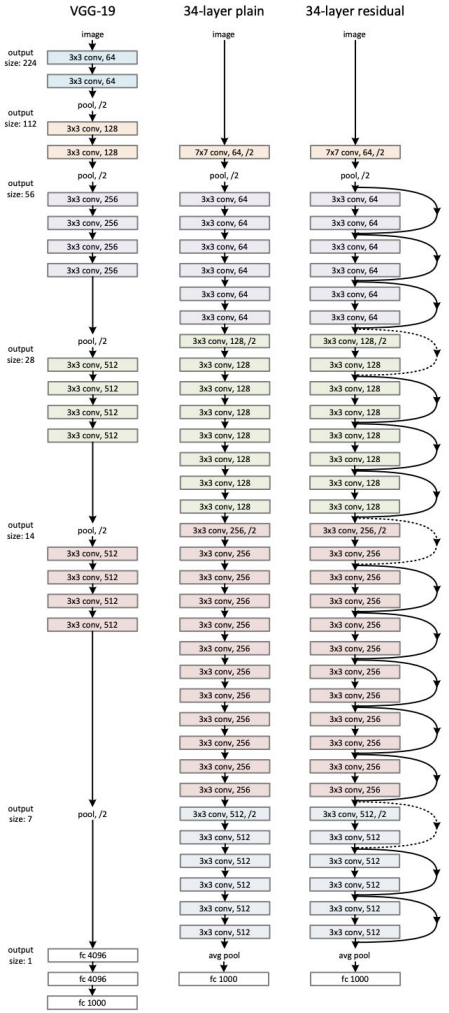




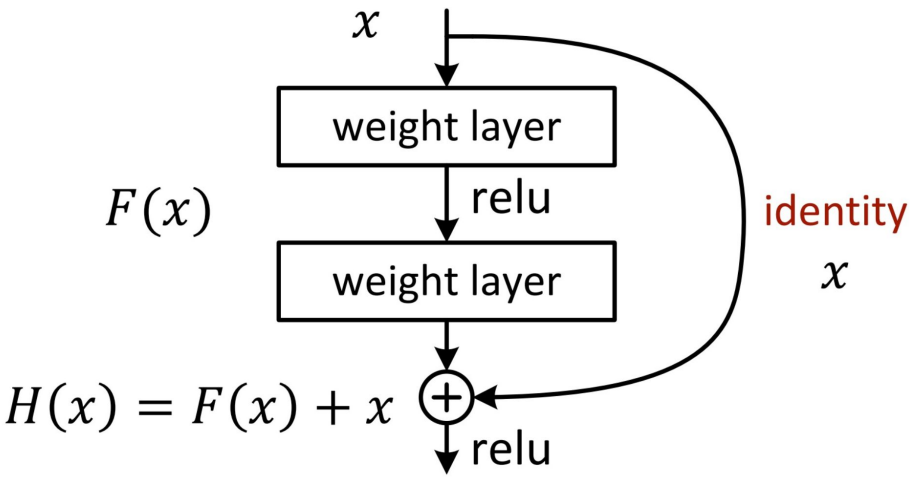
Is simply increasing the number of the layers the solution after all?

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016

Compare with VGG



ResNet

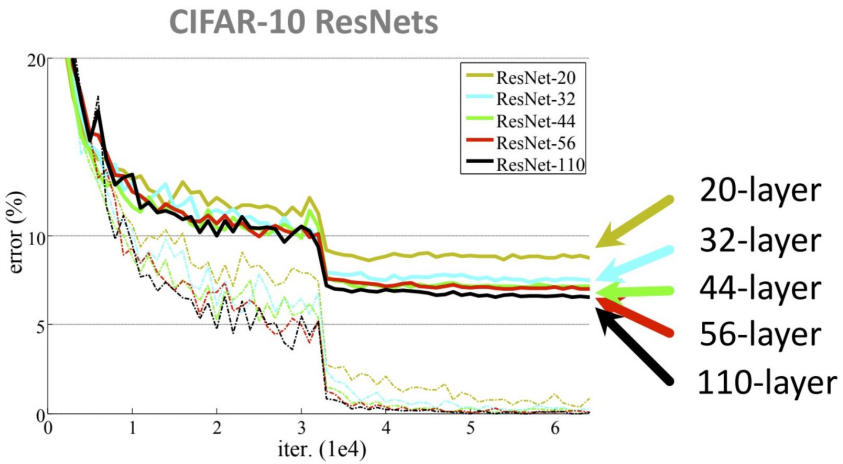


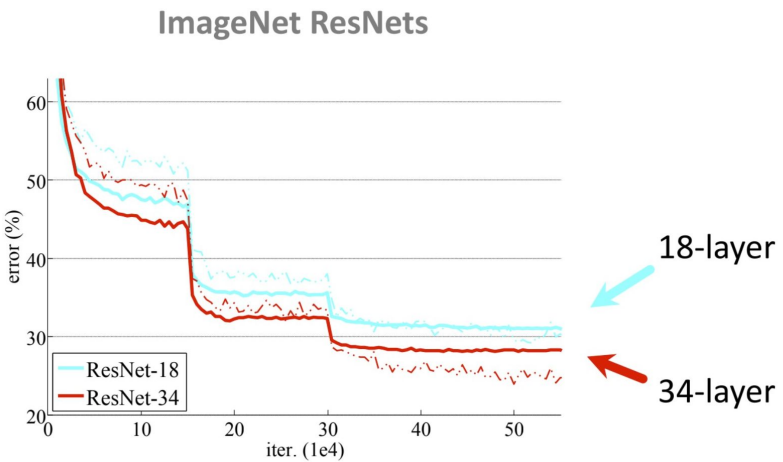
- $H(x)$ is any desired mapping, hope the small subnet fit $F(x)$
- If optimal mapping is closer to identity, easier to find small fluctuations

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016

PollEV how many extra parameters?

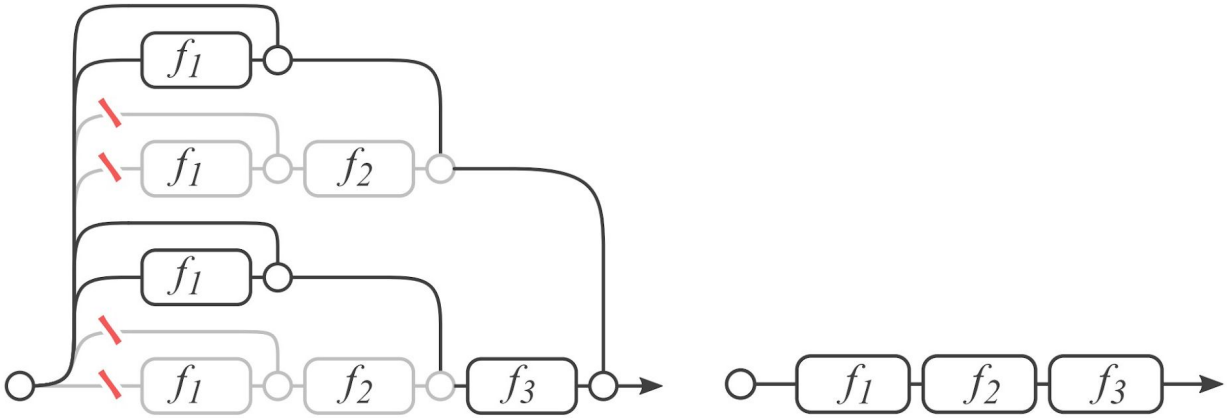
ResNet





Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016

Why ResNets work?

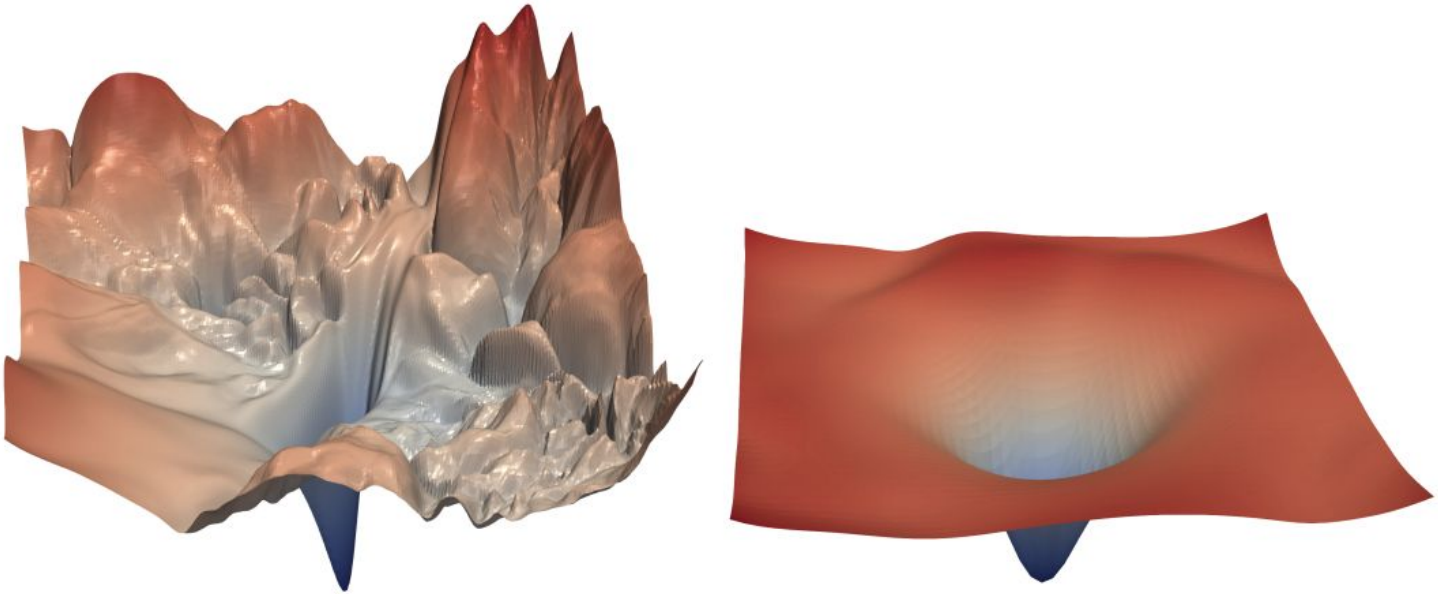


(a) Deleting f_2 from unraveled view (b) Ordinary feedforward network

Figure 2: Deleting a layer in residual networks at test time (a) is equivalent to zeroing half of the paths. In ordinary feed-forward networks (b) such as VGG or AlexNet, deleting individual layers alters the only viable path from input to output.

https://arxiv.org/pdf/1605.06431.pdf

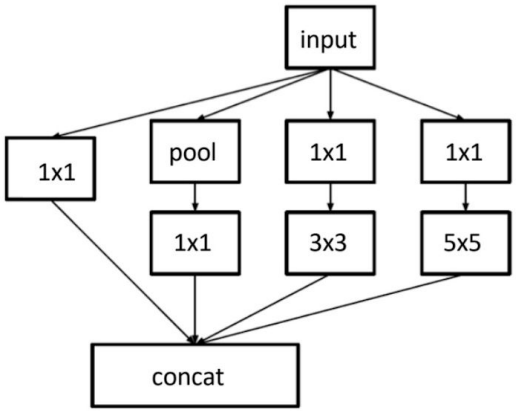
Loss landscapes look much better



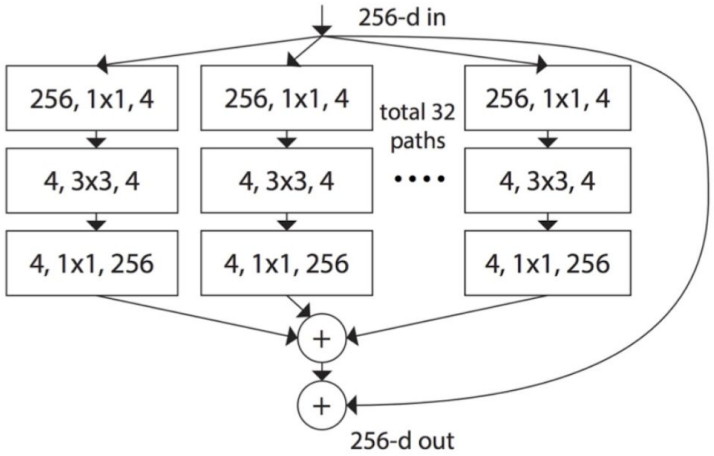
(a) without skip connections

(b) with skip connections

Combining Inception and ResNet - ResNeXt



Inception:
heterogeneous multi-branch



ResNeXt:
uniform multi-branch

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks". arXiv 2016 (CVPR 2017).

How does resnet implement identity function?

DenseNet

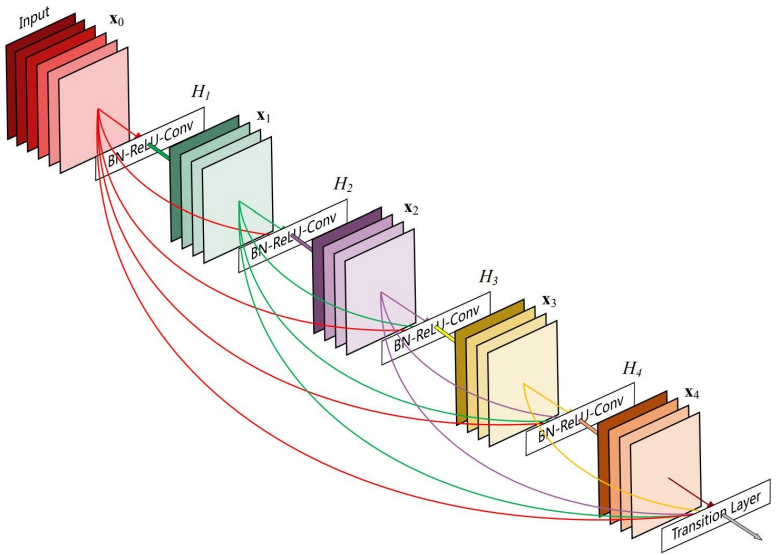
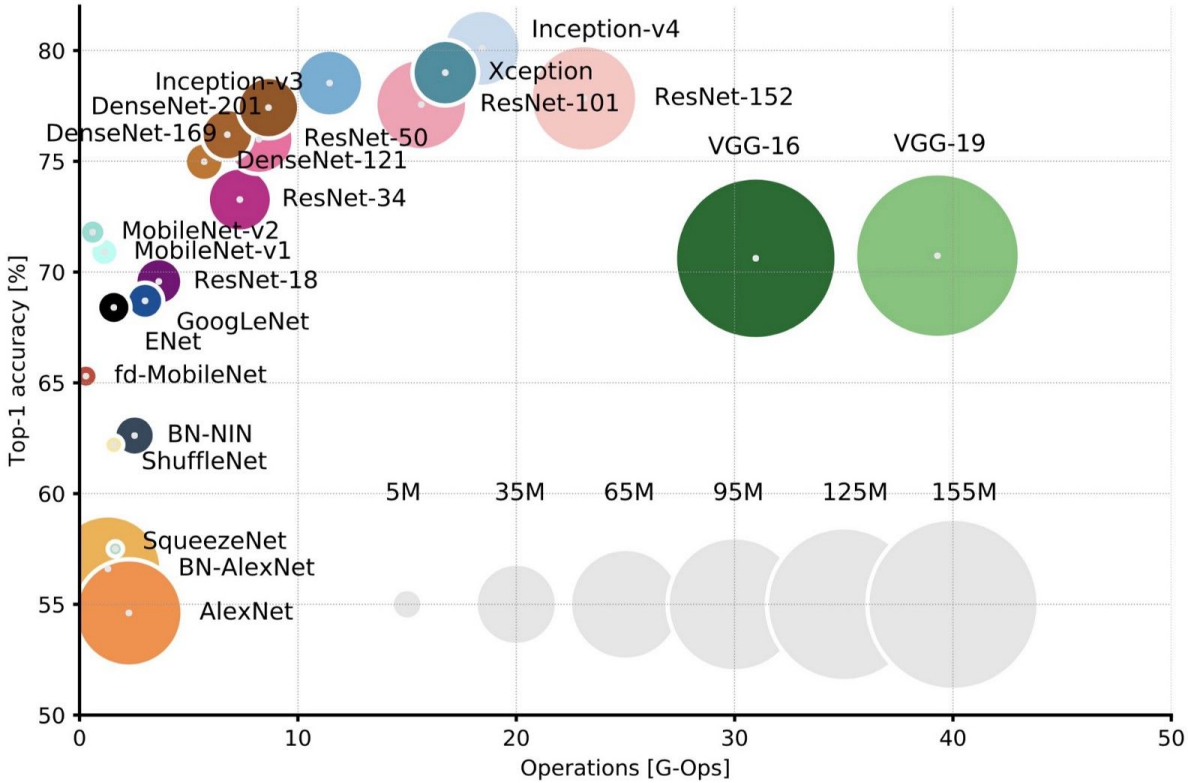


Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.

- Advantages:
  - Alleviate the vanishing-gradient problem
  - strengthen feature propagation
  - encourage feature reuse
  - substantially reduce the number of parameters

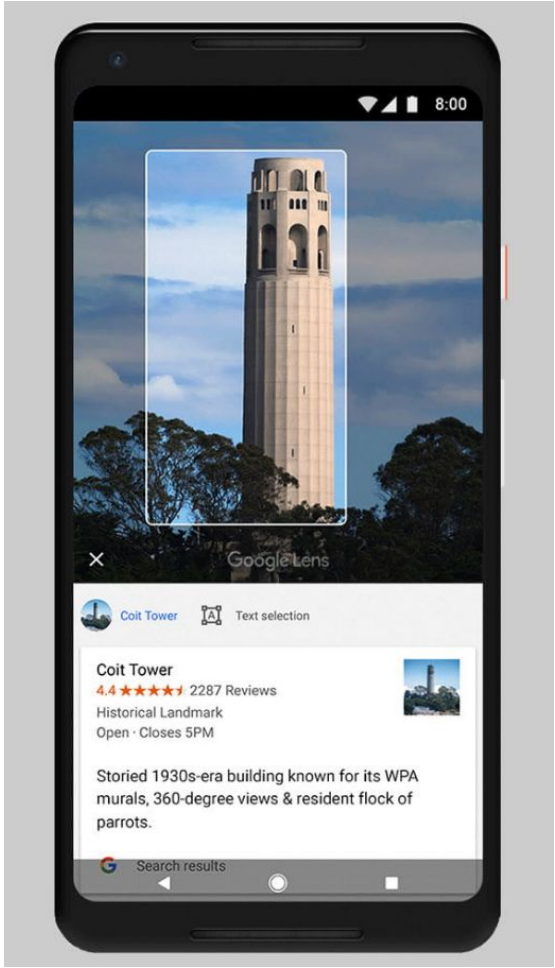
https://arxiv.org/pdf/1608.06993.pdf

Performance of convolutional architectures on ImageNet



Applications of CNNs

Image recognition



Style transfer

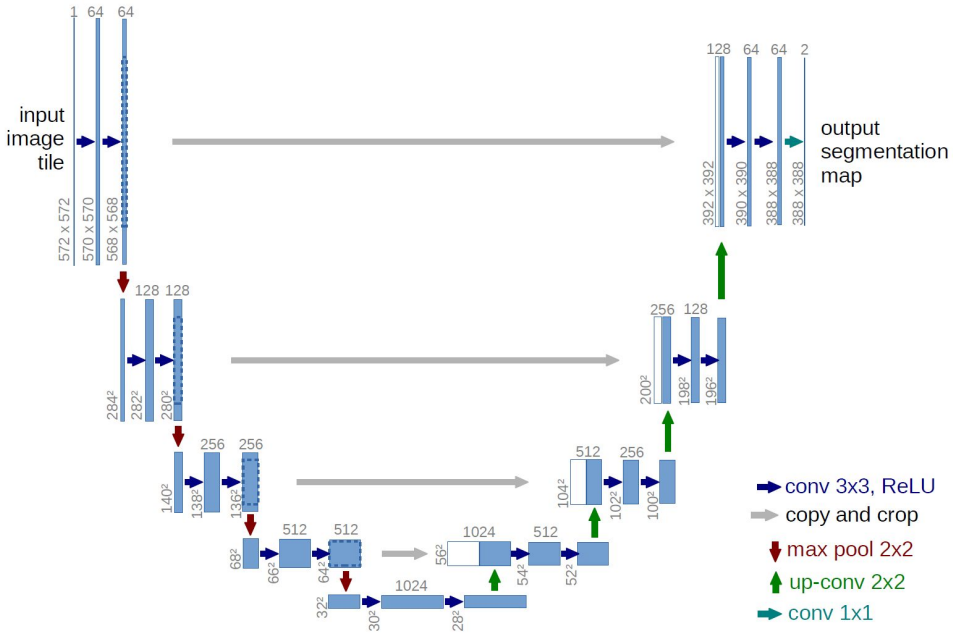
- Content is measured by deeper layers.
- Style is measured by the correlations between feature vectors at lower layers.

- Objective 1: Minimize content difference between new image and content template.
- Objective 2: Minimize style difference between new image and style template.

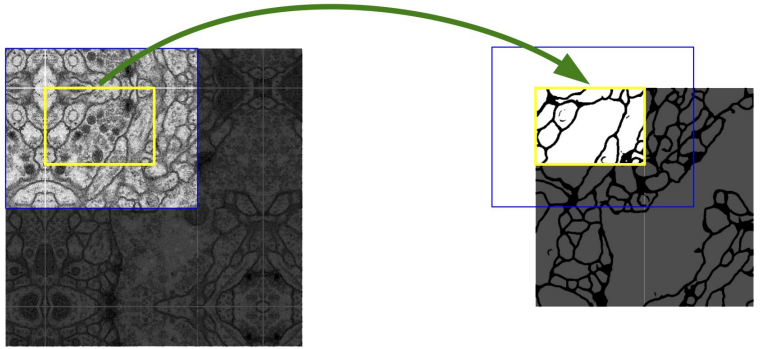
Gatys et al. (2016)

Image segmentation

U-Net (Ronneberger et al. 2015)



Segmenting biological cell membranes



Insights of U-Net

- "Up-convolution"
  - Fixes the problem of shrinking images with CNNs
  - One example of how to make "fully convolutional" nets: pixels to pixels
  - "Up-convolution" is just upsampling, then convolution
  - Allows for refinement of the upsample by learned weights
  - Goes along with decreasing the number of feature channels
  - Not the same as "de-convolution".
- Connections across the "U" in the architecture.

Object Detection (maybe skip)

Bounding-Box Object Detection

Object detection

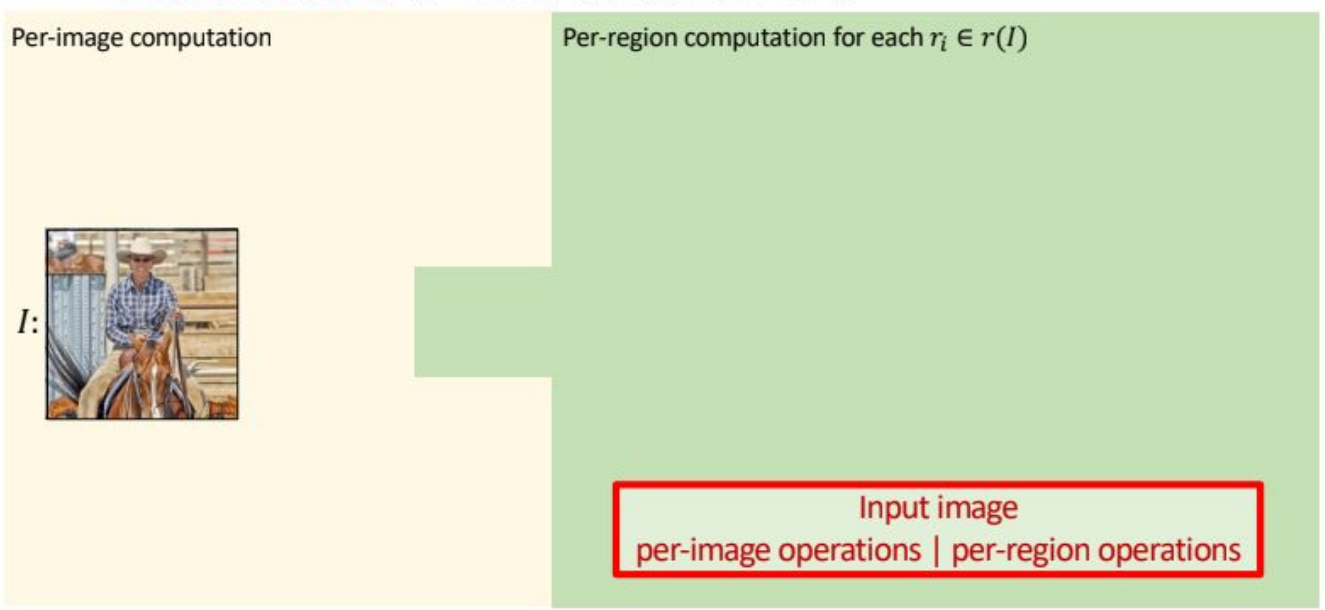
Object Detection

Detection Beyond Bounding Boxes

Mask R-CNN
[He, Gkioxari, Dollár, Girshick]

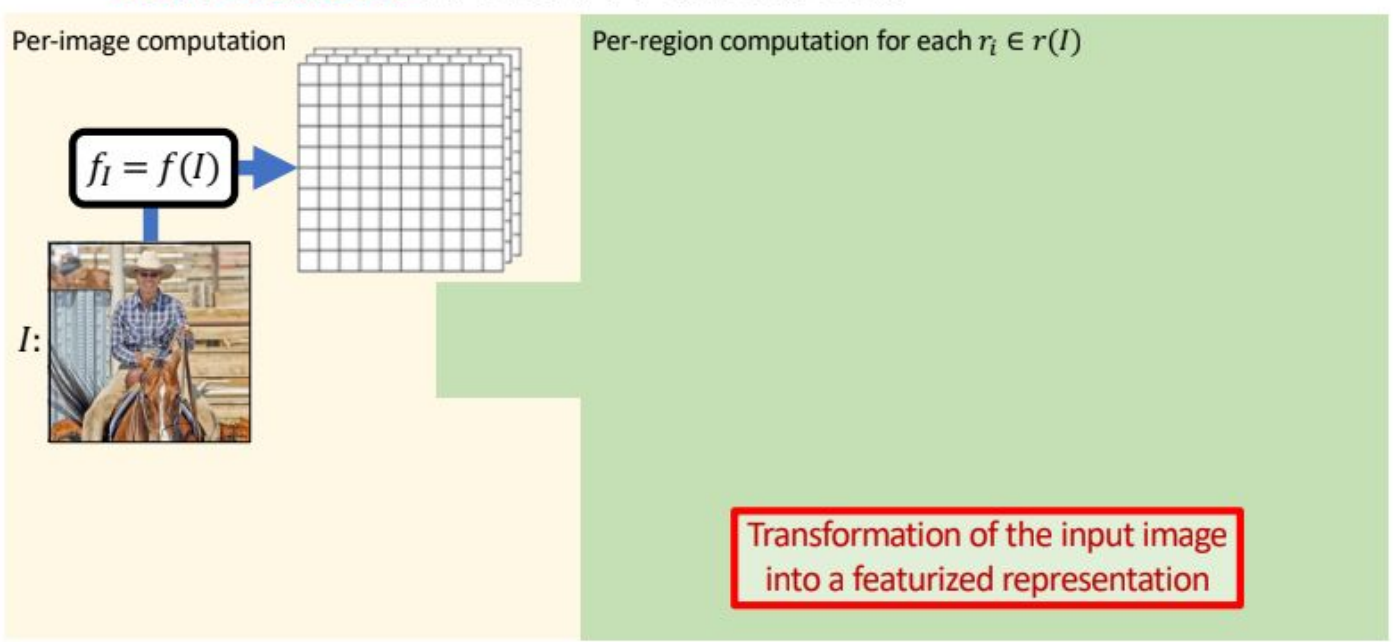
- Detects the objects.
- Does Semantic Segmentation for each object.

Generalized R-CNN framework for Object Detection



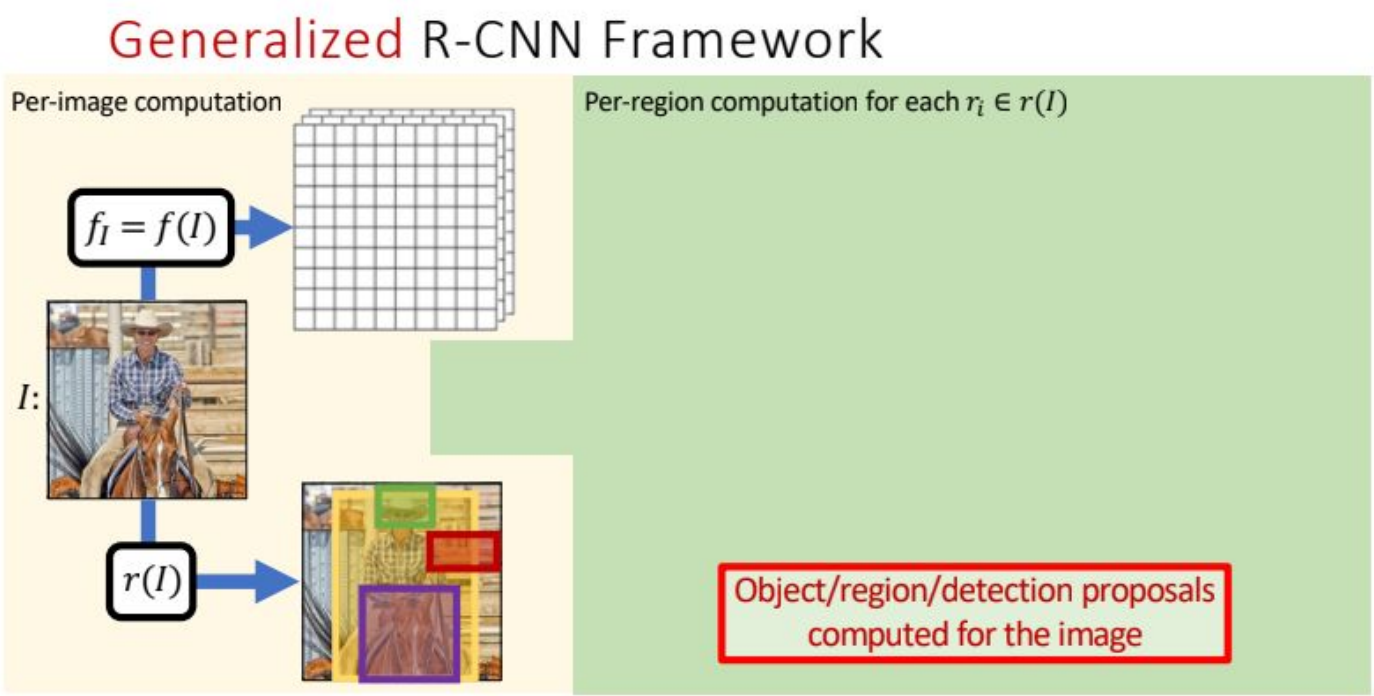
https://drive.google.com/file/d/1Z1i8mk8AqdtI_Q8avqtrzLQnLTG7QW6T/view

Generalized R-CNN framework for Object Detection



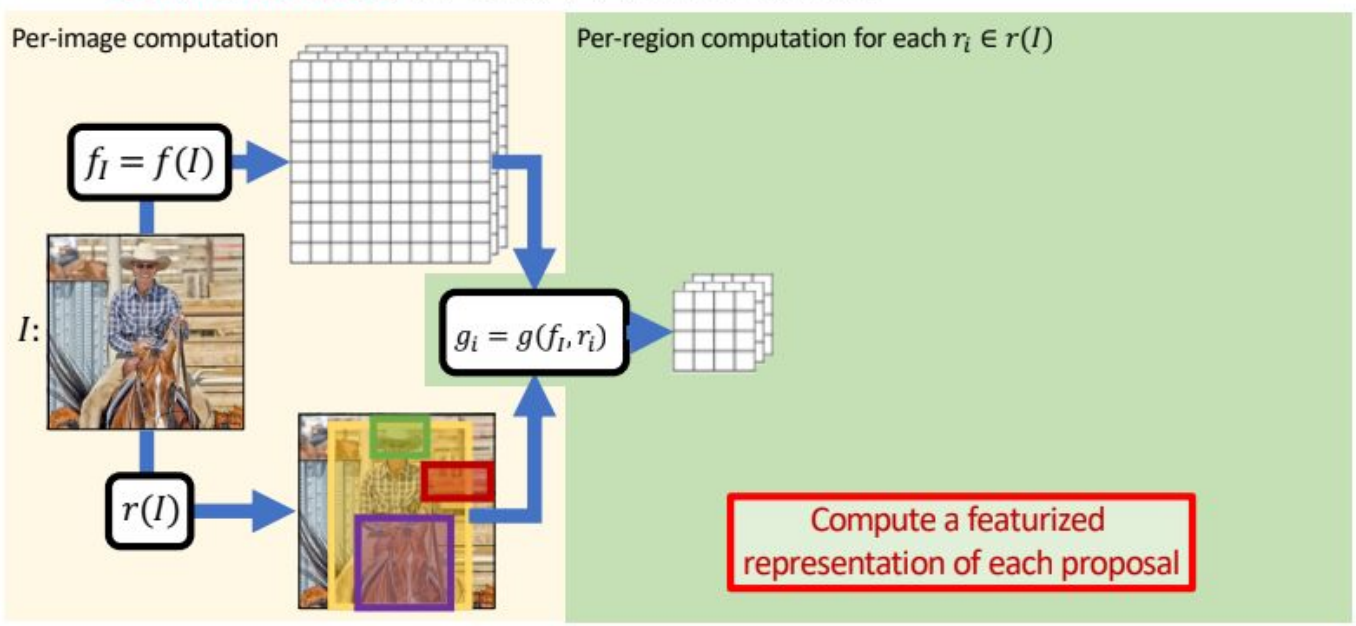
https://drive.google.com/file/d/1Z1i8mk8AqdtI_Q8avqtrzLQnLTG7QW6T/view

Generalized R-CNN framework for Object Detection



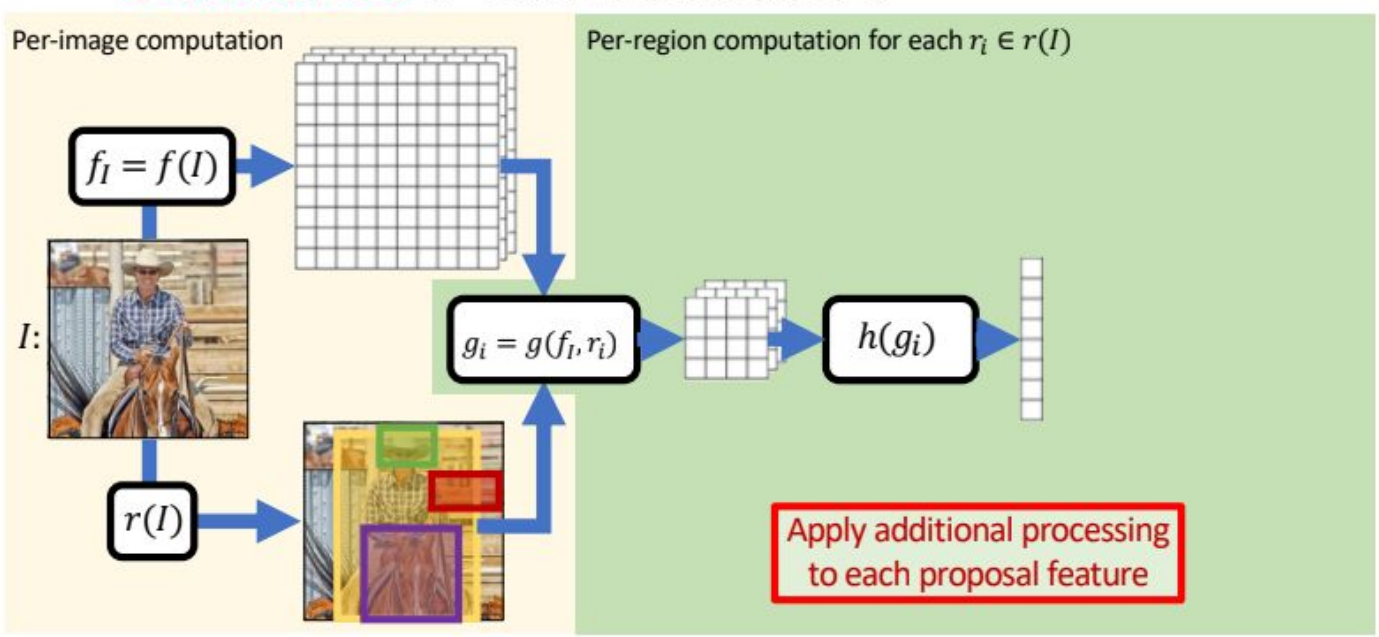
https://drive.google.com/file/d/1Z1i8mk8AqdtI_Q8avqtrzLQnLTG7QW6T/view

Generalized R-CNN framework for Object Detection



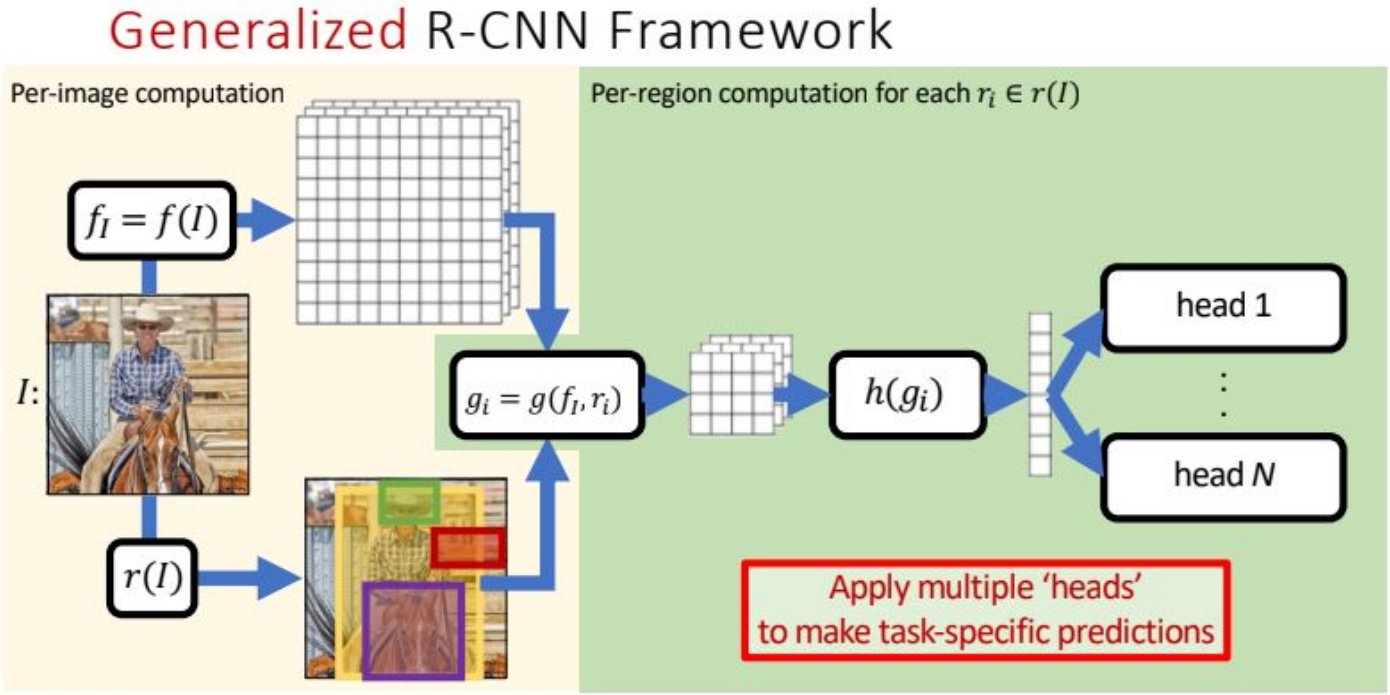
https://drive.google.com/file/d/1Z1i8mk8AqdtI_Q8avqtrzLQnLTG7QW6T/view

Generalized R-CNN framework for Object Detection



https://drive.google.com/file/d/1Z1i8mk8AqdtI_Q8avqtrzLQnLTG7QW6T/view

Generalized R-CNN framework for Object Detection



https://drive.google.com/file/d/1Z1i8mk8AqdtI_Q8avqtrzLQnLTG7QW6T/view

Face Recognition

- Database of K persons
- Get an input image
- Output the ID if image is of any of the K persons (or not recognized)

Face Recognition

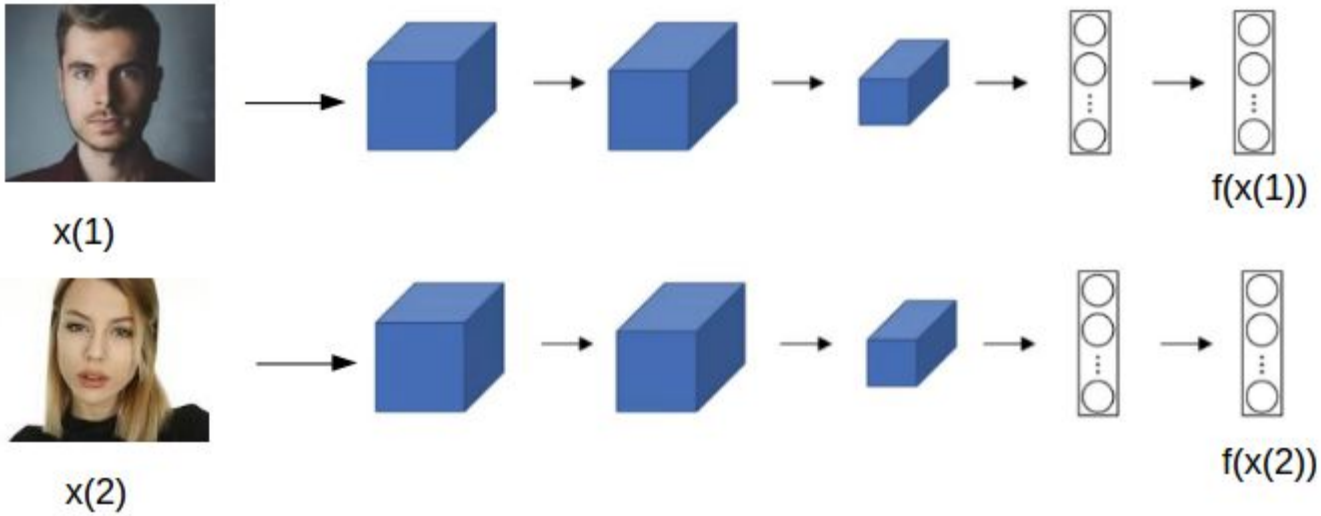
- One of the challenges that need to be addressed is one shot learning.
  - You need to recognize a person given just one example.
- If you train a network it won't be good enough to recognize a person from just one image.
- If a new person joins then you will have to retrain the network as well.
- How do you solve this problem?

Face Recognition

- Instead of learning how to recognize, the network is trained to learn a similarity function.
  - Given a new image, how similar it is to images in the database.
  - If similarity is less than threshold, then we say that the person is present in the database.
- We usually use networks called Siamese network to solve this issue.

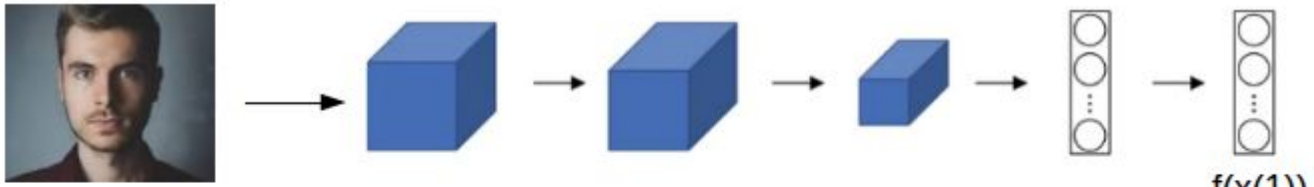
https://www.coursera.org/learn/convolutional-neural-networks/lecture/bjhmj/siamese-network

Face Recognition



$$d(x^{(1)}, x^{(2)}) = ||f(x^{(1)}) - f(x^{(2)})||_2^2$$

Goal of learning



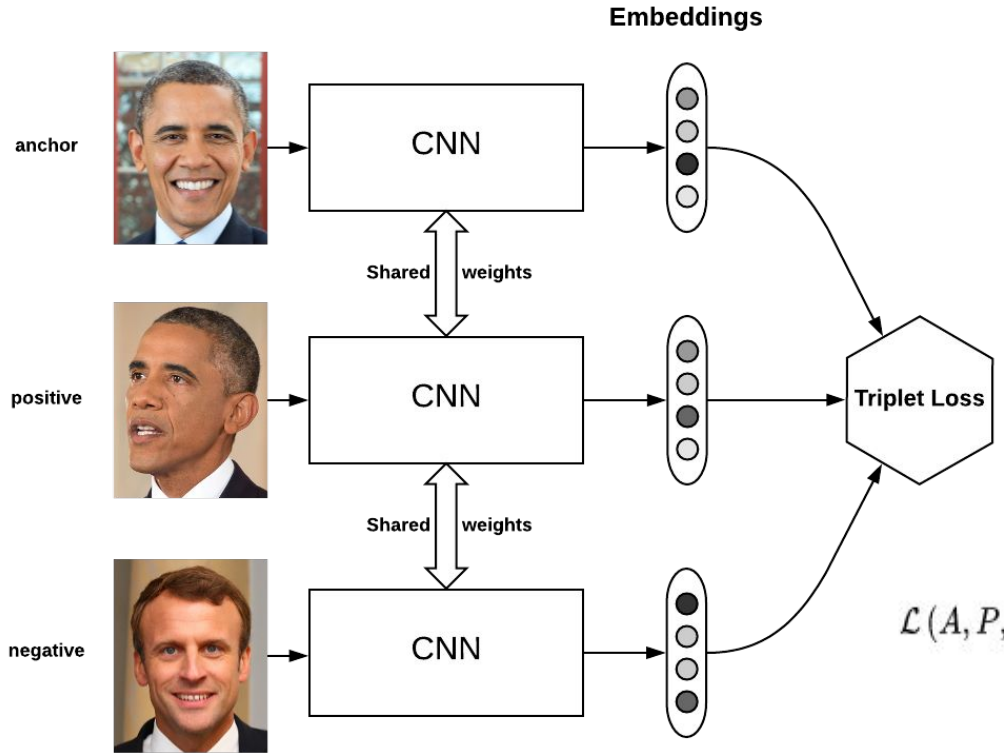
Parameters of NN define an encoding $f(x^{(i)})$

Learn parameters so that:

If $x^{(i)}, x^{(j)}$ are the same person, $\left\|f(x^{(i)}) - f(x^{(j)})\right\|^2$ is small.

If $x^{(i)}, x^{(j)}$ are different persons, $\left\|f(x^{(i)}) - f(x^{(j)})\right\|^2$ is large.

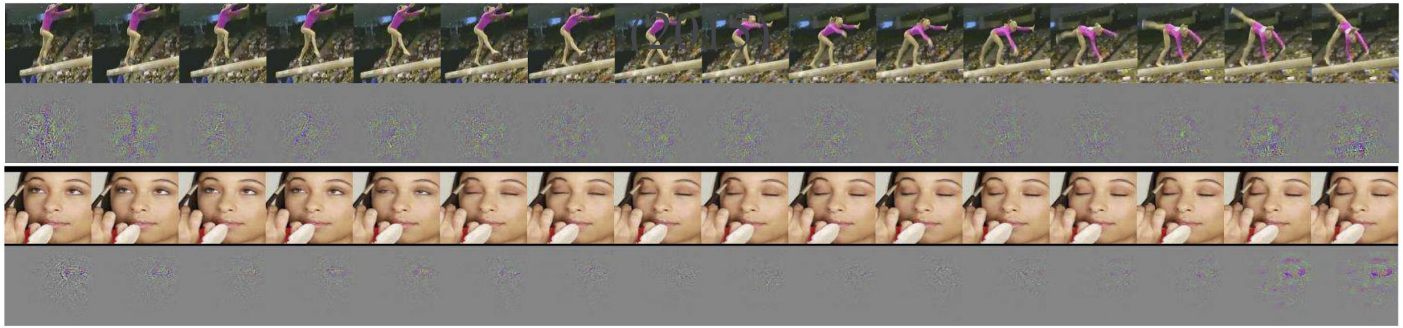
Face Recognition



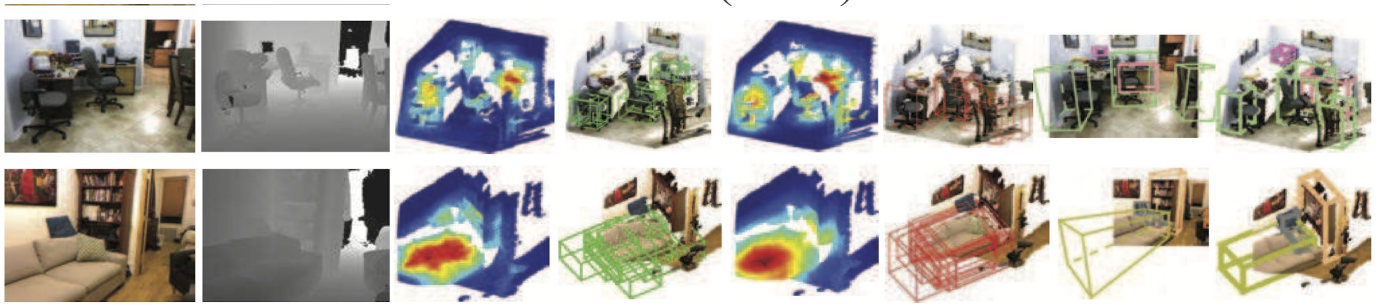
$$\mathcal{L}(A, P, N) = \max\left(\|\mathrm{f}(A) - \mathrm{f}(P)\|^2 - \|\mathrm{f}(A) - \mathrm{f}(N)\|^2 + \alpha, 0\right)$$

3D convolution

Motion in videos - Tran et al.

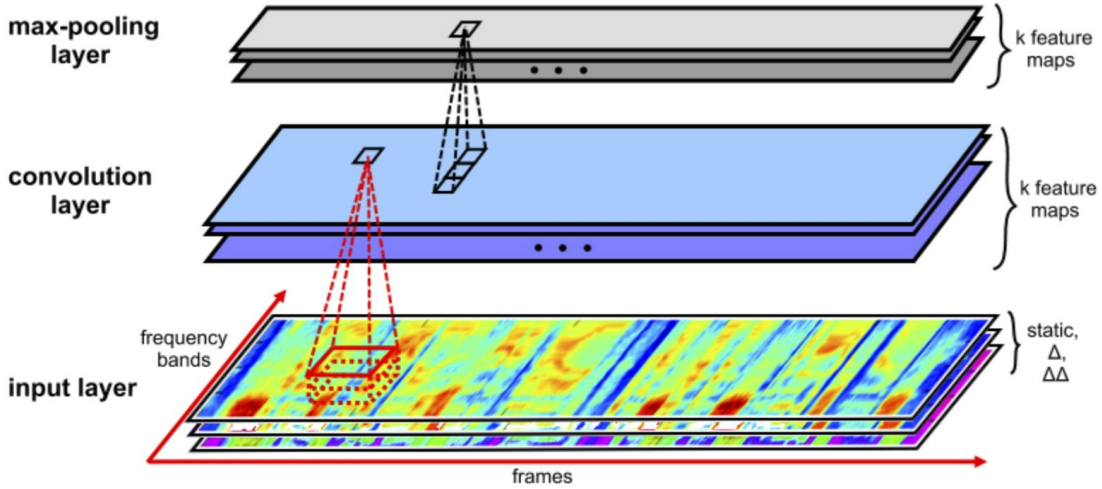


Objects from images w/ depths - Song & Xiao (2016)



Speech recognition

Zhang et al. 2017



- CNN competitive with RNNs (e.g. LSTMs)
- 10 layers, 3x5 conv, 3x1 pooling - deep enough for temporal dependencies
- TIMIT task, classifying phonemes

Myth: CNNs are for computer vision, and RNNs are for NLP.

CNNs are SOTA in many NLP tasks

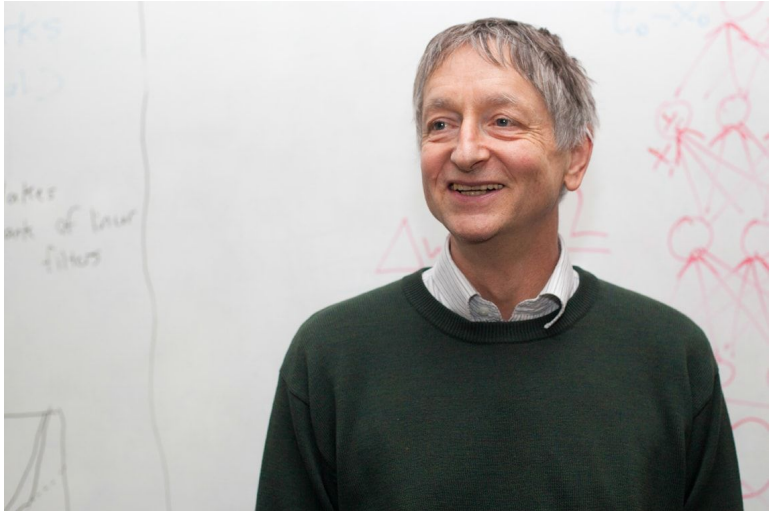
3. Text Classification

Research Paper	Datasets	Metric	Source Code	Year
Learning Structured Text Representations	Yelp	Accuracy: 68.6	• Tensorflow	2017
Attentive Convolution	Yelp	Accuracy: 67.36	• Theano	2017

https://github.com/bentrevett/pytorch-sentiment-analysis/blob/master/4%20-%20Convolutional%20Sentiment%20Analysis.ipynb

Geoffrey Hinton

- English-Canadian cognitive psychologist and computer scientist
- Popularized backpropagation
- The "Godfather of Deep Learning"
- Co-invented Boltzmann machines
- Contributed to AlexNet
- Advised Yann LeCunn, Ilya Sutskever, Radford Neal, Brendan Frey
- Works for google



What we learned today

Prehistoric convnets

How Alexnet works - i.e. how really good engineering looks like

How deeper networks win

How skip connections help

How CNNs are broadly used