SEOUL ARTIFICIAL INTELLIGENCE

seoulai.com

Dynamic Routing Between Capsules.

Geoffrey Hinton et al. (2017)

Dmitriy Khvan

Intro.

- **I am…**
- **Why Capsule Nets?**
- **Prerequisites**
- **My goal:**

⭐ encourage to read the paper
⭐ give intuitive explanation behind the idea
⭐ practice public speaking :) and mac Keynote… :(
⭐ make it fun for everyone!

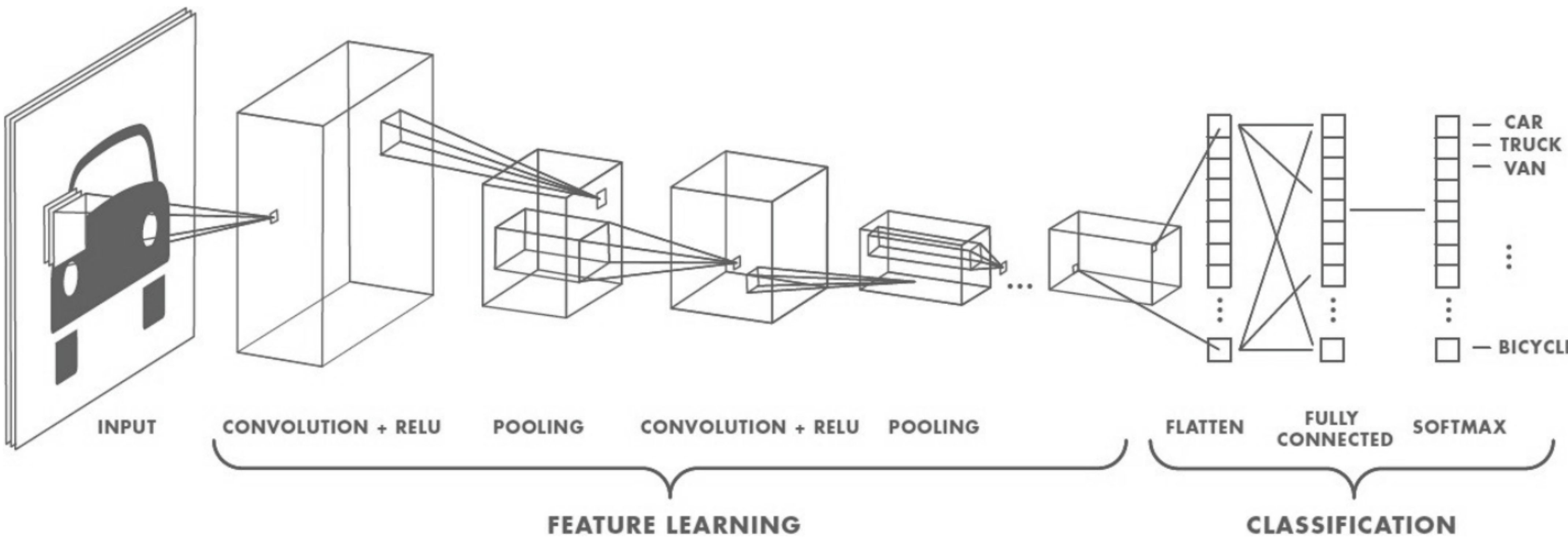
References.

- ✓ Dynamic Routing Between Capsules G. Hinton et al.
- ✓ Understanding Hinton’s Capsule Networks Max Pechyonkin
- ✓ Beginner's Guide to Capsule Networks Zafar (Kaggle)
- ✓ What is wrong with convolutional neural nets? G.Hinton
- ✓ Capsule Networks Tutorial Aurellien Geron

Motivation.

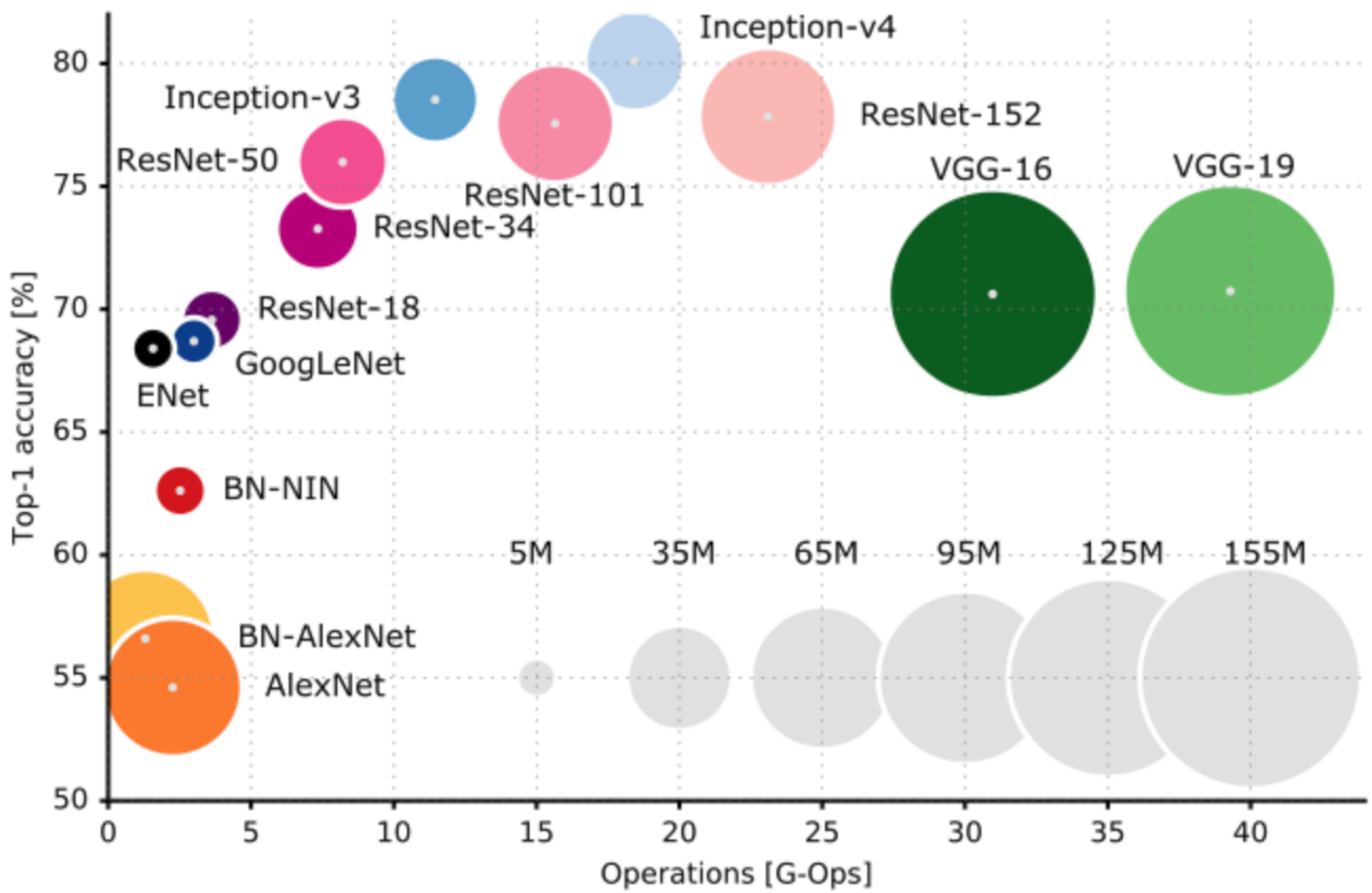
- ➡ Need for intuitive approach closely reflecting HVRS
- ➡ Reduce data needed for training
- ➡ Robustness to affine transformations
- ➡ Need for recognition of overlapping objects

CNN.



CAR
TRUCK
VAN
BICYCLE
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
FEATURE LEARNING
CLASSIFICATION

CNN.

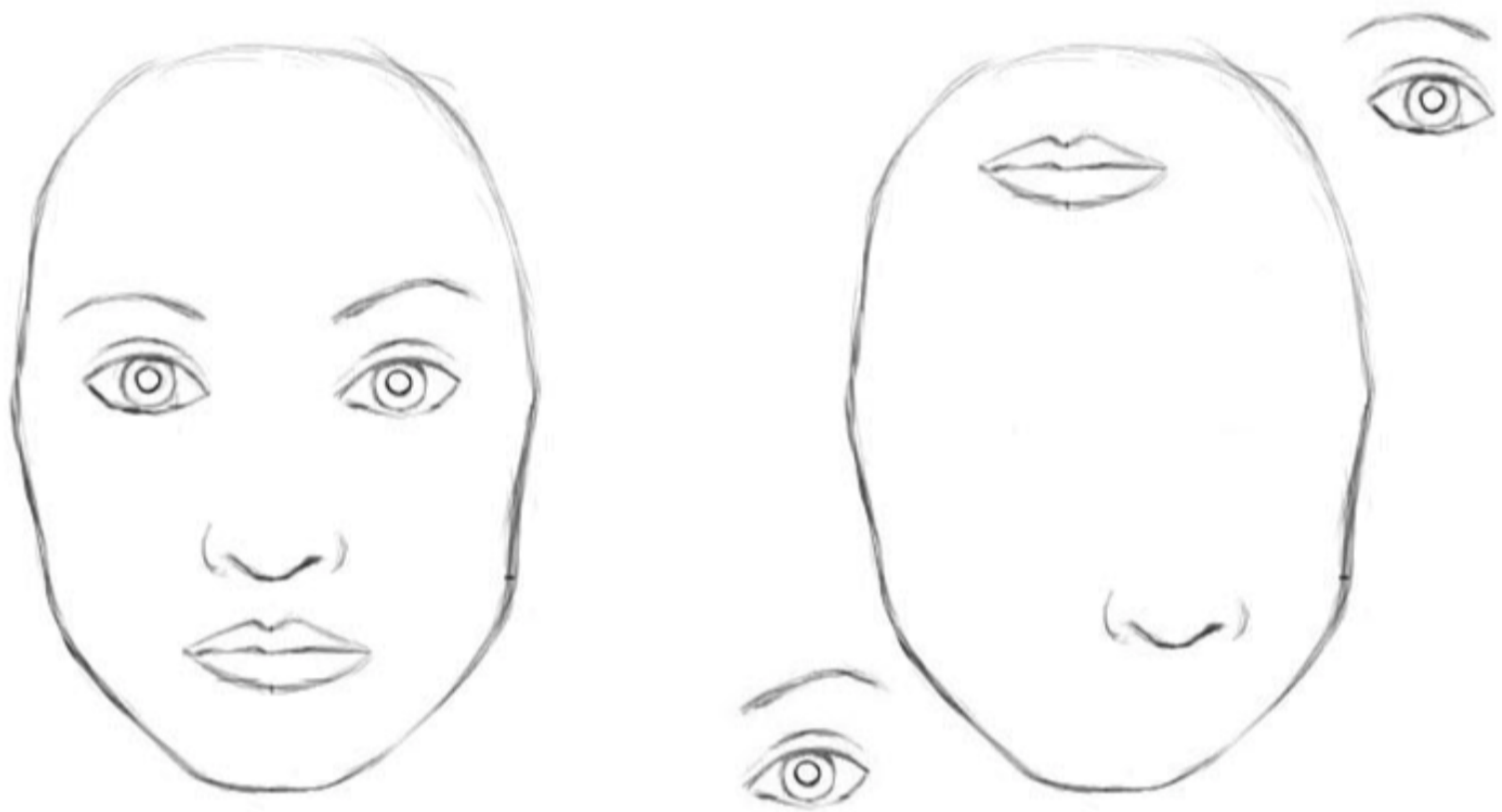


Inception-v4
Inception-v3
ResNet-152
ResNet-50
ResNet-101
VGG-16
VGG-19
ResNet-34
ResNet-18
GoogLeNet
ENet
BN-NIN
5M
35M
65M
95M
125M
155M
BN-AlexNet
AlexNet
80
75
70
65
60
55
50
Top-1 accuracy [%]
0
5
10
15
20
25
30
35
40
Operations [G-Ops]

What is wrong with convolutional neural networks?

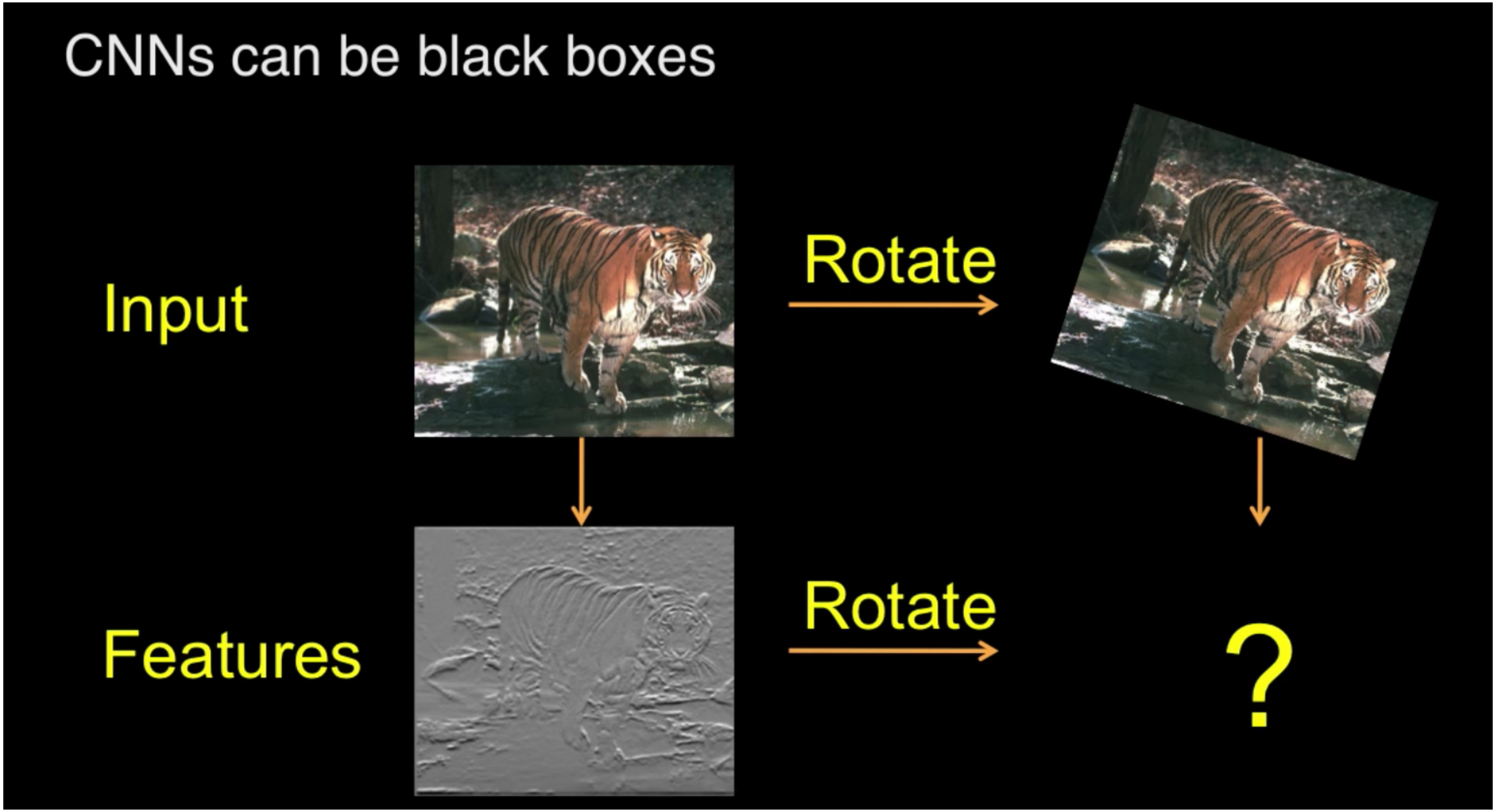
CNN fails to recognize variations of an image.

What is wrong with convolutional neural networks?



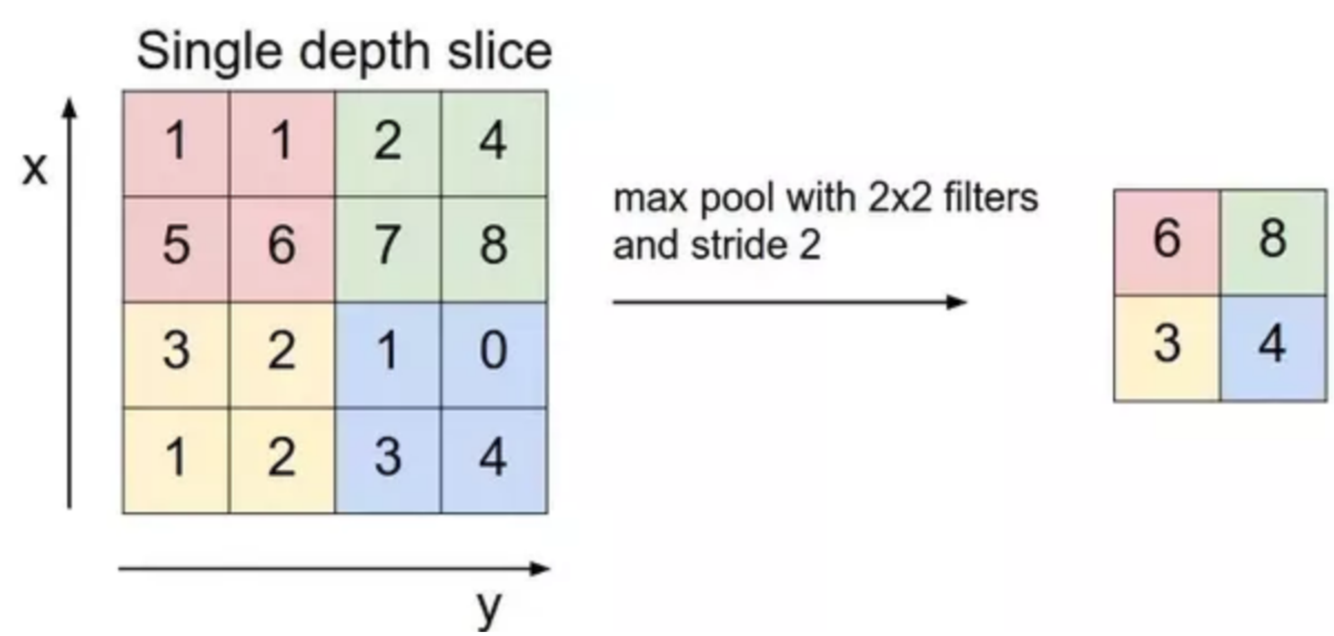
Max pooling takes highest activations and **forgets where they come from**! By doing this we lose information about spatial relations between different features.

What is wrong with convolutional neural networks?



What is wrong with convolutional neural networks?

- **Too few levels of structure (neuron, layer, nn).**
  - **Capsule** with 2 instantiation parameters: whether an entity present & entity properties.

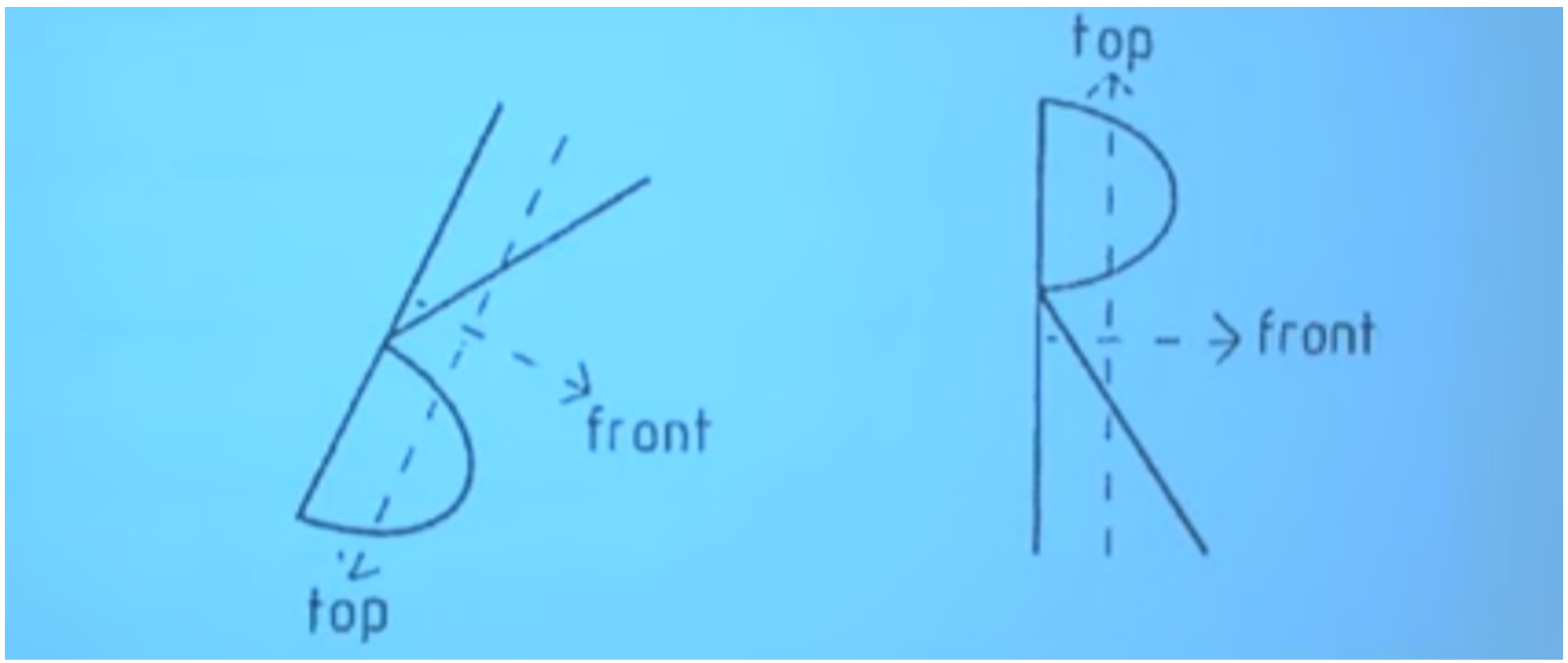


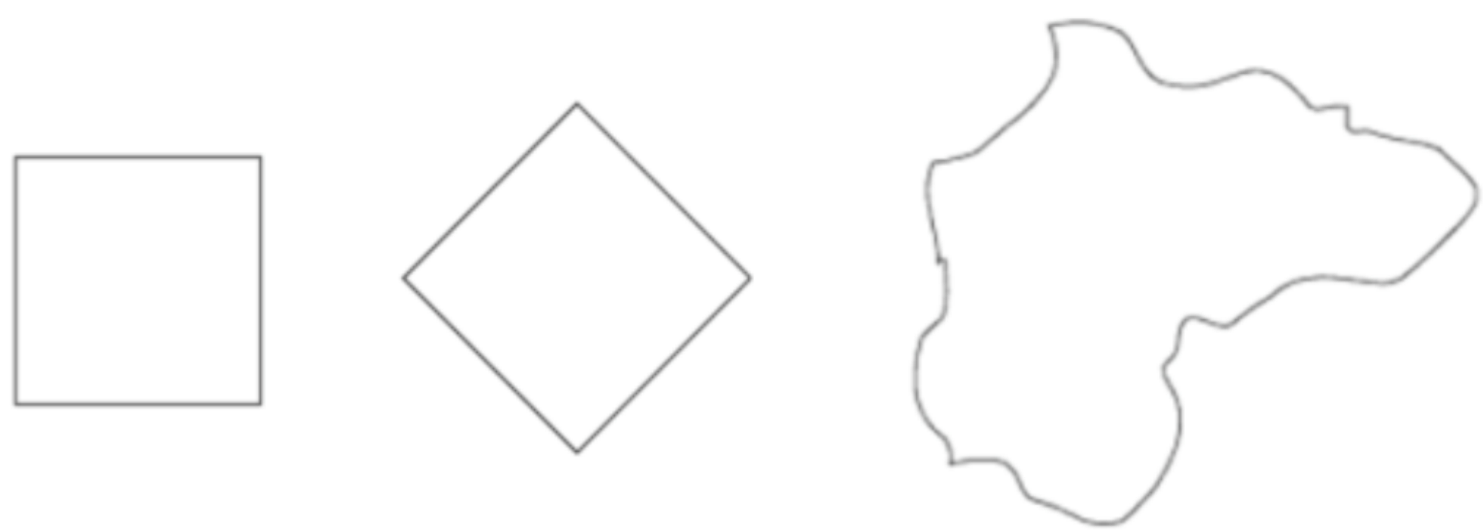
- **4 arguments against pooling:**
  - Does not reflect psychology of **shape perception**;
  - We want **equivariance** not invariance;
  - Ignores underlying **linear structure**;
  - Poor way of **routing**;

"I believe in convolution but I don't believe in pooling. The fact pooling works so well is a disaster." G.Hinton.

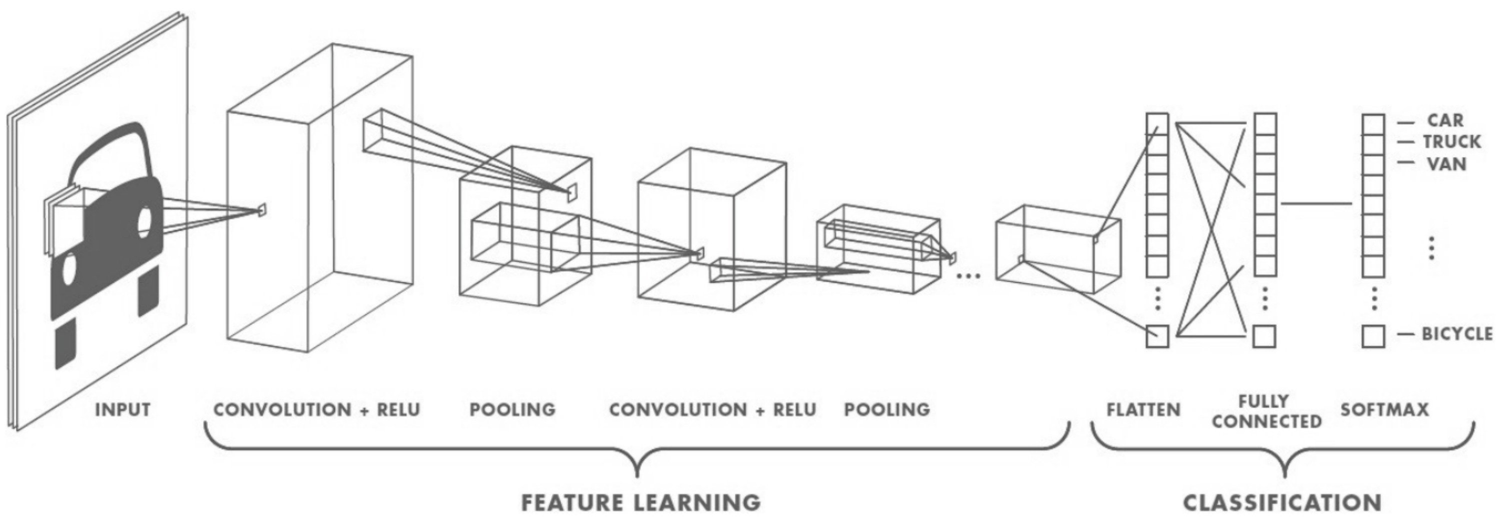
Argument 1: CNN does not reflect psychology of shape perception.

It is a bad fit to the psychology of shape perception: It does not explain why we **assign intrinsic coordinate frames to objects** and why they have such huge effects.





Argument 2: We want equivariance, not invariance.

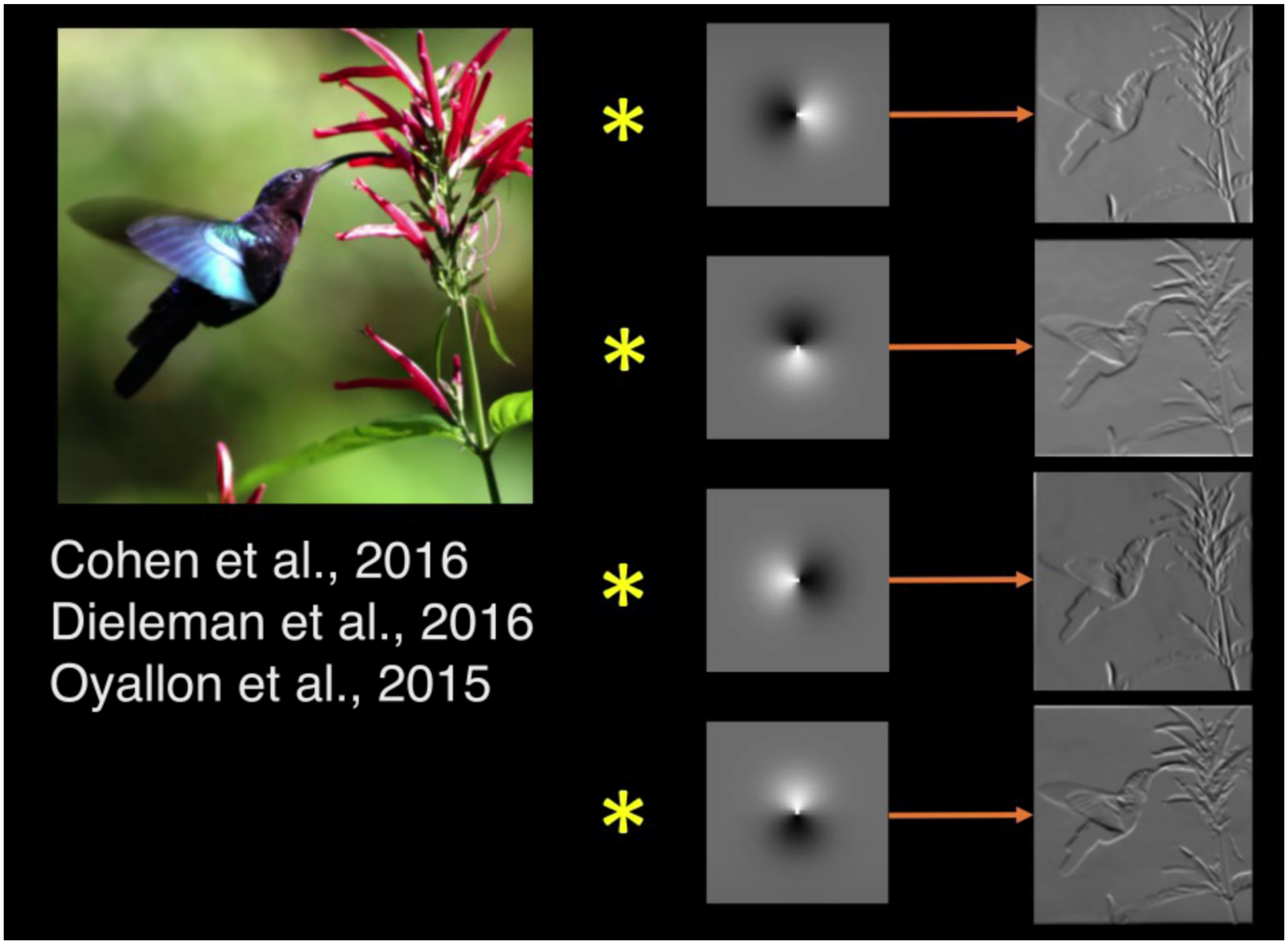


- [x] Translation invariant
- [ ] Rotation invariant
- [ ] Scale invariant

Several works aim to introduce rotation, scale invariance to CNNs.

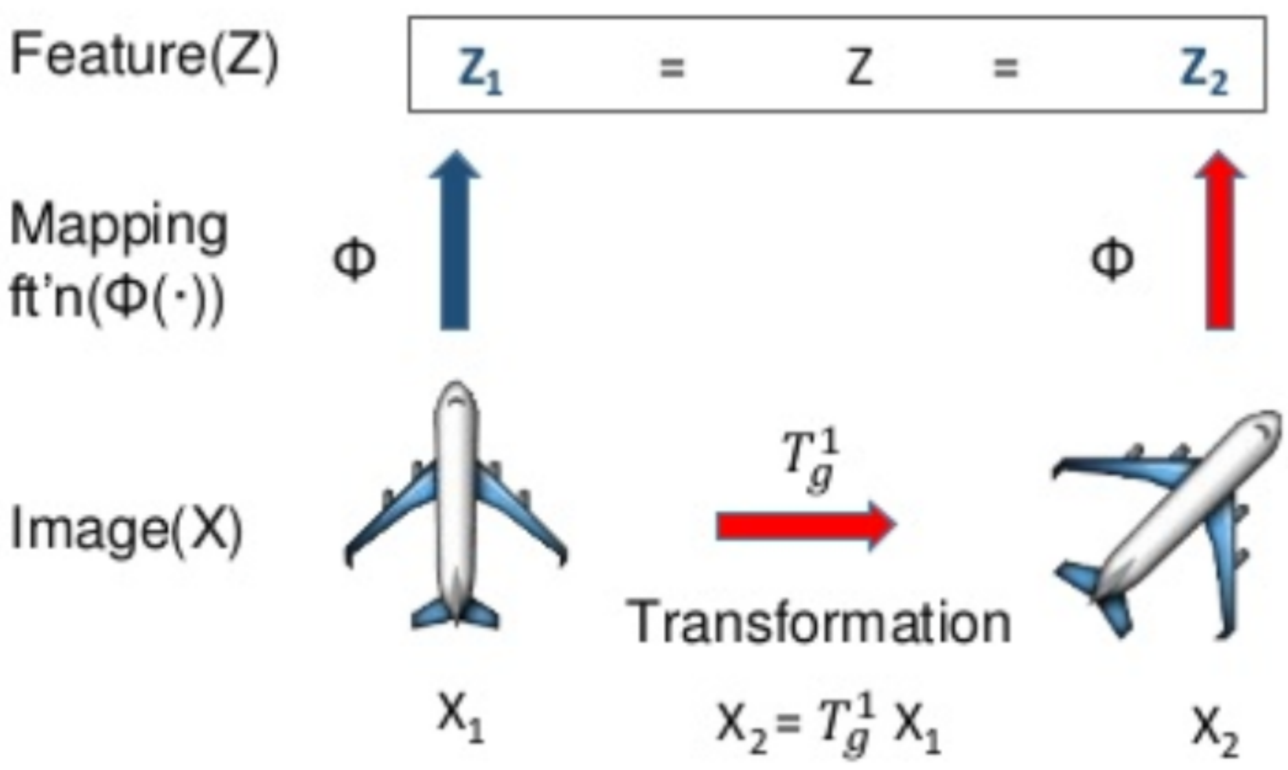
But they **trying to solve wrong problem** (invariance), we want equivariance instead!

Rotational Invariance.



Invariance.

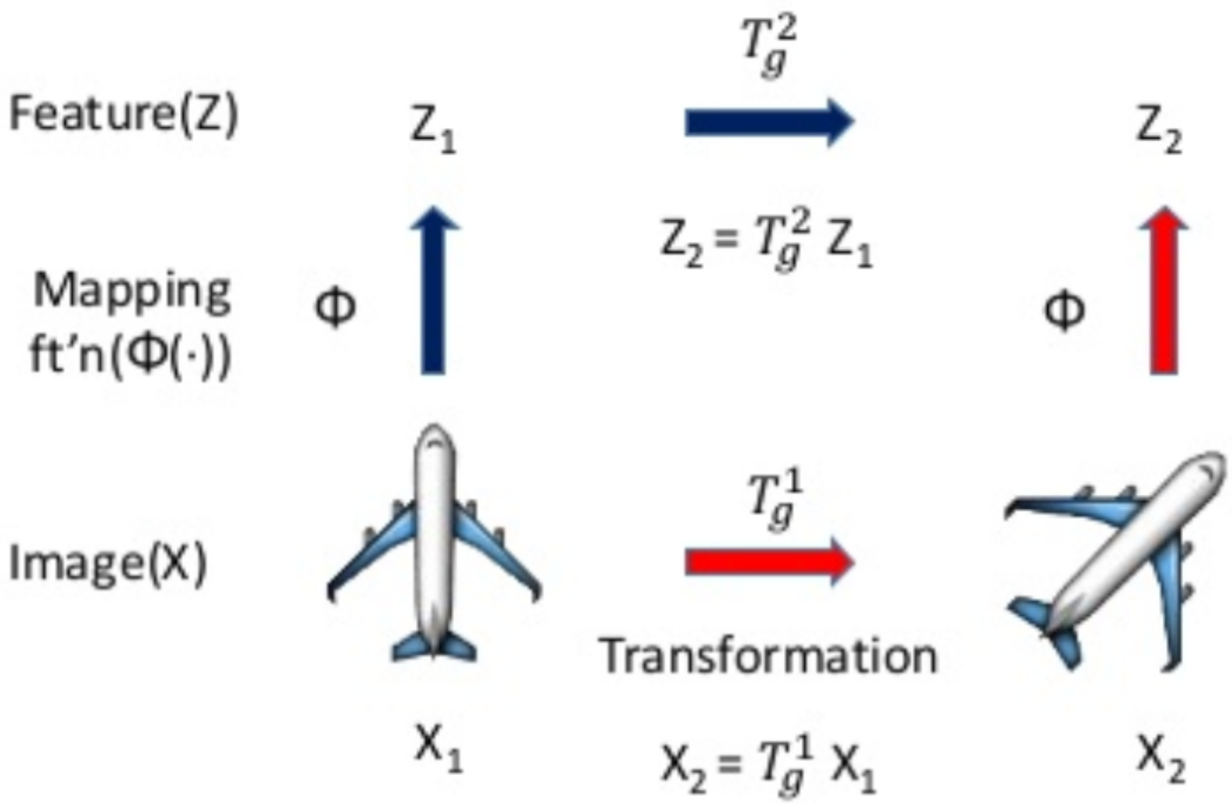
: Mapping independent of transformation, T_g, for all T_g



$$Z = Z_1 = \Phi(X_1) = Z_2 = \Phi(X_2) = \Phi(T_g^1 X_1)$$

Equivariance.

: Mapping preserves algebraic structure of transformation



$Z_1 \neq Z_2$ but keeps the relationship $\quad Z_2 = T_g^2 Z_1 = T_g^2 \Phi(X_1) = \Phi(T_g^1 \, X_1)$

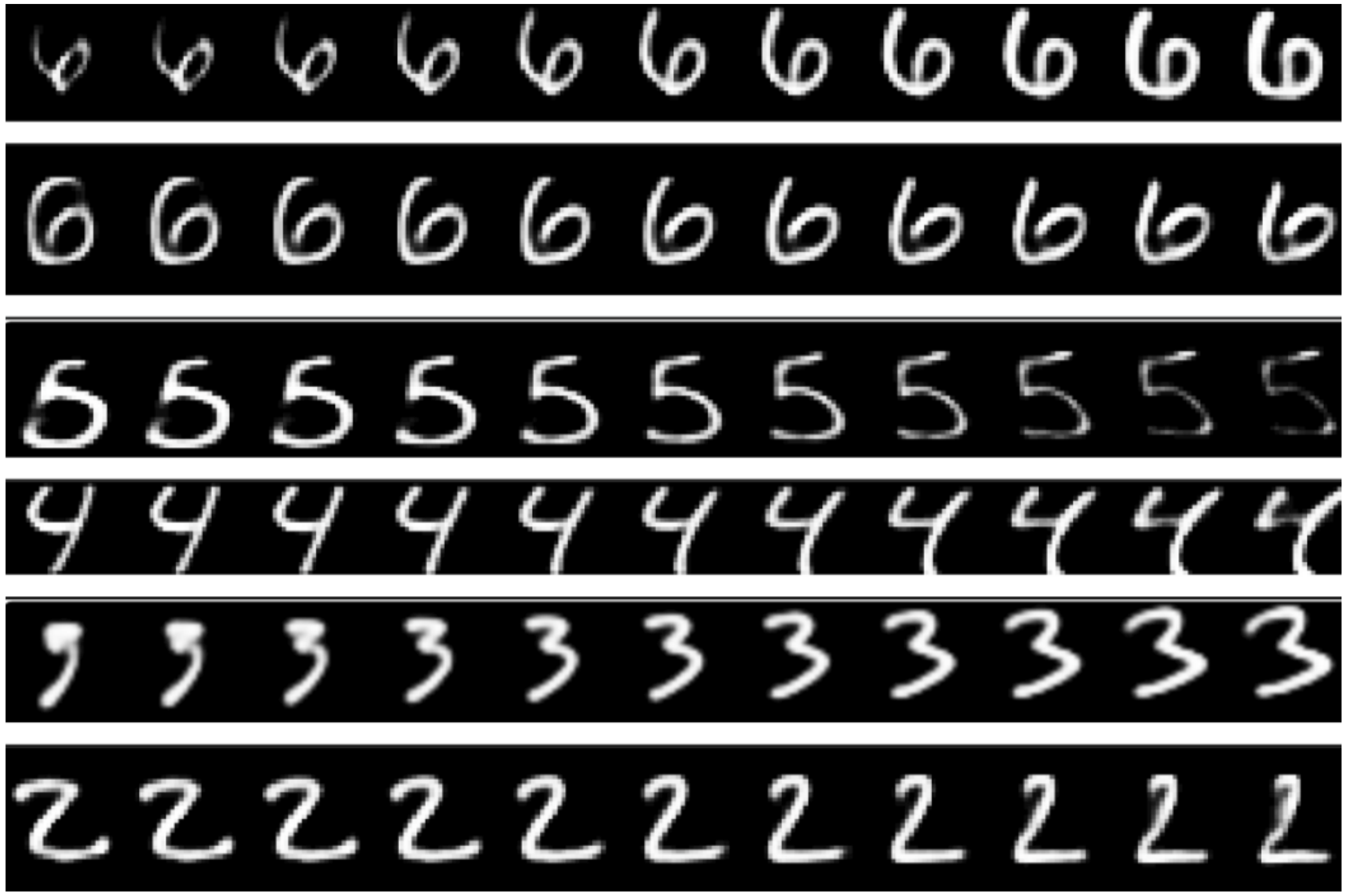
: Invariance is special case of equivariance where T_g^2 is the identity.

Viewpoint Equivariance.

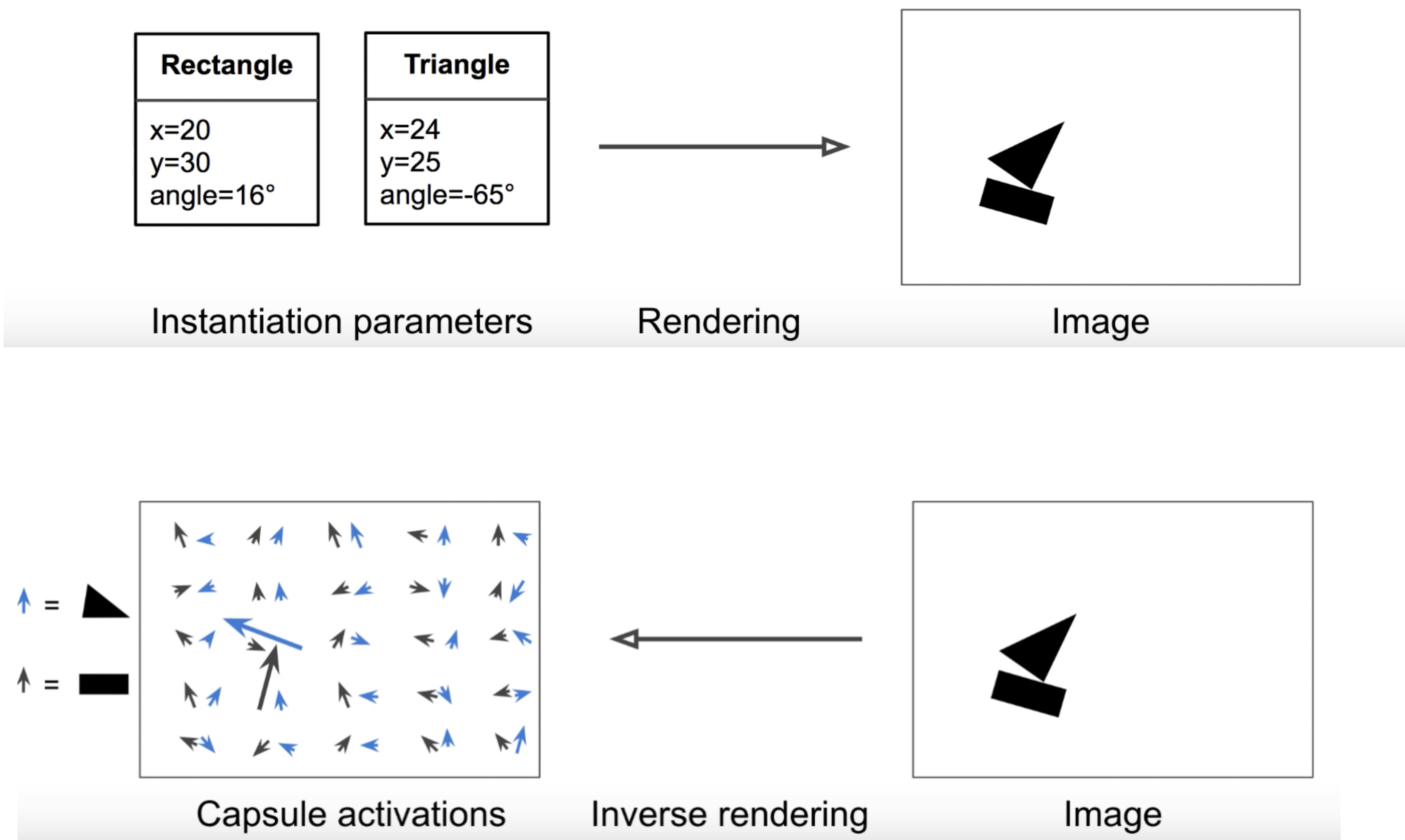
The capsule network is much better than other models at telling that images in top and bottom rows belong to the same classes, only the view angle is different.

Argument 3: Ignores underlying linear structure.

- CNNs try to conquer the variance of the viewpoint by feeding a lot of various images
- A better way to do that is to transform the image into a space in which the manifold is globally linear.
- Hinton proposed a study called **"inverse-graphics"** in order to reverse the 2D image into the desired space so that we can learn from a small amount of data and manipulate it linearly in that space.



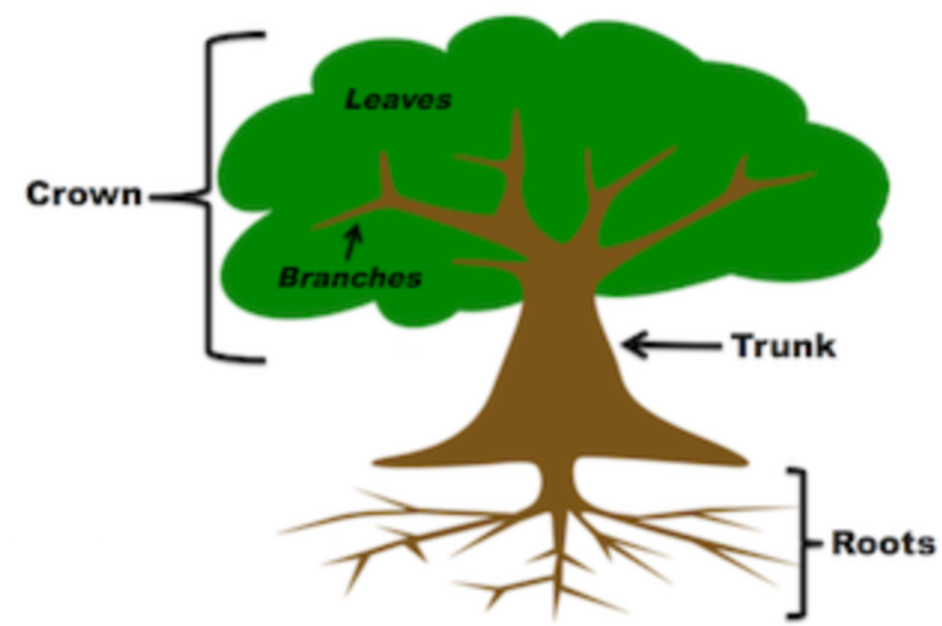
Inverse Rendering.

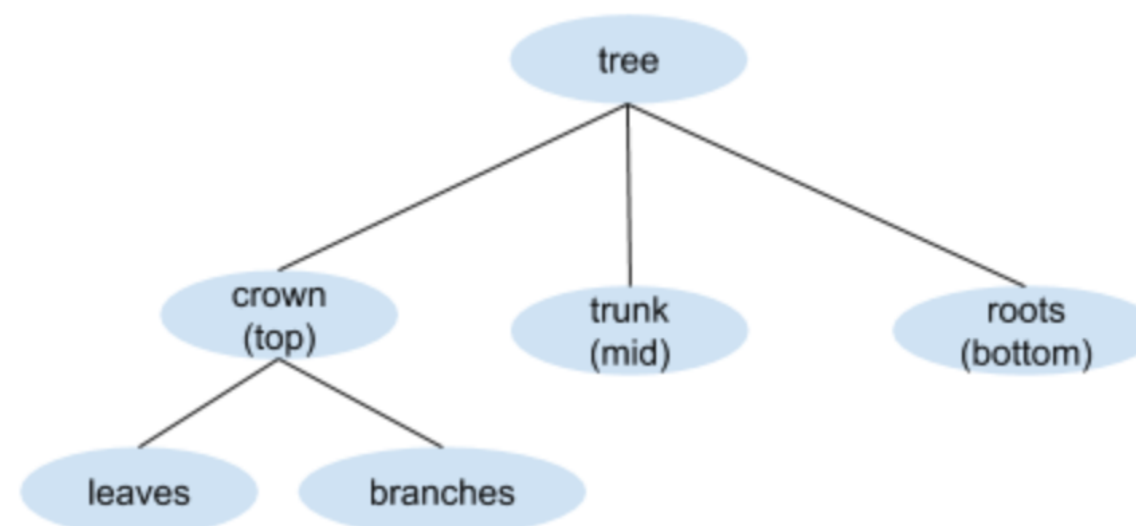


Human Visual Recognition.

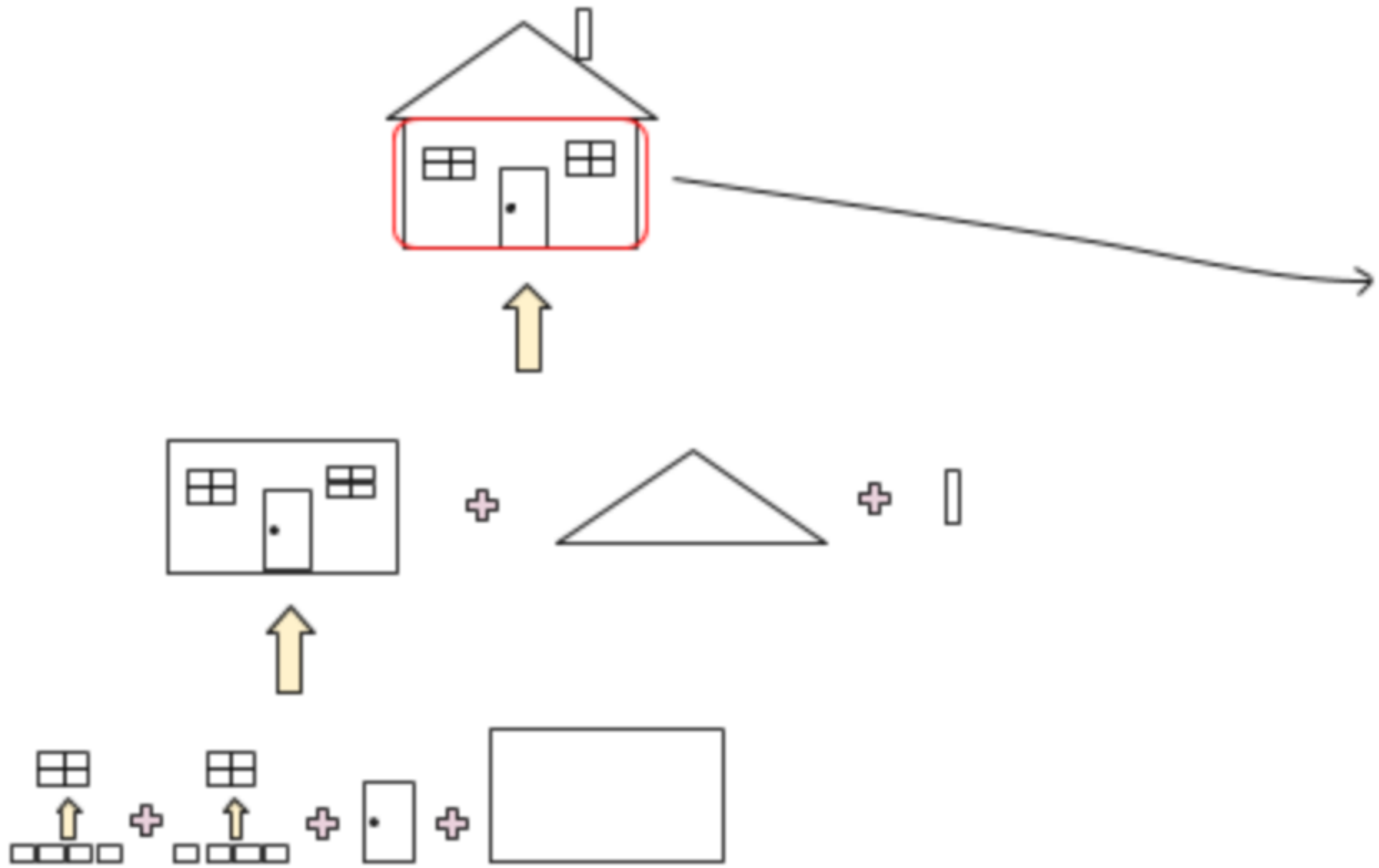
- Tree is composite of crown, trunk and roots;
- crown is composite of leaves and branches;
- HVS recognizes an object (entity) as composition of simpler (more primitive) entities, creating hierarchy of entities.

Q: Does CNN reflect HVS recognition?





Human Visual Recognition. Capsule.

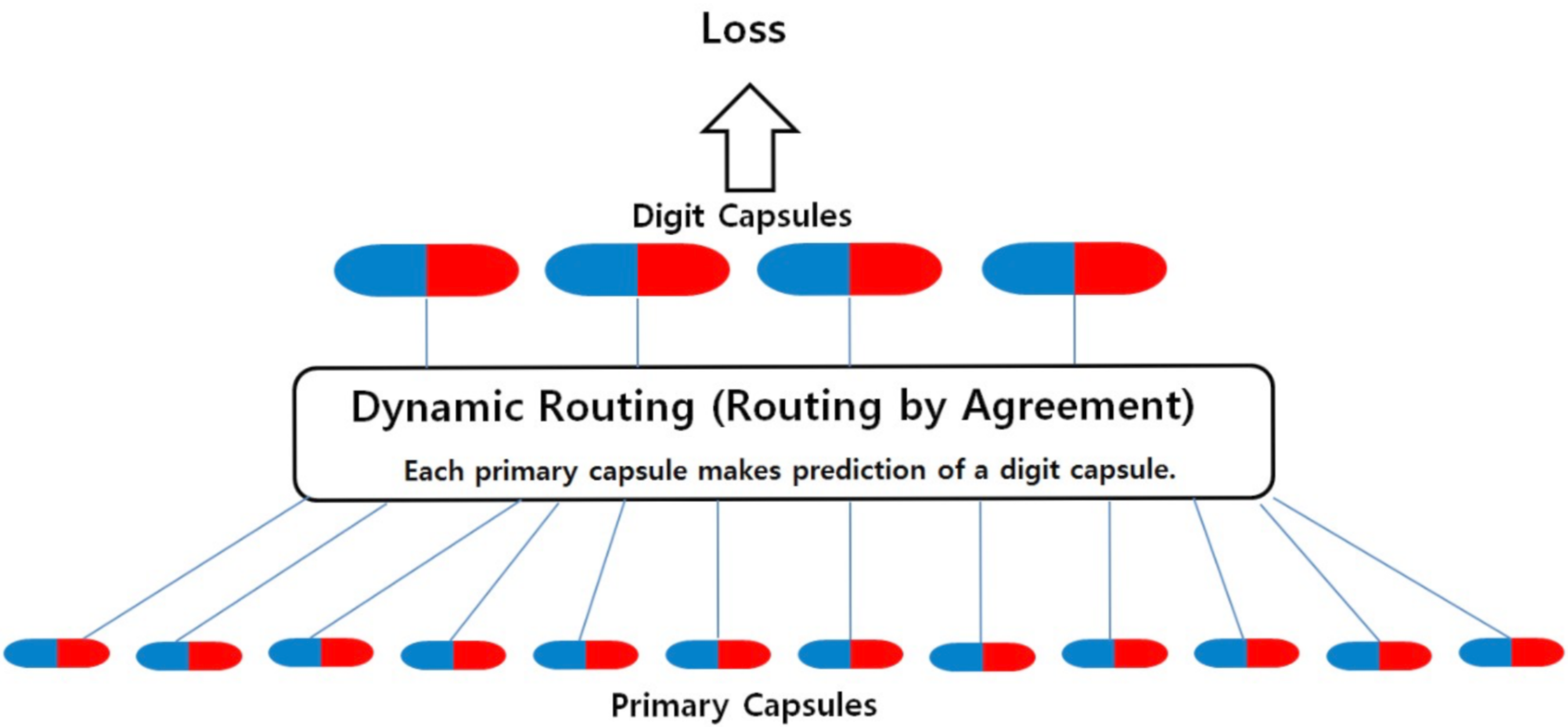


1.3	Color
0.6	Angle (x-axis)
7.4	Length
6.5	Breadth
0.5	Deformation
1.4	Texture

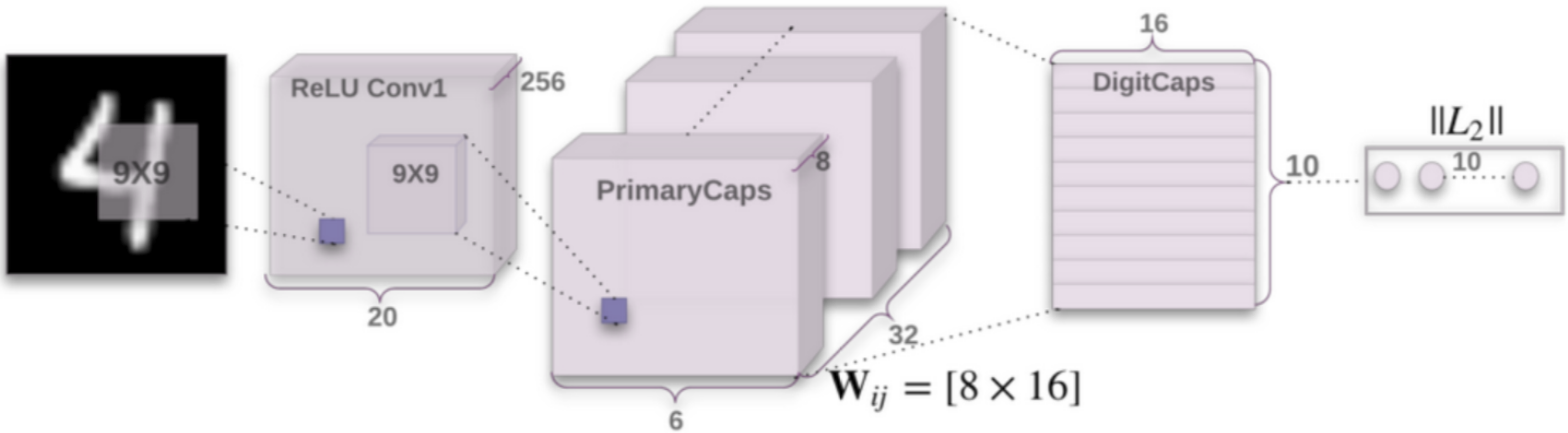
A 6-dimensional capsule corresponding to a rectangle

$$\sqrt{1.3^2 + 0.6^2 + 7.4^2 + 6.5^2 + 0.5^2 + 1.4^2} = 10.06$$

Capsule Hierarchy.



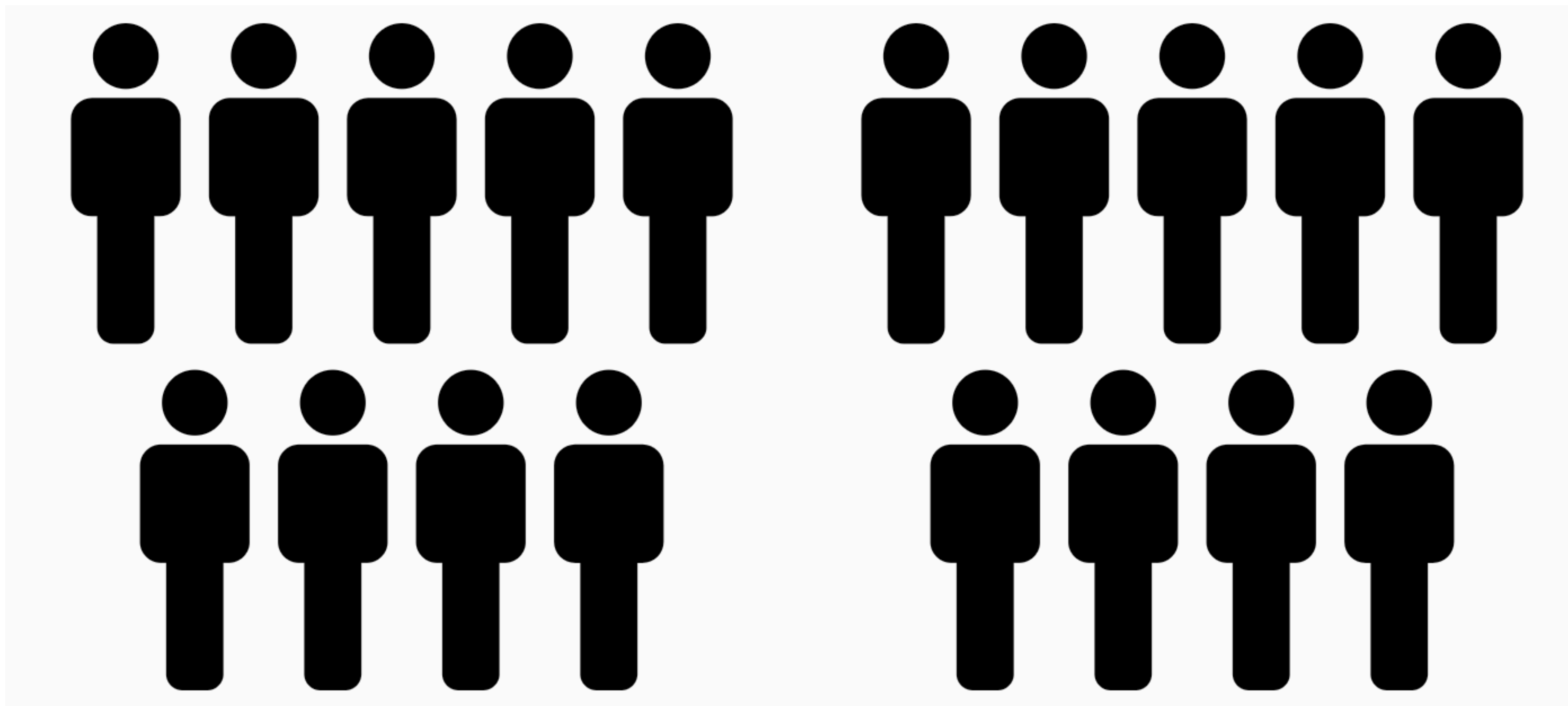
Capsule Network.



- Convolution (x 2)
- Reshape feature maps to 32 groups of 8 feature maps each of size 6 by 6 (**6x6x32=1152 primary capsules**)
- Dynamic Routing
- Digit Caps (higher level capsules of size 16x1)
- Compute Loss
- Backpropagate

Real Estate Agents Analogy.

1152 real estate agents



Geoffrey

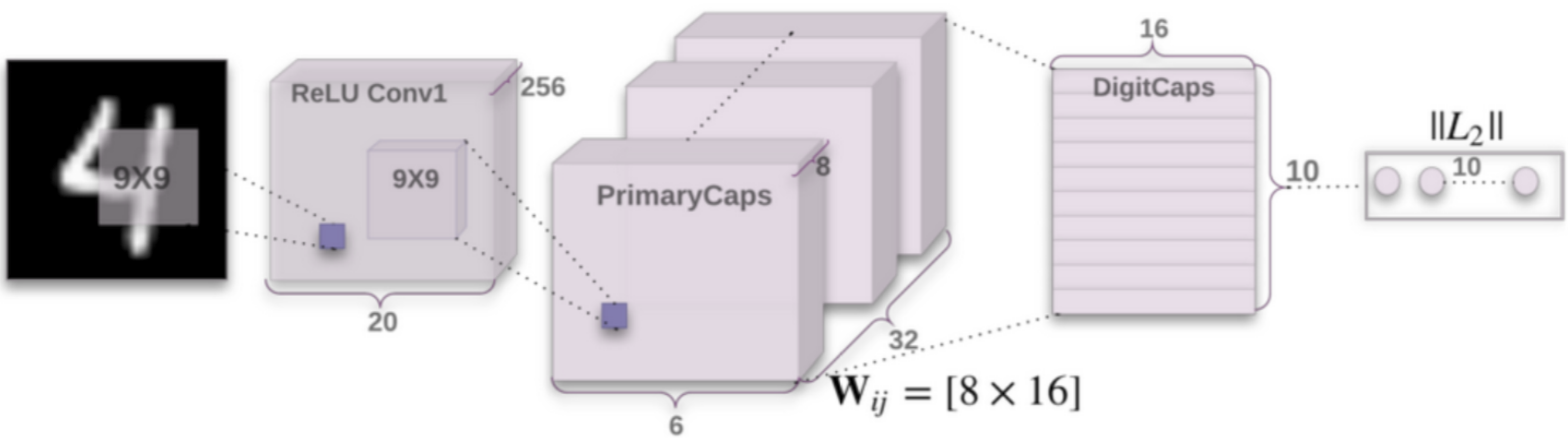
? ? ? ? ? ? ? ? ? ?

Dynamic Routing.

$$\mathbf{s}_j = \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}\ , \qquad \hat{\mathbf{u}}_{j|i} = \mathbf{W}_{ij}\mathbf{u}_i \qquad c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}$$

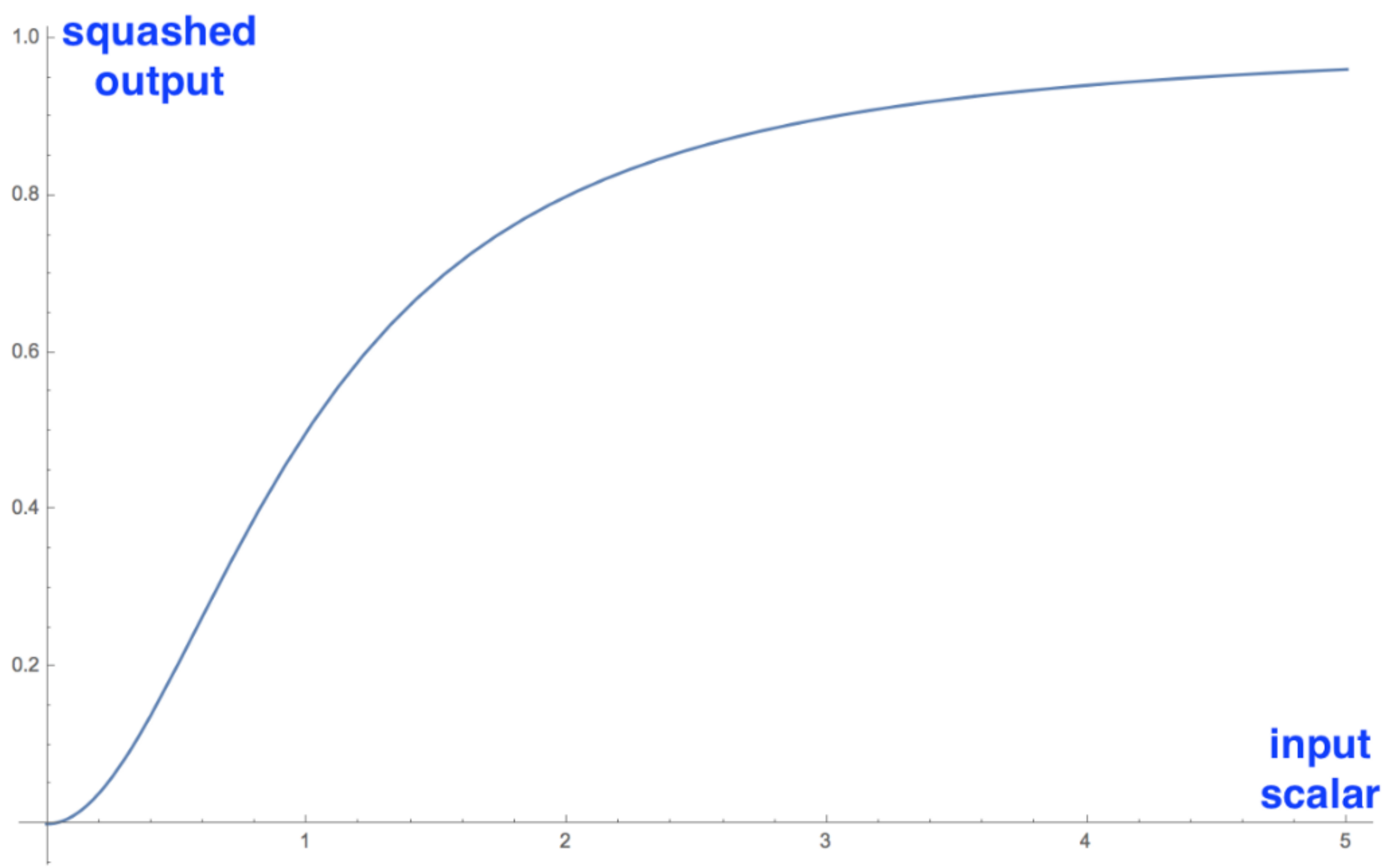
Procedure 1 Routing algorithm.

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: $\mathbf{c}_i \leftarrow \mathtt{softmax}(\mathbf{b}_i)$ ▷ `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 $(l+1)$: $\mathbf{v}_j \leftarrow \mathtt{squash}(\mathbf{s}_j)$ ▷ `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 $\mathbf{v}_j$



Squashing Function.

$$\mathbf{v}_j = \underbrace{\boxed{\frac{||\mathbf{s}_j||^2}{1+||\mathbf{s}_j||^2}}}_{\text{additional "squashing"}} \underbrace{\boxed{\frac{\mathbf{s}_j}{||\mathbf{s}_j||}}}_{\text{unit scaling}}$$



Dynamic Routing.

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**

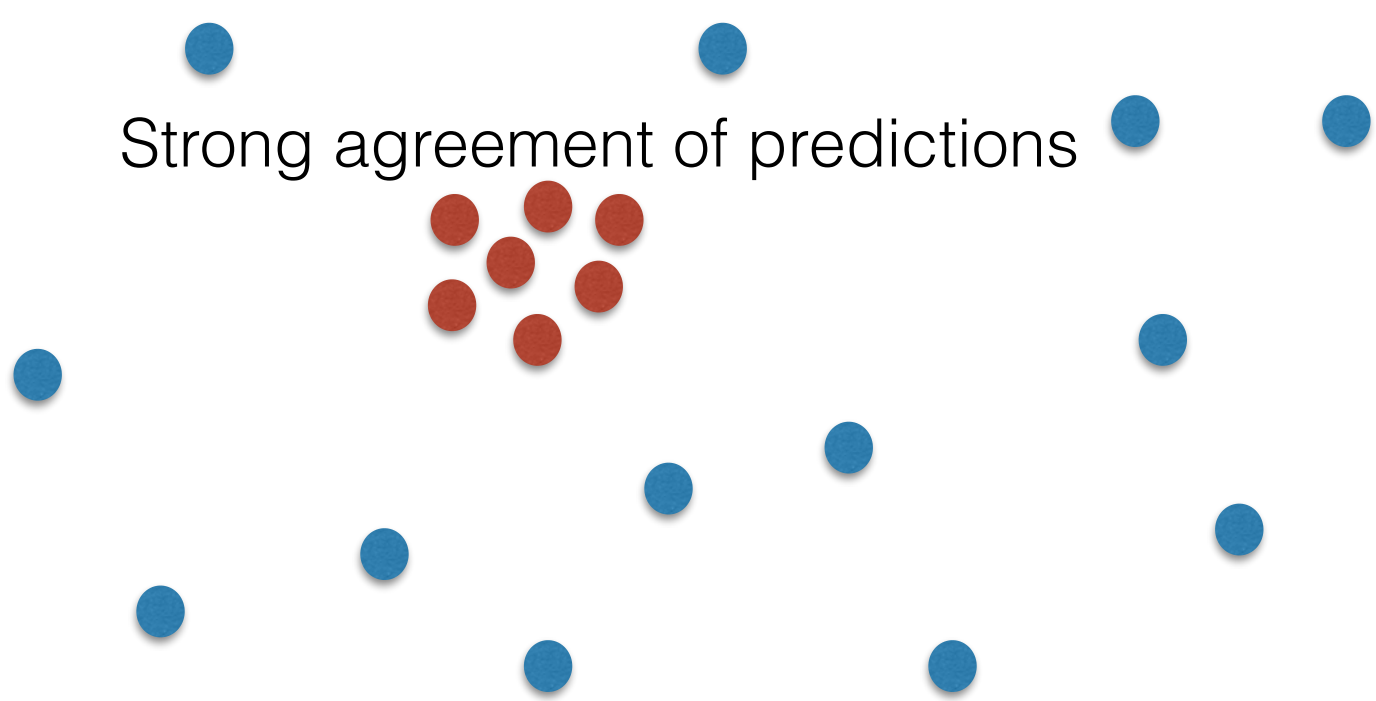
    for all capsule i in layer l: $\mathbf{c}_i \leftarrow \mathtt{softmax}(\mathbf{b}_i)$ ▷ `softmax` computes Eq. 3

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

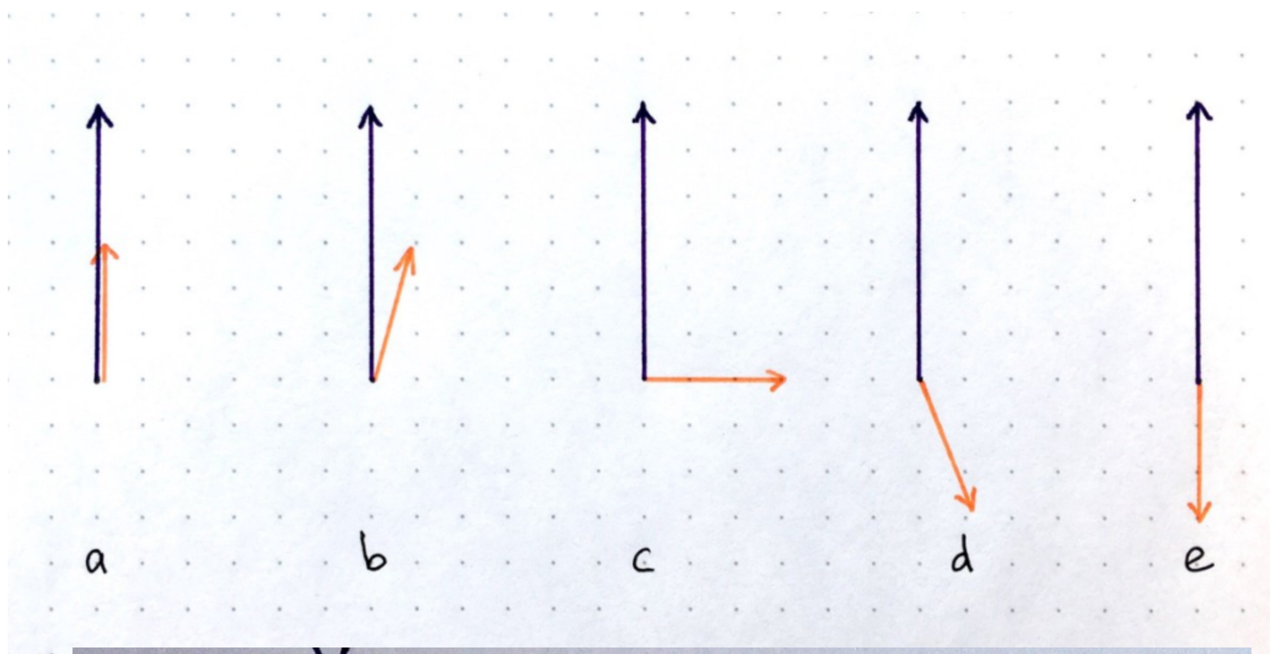
    for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow \mathtt{squash}(\mathbf{s}_j)$ ▷ `squash` computes Eq. 1

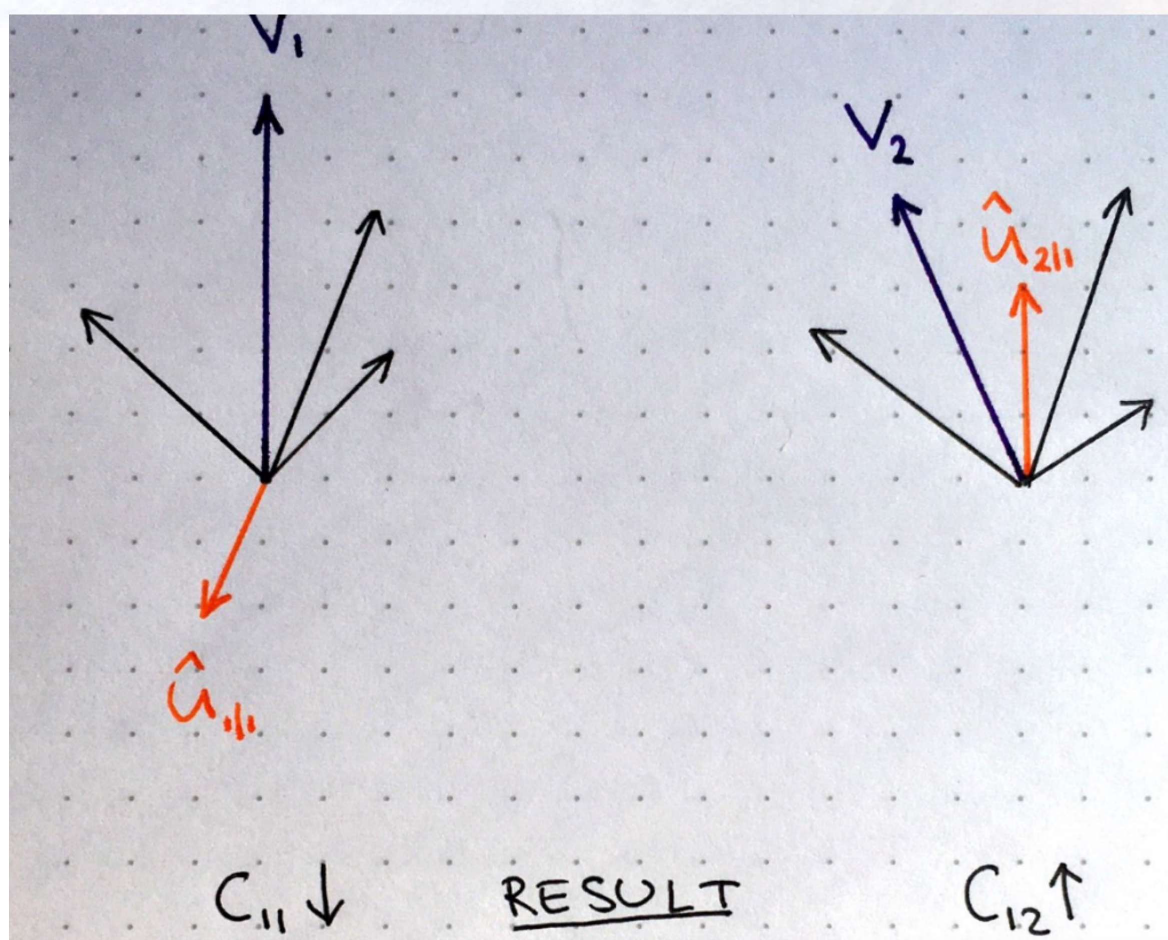
    for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.\mathbf{v}_j$

  return $\mathbf{v}_j$



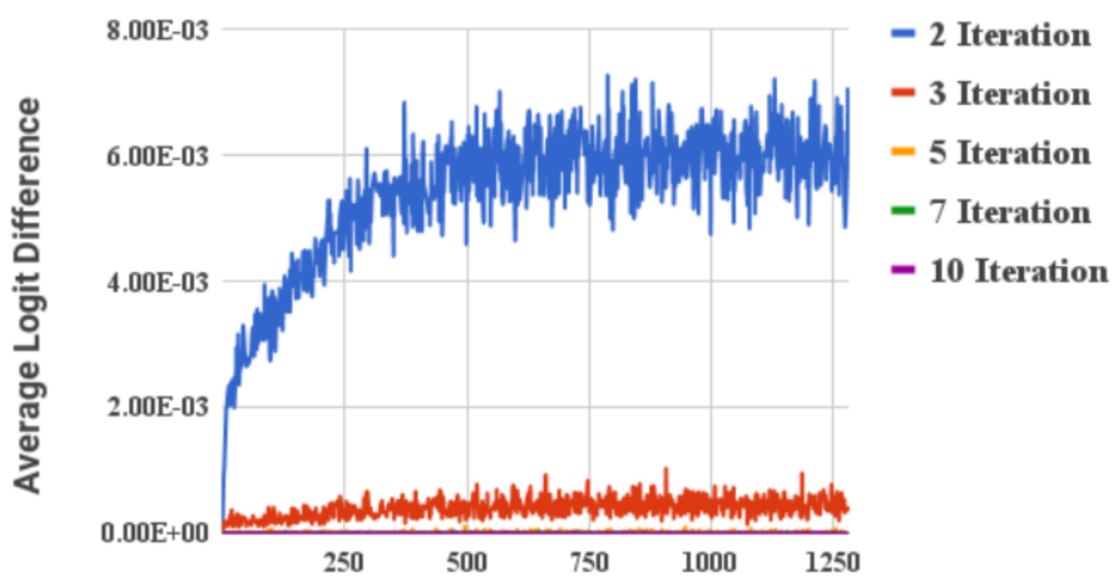
Dynamic Routing.



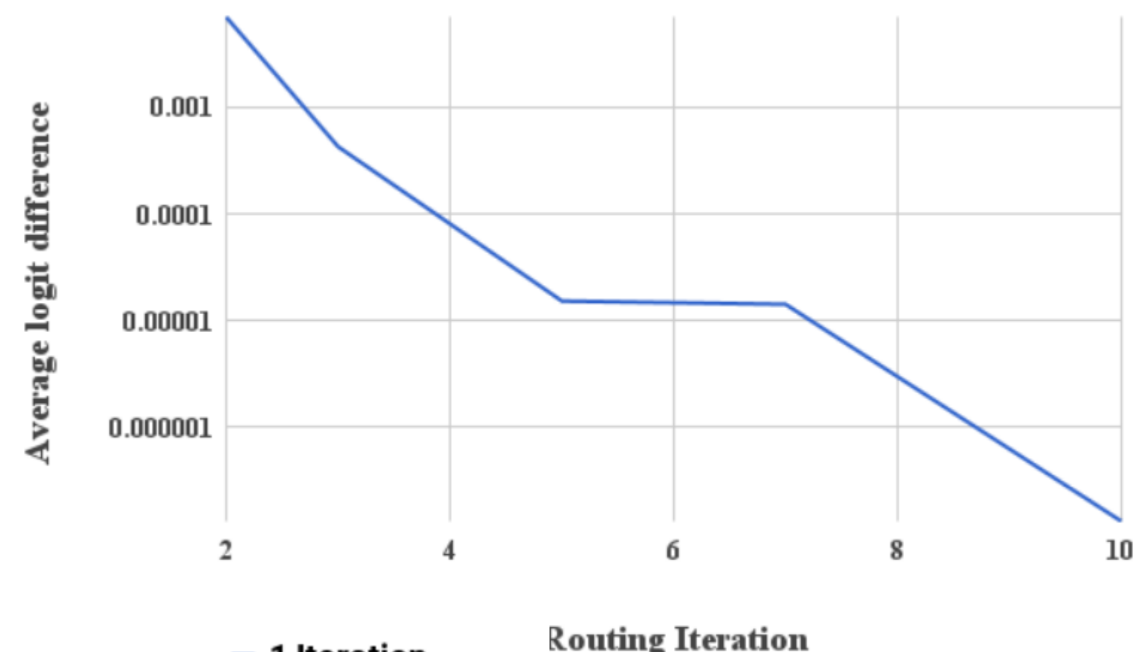


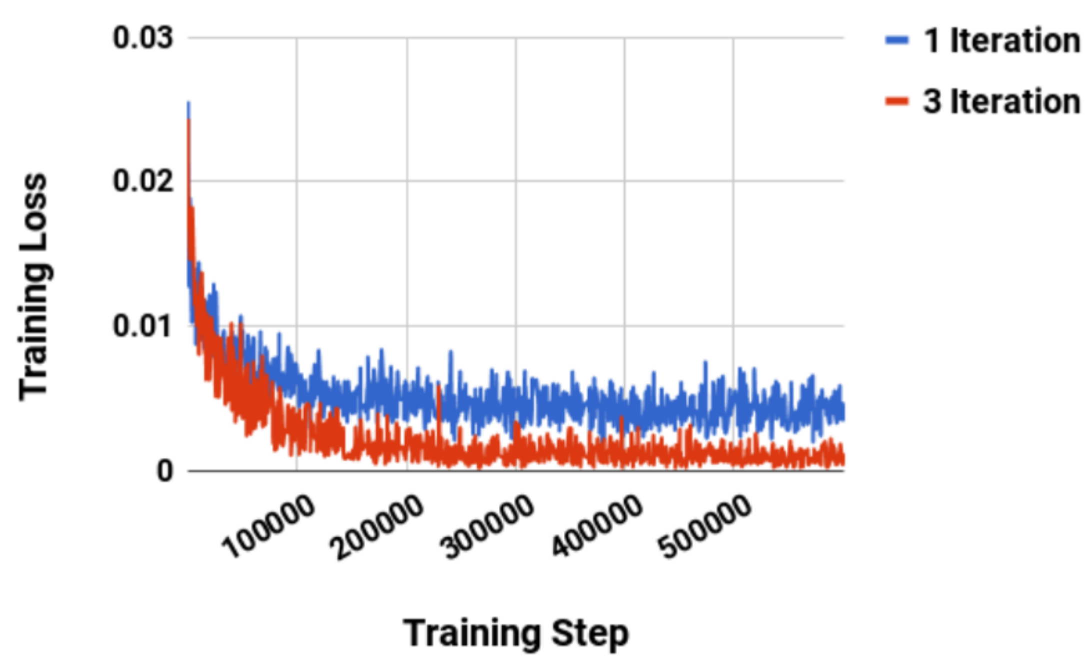
Training: How many routing iterations to use?

(a) During training.



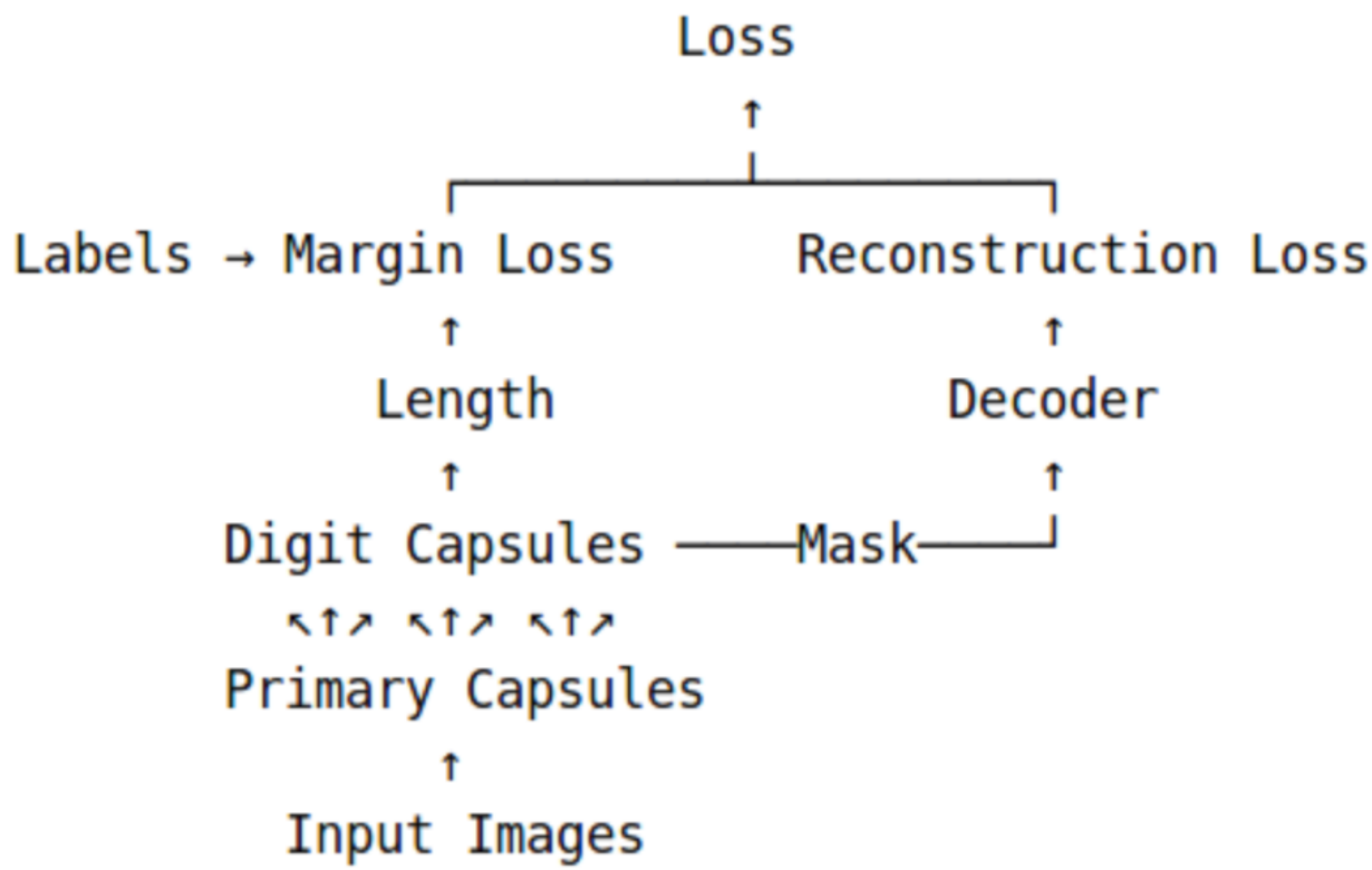
(b) Log scale of final differences.





we suggest 3 iteration of routing for all experiments.

Training.


Loss
Labels
Margin Loss
Reconstruction Loss
Length
Decoder
Digit Capsules
Mask
Primary Capsules
Input Images

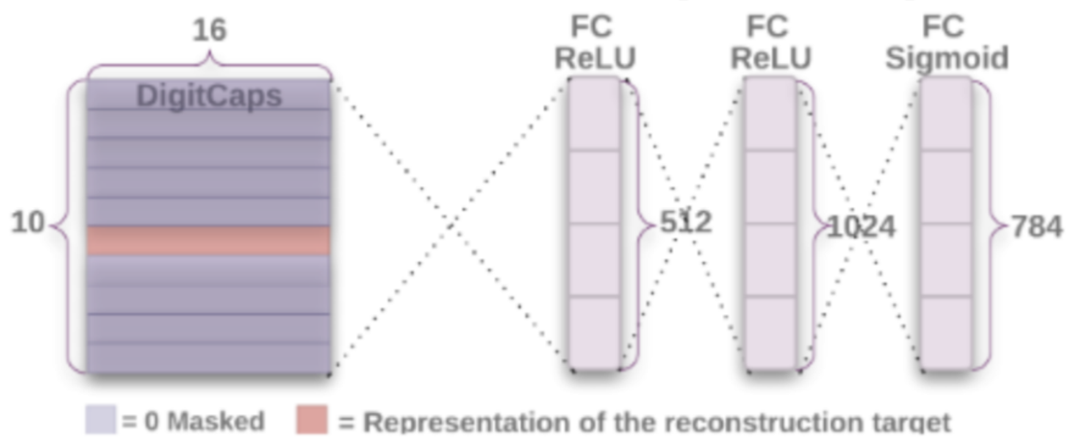
Training.

Loss: Marginal + Reconstruction

Marginal loss:

$$L_k = T_k \ \max(0, m^+ - ||\mathbf{v}_k||)^2 + \lambda \ (1 - T_k) \ \max(0, ||\mathbf{v}_k|| - m^-)^2$$

Reconstruction (MSE) loss:



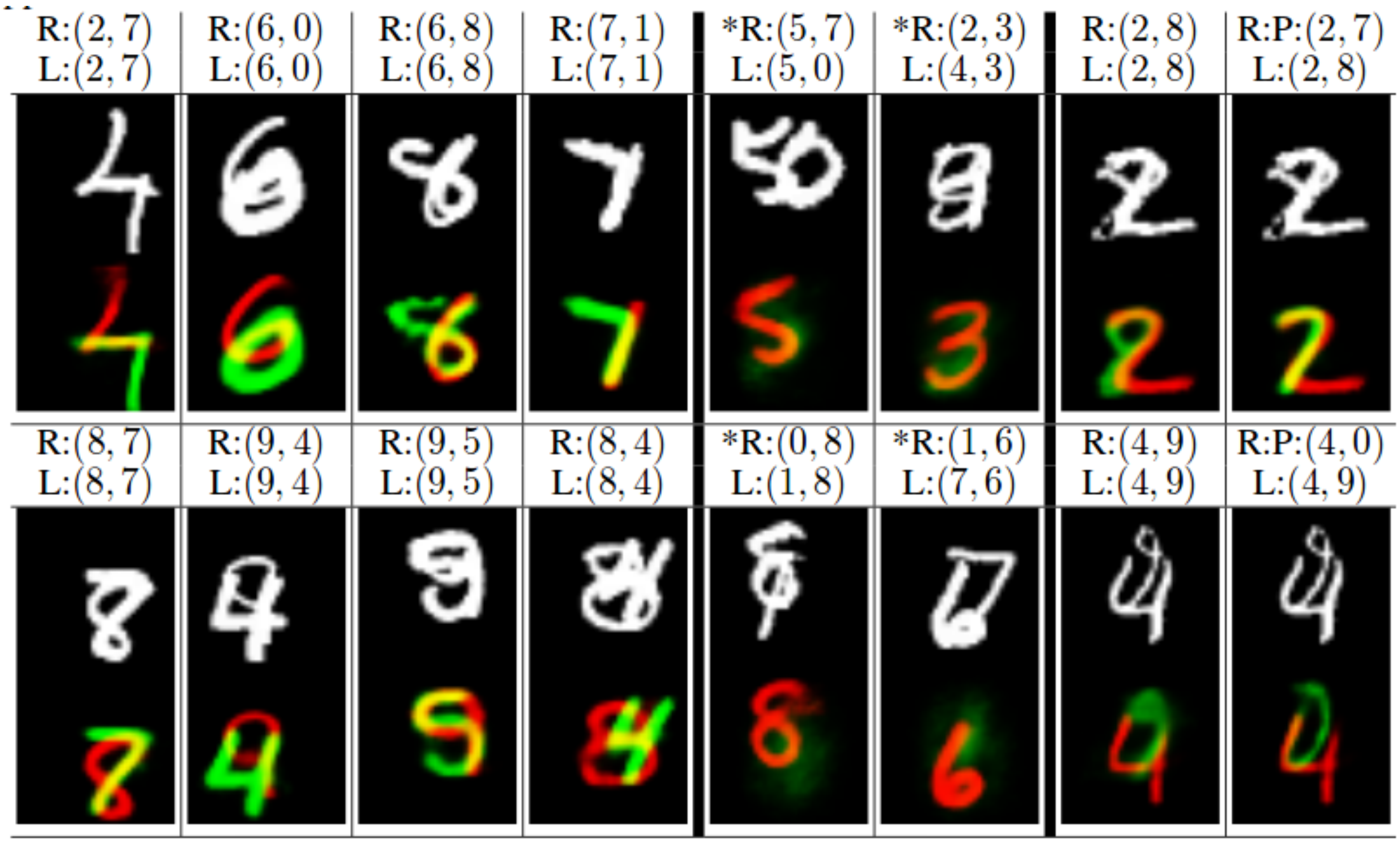
total_loss = marginal + 0.0005*recon_loss.

What do individual dimensions of capsule represent?

Scale and thickness	
Localized part	
Stroke thickness	
Localized skew	
Width and translation	
Localized part	

Each row shows the reconstruction when one of the16 dimensions in the DigitCaps representation is tweaked by intervals of 0.05in the range [–0.25,0.25].

Segmenting highly overlapping digits.



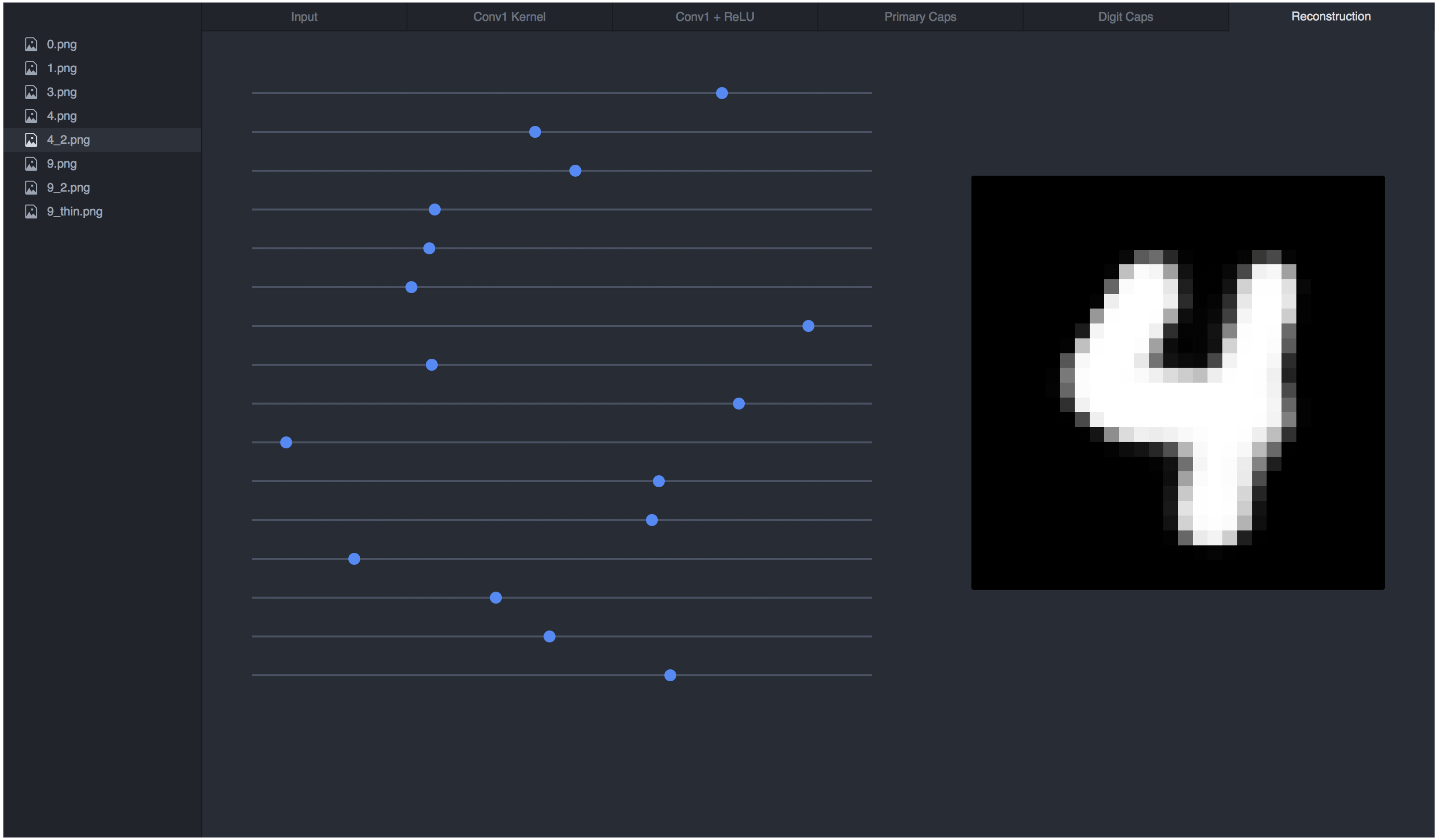
PROS.

- ☑ Reaches high accuracy in MNIST, promising results for CIFAR10;
- ☑ Requires less training data
- ☑ Position and pose information is preserved
- ☑ Promising for image segmentation and detection
- ☑ Dynamic Routing works well for overlapping objects
- ☑ Capsule activations nicely map the hierarchy of parts

Arguments against Capsule Networks.

- How about more complex data?
- Complexity
- No comparison with other architectures.

Visualization.



https://github.com/bourdakos1/CapsNet-Visualization