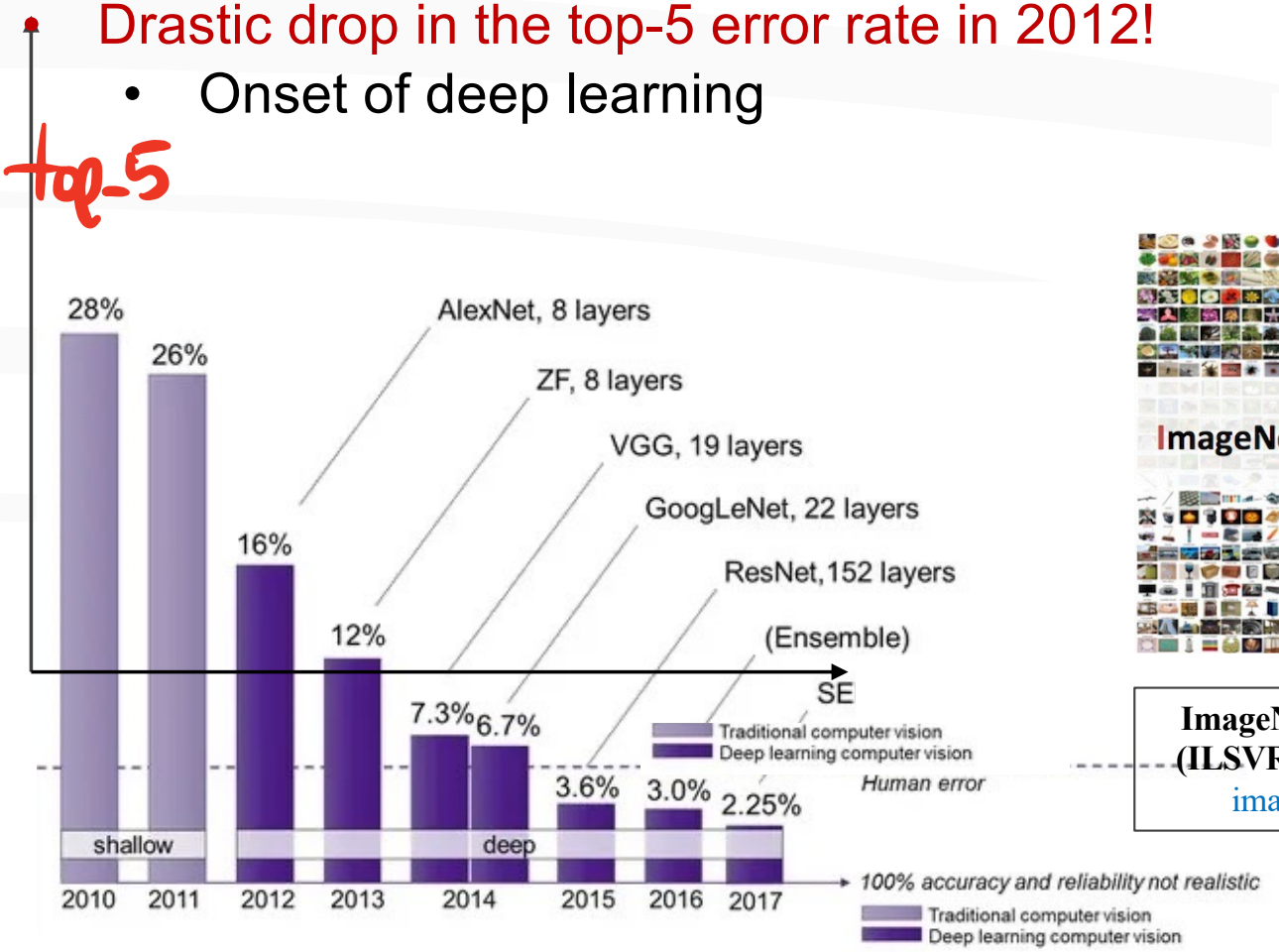
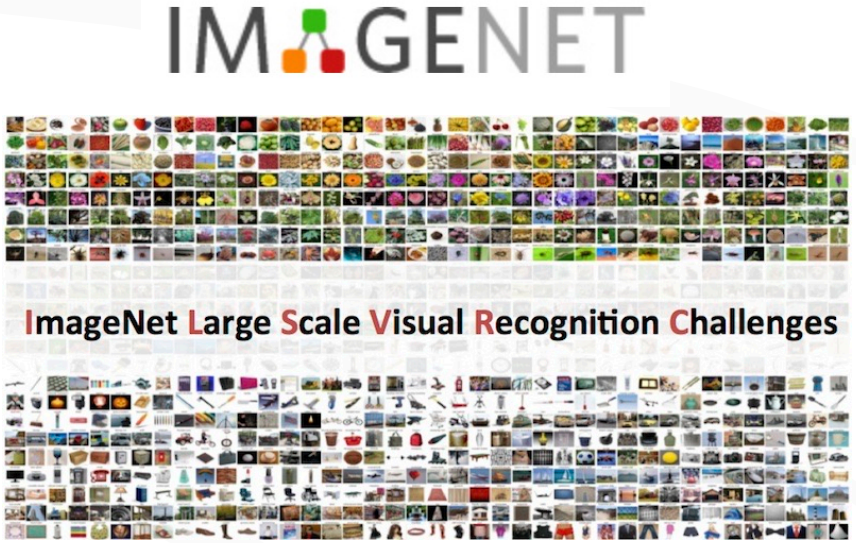
Deep Learning w/ Convolutional Neural Networks

Machine Learning – Timeline

- ILSVRC competition has been organized to benchmark progress in **large scale object classification**, using the **ImageNET** dataset
- Drastic drop in the top-5 error rate in 2012!
  - Onset of deep learning

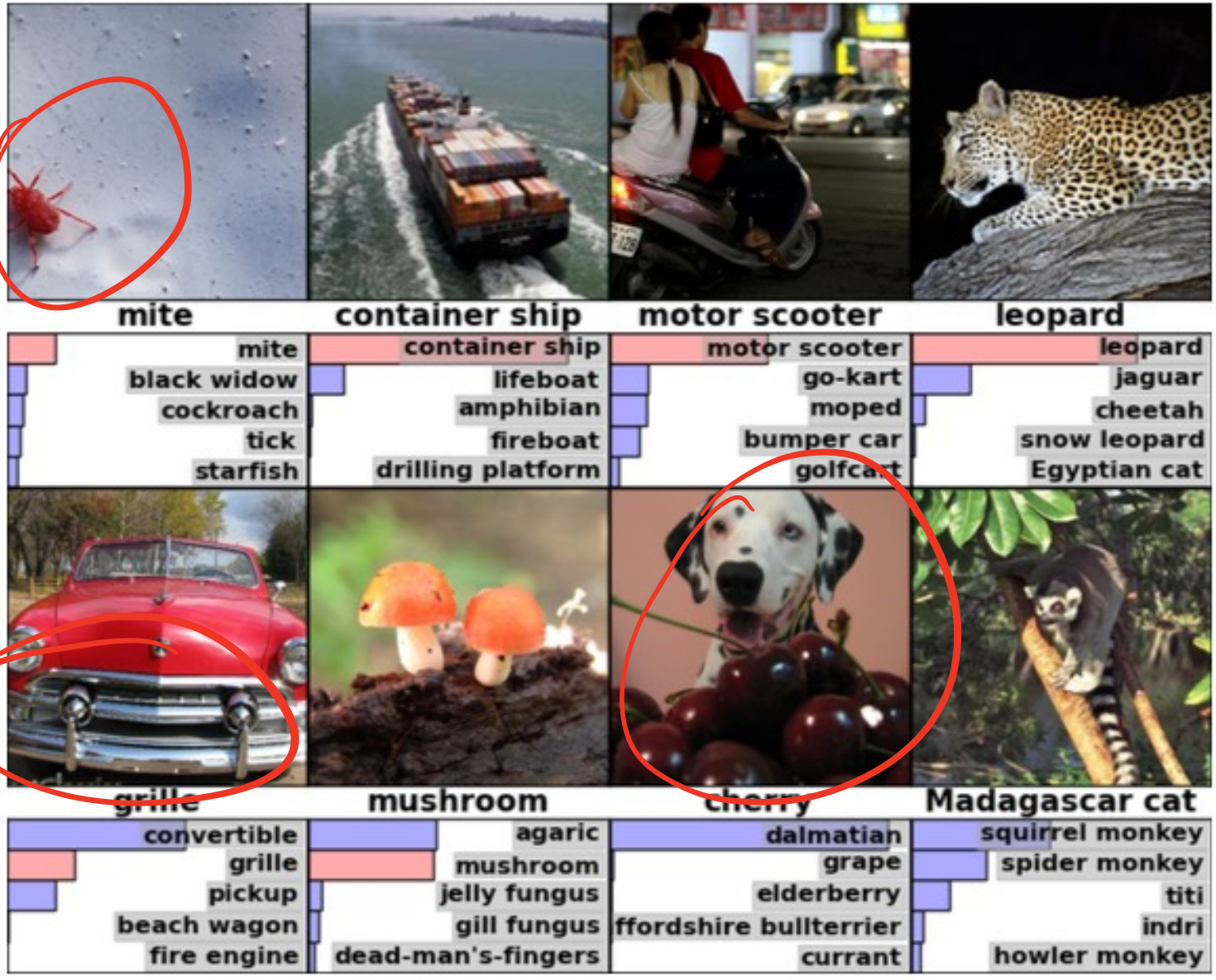




ImageNet Large Scale Visual Recognition Challenge (ILSVRC) benchmarks progress in object detection and image classification at large scale (2010-2017).

Image adapted from: https://semiengineering.com/new-vision-technologies-for-real-world-applications/

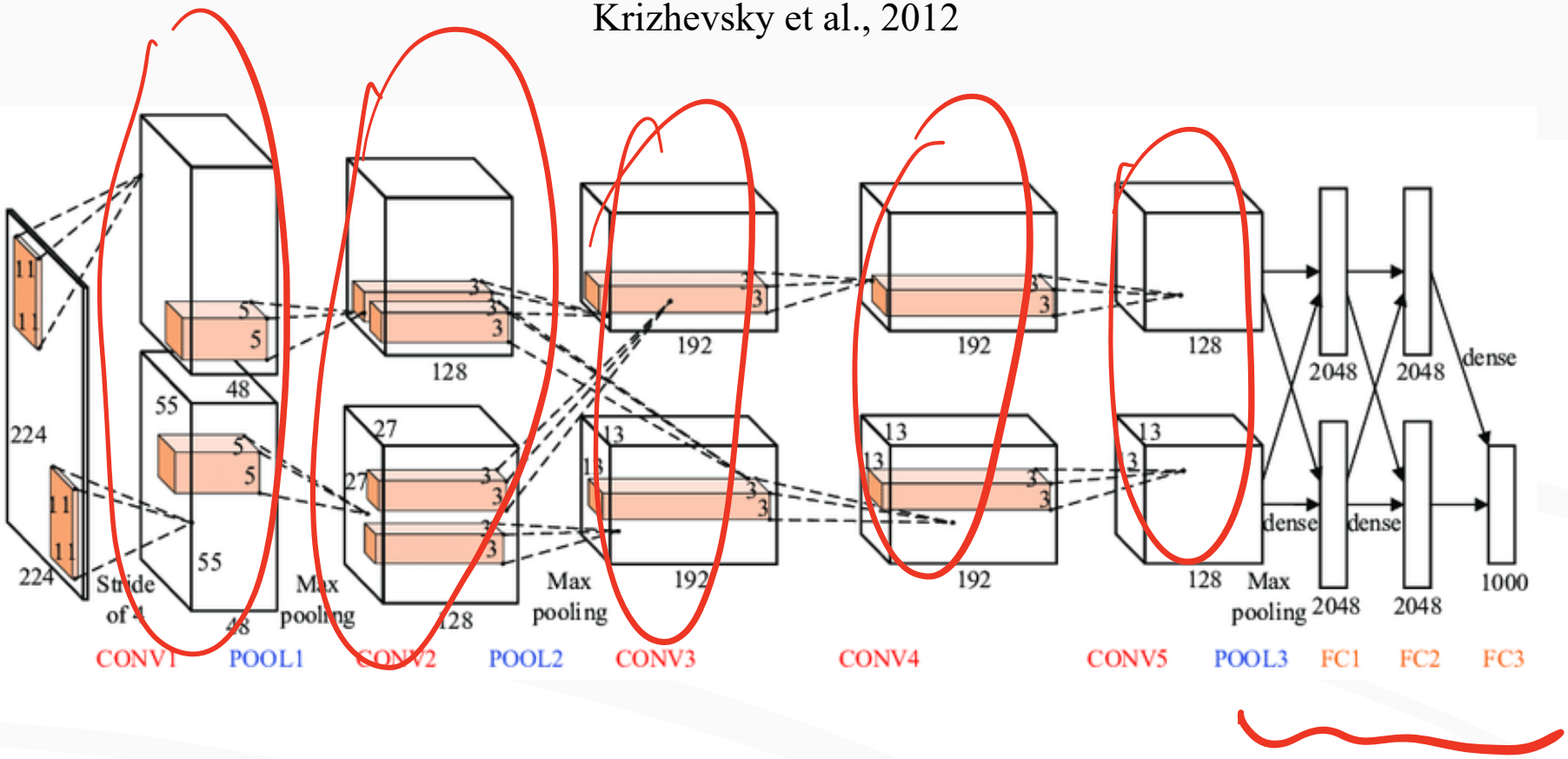
ILSVRC-2010 images



1000 classes

AlexNet

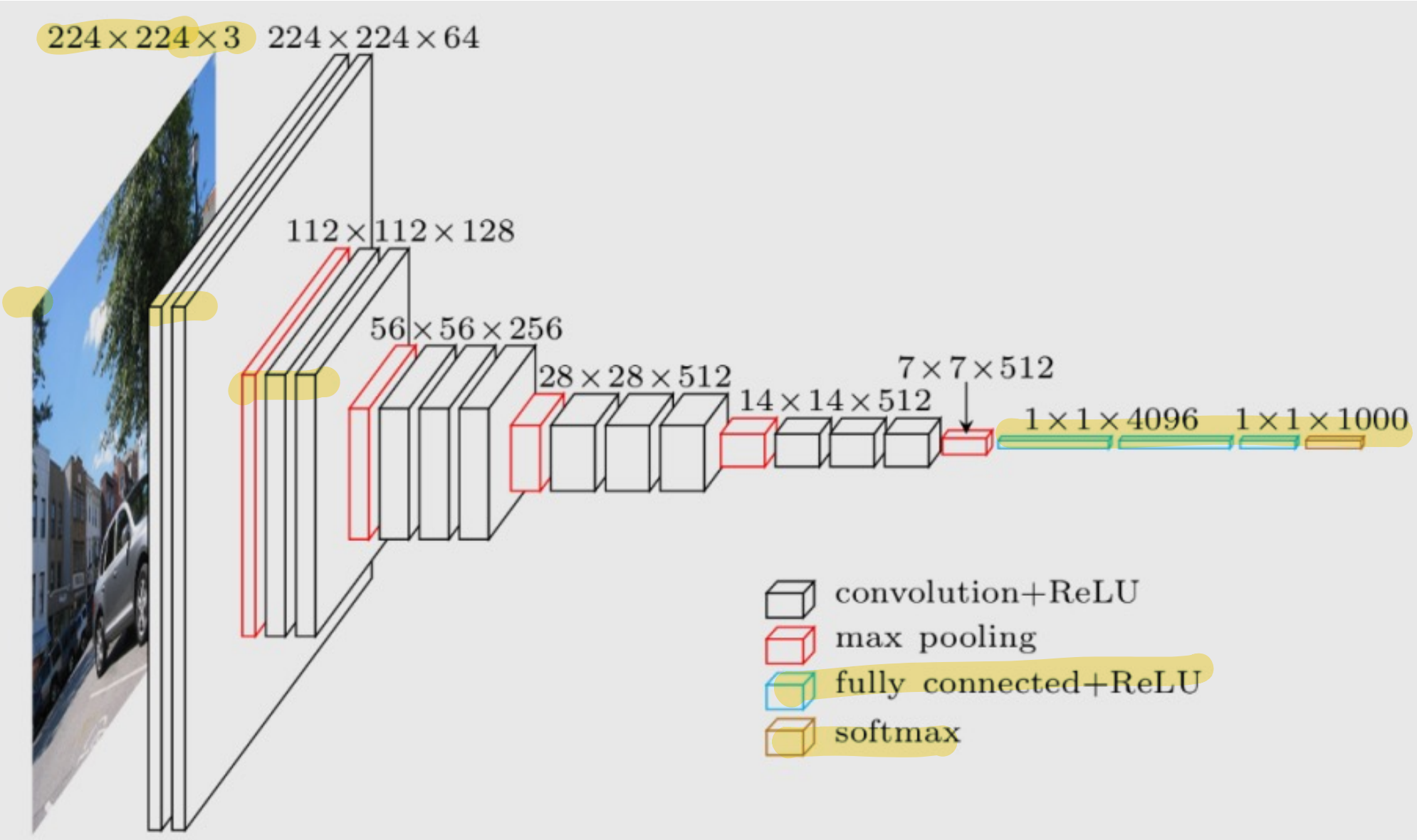
Krizhevsky et al., 2012



https://www.researchgate.net/figure/Network-architecture-of-AlexNet-CONV-convolutional-layer-FC-fully-connected-layer_fig1_331603323

VGG-16 Network

Simonyan & Zisserman, 2014



Shallow vs Deep Networks

Shallow neural networks: Typically 1-3 layers of weights

Deep neural networks: Many layers (often 100s of layers) of weights.

Underlying work since 1980s, but **new progress thanks to:**

- **Data**
- **Computing power (GPUs)**
- **Theoretical novelties**
  - ReLU
  - Dropout
  - Maxpool
  - BatchNorm
  - …

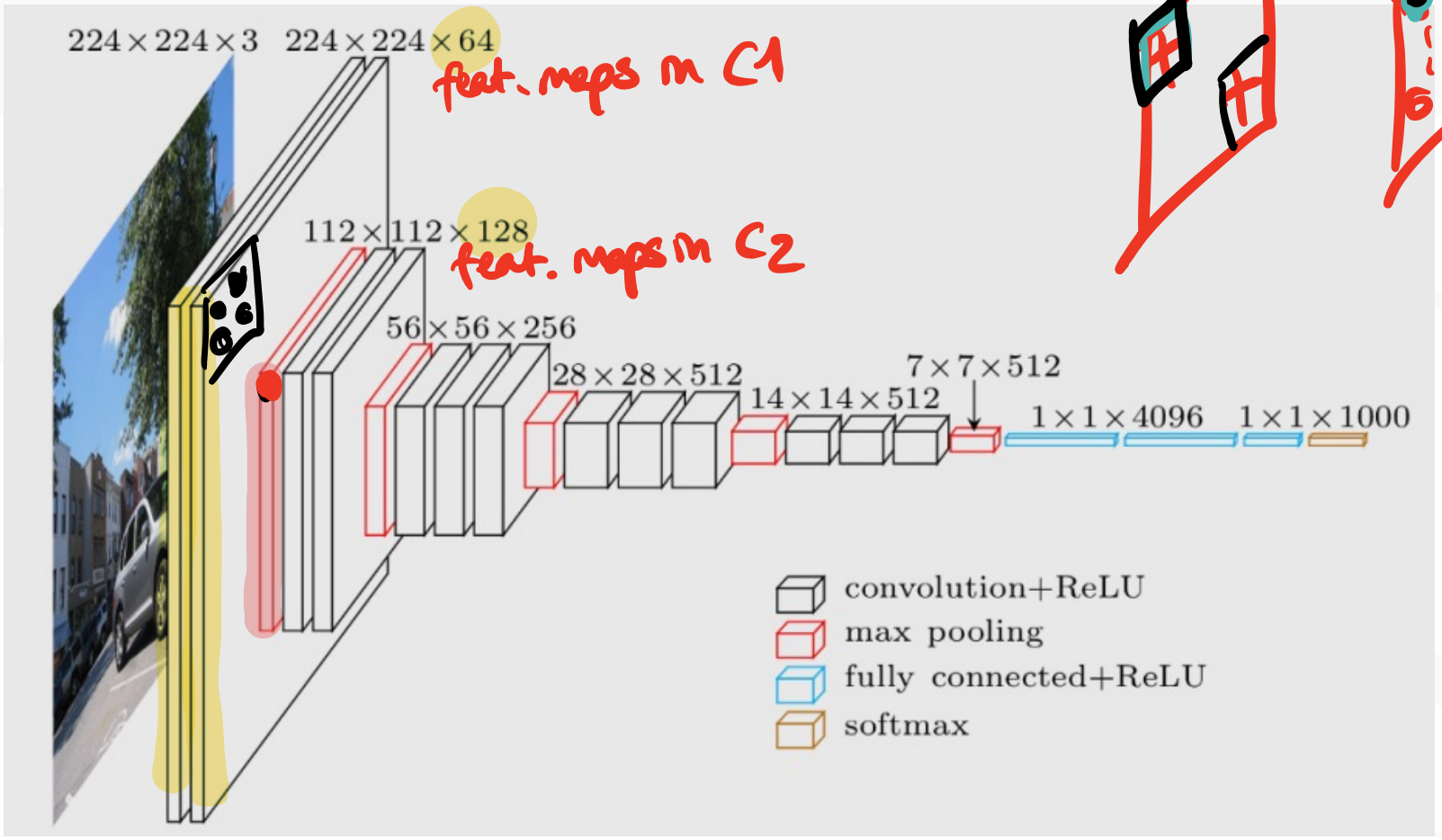
Image: http://cs231n.stanford.edu/

Convolutional Neural Networks

Receptive Fields, Convolution Operation Layers and Feature Maps

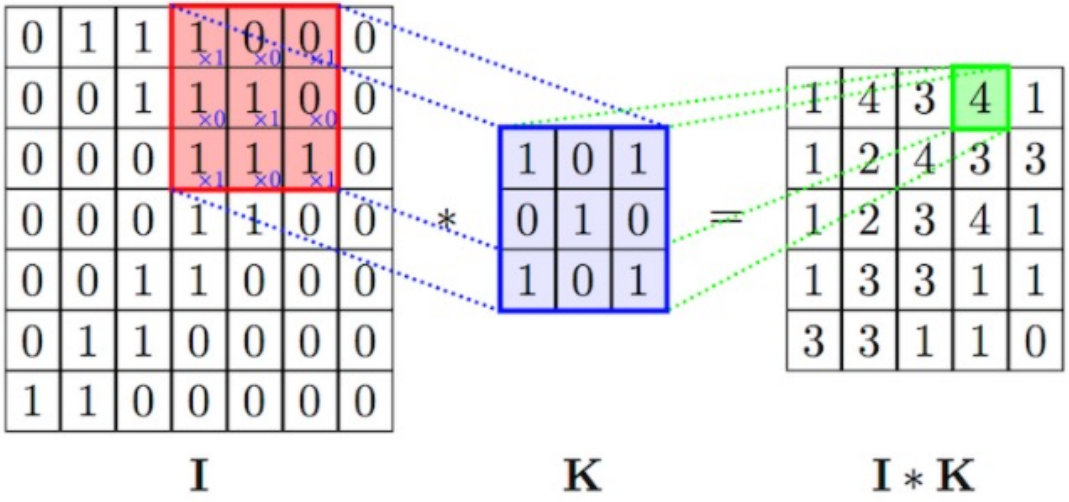
Convolutional Neural Networks (CNNs)

- Cells in a layer take input from small sub-regions of the visual field (receptive field).
- Nearby neurons are connected to nearby regions and respond to spatially local input patterns.
- Receptive fields are tiled to cover the entire visual field.



Convolution

$$(I * K)_{xy} = \sum_{i=1}^{h} \sum_{j=1}^{w} K_{ij} \cdot I_{x+i-1,y+j-1}$$

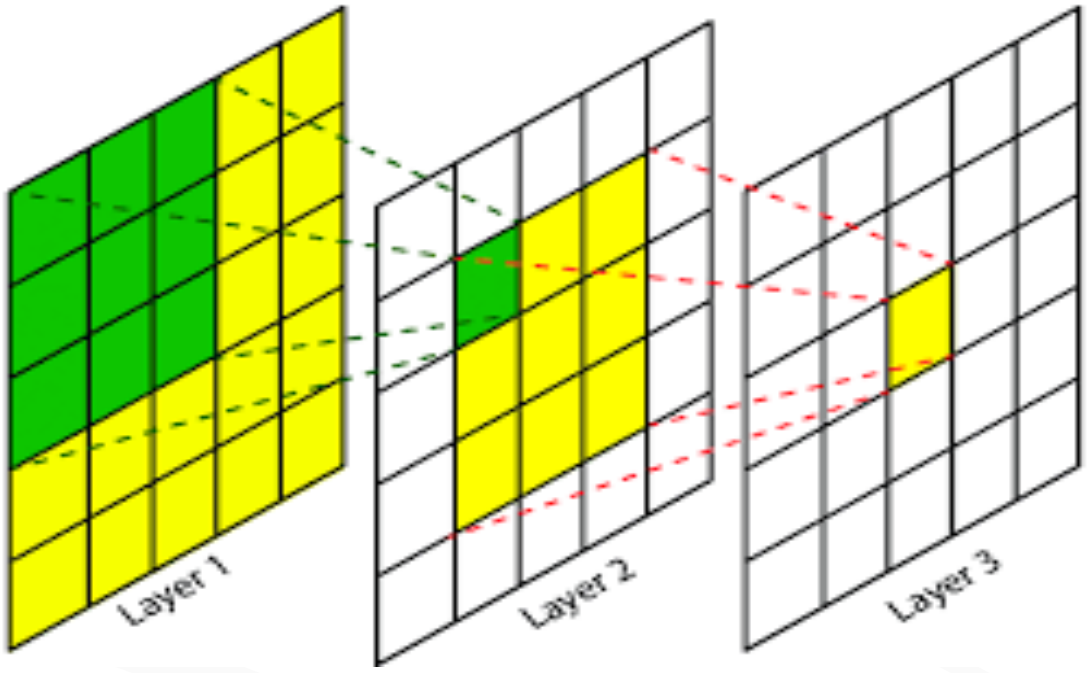


Output of this neuron is 4

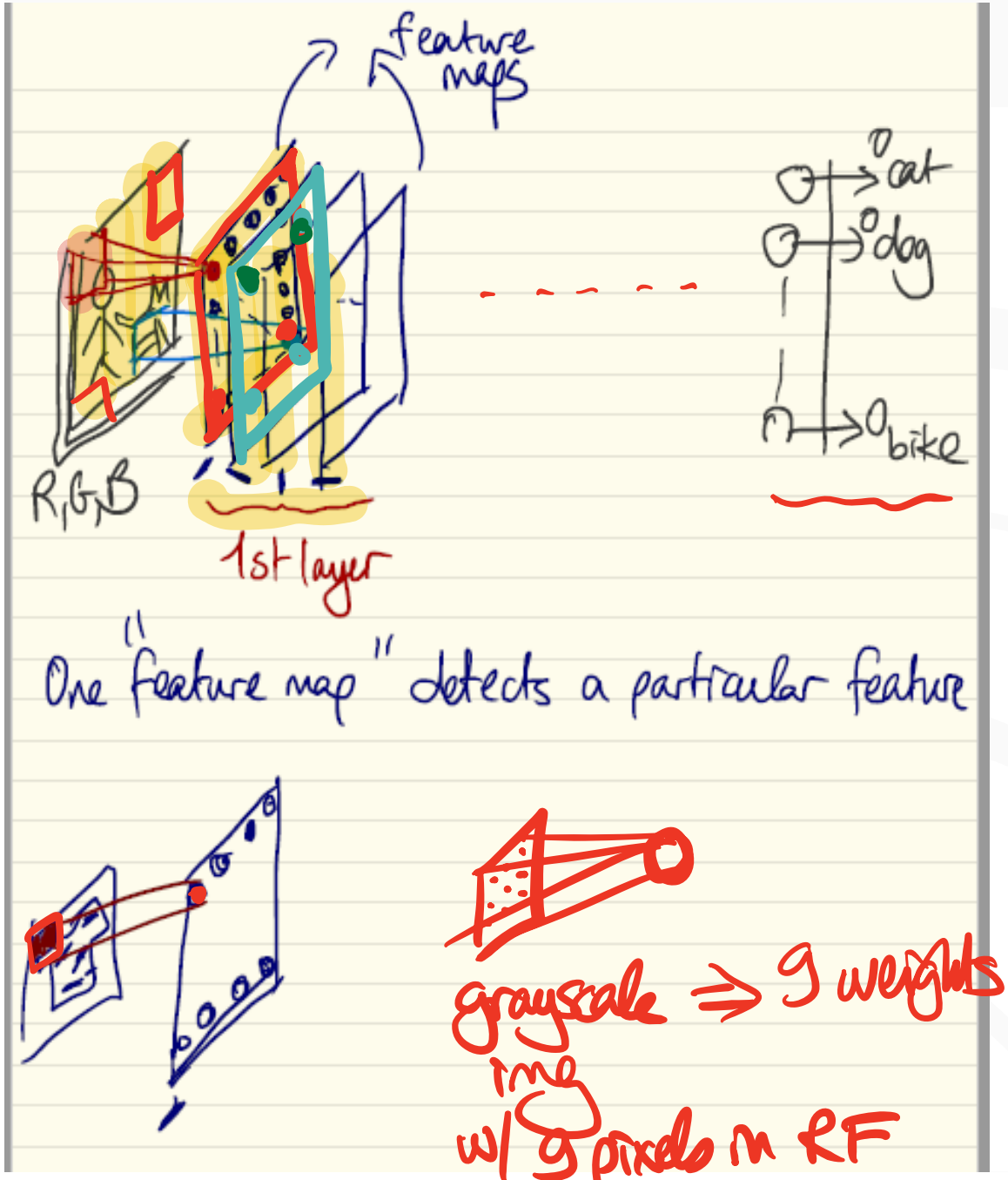
- Dot product between pixels in the receptive field (subregion of I) and the kernel (K).

Receptive Fields

- Each neuron computes its output by computing the dot product of its weight matrix and its receptive field.



$$(I * K)_{xy} = \sum_{i=1}^{h} \sum_{j=1}^{w} K_{ij} \cdot I_{x+i-1,y+j-1}$$



- Each filter is replicated across the entire visual field (image).

- These replicated units share the same weights and form a feature map.

- Replicating units in this way allows for features to be detected regardless of their position in the visual field.

- Additionally, weight sharing greatly reduces the number of free parameters being learnt

- Each neuron that share the same position (in different feature maps) take input from the same cells in the previous layer.