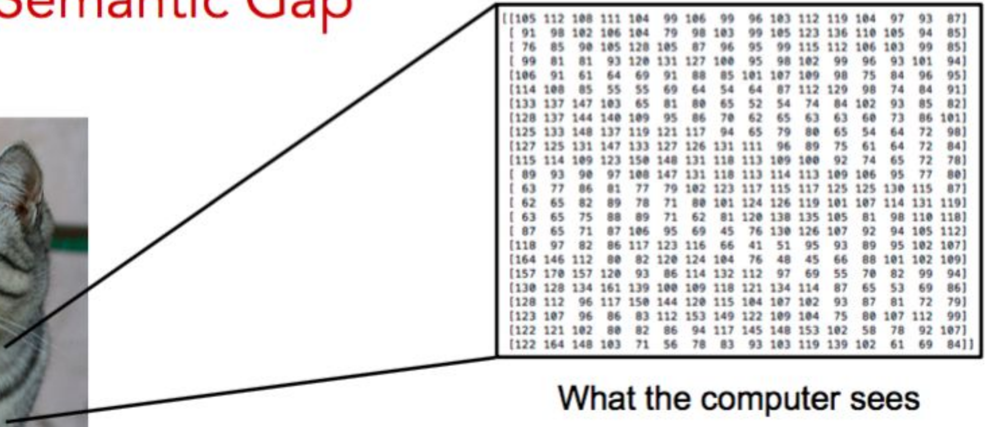
# Deep Learning for Images

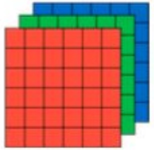
## The Problem: Semantic Gap



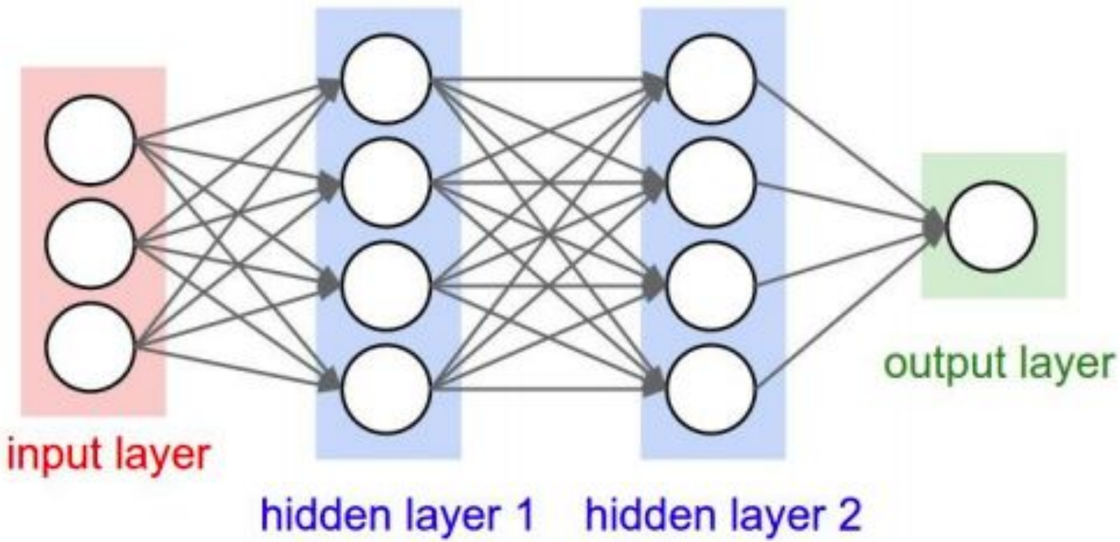

What the computer sees

An image is just a big grid of numbers between [0, 255]:

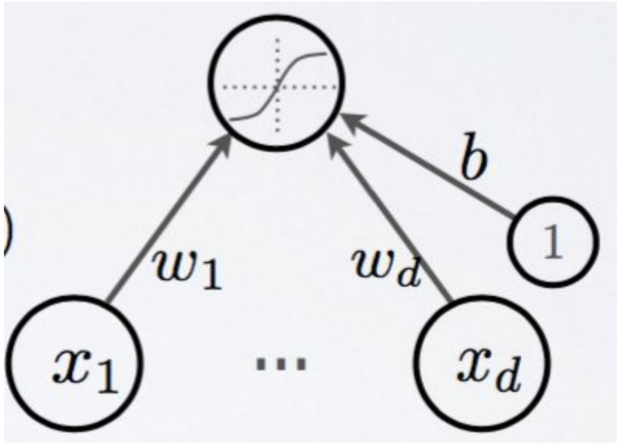
Example:
800 x 600 x 3
(3 channels RGB)



h
x
w
x
3




“3-layer Neural Net”, or
“2-hidden-layer Neural Net”



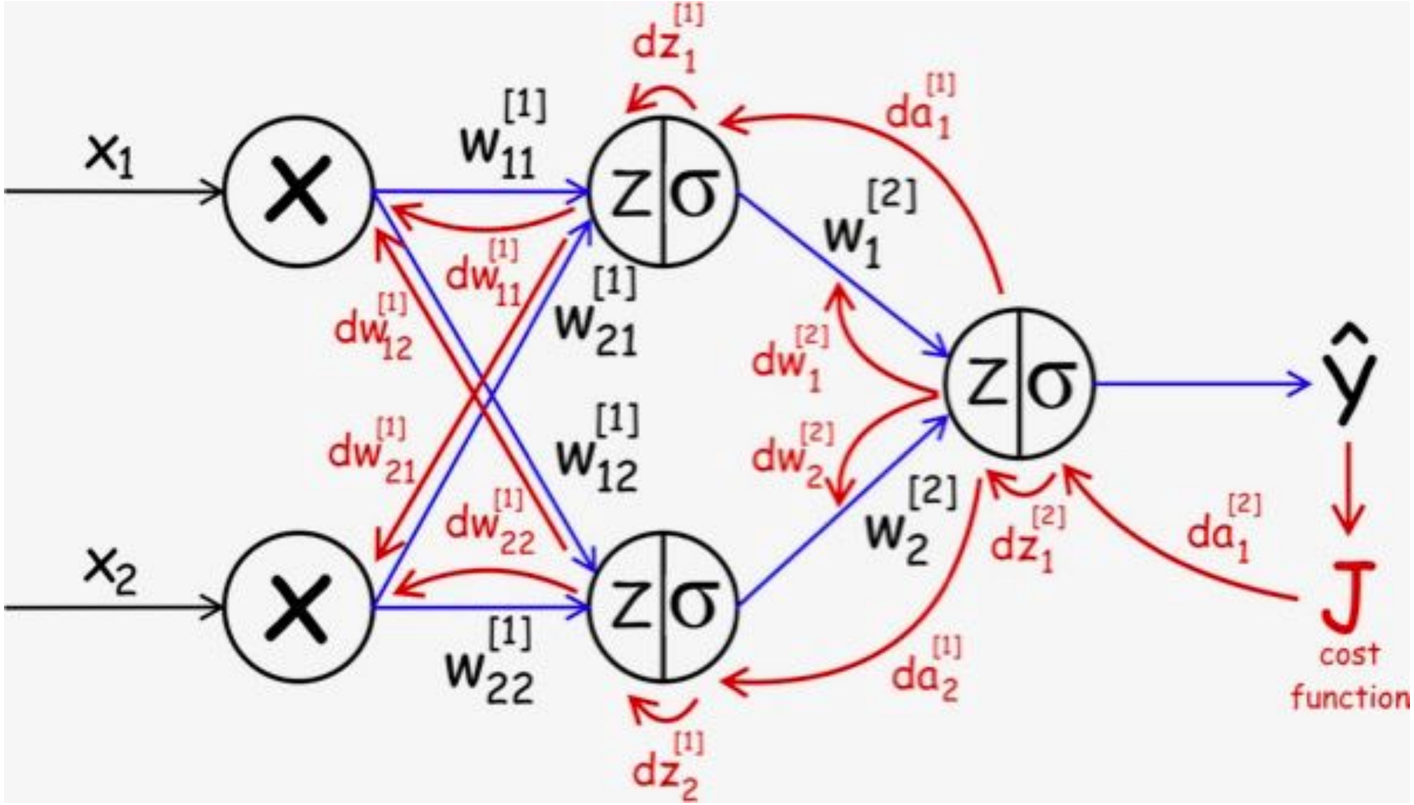

$$a(\mathbf{x}) = b + \sum_i w_i x_i$$

Suppose: 3 training examples, 3 classes.
With some W the scores $f(x,W) = Wx$ are:

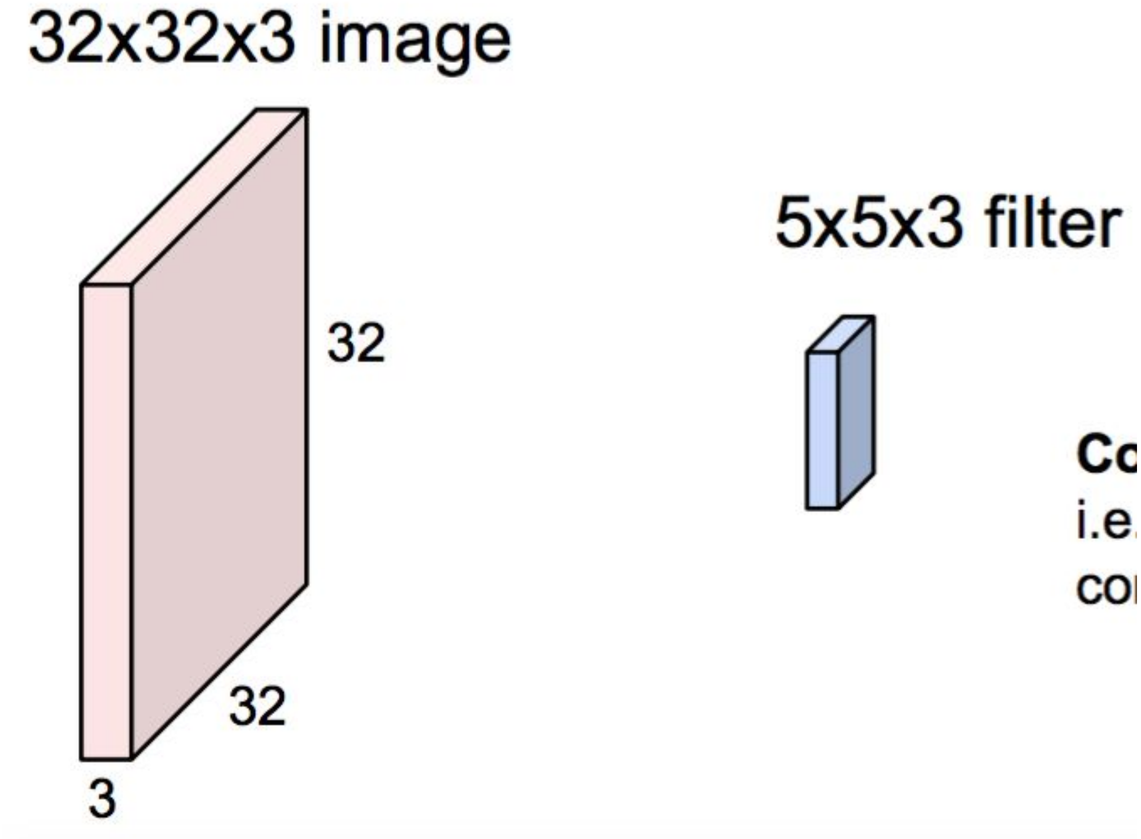


| | | | |
|---|---|---|---|
| cat | **3.2** | 1.3 | 2.2 |
| car | 5.1 | **4.9** | 2.5 |
| frog | -1.7 | 2.0 | **-3.1** |
| Losses: | 2.9 | 0 | 12.9 |

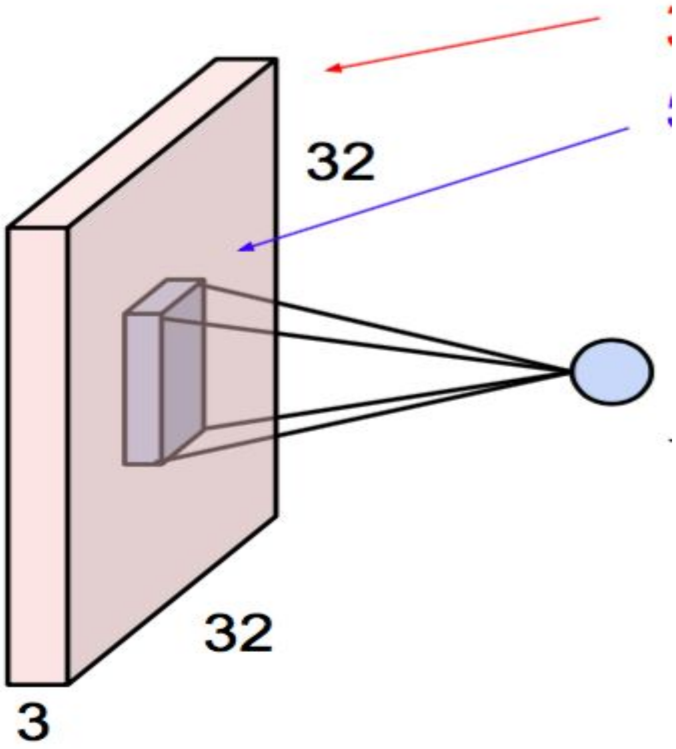
# Backpropagation Algorithm



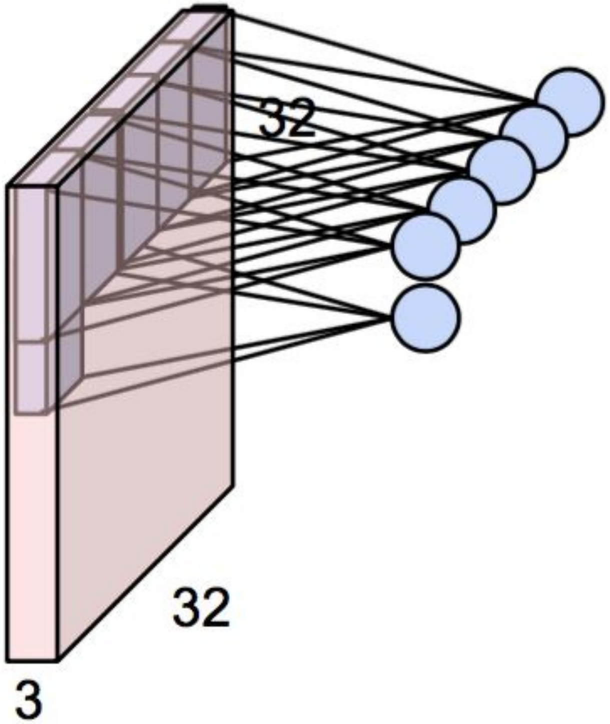
# Convolution

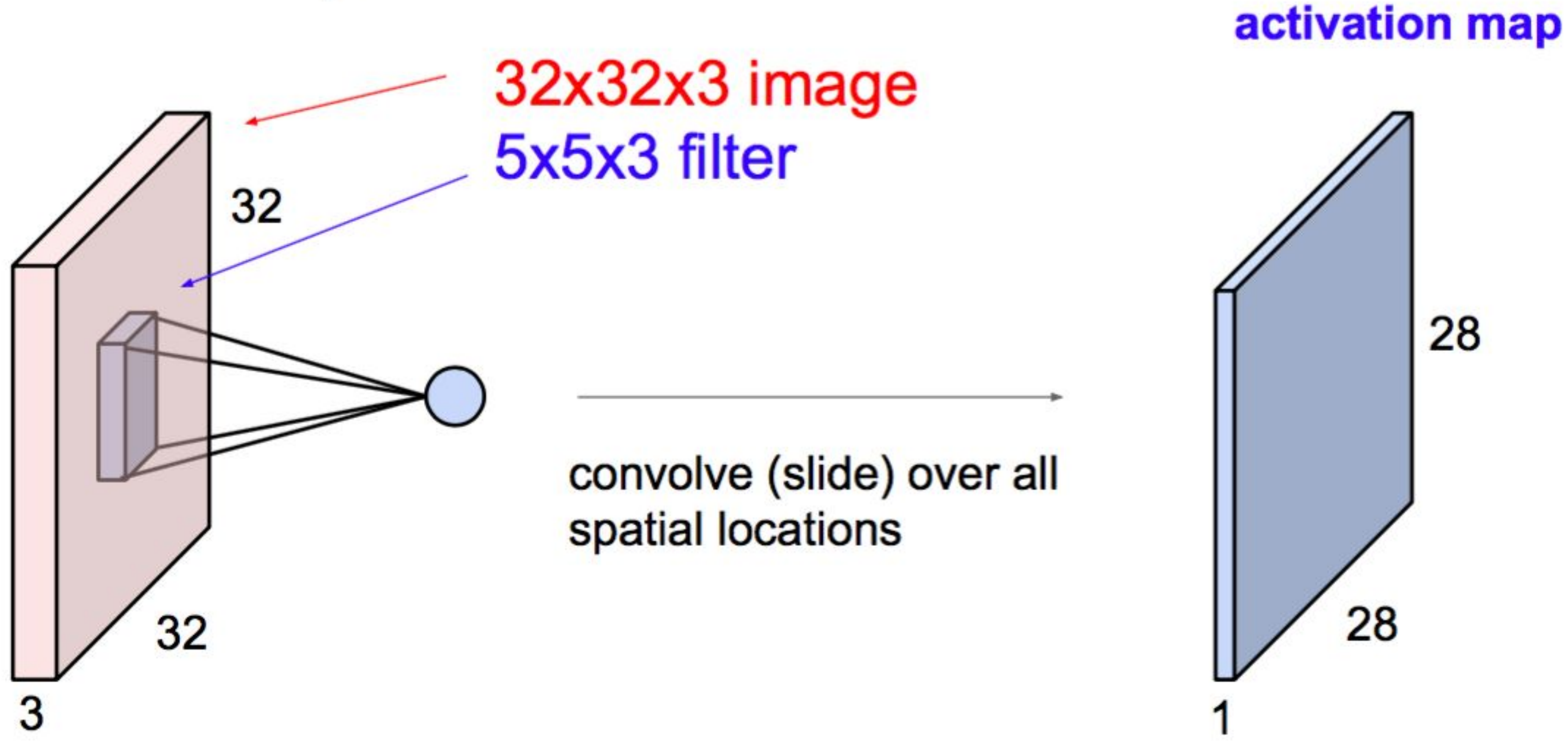


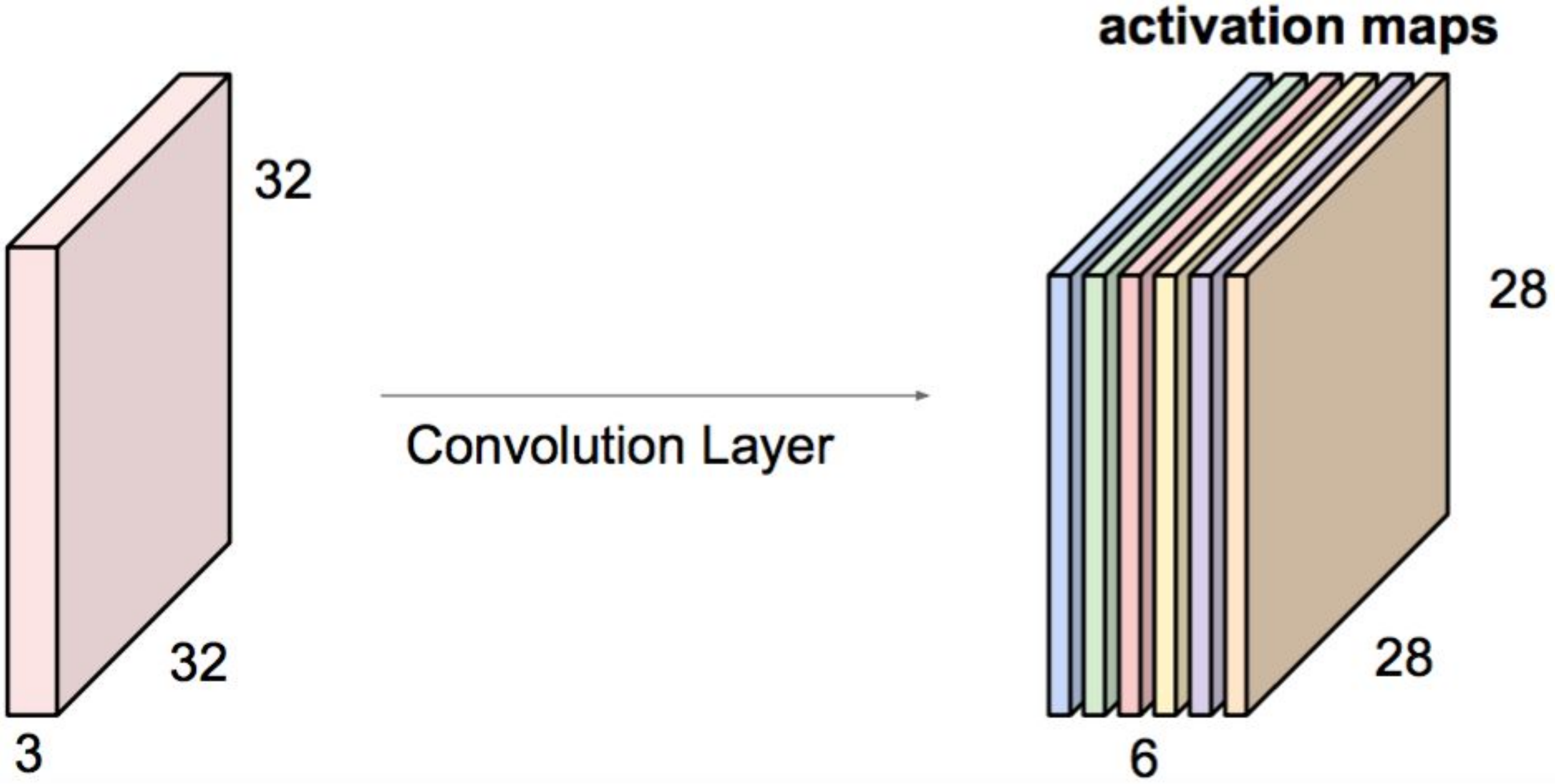
# Convolution Layer

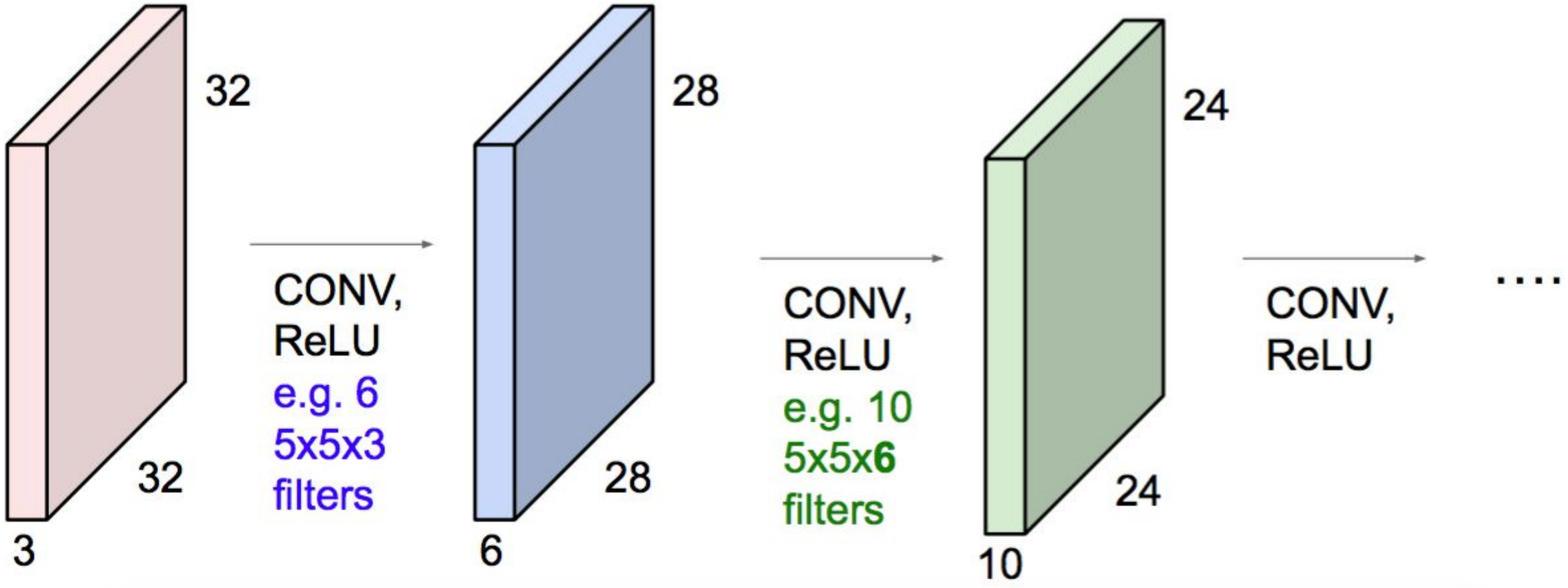


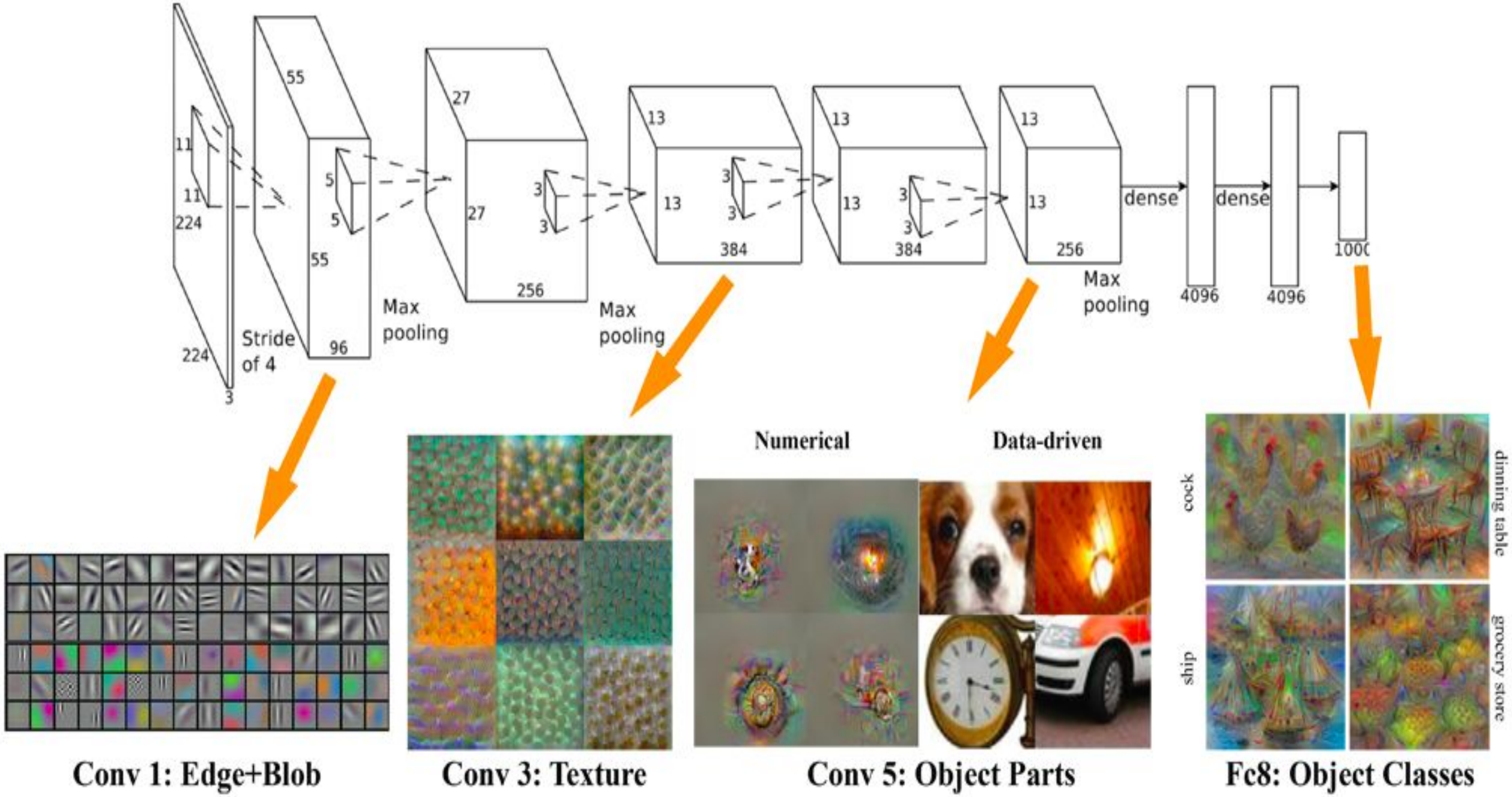

$$w^T x + b$$




activation map
32x32x3 image
5x5x3 filter
32
32
3
convolve (slide) over all spatial locations
28
28
1


activation maps
32
32
3
Convolution Layer
28
28
6


32
32
3
CONV,
ReLU
e.g. 6
5x5x3
filters
28
28
6
CONV,
ReLU
e.g. 10
5x5x**6**
filters
24
24
10
CONV,
ReLU
....

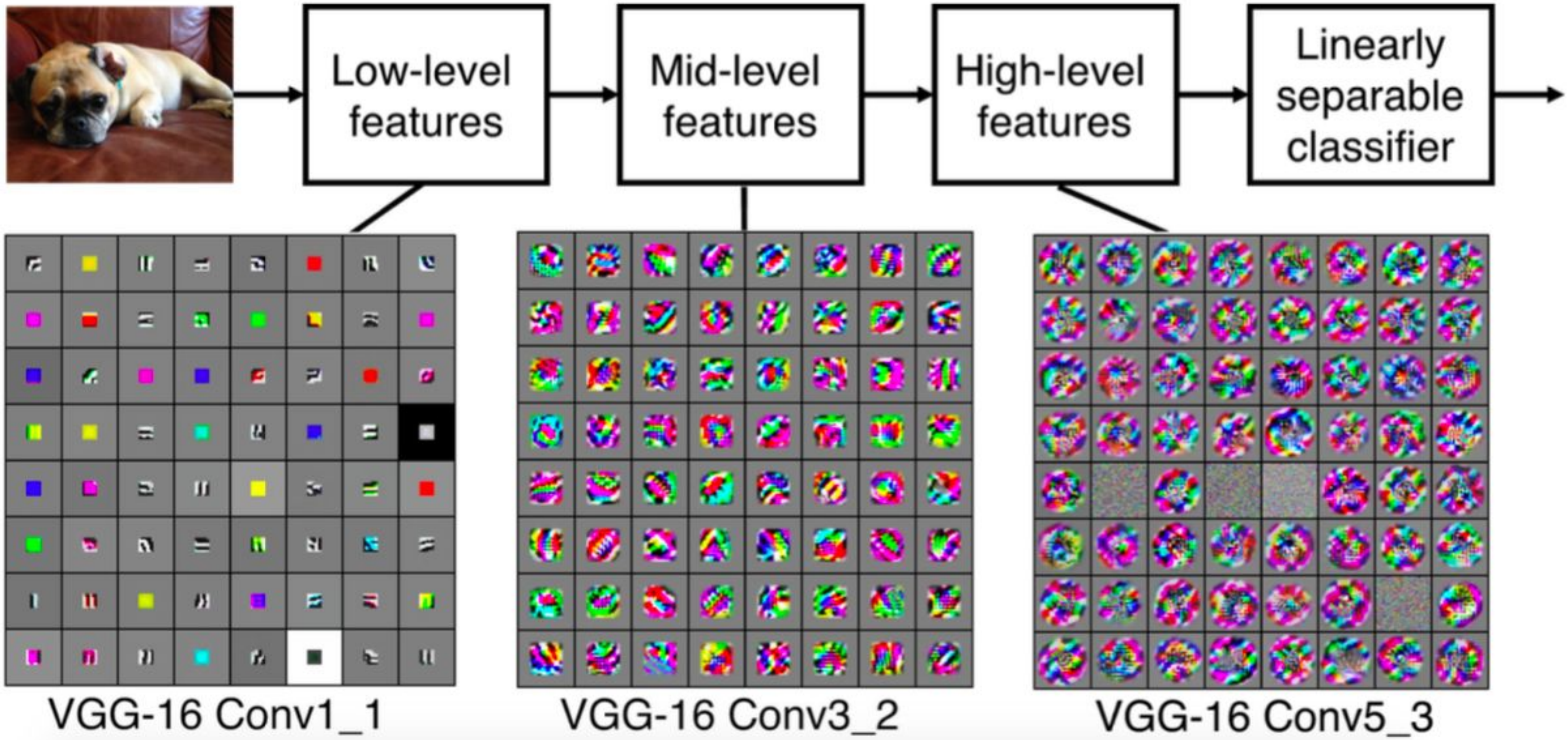



AlexNet / VGG-F network visualized by **mNeuron**.

# Preview

*[Zeiler and Fergus 2013]*

Visualization of VGG-16 by Lane McIntosh. VGG-16 architecture from [Simonyan and Zisserman 2014].

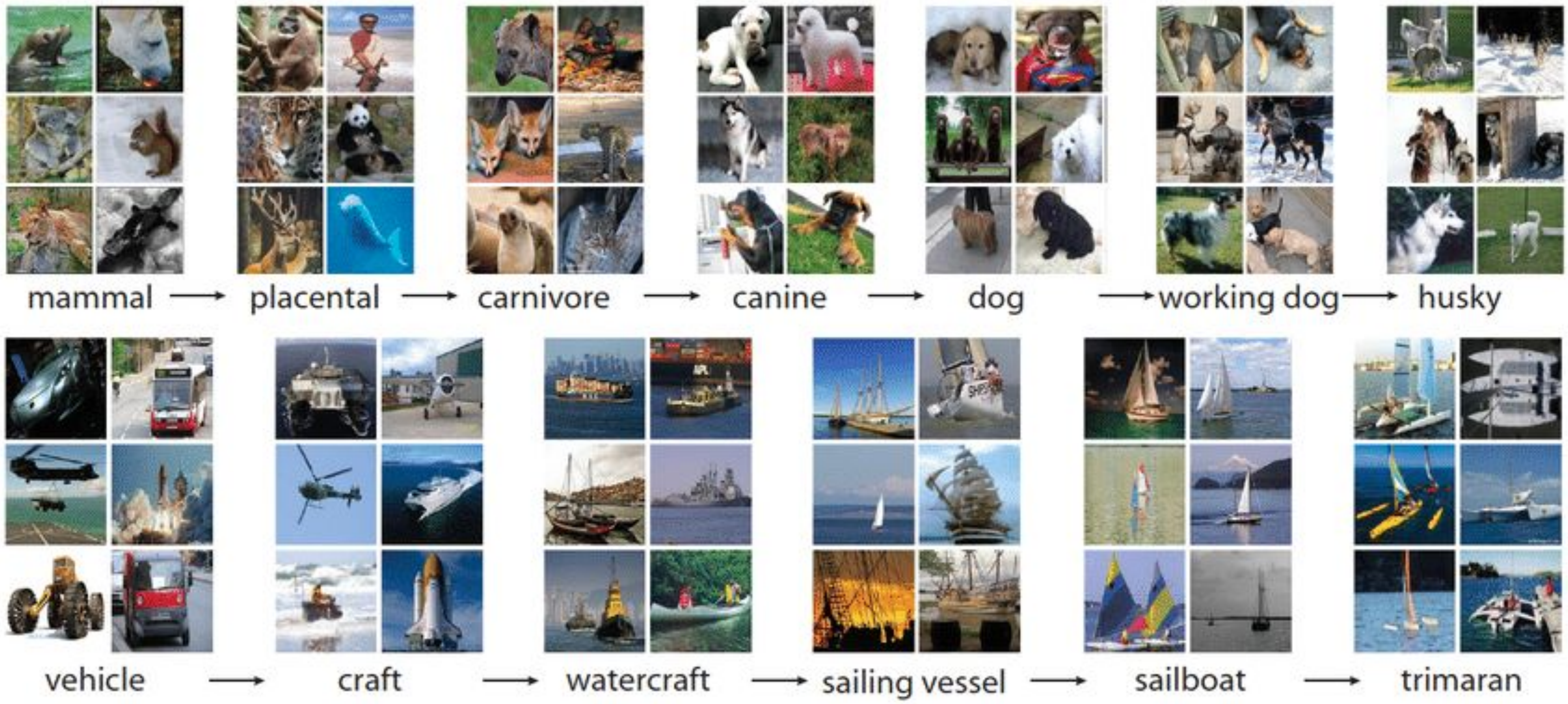


# Deep Learning Consists of two characteristics

1. End to End Training due to Backpropagation Algorithm.
2. Hierarchical Learning of Features.

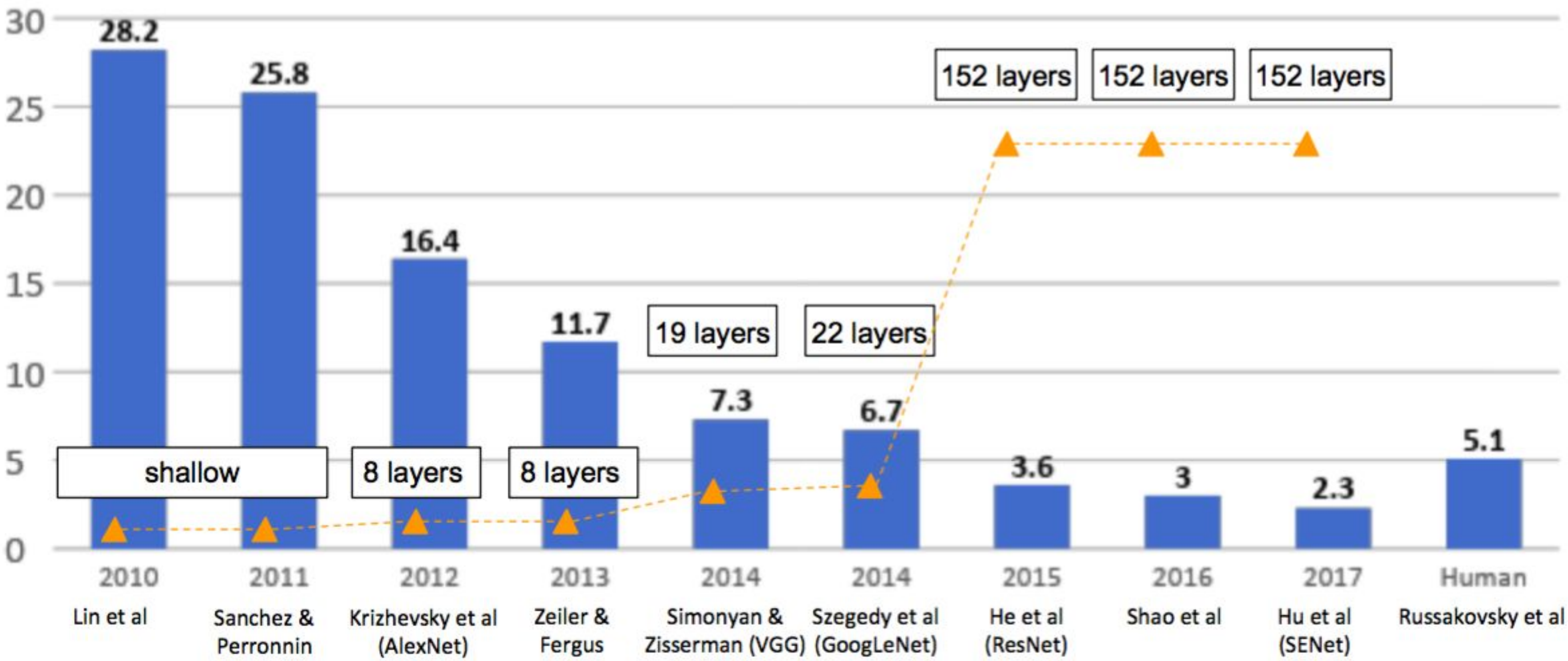
Deep Learning become popular recently due to three reasons :-

1. More Data
2. More Computational Power due to GPU(Parallel Processing)
3. Better Algorithms

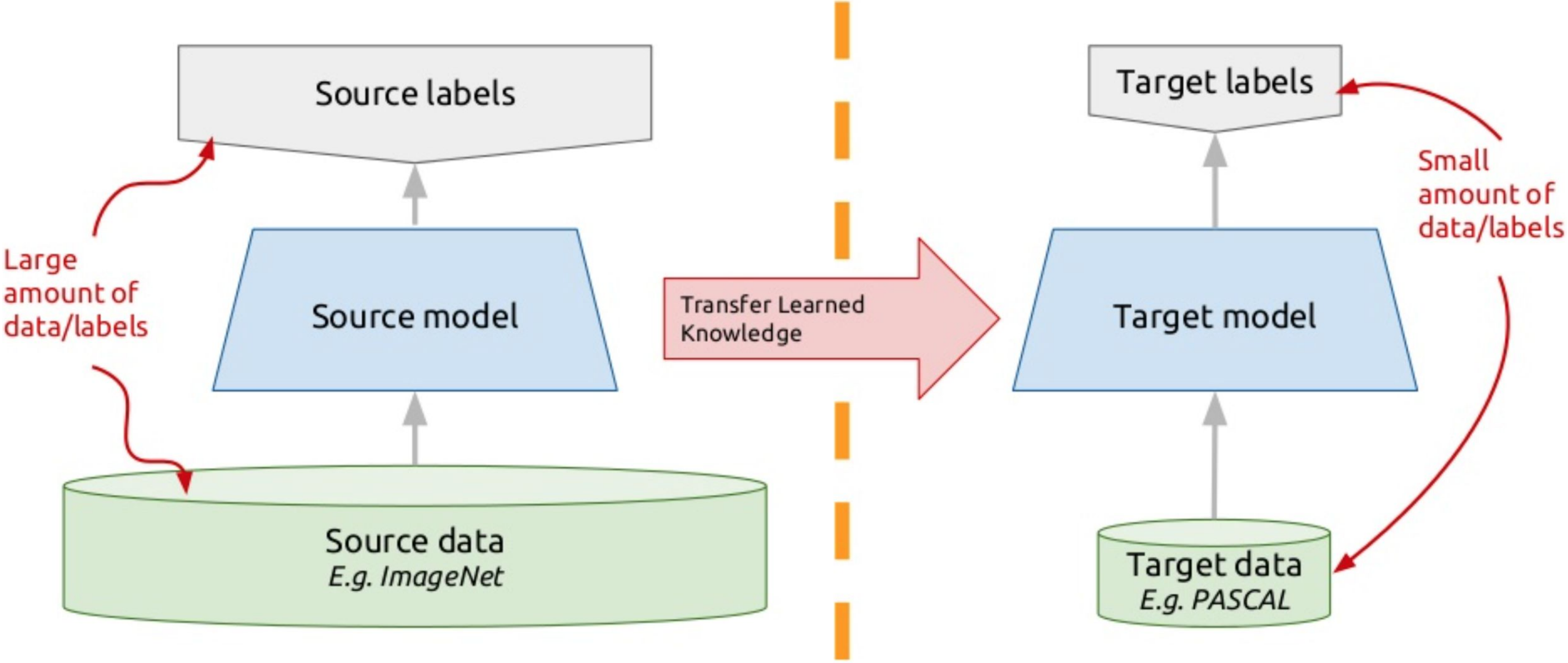



Imagenet Challenge(Dataset)- 1.2 million images and 1000 categories

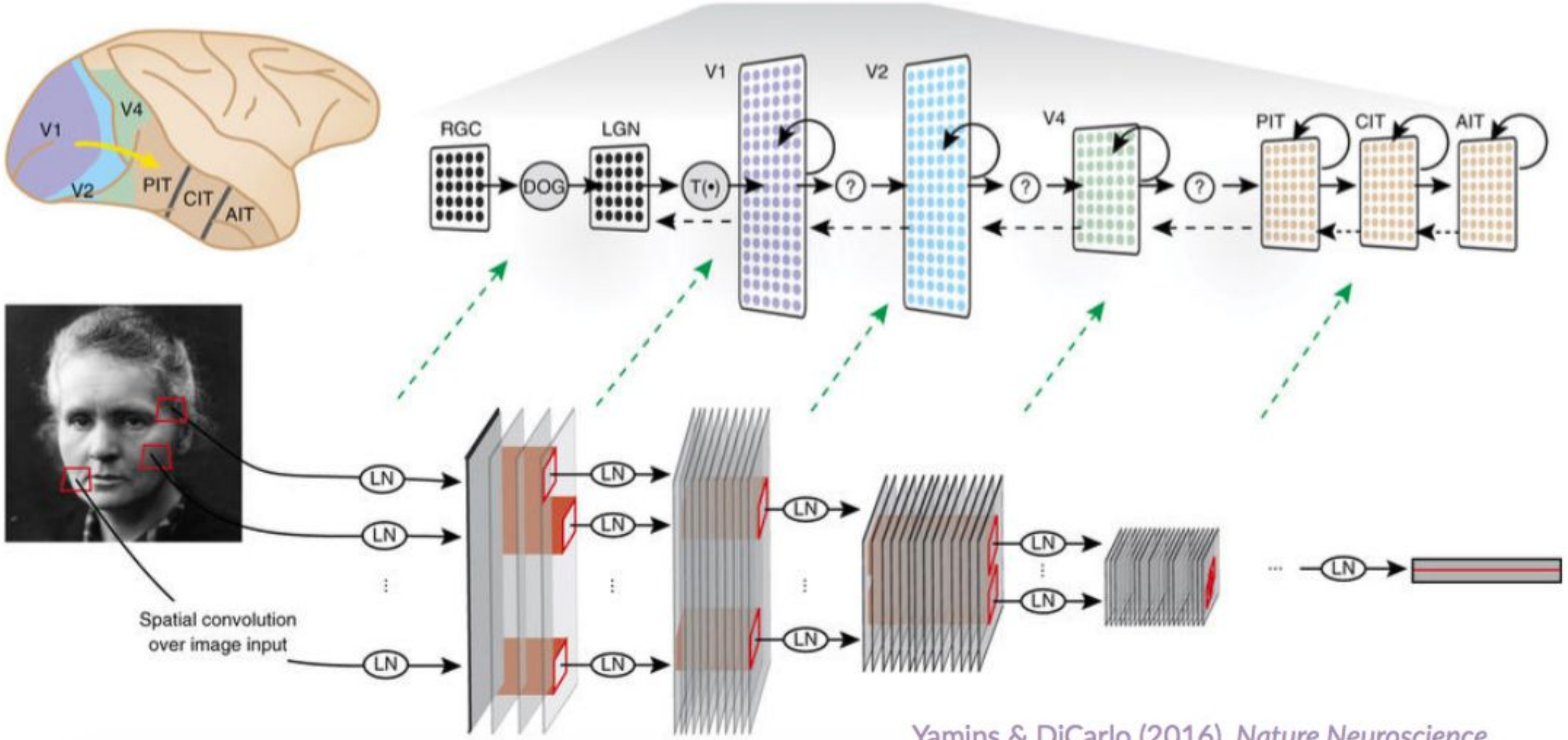
# ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



# Transfer learning: idea



# At an abstract level, deep neural networks operate with some similar principals to the real brain (though there are some important differences!)




Yamins & DiCarlo (2016), *Nature Neuroscience*

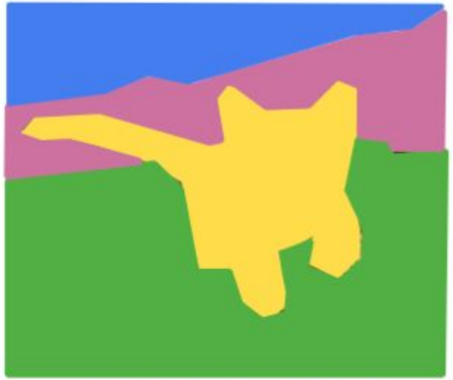
# Computer Vision Tasks

**Classification**

**CAT**

No spatial extent

**Semantic Segmentation**



**GRASS, CAT, TREE, SKY**

No objects, just pixels

**Object Detection**

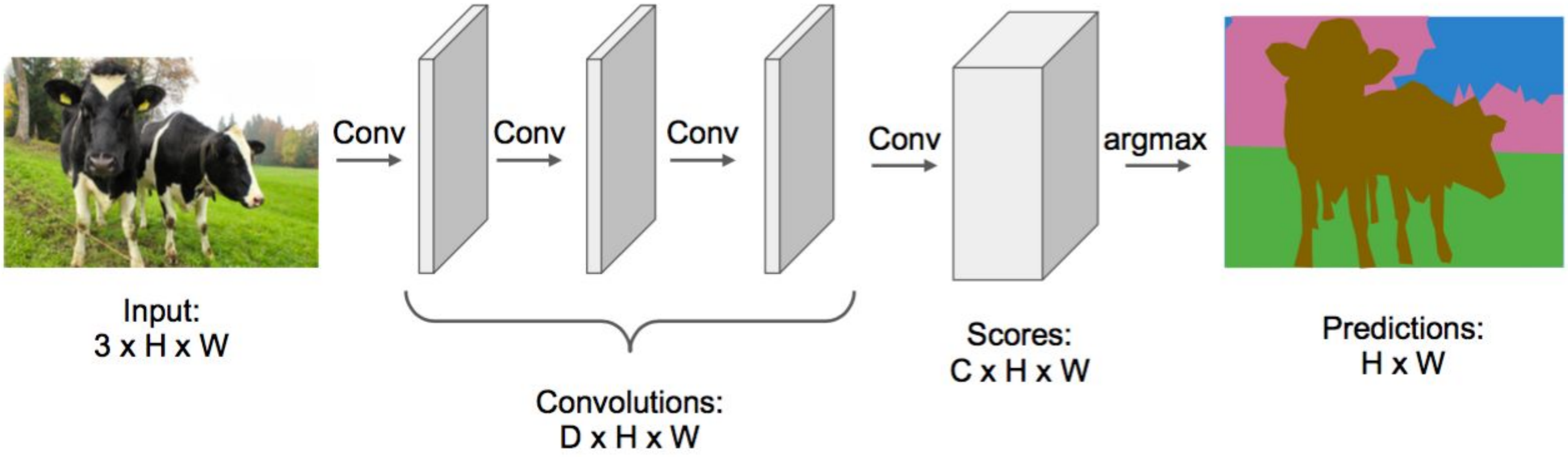
**DOG, DOG, CAT**

**Instance Segmentation**

**DOG, DOG, CAT**

Multiple Object

# Image Segmentation

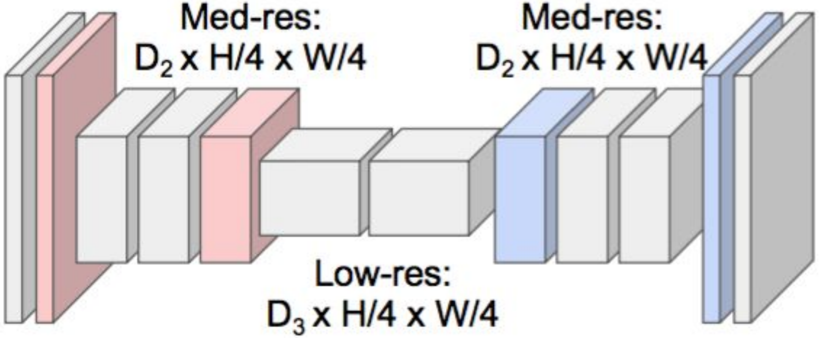


**Downsampling**:
Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

**Upsampling**:
???

Input:
3 x H x W



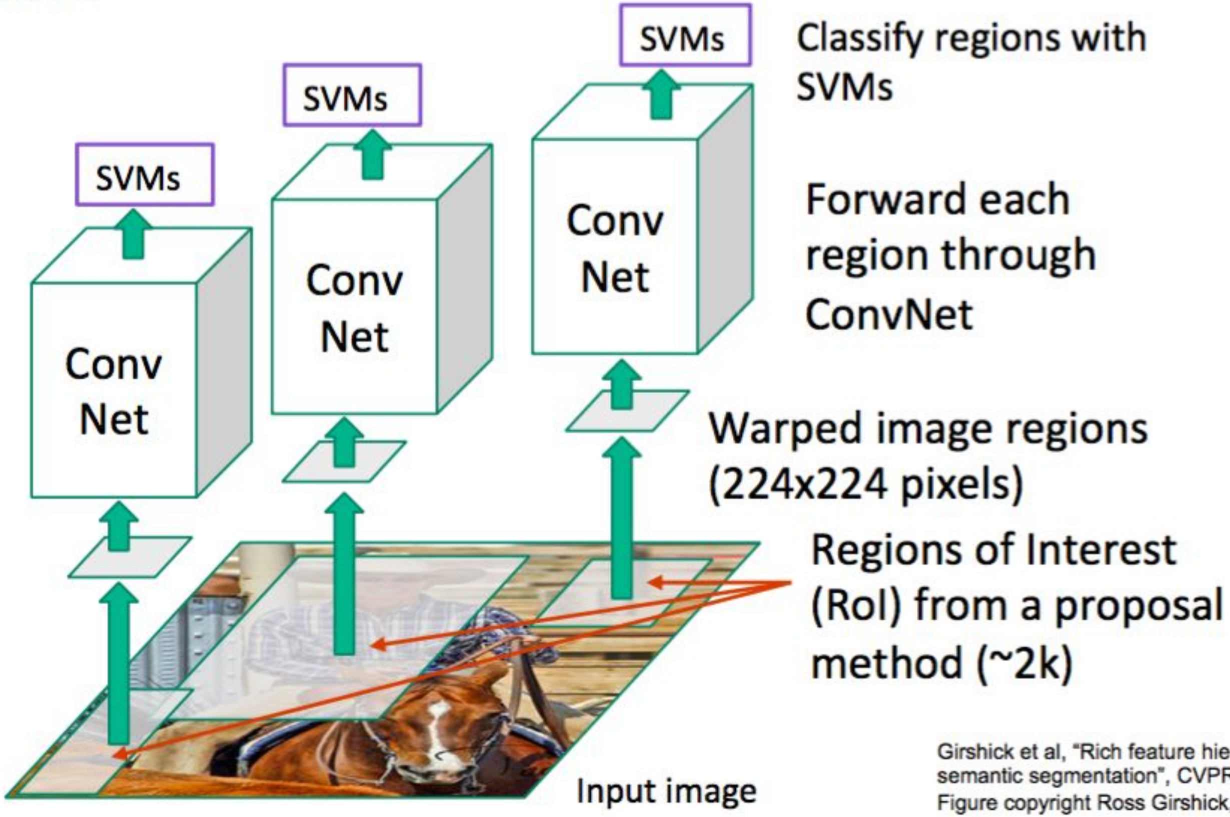

High-res:
$D_1$ x H/2 x W/2

High-res:
$D_1$ x H/2 x W/2

Predictions:
H x W

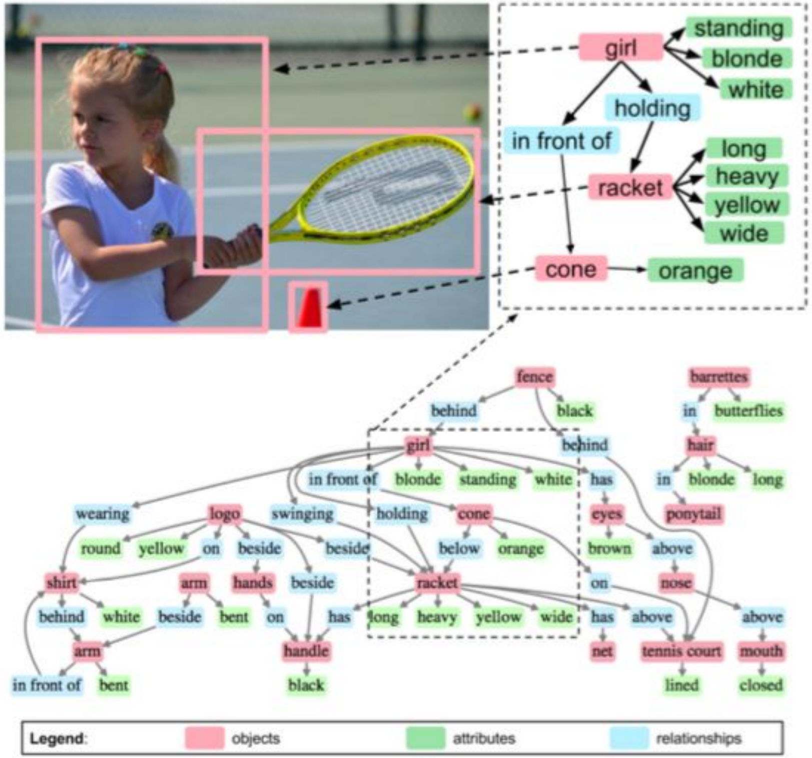
# Object Detection

# R-CNN




Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

# Objects + Relationships = Scene Graphs



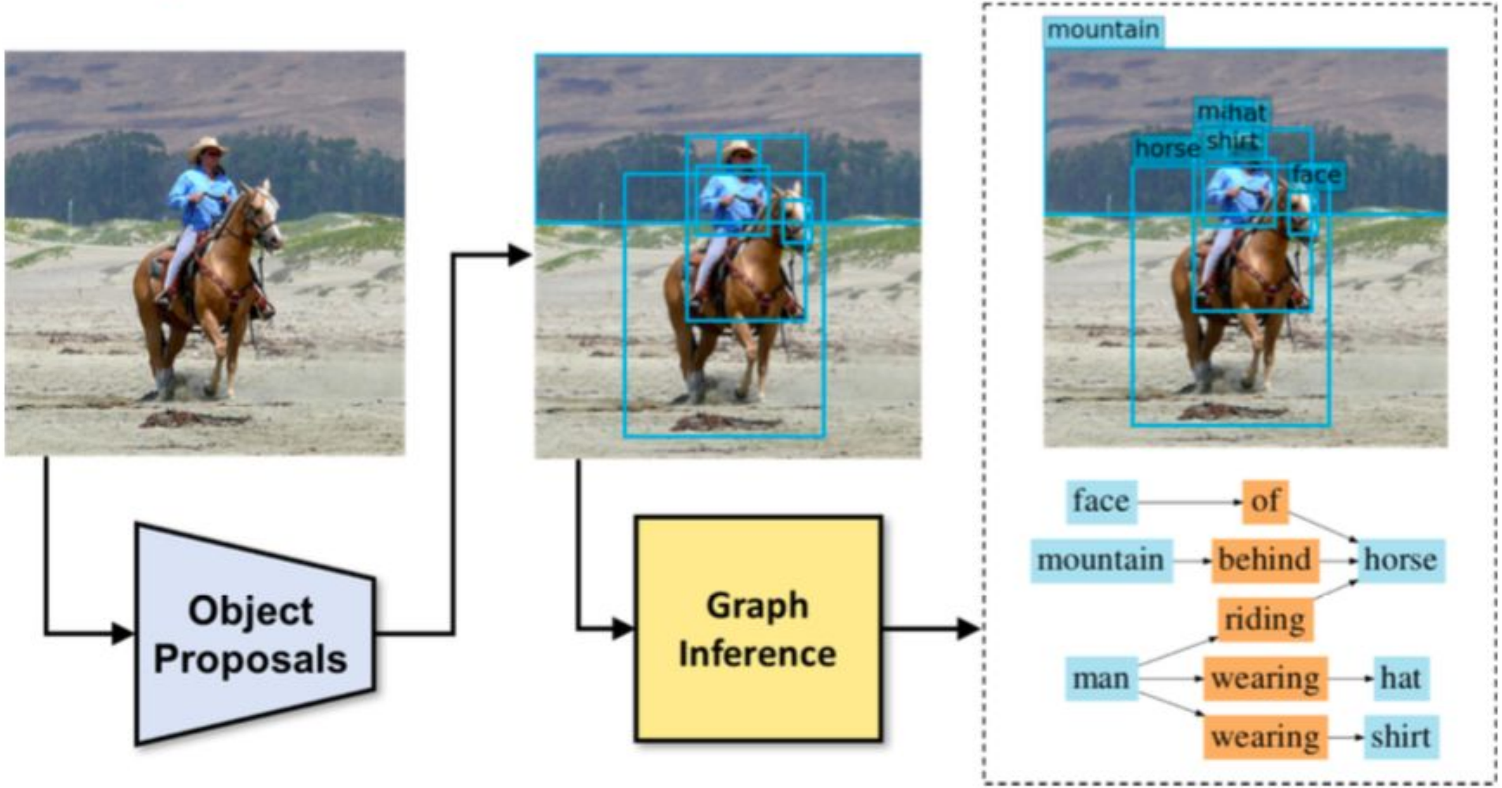

108,077 Images
5.4 Million Region Descriptions
1.7 Million Visual Question Answers
3.8 Million Object Instances
2.8 Million Attributes
2.3 Million Relationships
Everything Mapped to Wordnet Synsets

VISUALGENOME

Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International Journal of Computer Vision 123, no. 1 (2017): 32-73.

Slide from Stanford CS231n [Johnson, Yeung, and Fei-Fei]

# Scene Graph Prediction



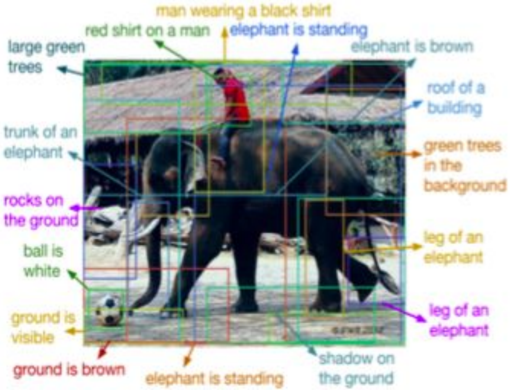
# Language and vision

## Captioning

"man in black shirt is playing guitar."

Karpathy and Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions", CVPR 2015

## Dense Captioning




Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning", CVPR 2016

## Referring Expressions

***largest elephant standing behind baby elephant.***

Zhang, Liu, and Chang, "Grounding Referring Expressions in Images by Variational Context", CVPR 2018

# Visual Question Answering (VQA)

**VQA**


"VQA: Visual Question Answering"
[Agrawal et al, ICCV 2015]

- Understanding of visual input
- Understanding of language
- World knowledge
- Reasoning

- Can ask about anything
- Easier to evaluate (at least for multiple choice)

# Visual Question Answering (VQA)

## VQA

Who is wearing glasses?

man woman

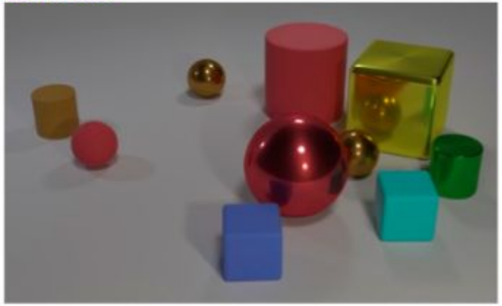
Is the umbrella upside down?

yes no

"Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering"
[Goyal et al, CVPR 2017]

## CLEVR

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.



**Q:** Are there an **equal number** of **large things** and **metal spheres**?
**Q:** **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
**Q:** There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
**Q:** **How many** objects are **either** **small cylinders** or **red** things?

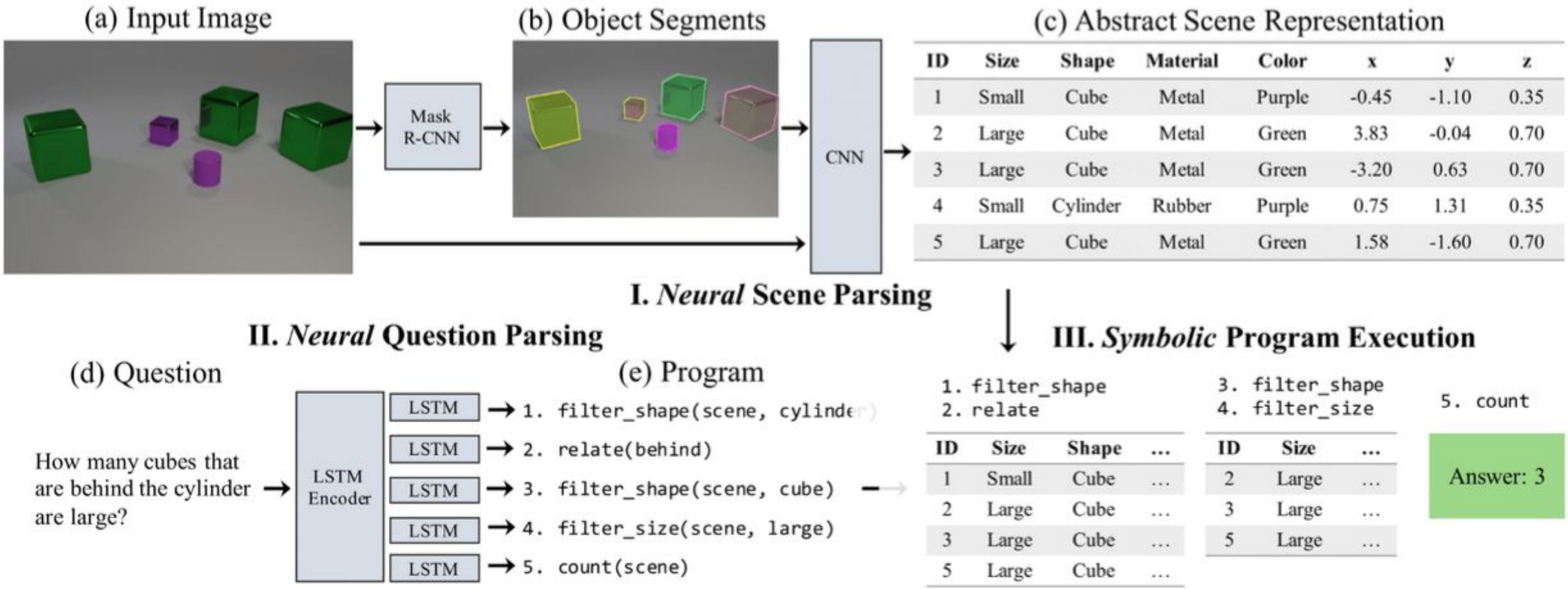
"CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning"
[Johnson et al, CVPR 2017]

## GQA


*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

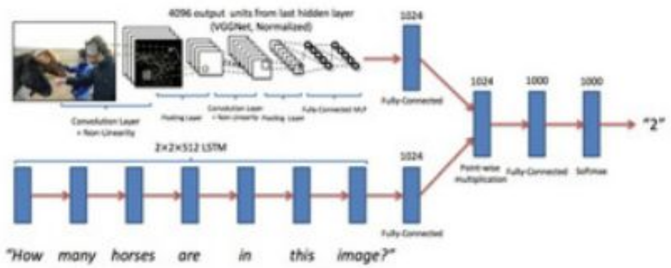
"GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering"
[Hudson and Manning, CVPR 2019]

# Reasoning




"Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding"
[Yi, Wu, Gan, Torralba, Kohli, and Tennebaum, NeurIPS 2018]

# Task- and dataset-specific models




Visual Question Answering [Antol et. al. 2015]

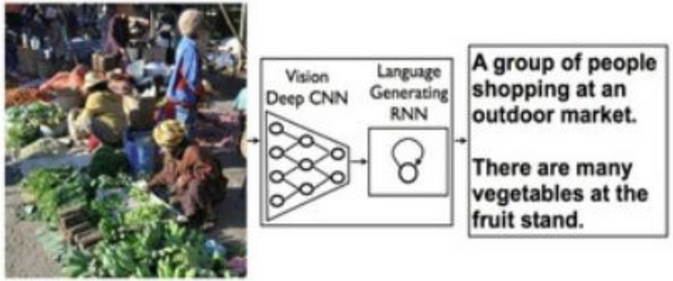
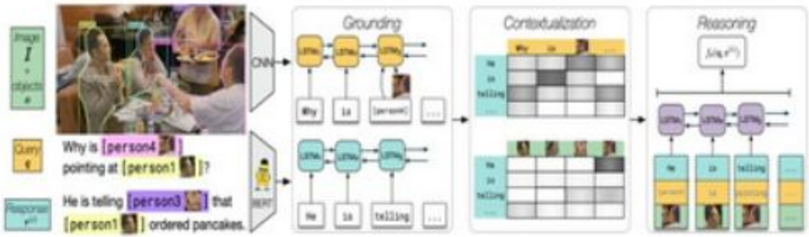
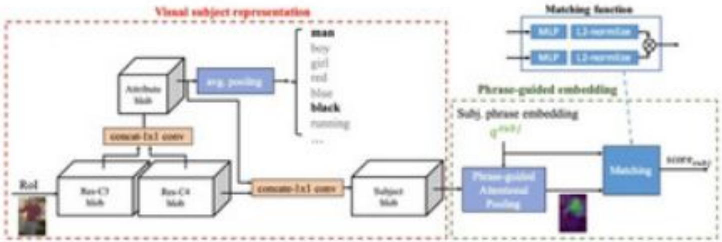



Image Captioning [Vinyals et. al. 2015]

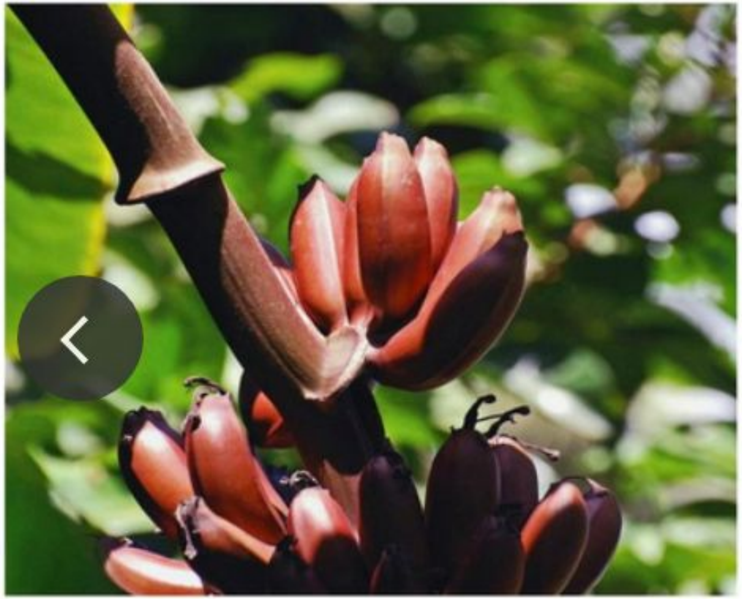



Visual Commonsense Reasoning [Zellers et. al. 2018]




Refer Expression [Yu et. al 2018]

# Task- and dataset-specific models



VQA model:
Q: What type of plant is this?
A: Banana

Captioning model:
A bunch of red and yellow flowers on a branch.

Common model for visual grounding
Leverage for a variety of vision-and-language tasks

# ViLBERT Multi-Task [CVPR 2020]

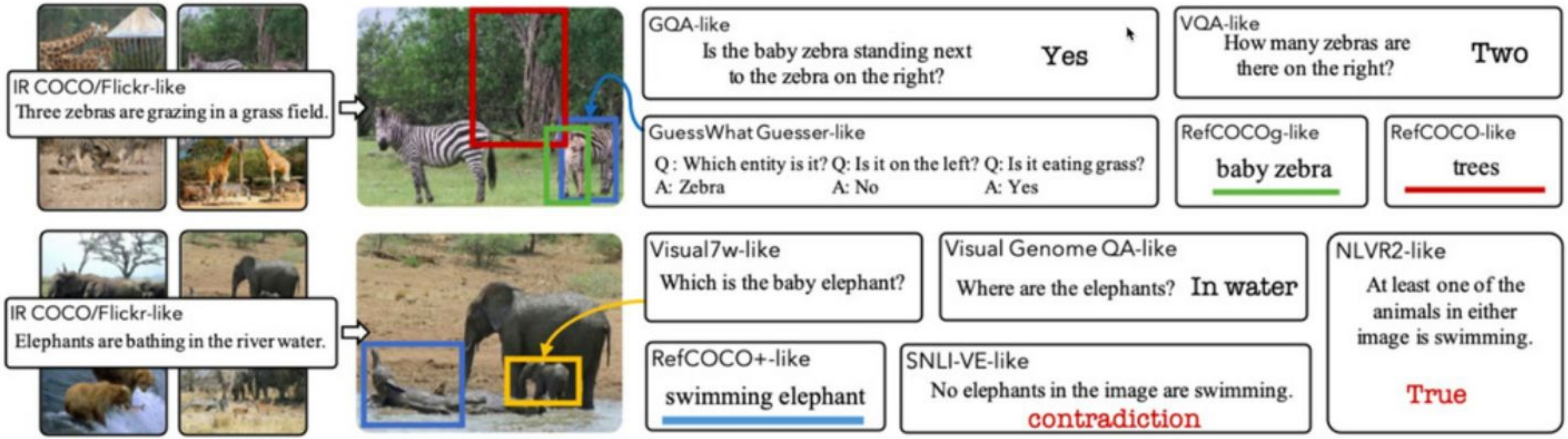
Jiasen Lu

Vedanuj Goswami

1 model for 12 tasks!

Higher performance, 1/12$^{th}$ the model size!

SOTA on 7 after fine-tuning



# Extending to 3D

**Classification**

**Apartment**

**Semantic Segmentation**

**Table, Bed, Couch, Cabinet**

**Object Detection**

**Bed, Couch, Cabinet, Cabinet**

**Instance Segmentation**

**Bed, Couch, Cabinet, Desk**

Single label

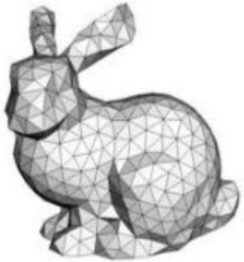
No objects

Multiple Object

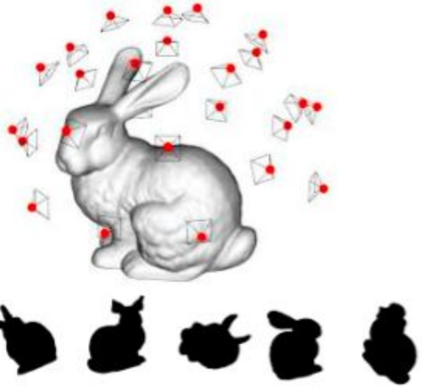
"ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes"
[Dai et al, CVPR 2017]

# Extending to 3D - Representation

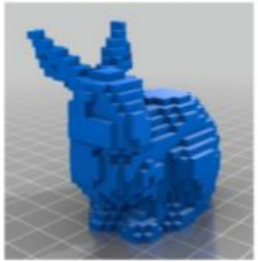
**Surface**:
Triangle Mesh



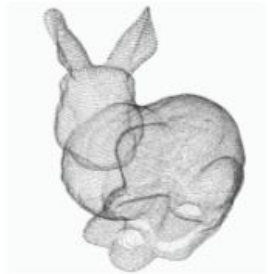
**Multi-View**:
Set of Images



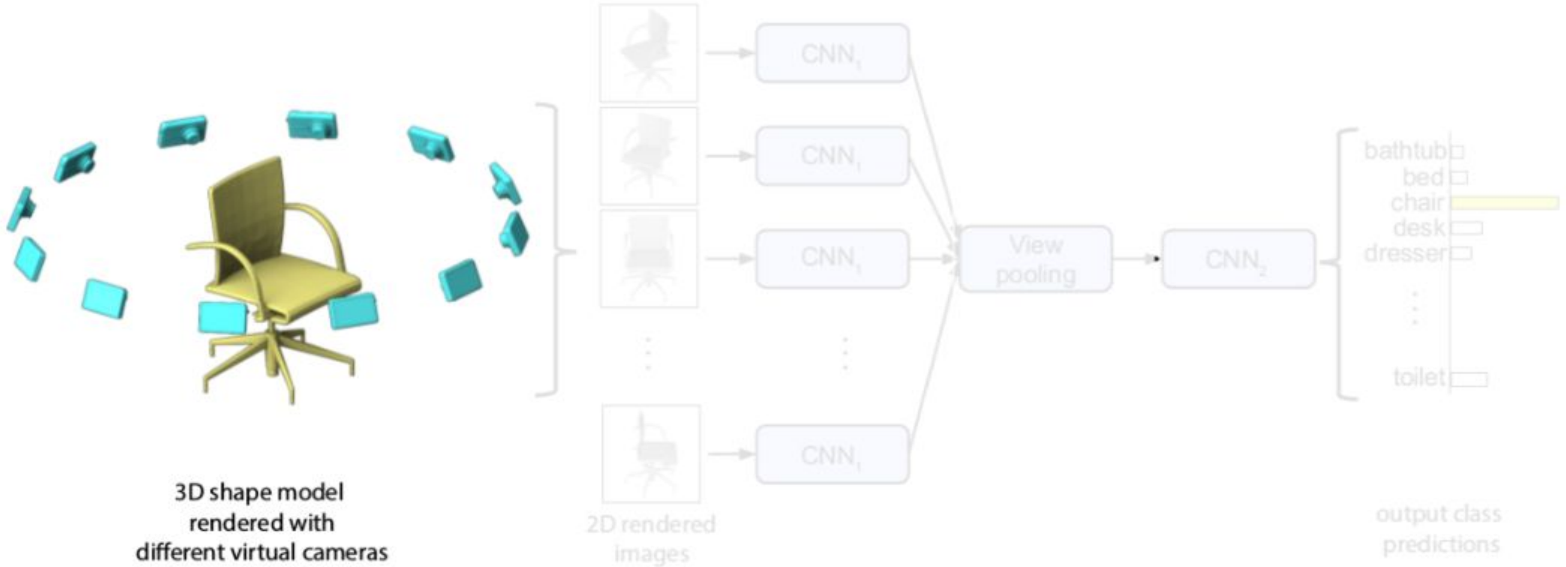
**Volumetric**:
Voxels



**Pointcloud**:
Set of points

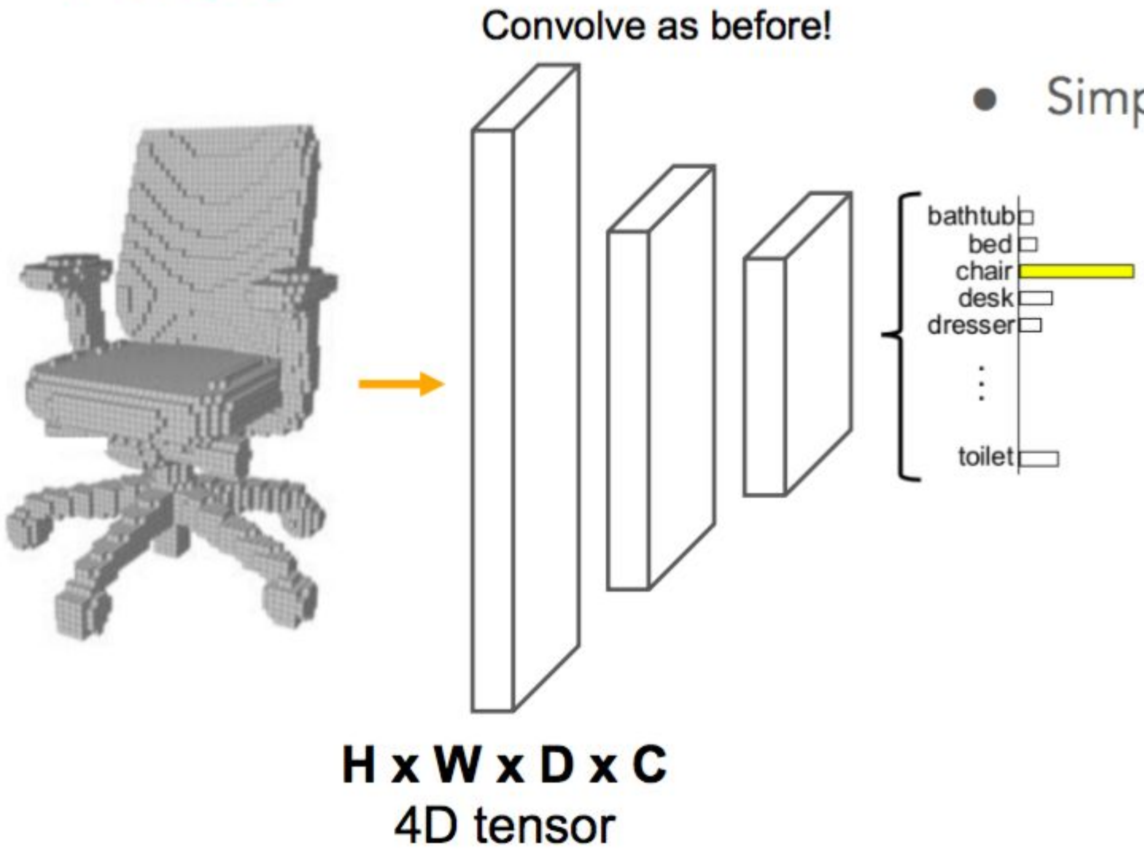


# Multiview



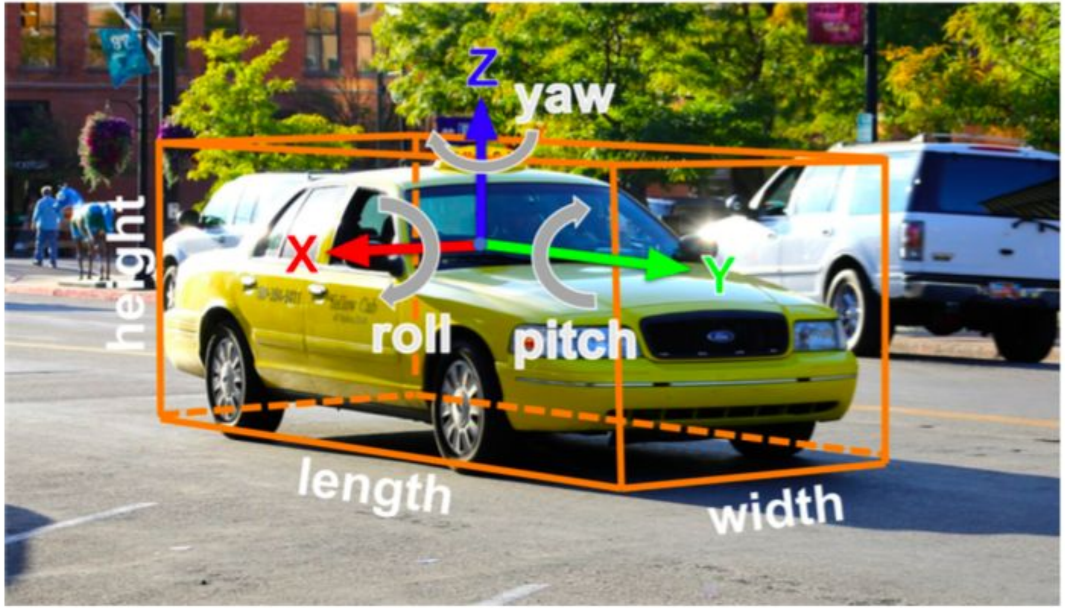

"Multi-view Convolutional Neural Networks for 3D Shape Recognition"
[Su, Maji, Kalogerakis, Learned-Miller, ICCV 2015]

# Voxels




- Simple extension of pixel

# 3D Object Detection



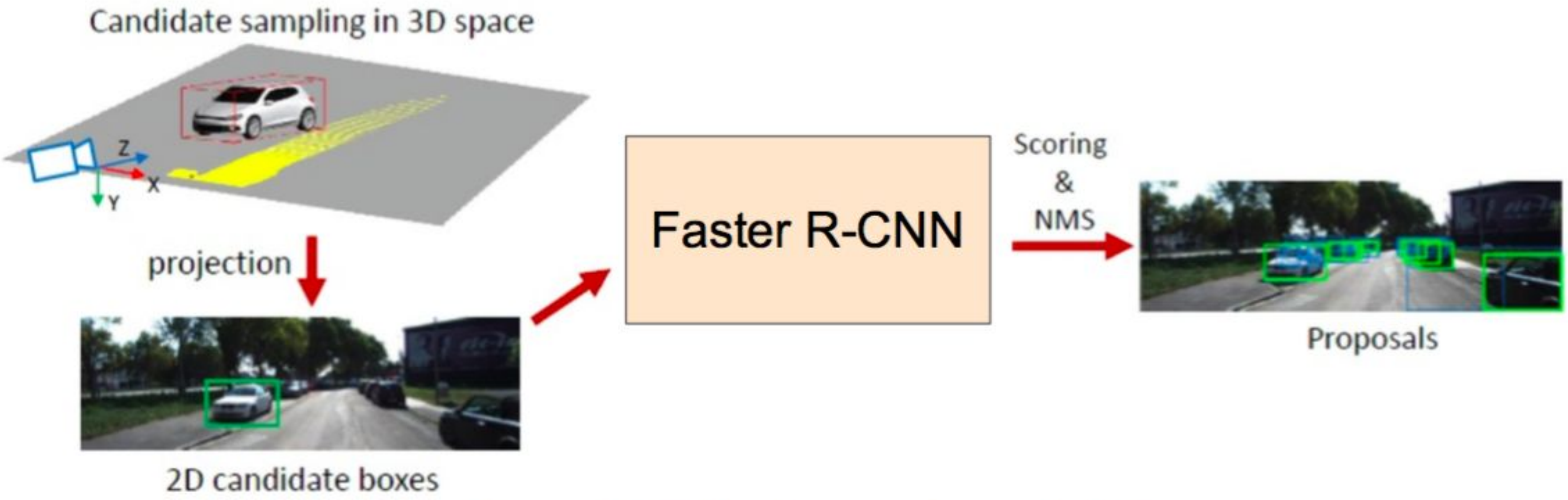

2D Object Detection:
2D bounding box
(x, y, w, h)

3D Object Detection:
3D oriented bounding box
(x, y, z, w, h, l, r, p, y)

Simplified bbox: no roll & pitch

Much harder problem than 2D object detection!

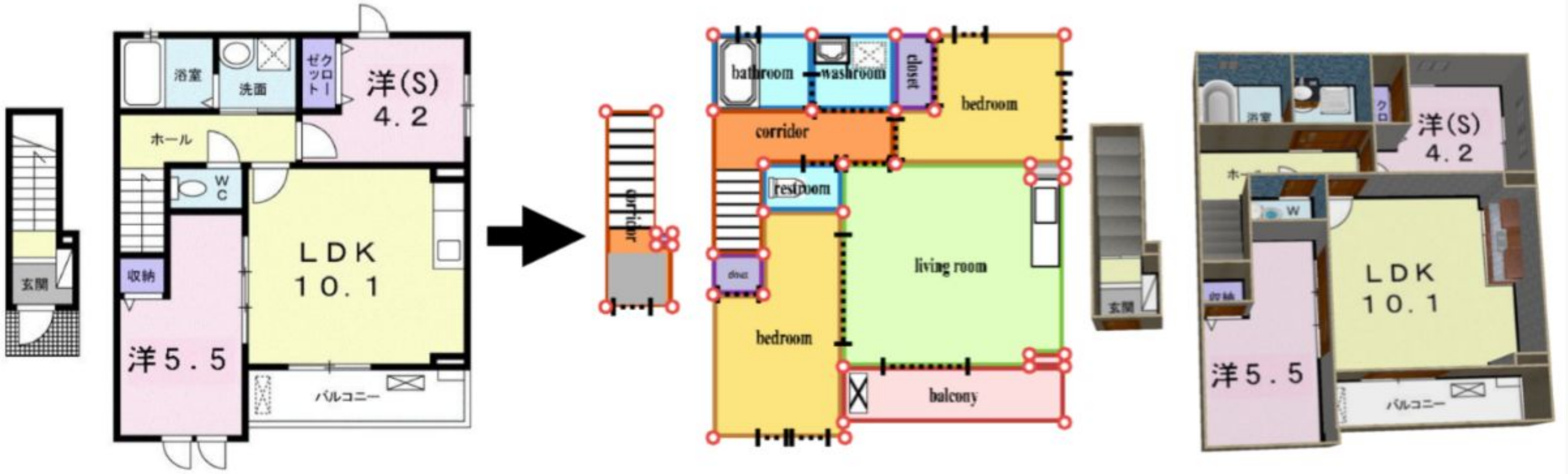
# 3D Object Detection: Monocular Camera



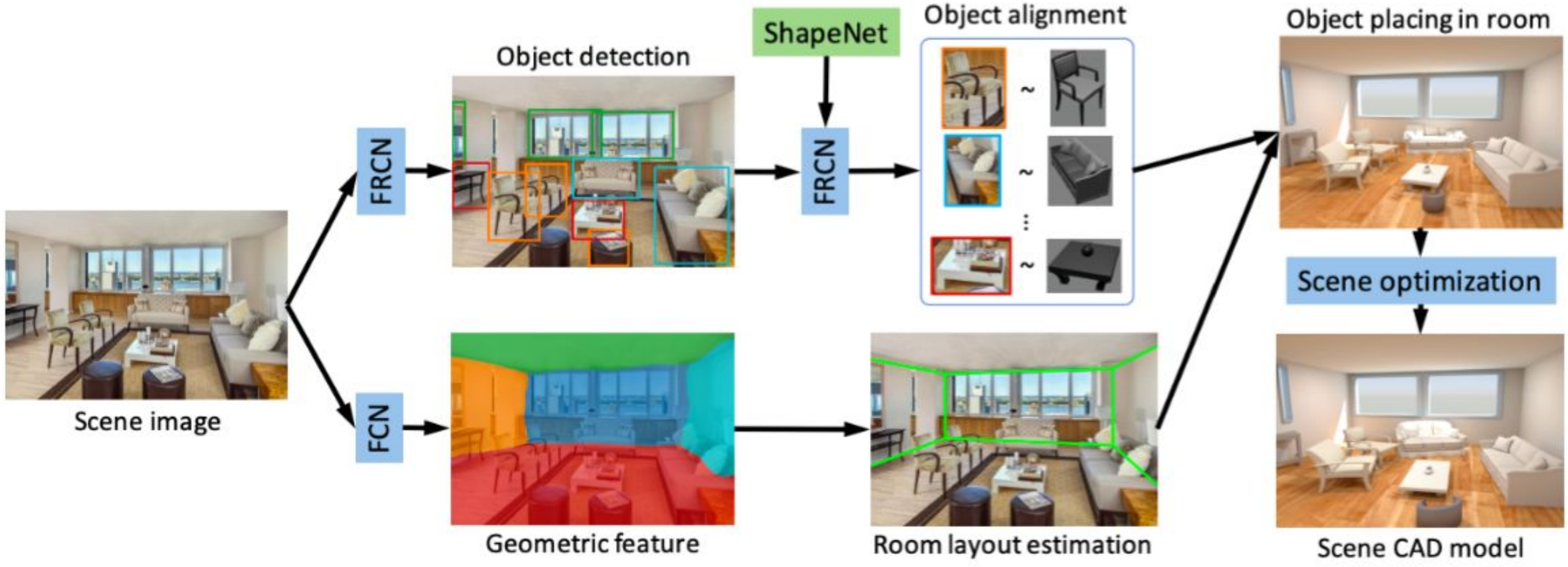

- Same idea as Faster RCNN, but proposals are in 3D
- 3D bounding box proposal, regress 3D box parameters + class score

Chen, Xiaozhi, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. "Monocular 3d object detection for autonomous driving." CVPR 2016.

# 2D floorplan to 3D model




"Raster-to-Vector: Revisiting Floorplan Transformation"
[Liu, Wu, Kohli, Furukawa, ICCV 2017]




"IM2CAD"
[Izadinia et al, CVPR 2017]

# Can we generate 3D scenes from (almost) scratch?

Empty Room

Select and arrange 3D models

Nicely arranged Living Room

3D model database