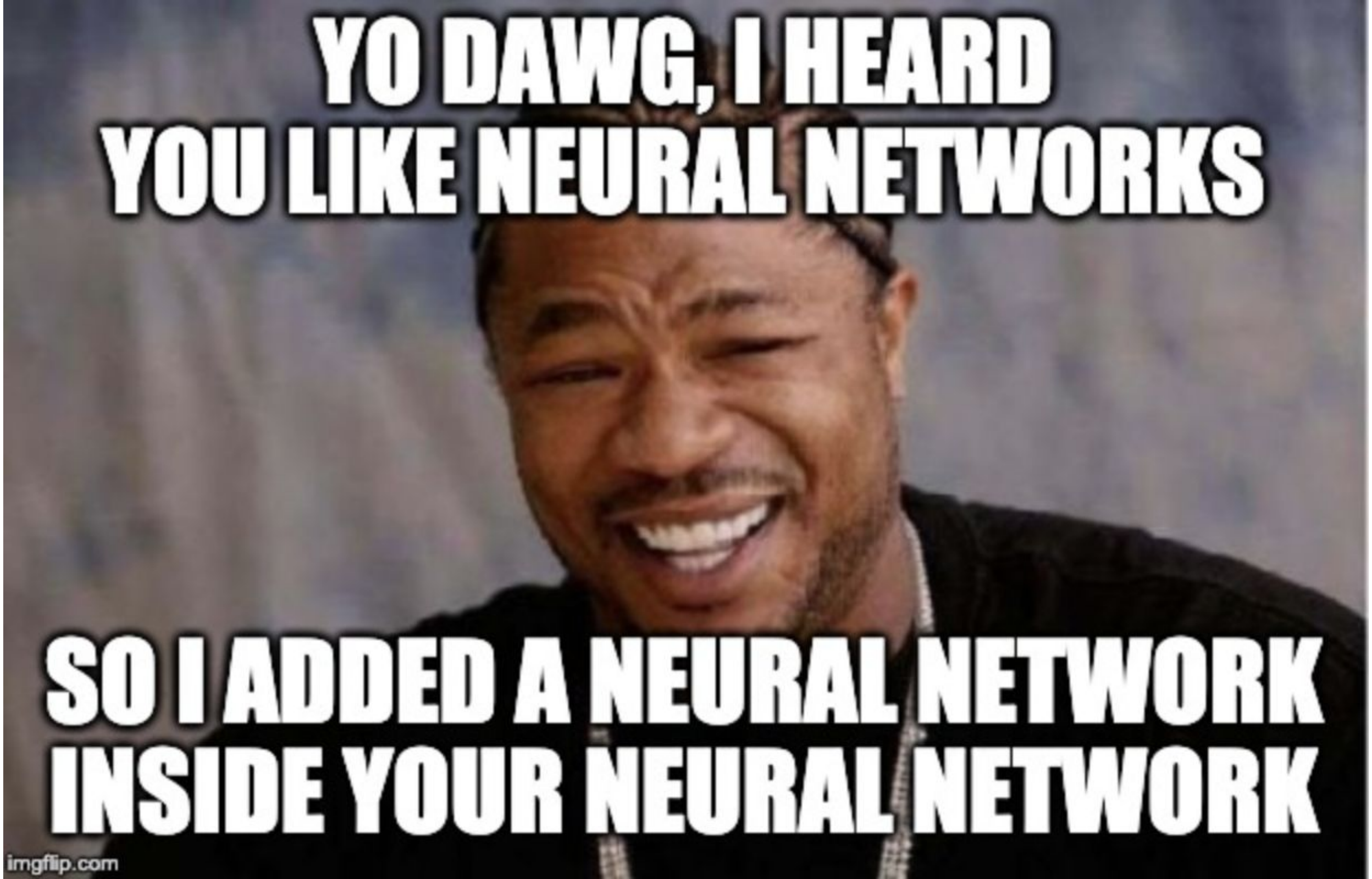
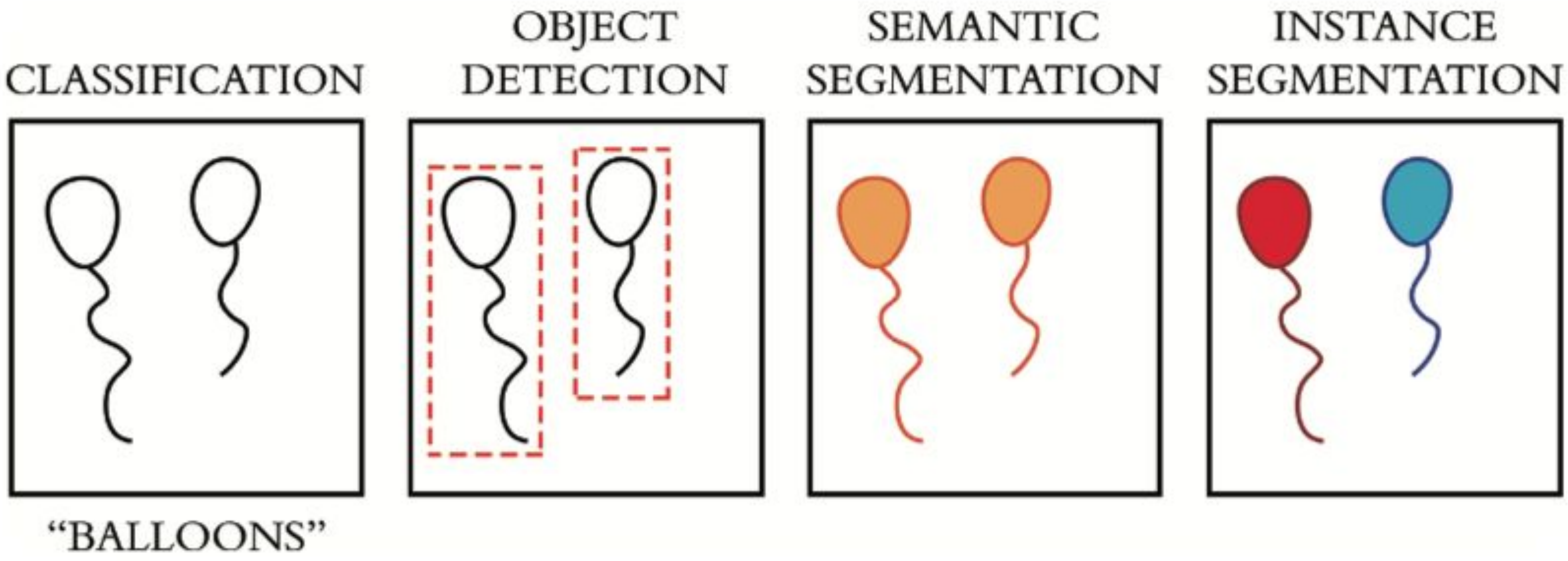
Capsule Networks

What’s the big idea?

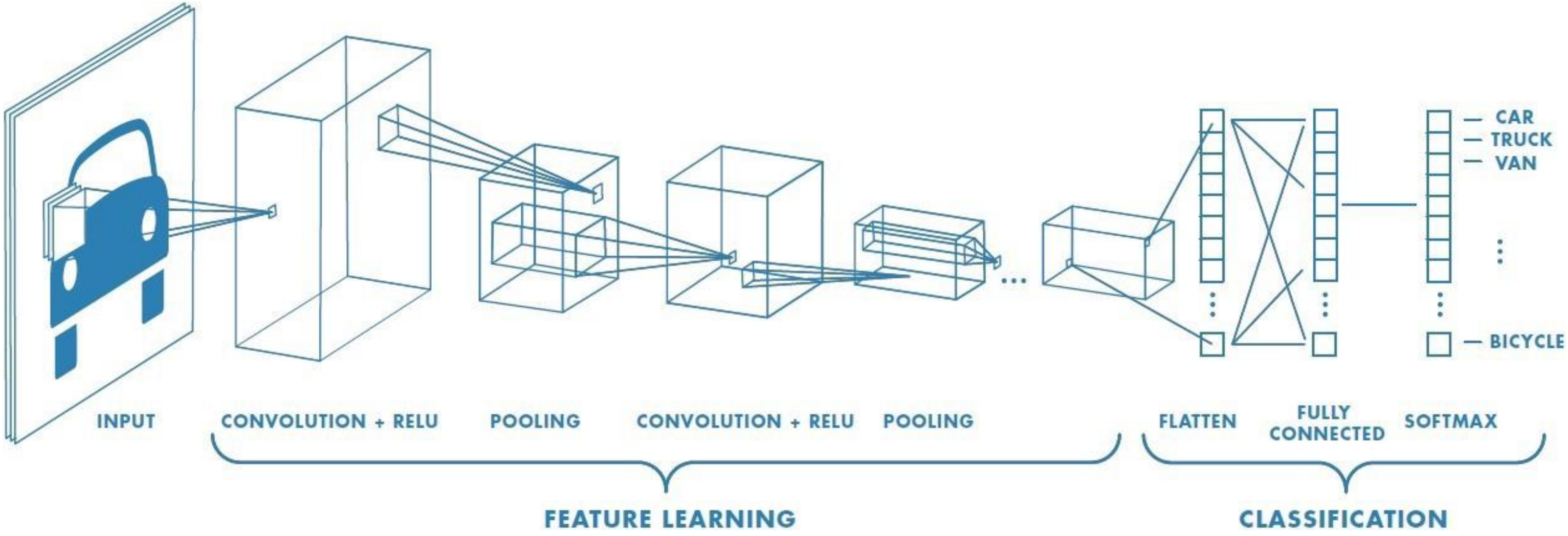

YO DAWG, I HEARD YOU LIKE NEURAL NETWORKS
SO I ADDED A NEURAL NETWORK INSIDE YOUR NEURAL NETWORK
imgflip.com

Goals of Machine Vision



What Deep Convolutional Neural Networks Do

- Recognize an object’s appearance anywhere in an image
- Optimize weights
- Create a hierarchy of object recognition

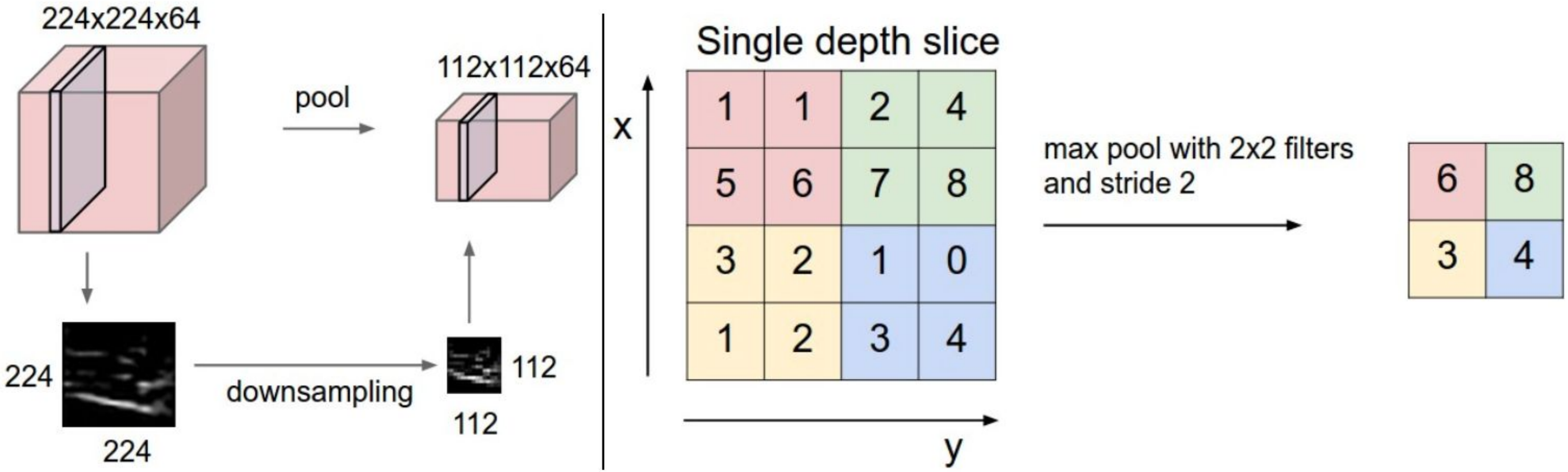


(Image: Toward Data Science)

What Are They Missing?

- “The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster.” - Geoff Hinton
- Networks learn to identify patterns but not to identify connections between the objects present in the image.
- Max-pooling is information loss

Max Pooling is Evil

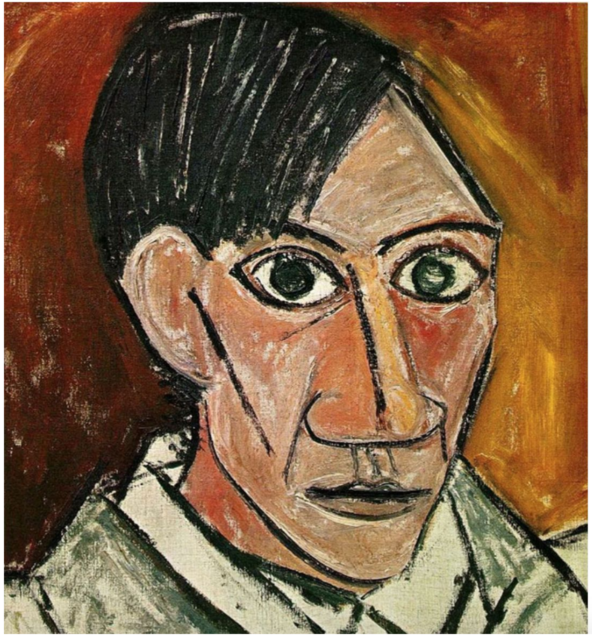


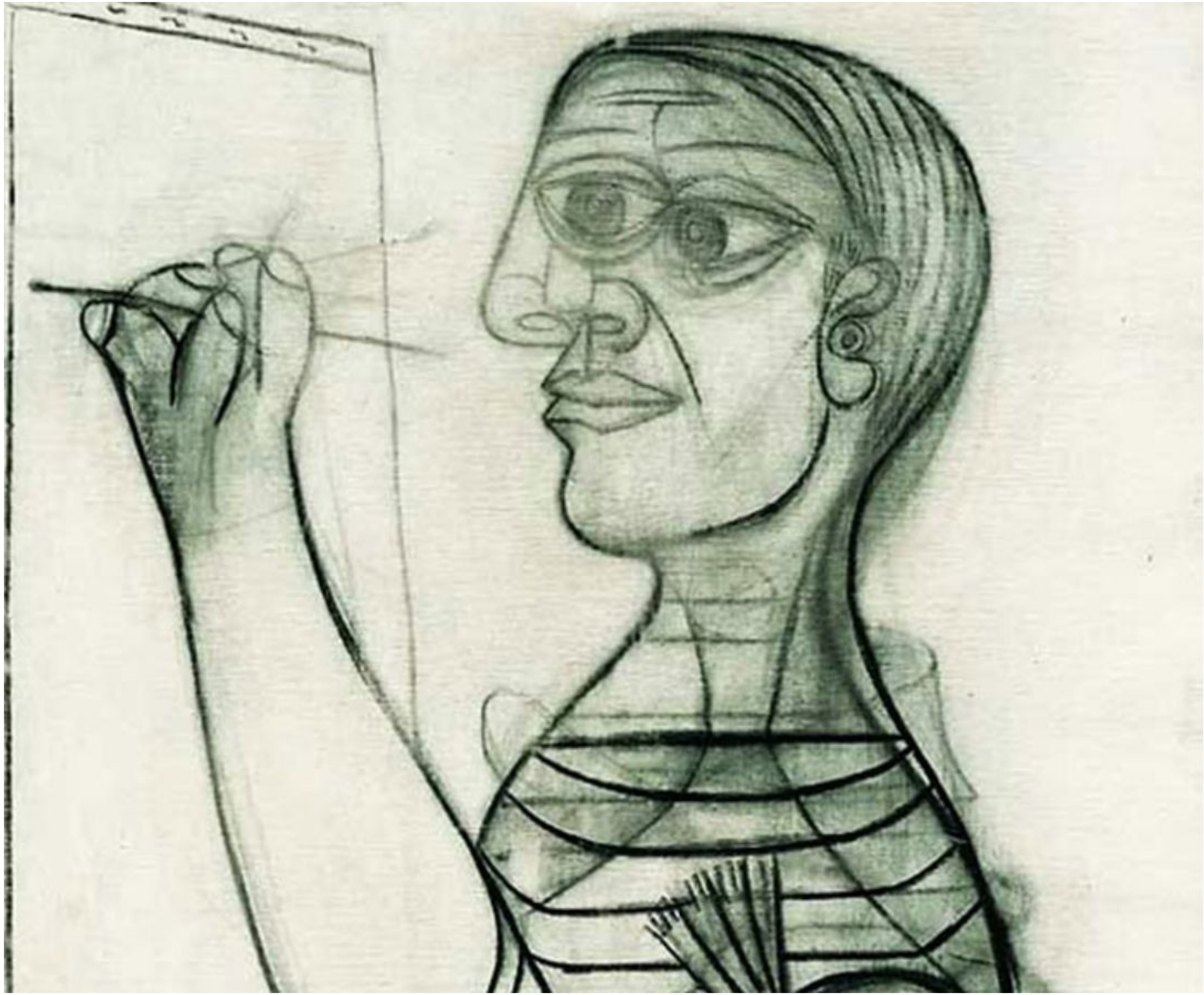
What Are They Missing Part 2

- “Orientational and relative spatial relationships between these components are not very important to a CNN.” - Max Pechonkin
- Tried to make up for the issues by introducing data set augmentation
    - Rotate the images in every which way to train your network to recognize rotated cats.
    - Requires a ton more data
- The resulting neural nets have no notion of coordinate systems, rotating pixels and rearranging pixels can fool them easily.

FACE

FACE





FACE

What Do Humans Have?

- Humans naturally apply transformations on objects when trying to recognize them in an image
- Humans use coordinate system + pattern recognition, not just pattern recognition

- They have a cortical column, also known as the minicolumn, and it is considered to be the basic functional unit of the cerebral cortex

Related Brain Research

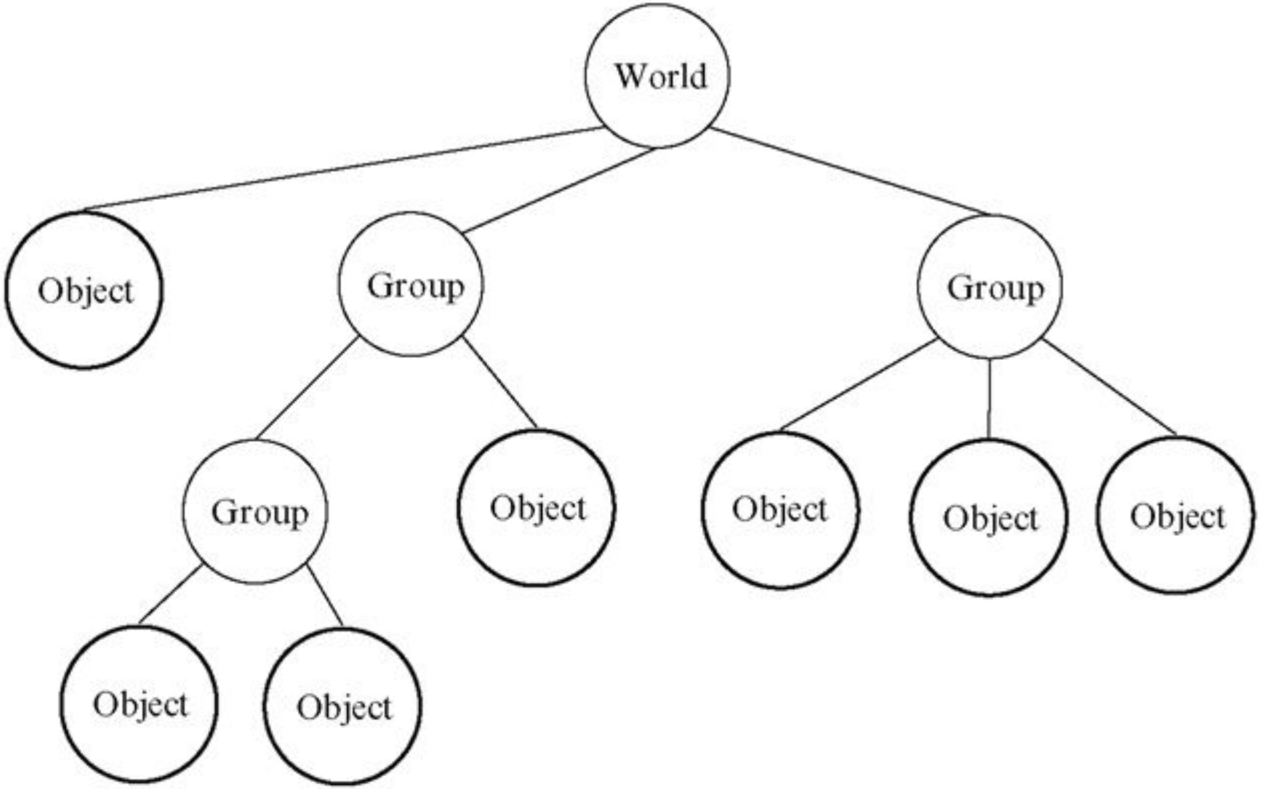
- Neurons within a minicolumn (microcolumn) encode similar features [1]
- Vernon Mountcastle maps the brain’s response to touch to find that the cortical columns: “code for both location and quality of stimulation”[4]
- Folks at Numenta published about columns in the neocortex enable learning the structure of the world [2]
- “The output layer learns complete models of objects as a set of features at locations. This is analogous to how computer-aided-design programs represent multi-dimensional objects.” - John Hawkins [3]

What Capsule Networks Add

- 3D orientation between objects in an image with CapsNet is much easier to retrieve because these relationships are explicitly modeled
- Training needs a fraction of the data set that was required for a ConvNet*
- Each capsule learn a vector that describes object and it’s orientation adding pose information to the object.
- They build a graph of the objects in the image
- Regularized by an auto encoder which promotes keeping relevant information around

How Do They Do That?!

- Enter the field of computer graphics: object rendering and **scene graphs.**
- “Representation of objects in the brain does not depend on view angle.”
- Learn about objects **regardless of their orientation** to the viewer
- Learn about objects and their **relative orientation** to each other


World
Object
Group
Group
Group
Object
Object
Object
Object
Object
Object

How Do They Do That?!

- Capsules vote among each other to determine who is the best to handle information
    - Lower level capsule place bets, higher level capsules take winning bets
- This is rerouting based on agreement
- They are learning a graphics system instead of only weights for a neural network

- Use a mix of unsupervised and supervised methods
- Difficulties: De-rendering in the early level to get pose information in the higher level.

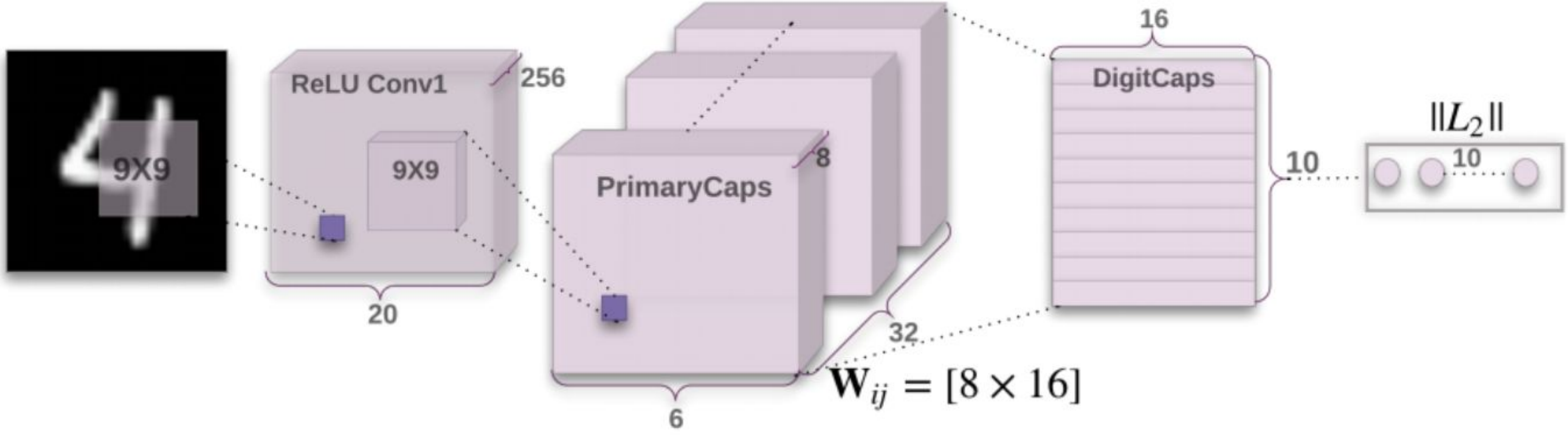


image 3.0: CapsNet Architecture

(Image: Dynamic Routing Between Capsules)

Training a Capsule

Procedure 1 Routing algorithm.

procedure ROUTING($\hat{\boldsymbol{u}}_{j|i}$, r, l)

  for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$.

  for r iterations **do**

    for all capsule i in layer l: $\mathbf{c}_i \leftarrow \texttt{softmax}(\mathbf{b}_i)$ ▷ `softmax` computes Eq. 3

    for all capsule j in layer $(l+1)$: $\mathbf{s}_j \leftarrow \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}$

    for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow \texttt{squash}(\mathbf{s}_j)$ ▷ `squash` computes Eq. 1

    for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.\mathbf{v}_j$

  return $\mathbf{v}_j$

[Let's look at some PyTorch code!]

Results and Accuracy

Table 1: CapsNet classification test accuracy. The MNIST average and standard deviation results are reported from 3 trials.

Method	Routing	Reconstruction	MNIST (%)	MultiMNIST (%)
Baseline	-	-	0.39	8.1
CapsNet	1	no	$0.34_{\pm 0.032}$	-
CapsNet	1	yes	$0.29_{\pm 0.011}$	7.5
CapsNet	3	no	$0.35_{\pm 0.036}$	-
CapsNet	3	yes	$\mathbf{0.25}_{\pm 0.005}$	**5.2**

Results In Face and Object Recognition

Table 3. Comparison of classification results.

			Baselines			**CapsNet**	
Dataset	**Classes**	**Instances**	**Algorithm**	**Avg. training time**	**Test accuracy**	**Avg. training time**	**Test accuracy**
Yale Face Database B	38	5850	Fisherface	~5 minutes*	98.2%**	~24 hours***	95.3%
MIT CBCL (faces)	10	5240	Fisherface	~1 minute*	98.3%**	~14 hours***	99.87%
BelgiumTS (traffic signs)	62	7000	Modified LeNet	<1minute*	98.2%	16 hours***	92% (40 epochs)
CIFAR-100 (objects)	100	60000	Resnet 50	20 hours (200 epochs)	65.5%	18 hours***	18% (35 epochs)

Results on CIFAR 10

Table 1: Accuracy Results for Various Models

	Validation Accuracy	
Models	25 Epochs	50 Epochs
MNIST Model Baseline	67.51%	68.93%
64 Capsule Layers	60.54%	64.67%
4-Model Ensemble (4 Ensemble)	68.97%	70.78%
2-Convolution Layers (2 Conv)	68.14%	69.34%
4 Ensemble + 2 Conv	70.34%	71.50%
7 Ensemble + 2 Conv	70.50%	______
4 Ensemble + 2 Conv + 0.0001 Reconstruction Scaling	69.21%	______
Stack Additional Capsule Layer	10.11%	______

Resources

Geoff Hinton talks about capsule networks

Capsule networks Tutorial by Aurelien Geron

Understanding Hintons Capsule Networks by Max Pechyonkin

Capsule networks overview,

Expressing Pose

Scene graph

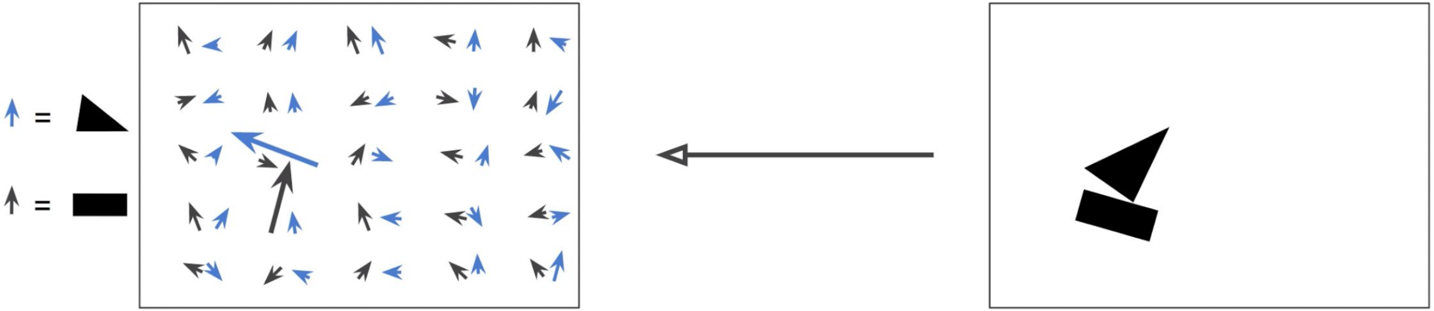
Secret to Strong AI by Jeff Hawkins

Columns in the neocortex enable learning the structure of the world by Jeff Hawkins

CapsNet comparative performance evaluation for image classification

Aurélien Géron, 2017

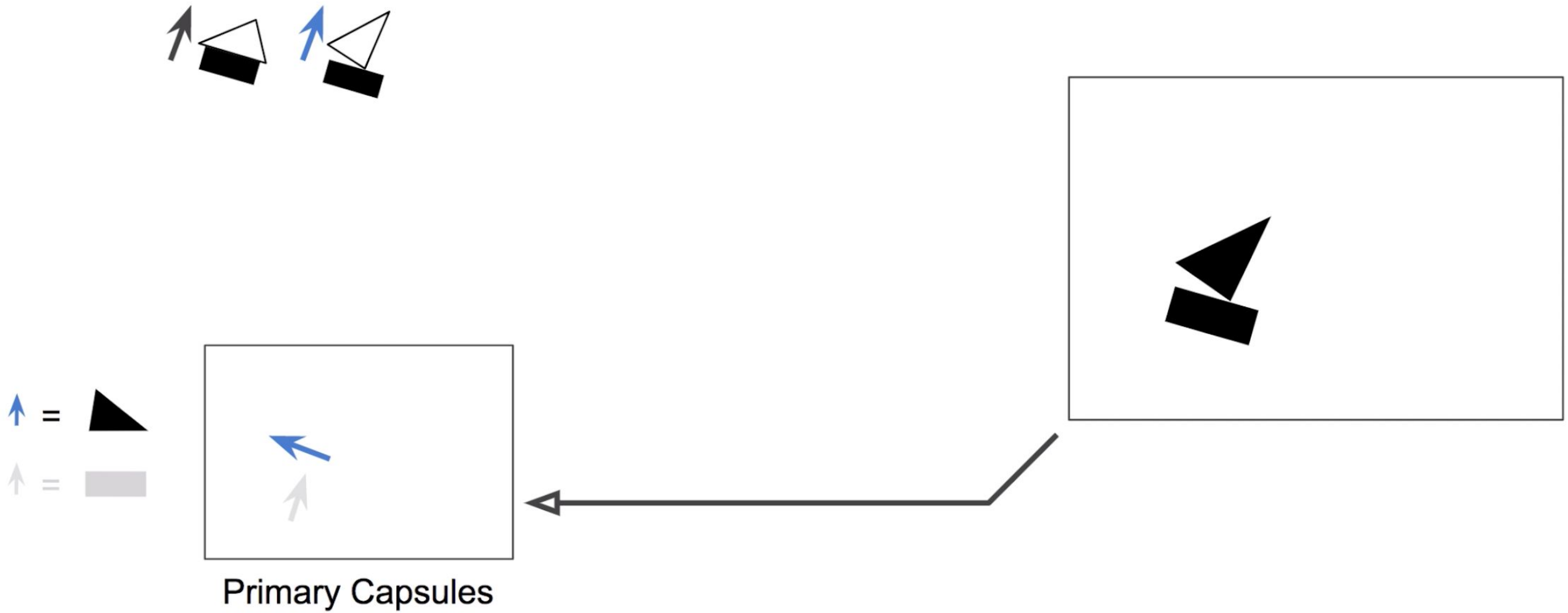
Capsules



Activation vector: **Length** = estimated probability of presence
Orientation = object's estimated pose parameters

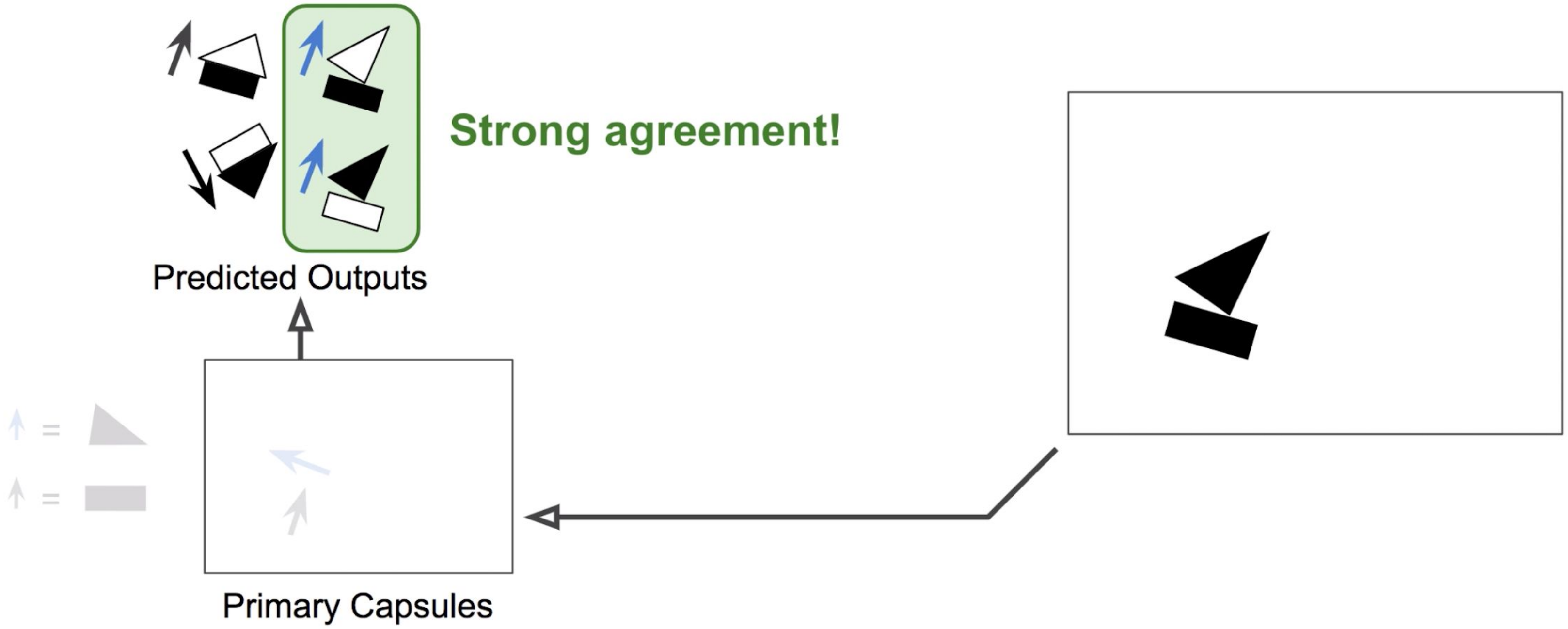
Aurélien Géron, 2017

Predict Next Layer's Output



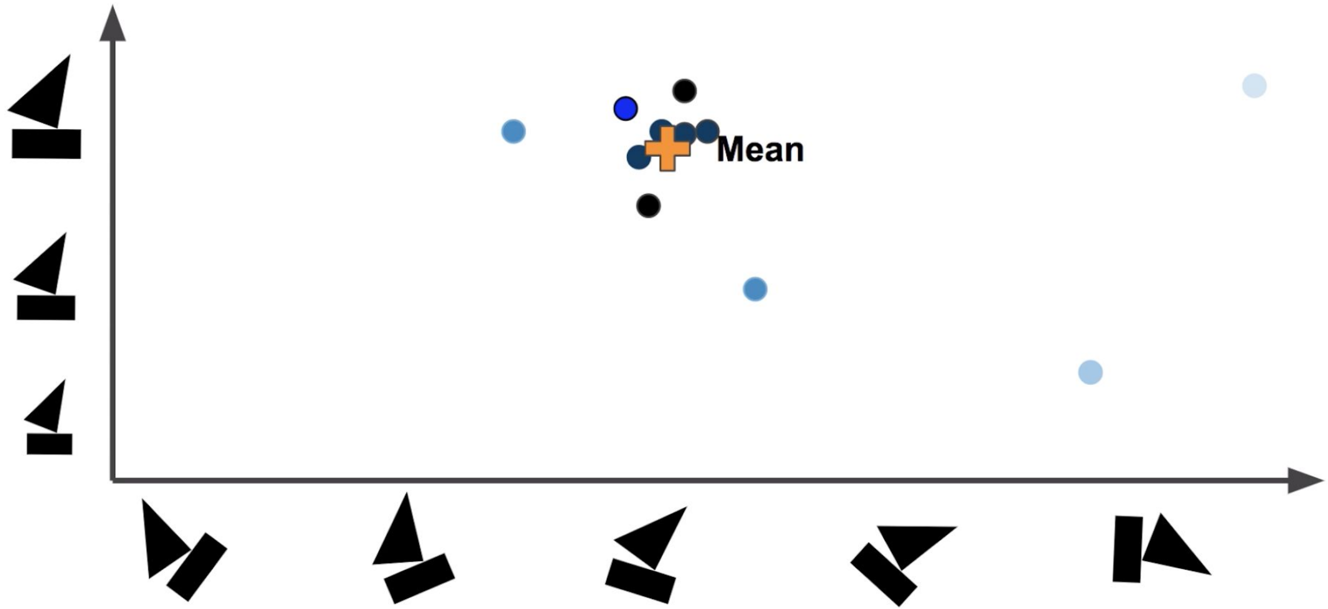
Aurélien Géron, 2017

Routing by Agreement



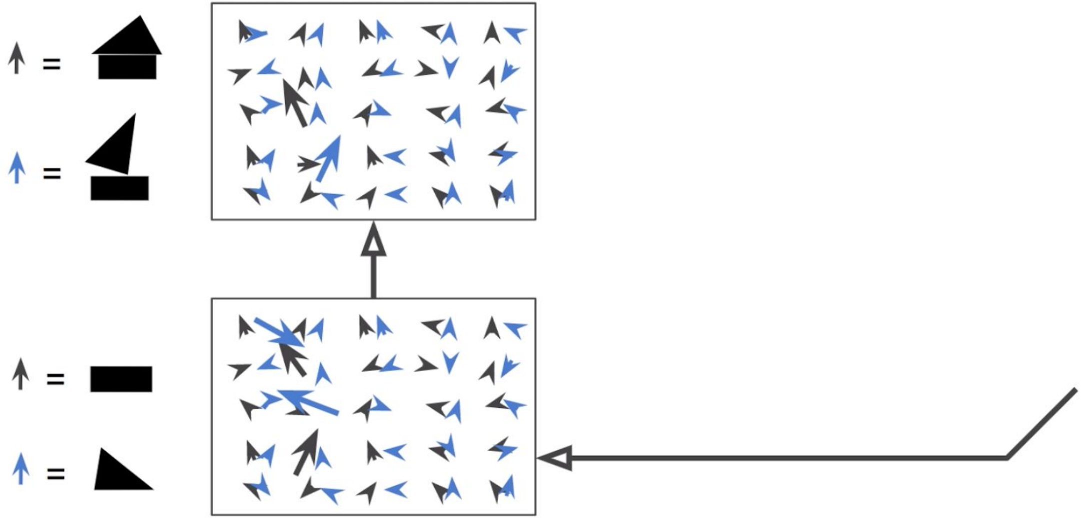
Aurélien Géron, 2017

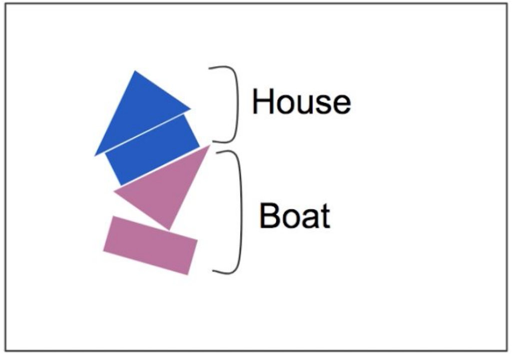
Clusters of Agreement



Aurélien Géron, 2017

Handling Crowded Scenes





Thanks to routing by agreement, the ambiguity is quickly resolved (explaining away).