Convolutional Neural Network

Phong Nguyen
Kien Le

Is Deep Neural Network a magic?

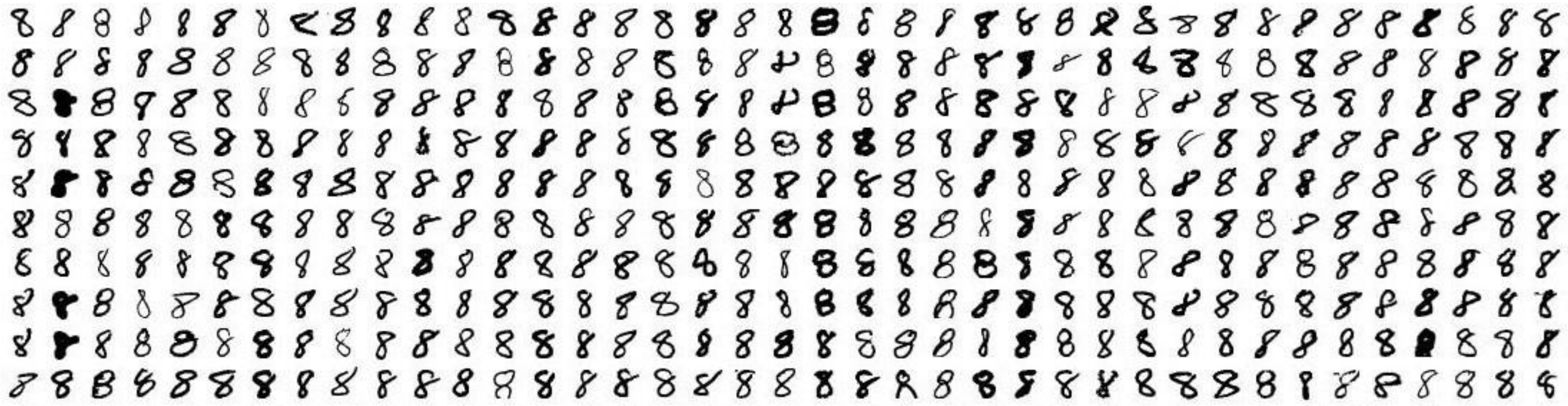
- We have learned about the basic of Deep Neural Network in the previous lectures.
- They are really magical, right?
- Let's use them again to train a model to recognize hand-written letters (MNIST dataset)

U.S. EMBASSY
HANOI, VIETNAM

VIETNAMESE ALUMNI
U.S. STATE DEPARTMENT

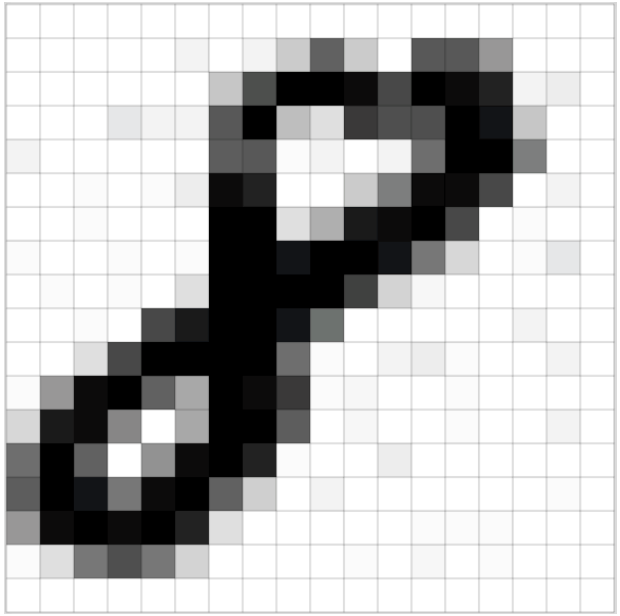
MNIST Dataset

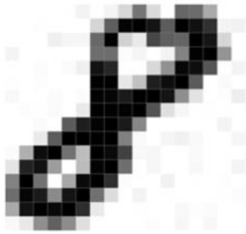
- MNIST provides 60,000 images of handwritten digits, each as an 18x18 image.
- Let’s just say we need to recognize a digit is “8” or not.



Just a simple classification problem

- You would think the power of ML can do it very well?
- Let's apply the DNN we have learned before.



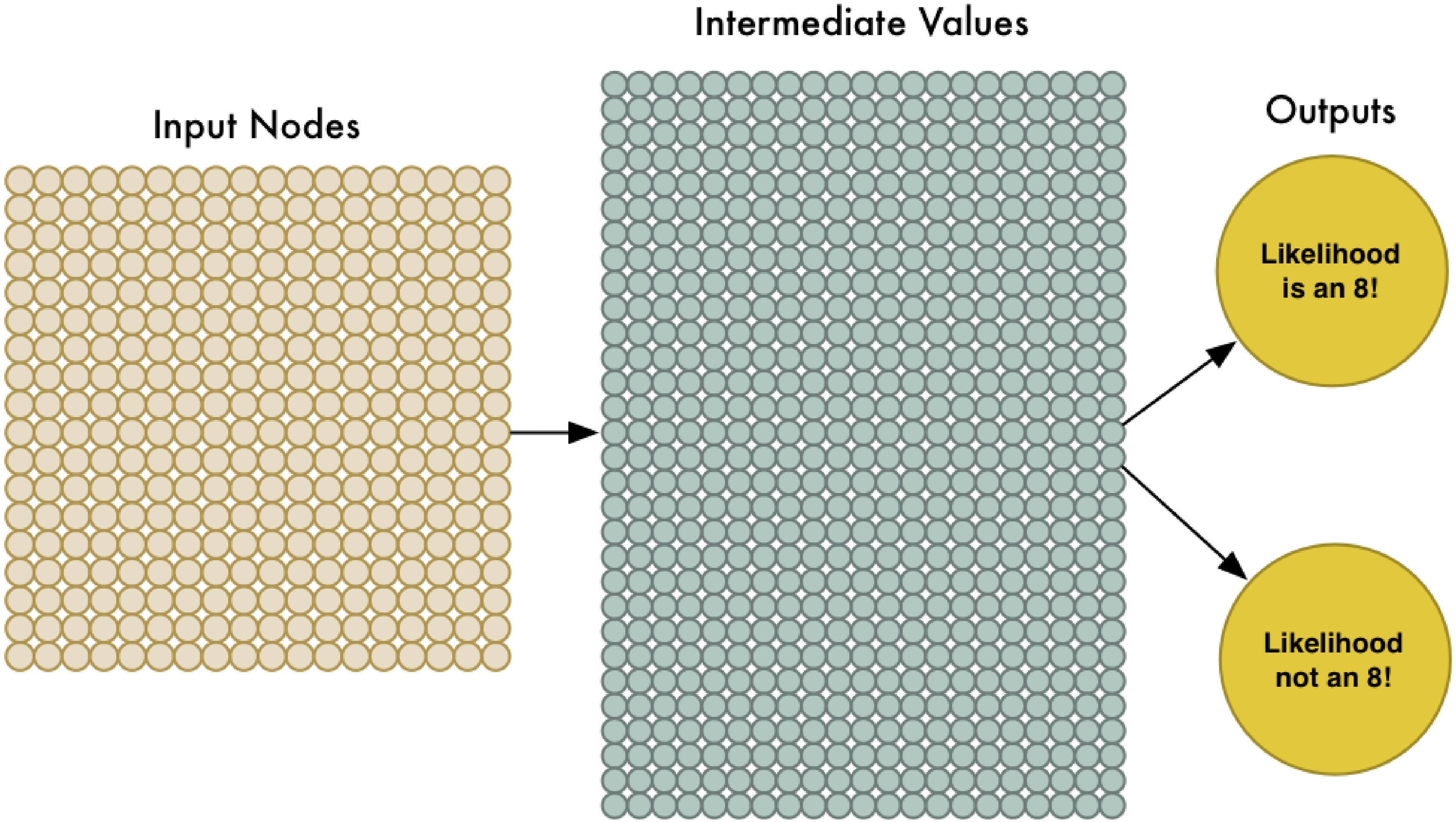


=


```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 12, 0, 11, 39, 137, 37, 0, 152, 147, 84, 0, 0, 0, 0,
0, 1, 0, 0, 0, 41, 160, 250, 255, 235, 162, 255, 238, 206, 11, 13, 0, 0, 0, 0, 16, 9, 9, 150, 251, 45, 21, 184, 159, 154, 2
55, 233, 40, 0, 0, 10, 0, 0, 0, 0, 0, 145, 146, 3, 10, 0, 11, 124, 253, 255, 107, 0, 0, 0, 0, 3, 0, 4, 15, 236, 216, 0, 0,
38, 109, 247, 240, 169, 0, 11, 0, 1, 0, 2, 0, 0, 0, 253, 253, 23, 62, 224, 241, 255, 164, 0, 5, 0, 0, 6, 0, 0, 4, 0, 3, 252
, 250, 228, 255, 255, 234, 112, 28, 0, 2, 17, 0, 0, 2, 1, 4, 0, 21, 255, 253, 251, 255, 172, 31, 8, 0, 1, 0, 0, 0, 0, 0, 4,
 0, 163, 225, 251, 255, 229, 120, 0, 0, 0, 0, 0, 11, 0, 0, 0, 0, 21, 162, 255, 255, 254, 255, 126, 6, 0, 10, 14, 6, 0, 0, 9
, 0, 3, 79, 242, 255, 141, 66, 255, 245, 189, 7, 8, 0, 0, 5, 0, 0, 0, 0, 26, 221, 237, 98, 0, 67, 251, 255, 144, 0, 8, 0, 0
, 7, 0, 0, 11, 0, 125, 255, 141, 0, 87, 244, 255, 208, 3, 0, 0, 13, 0, 1, 0, 1, 0, 0, 145, 248, 228, 116, 235, 255, 141, 34
, 0, 11, 0, 1, 0, 0, 0, 1, 3, 0, 85, 237, 253, 246, 255, 210, 21, 1, 0, 1, 0, 0, 6, 2, 4, 0, 0, 0, 6, 23, 112, 157, 114, 32
, 0, 0, 0, 0, 2, 0, 8, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```


- This is basically the input of the DNN.

The model we have learned so far



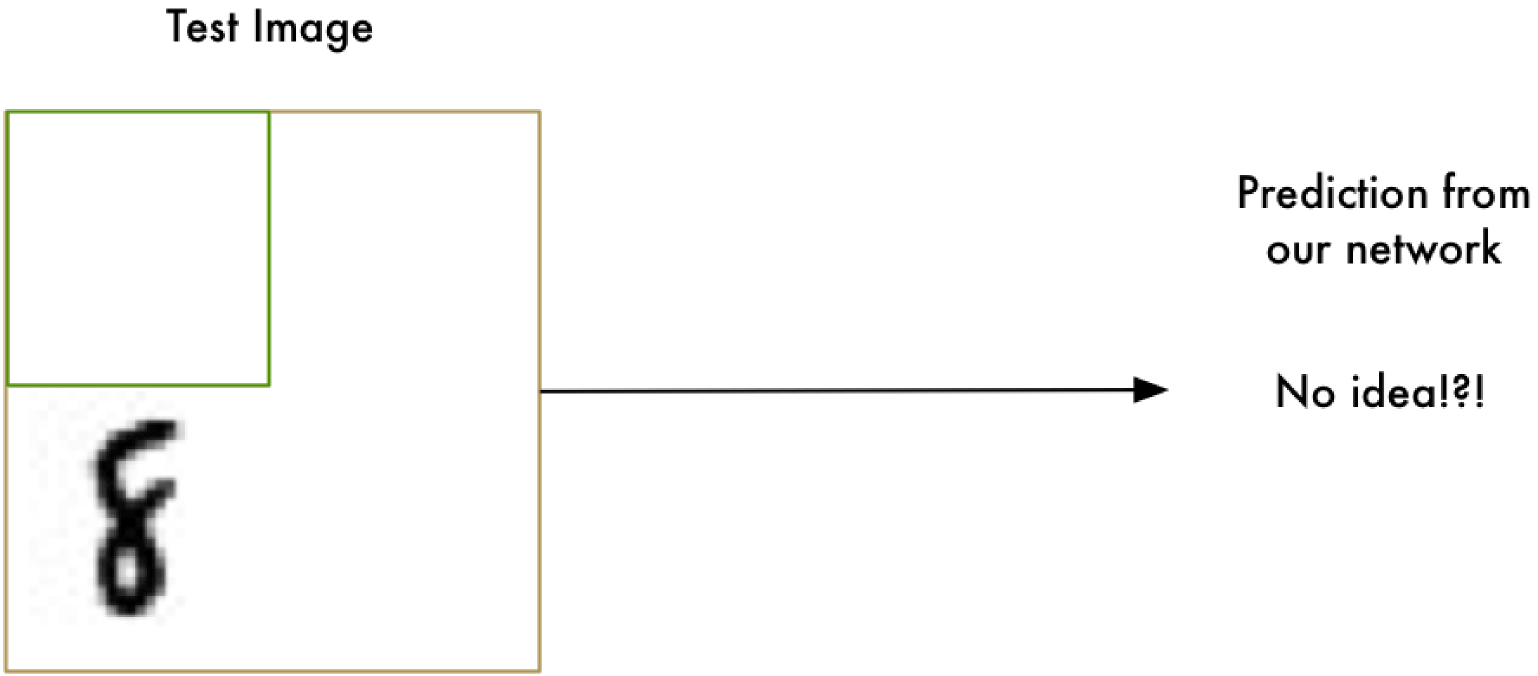
Too easy for DNN???

- Good news!

- But now, the really bad news!

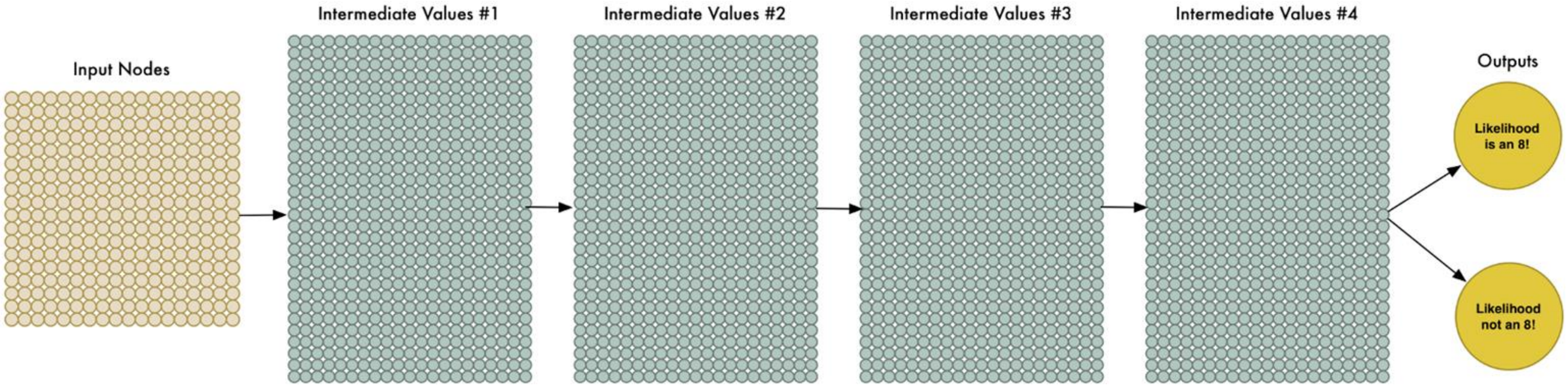
Call for ideas!!!!

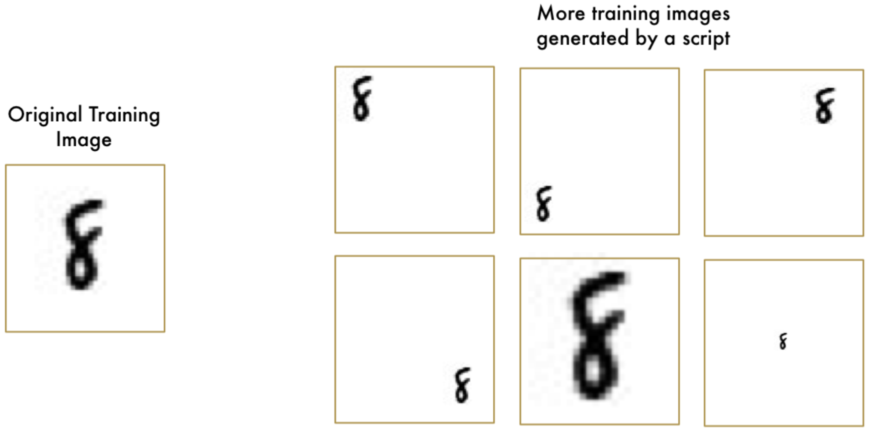
- Brute force is the king. But that's not the point we learn ML.



Call for ideas!!!!

- More ML way, more data for training, deeper network.





But does it make sense?

- First, It doesn't make sense to train a network to recognize an "8" at the top of a picture separately from training it to recognize an "8" at the bottom of a picture as if those were two totally different objects.
- And what about different way to write number "8"? It's not ML anymore if we need to cover for all such situations.

And magic does exist!

- And the elegant solutions for such situation is: CONVOLUTION.
- Anything in this world has a hierarchy or conceptual structure.
- For example: Human
  - Body
    - Hands
      - Fingers
    - Foot
      - Toes
    - Torso
  - Head
    - Eyes
    - Noses
    - Ears

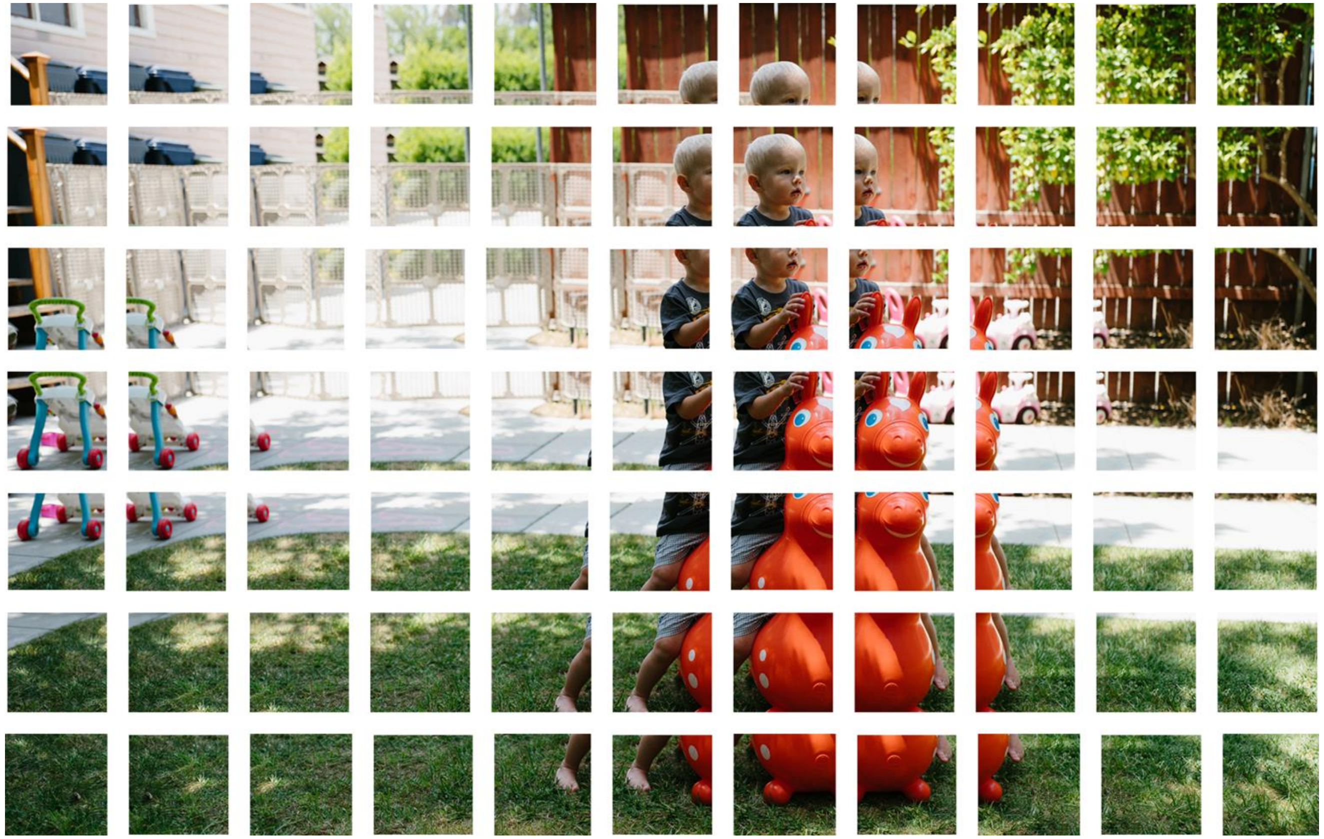
Let’s consider this picture

What your brain sees?

- As a human, you instantly recognize the hierarchy in this picture:
  - The ground is covered in grass and concrete
  - There is a child
  - The child is sitting on a bouncy horse
  - The bouncy horse is on top of the grass
- More importantly, we recognize the child no matter where the background is, without “re-learning”.
- But our DNN can’t do this right now.

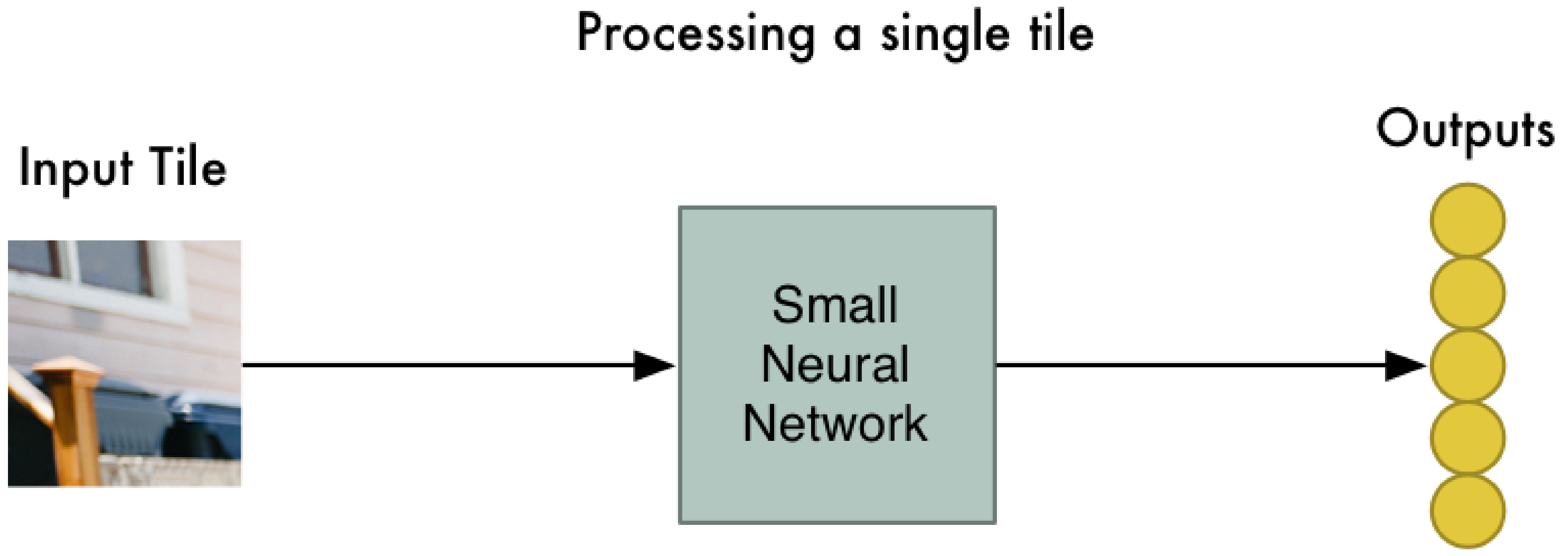
How Convolution Works

- Step 1: Break the image into overlapping images tiles



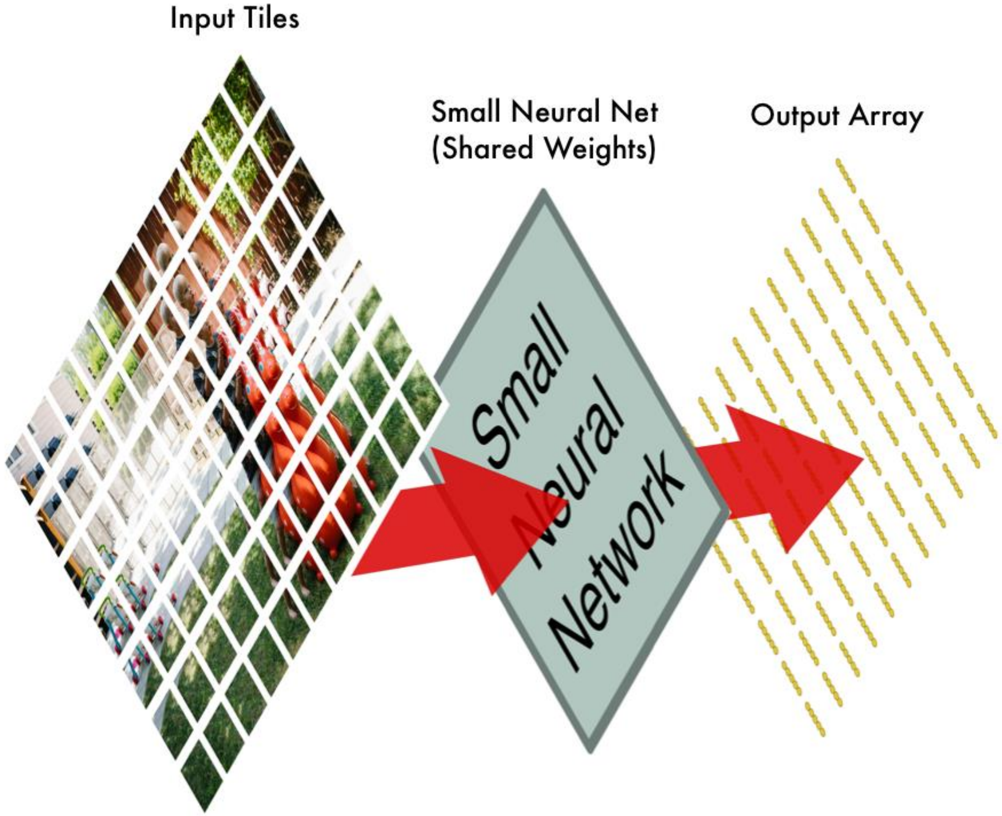
How Convolution Works

- Step 2: Feed each image tile into a small neural network



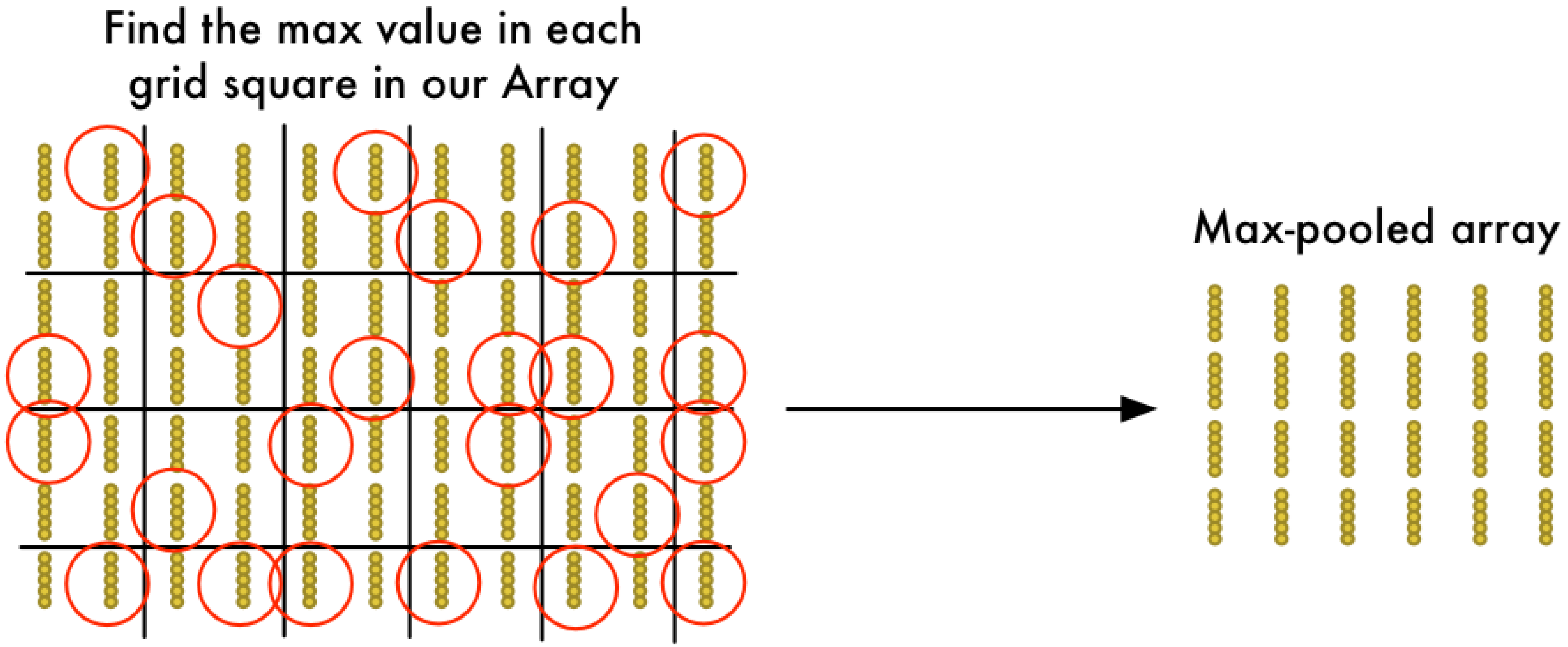
How Convolution Works

- Step 3: Save the results from each tile into a new array



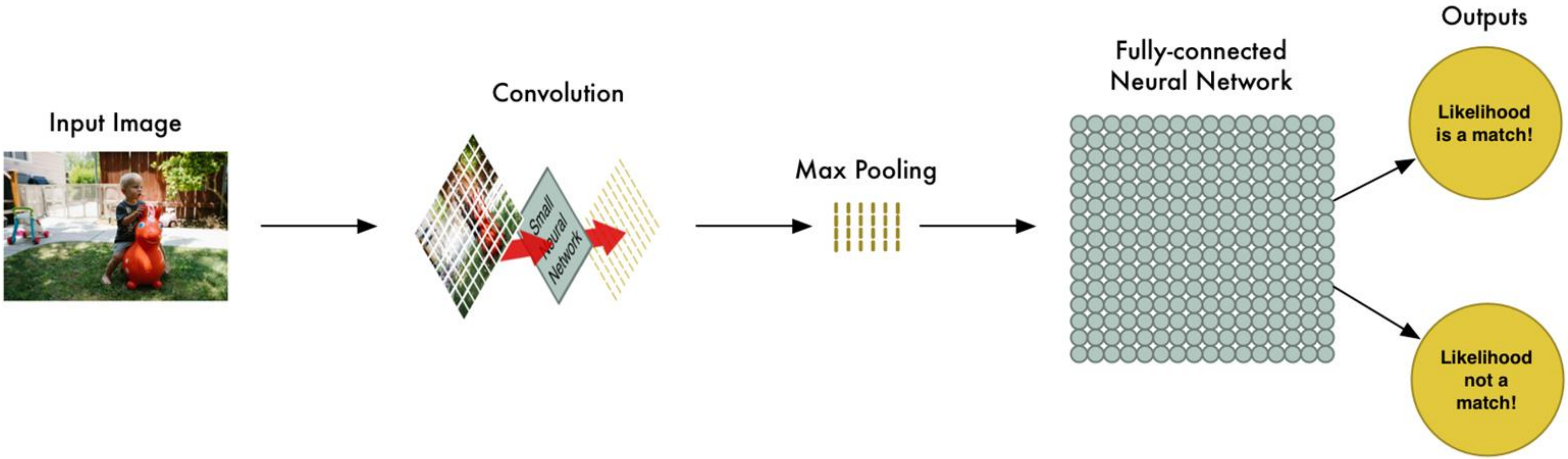
How Convolution Works

- Step 4: Downsampling

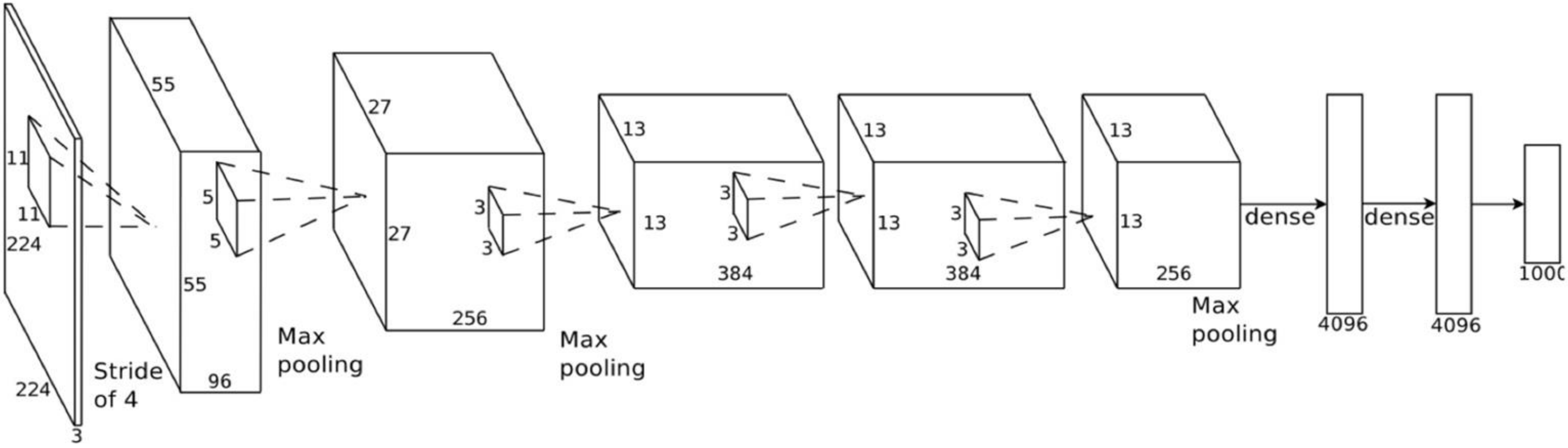


How Convolution Works

- Step 5: Make a prediction



A realistic Convolutional Neural Network



U.S. EMBASSY HANOI, VIETNAM

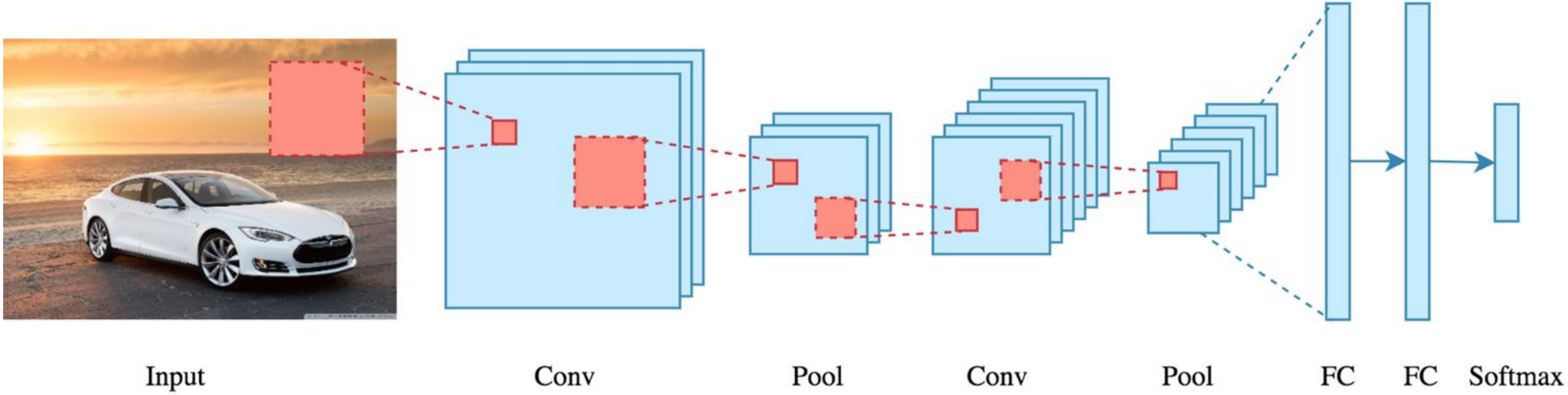
VIETNAMESE ALUMNI U.S. STATE DEPARTMENT

I’m sorry, but time for history!

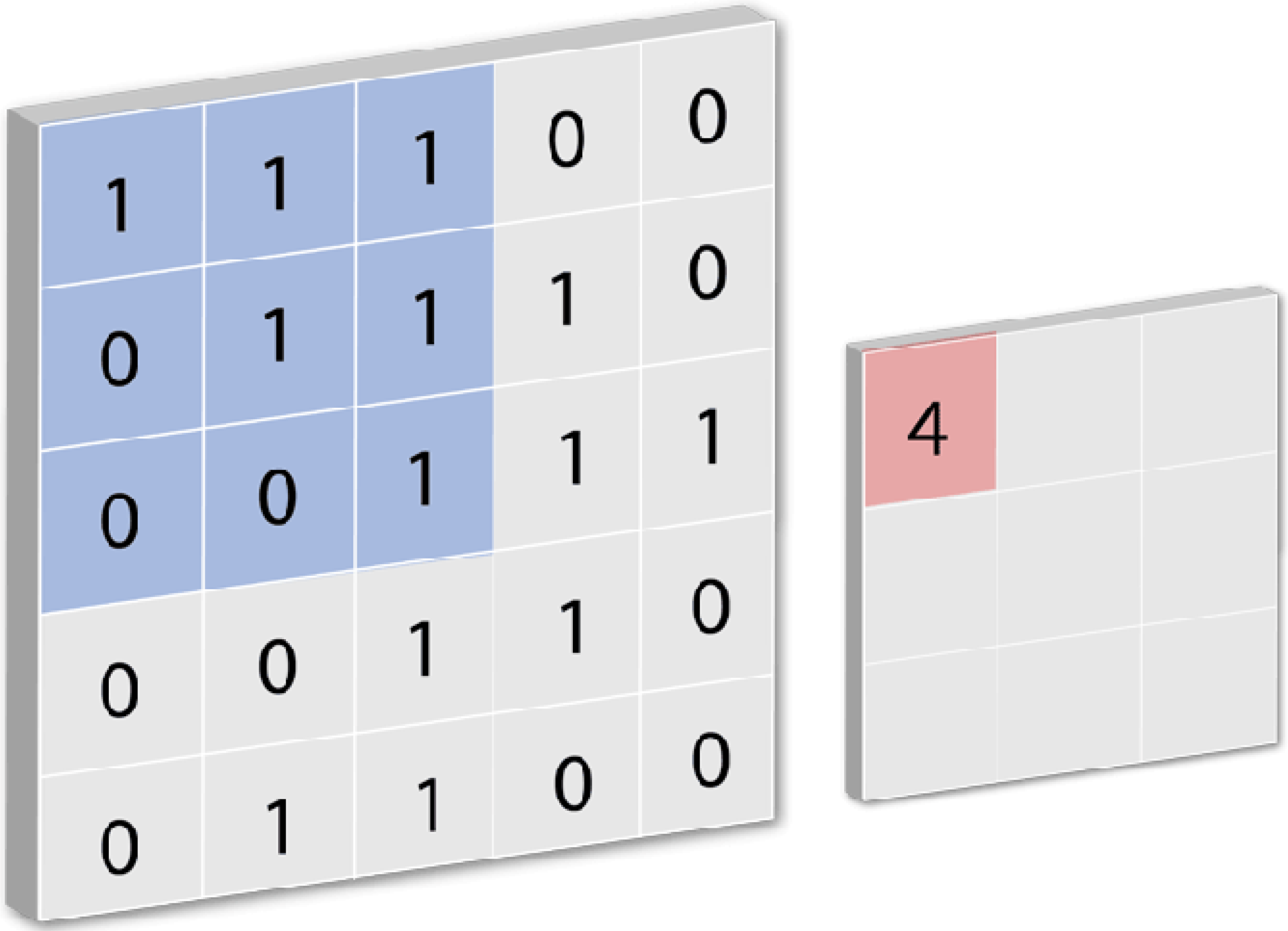
- CNNs was popularized mostly thanks to [Yann LeCun](), now the Director of AI Research at Facebook.
- In early 1990s, LeCun worked at Bell Labs, one of the most pretigous research labs in the world at the time.
- Check out this video demo in 1993. Again, 1993.

https://www.youtube.com/watch?v=FwFduRA_L6Q

Let’s recap

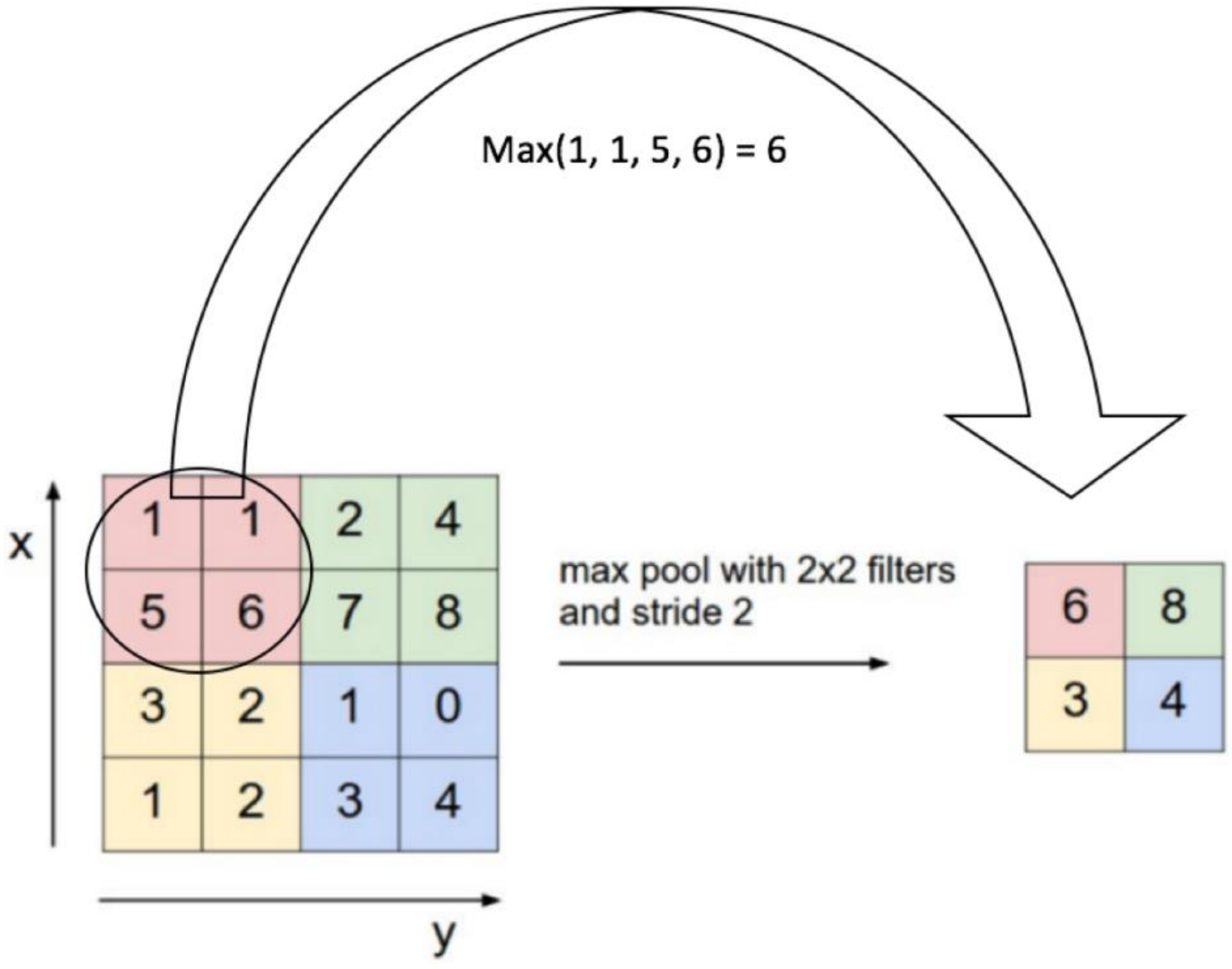


1) Convolution Layer



- Beware of the size of the “filters” or “kernels”
- The smaller matrix is called “Feature Map”
- Size of feature map is controlled by “depth” (the number of filters used), and “stride” (the number of pixels slid over)
- ReLu activation function.

2) Pooling Layer



Rectified Feature Map

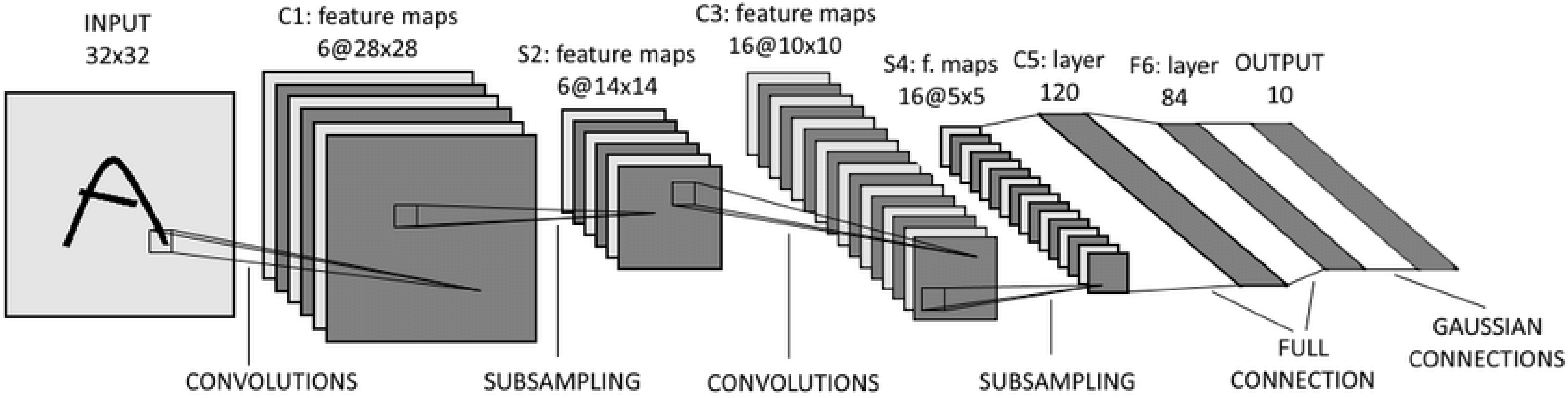
- To reduce dimensionality of each feature map.

=> reduce # of parameters, computations & avoid overfitting.

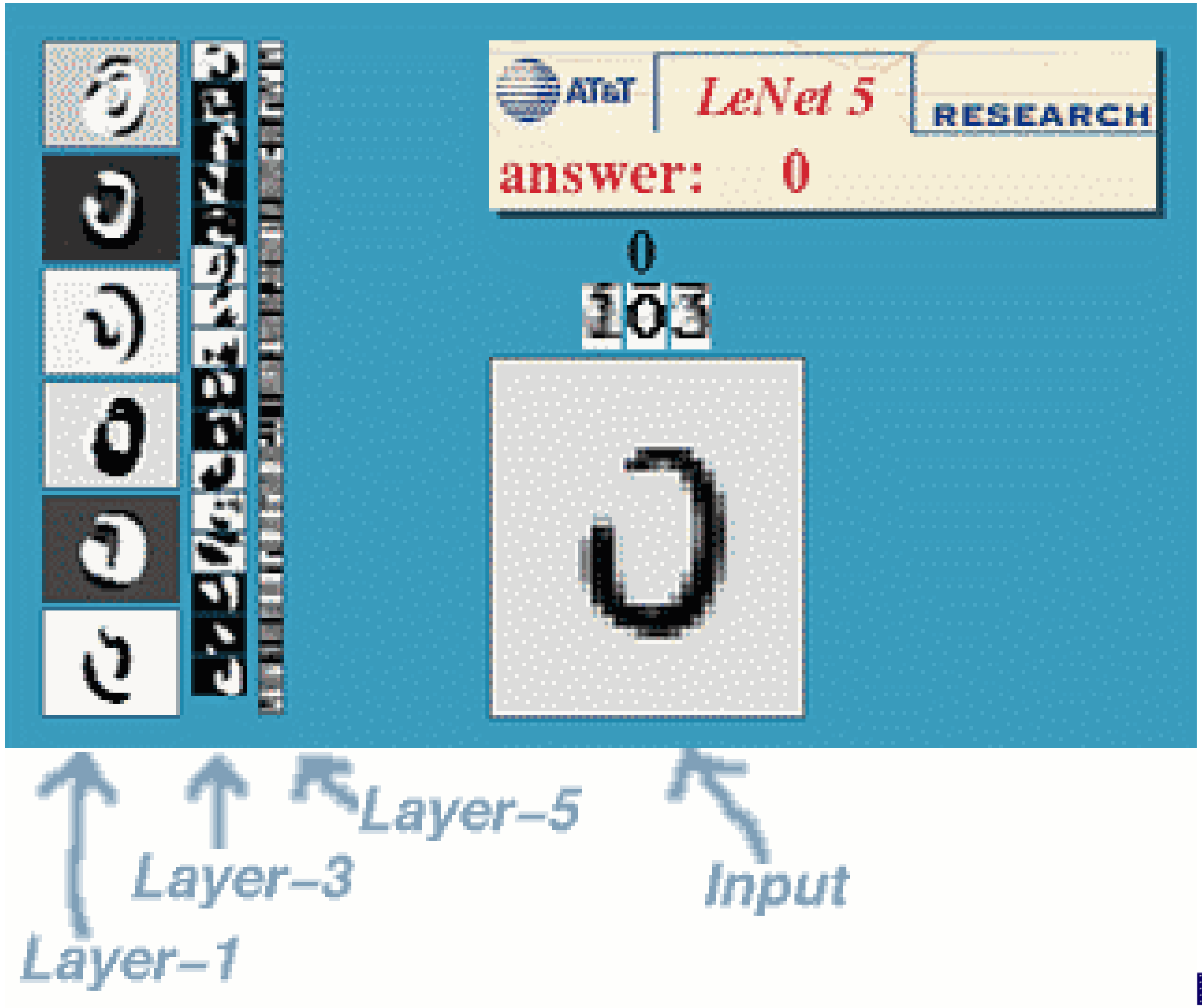
- Max pool vs. average pool.
- invariant to small transformations, distortions, and translations in the input

Discussion of some CNN architecture

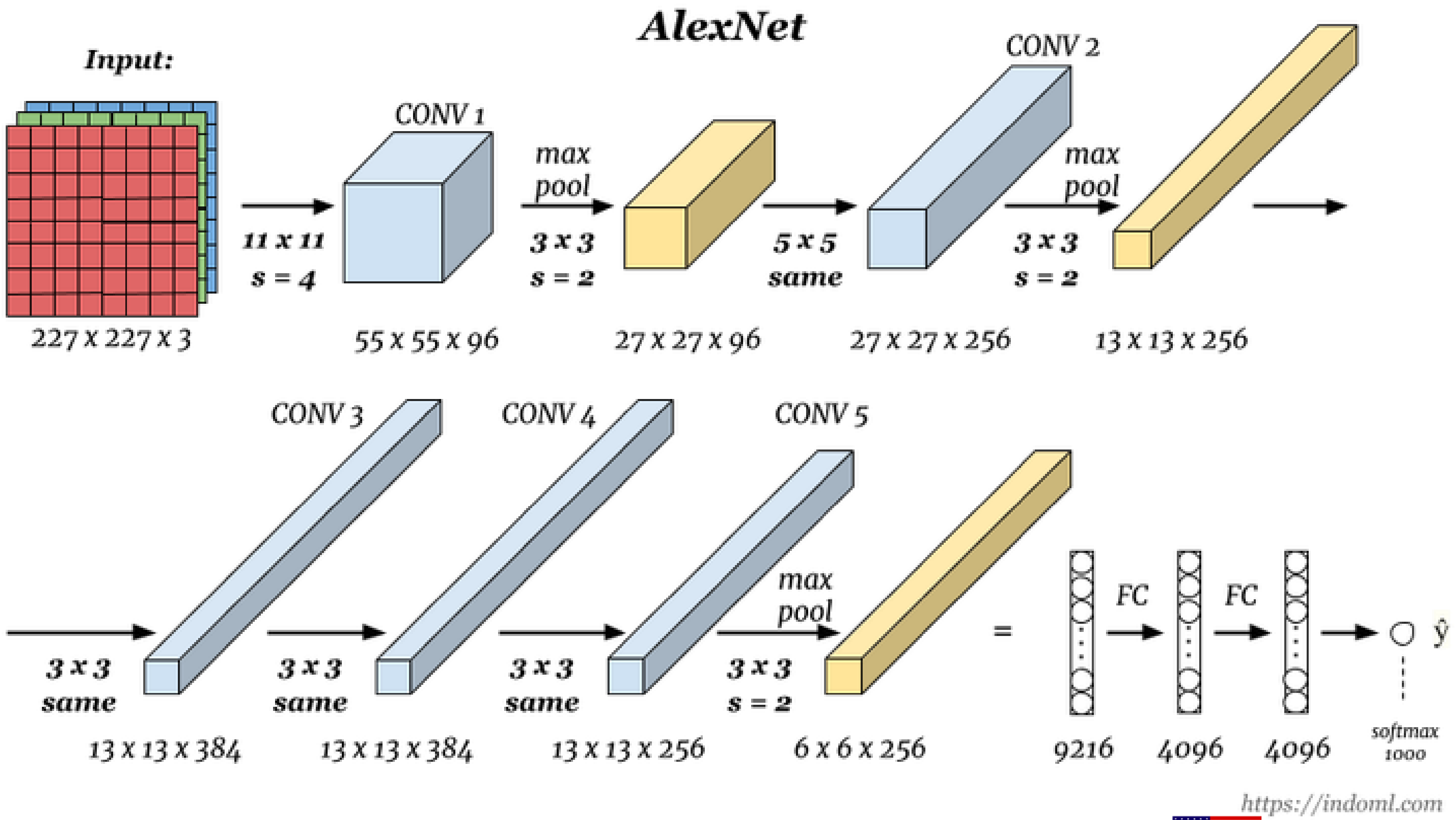
LeNet



LeNet results

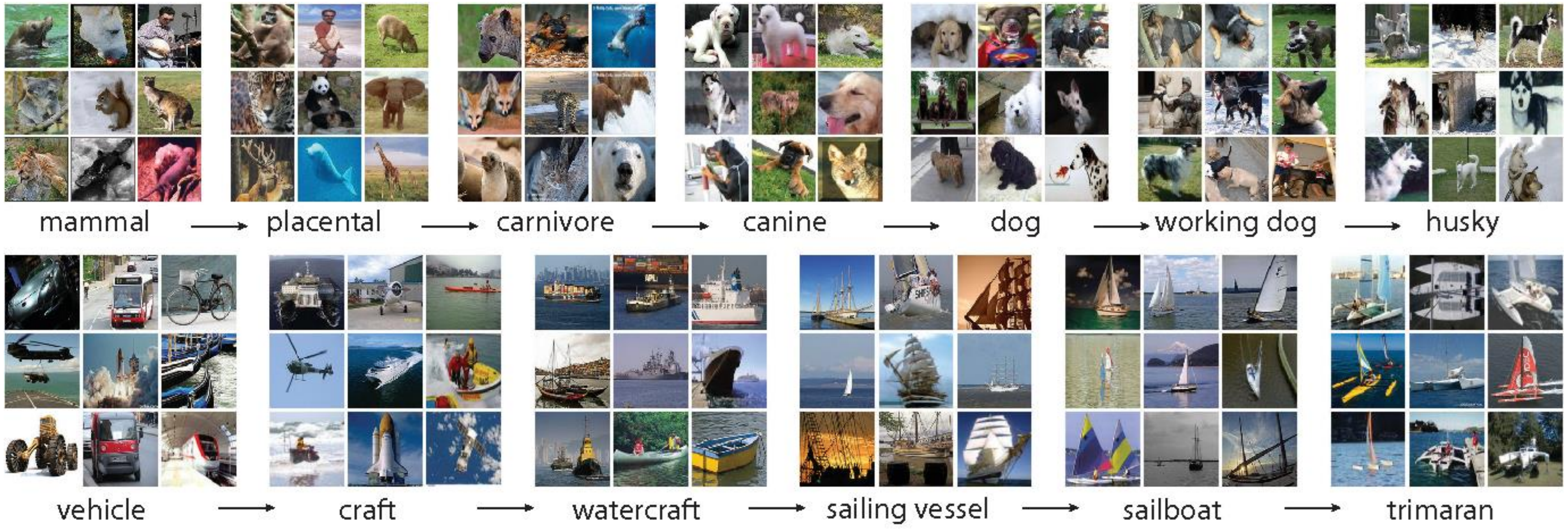


AlexNet - 2012

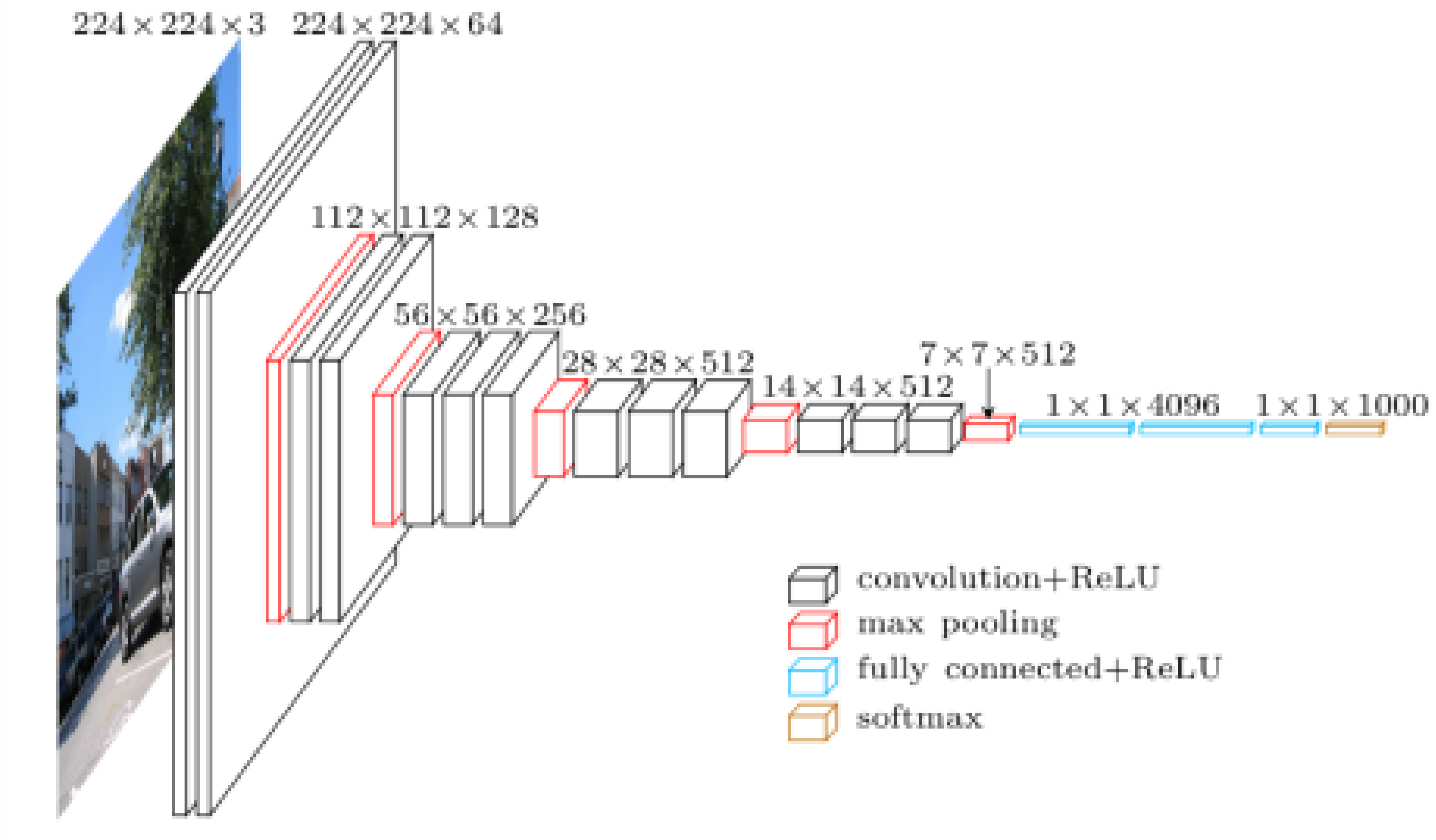


AlexNet results

- Winner of the first ILSVRC-2012 competition (ImageNet Large Scale Visual Recognition Challenge).
- Pioneering “deep” CNN with 84.6% accuracy, while the second-place model only achieved 73.8% accuracy rate.



VGG Net - 2014



https://arxiv.org/pdf/1409.1556.pdf

Varients

6 different configurations for vgg net. The VGG16 produced the best result.

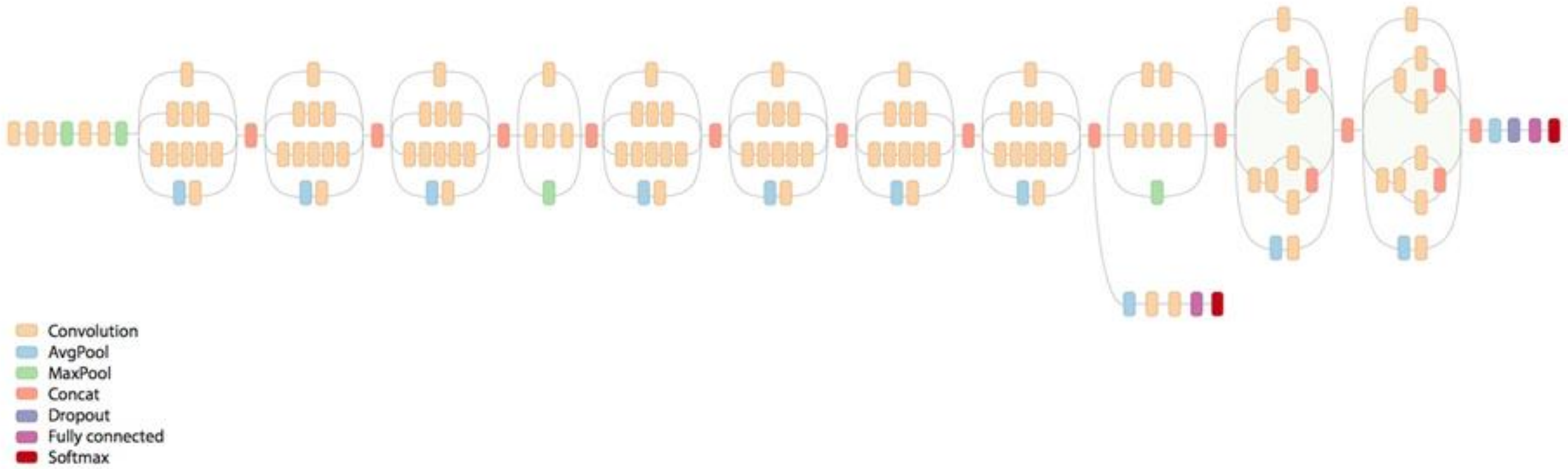
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	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					

The 6 different architecures of VGG Net. Configuration D produced the best results

VGG Net results

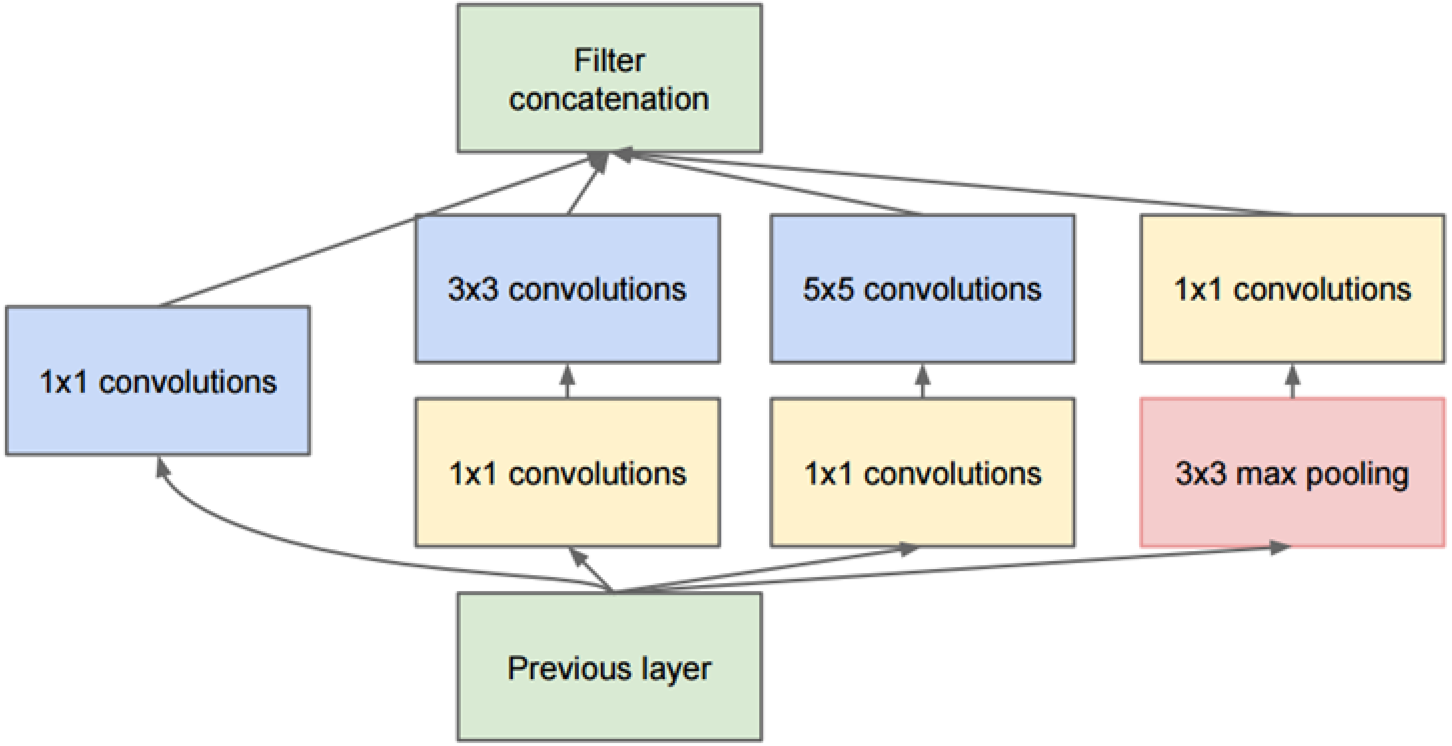
- 1st runner-up of ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2014 (Not the winner, winner was GoogleNet)
- The first model reached a error rate below 10% (approx 7% on final model)
- Very beauty and straightforward design

Inception Net - 2014



Another view of GoogLeNet's architecture.

Inception Module

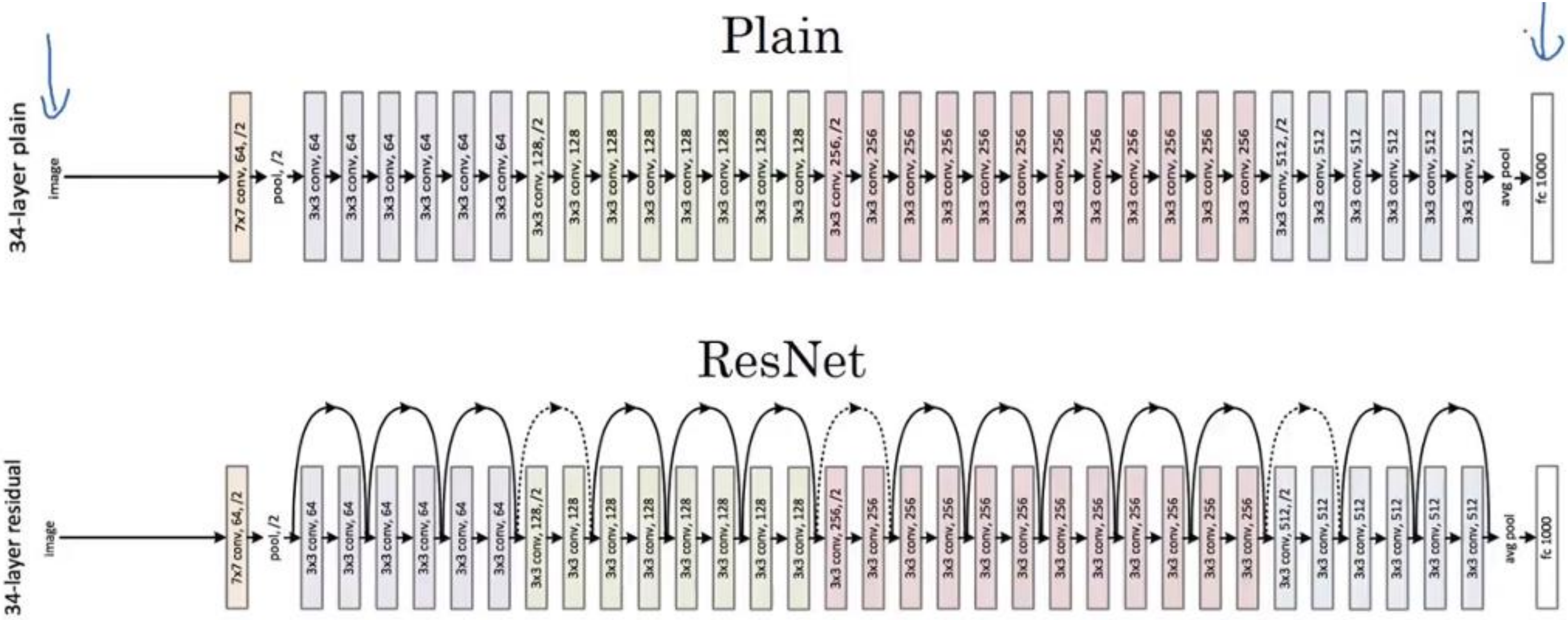


Full Inception module

Inception Net results

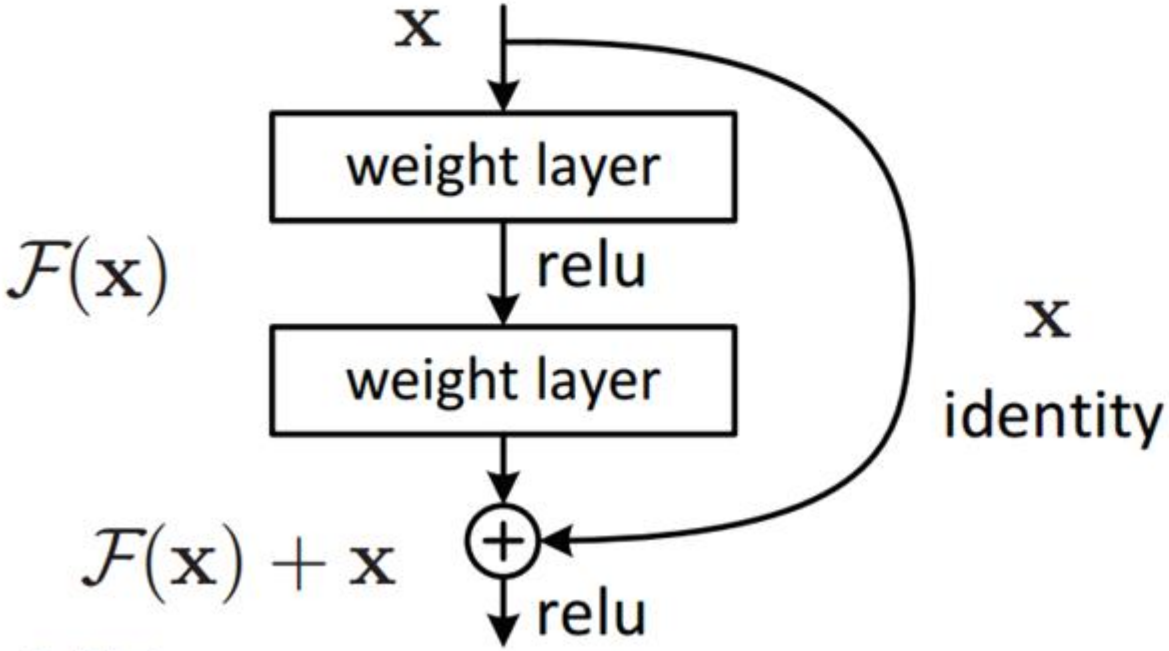
- Winner of ILSVRC 2014 with error rate of 6.7%
- The first model didn’t follow the “conv-pool” stacking rule.
- Doing pooling and conv with different filter size simutaneously
- Uses 12x fewer parameters than AlexNet

Microsoft Resnet - 2015



https://arxiv.org/pdf/1512.03385.pdf

Resnet - Residual block



A residual block

U.S. EMBASSY HANOI, VIETNAM

VIETNAMESE ALUMNI U.S. STATE DEPARTMENT

Resnet results

- New records in # of layers : 152 layers
- ResNet won ILSVRC 2015 with an error rate of 3.6%

Applications

- Object Detection
- Object tracking
- Object Recognition
- Semantic and Segmentation
- Video Image Captioning

Object Detection

- R-CNN, Fast R-CNN, Faster R-CNN

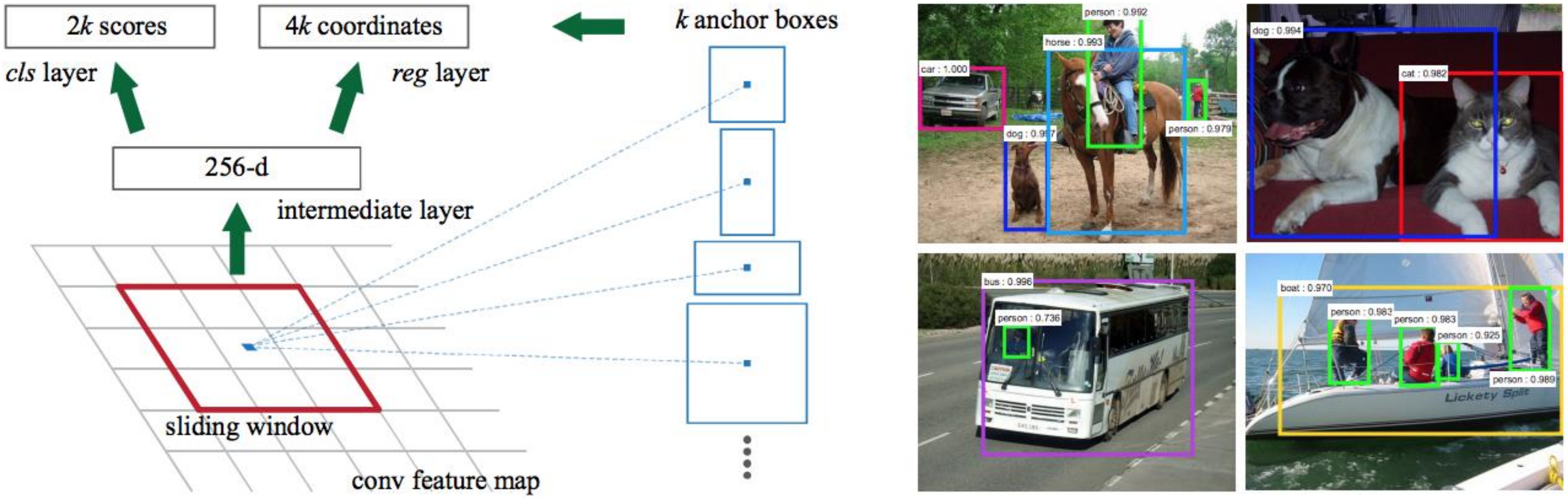


Figure 3: **Left**: Region Proposal Network (RPN). **Right**: Example detections using RPN proposals on PASCAL VOC 2007 test. Our method detects objects in a wide range of scales and aspect ratios.

Object Tracking

- DeepTrack

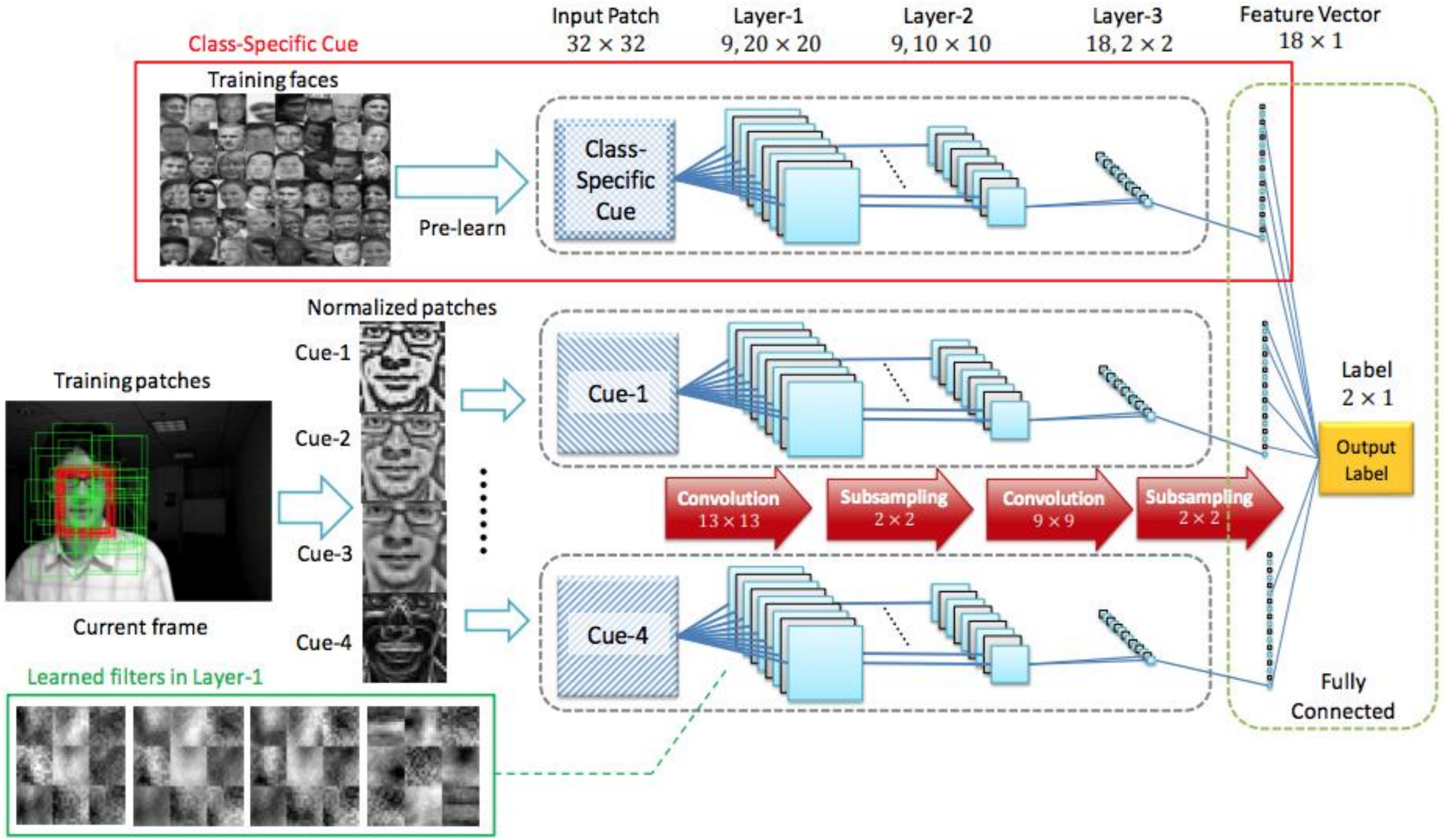


Figure 1: Overall architecture with (red box) and without (rest) the class-specific version.

Object Recognition

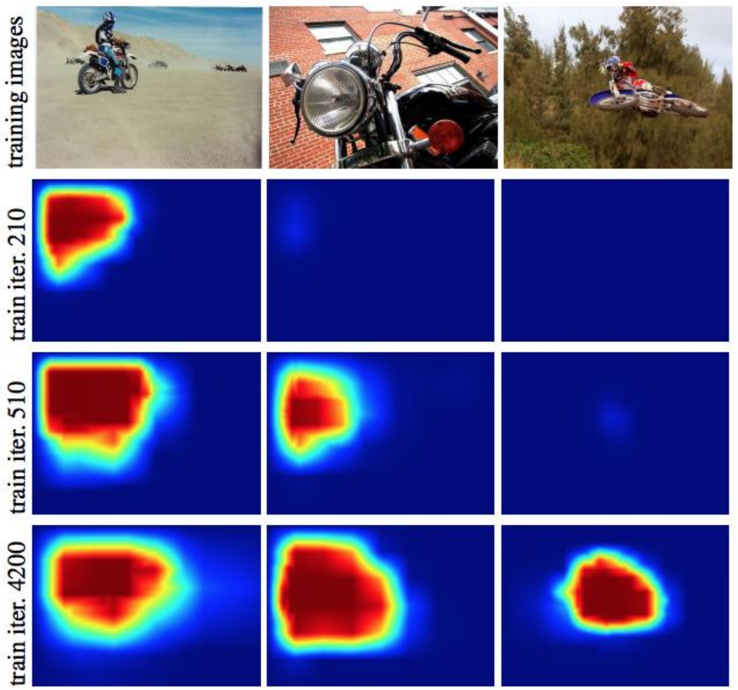


Figure 1: Evolution of localization score maps for the `motorbike` class over iterations of our weakly-supervised CNN training. Note that the network learns to localize objects despite having no object location annotation at training, just object presence/absence labels. Note also that locations of objects with more usual appearance (such as the motorbike shown in left column) are discovered earlier during training.

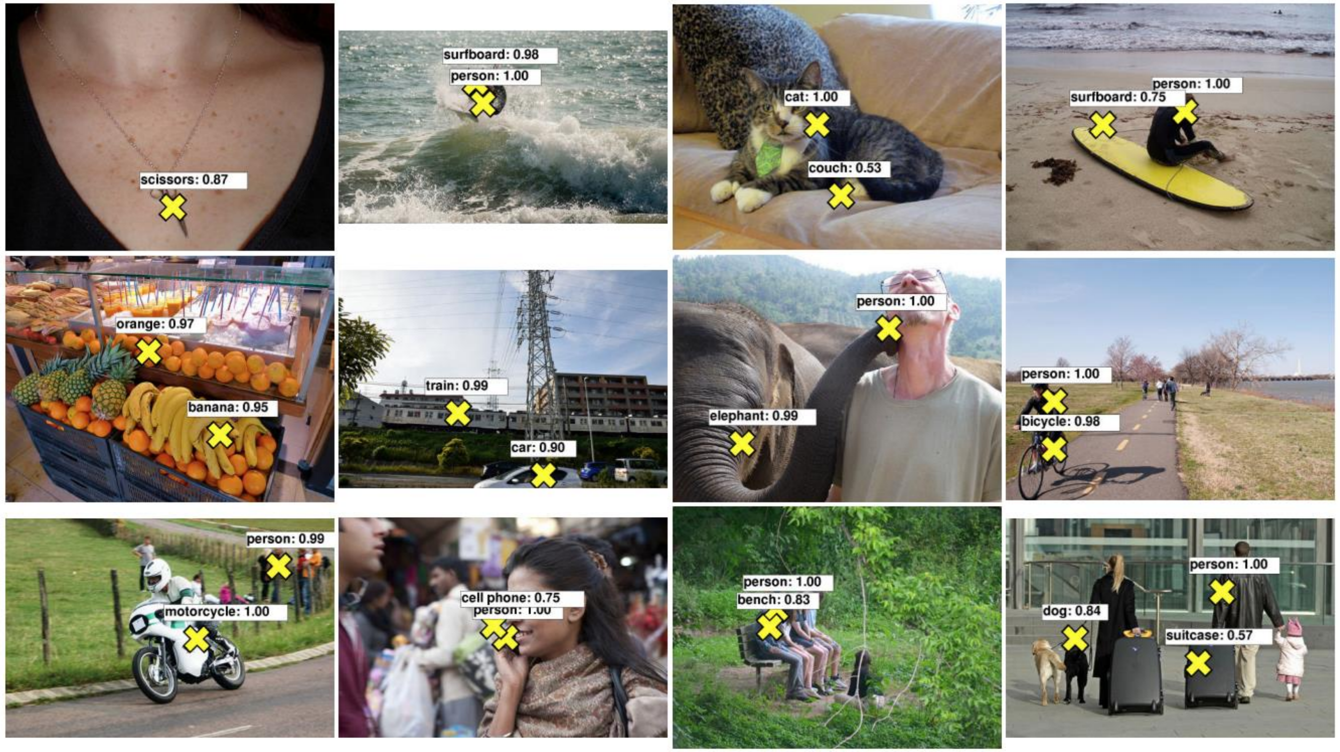


Figure 6: Example location predictions for images from the Microsoft COCO validation set obtained by our weakly-supervised method. Note that our method does not use object locations at training time, yet can predict locations of objects in test images (yellow crosses). The method outputs the most confident location per object per class. **Please see additional results on the project webpage[1].**

Semantic Segmentation

4

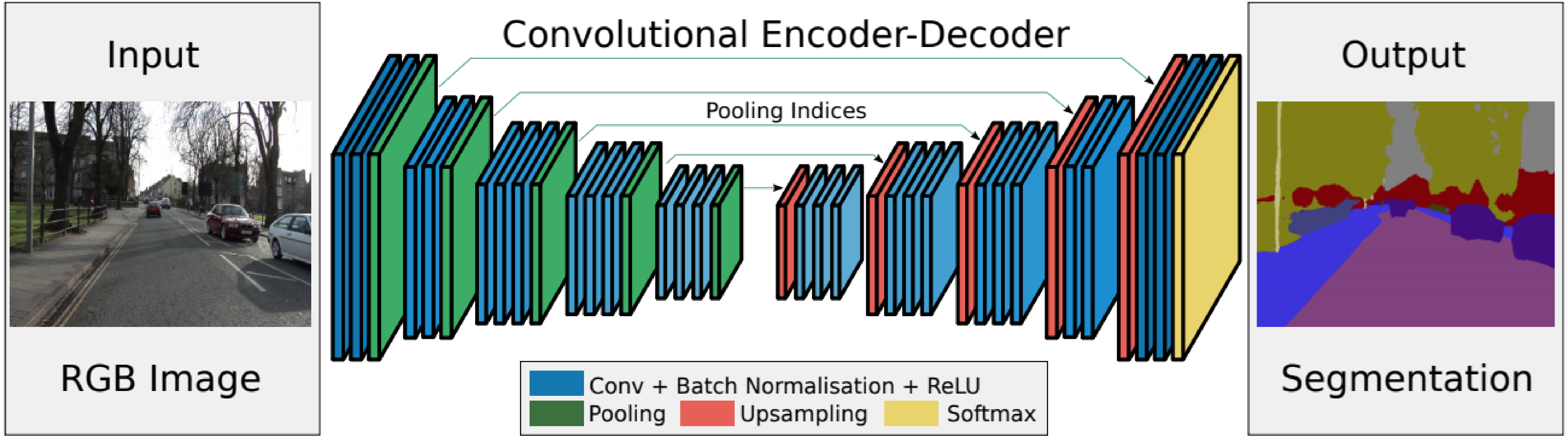


Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

Video Image Captioning