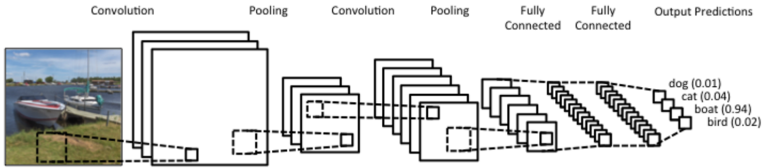
Dynamic Routing Between Capsules

Sara Sabour, Nicholas Frosst and Geoffrey E. Hinton
(Google Brain, Toronto)

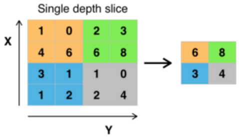
presented by
William Weimin Yoo
(Leiden University)

National NIPS Debriefing, 2018

What is wrong with Convolutional Neural Nets?



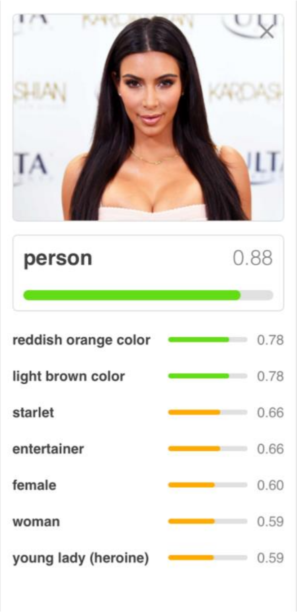
1. Does not encode spatial relationships, e.g., orientation, rotations etc.
2. Max pooling looks only at most active neuron, location information is not recorded ⇒ positional invariance



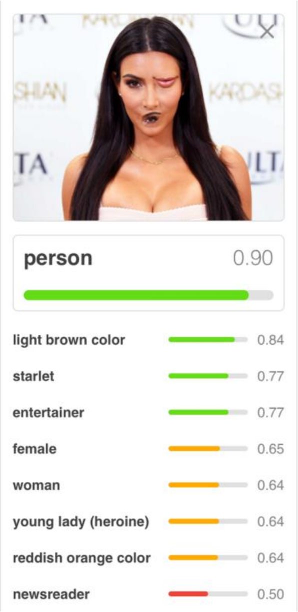
1	0	2	3
4	6	6	8
3	1	1	0
1	2	2	4

→

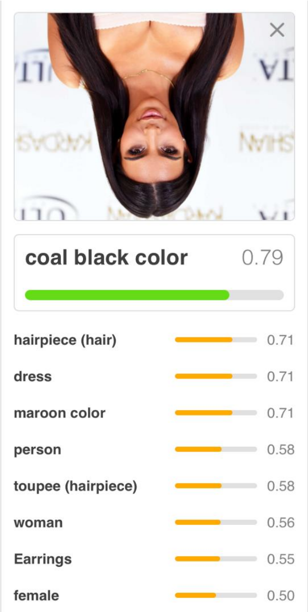
6	8
3	4



person	0.88
reddish orange color	0.78
light brown color	0.78
starlet	0.66
entertainer	0.66
female	0.60
woman	0.59
young lady (heroine)	0.59



person	0.90
light brown color	0.84
starlet	0.77
entertainer	0.77
female	0.65
woman	0.64
young lady (heroine)	0.64
reddish orange color	0.64
newsreader	0.50

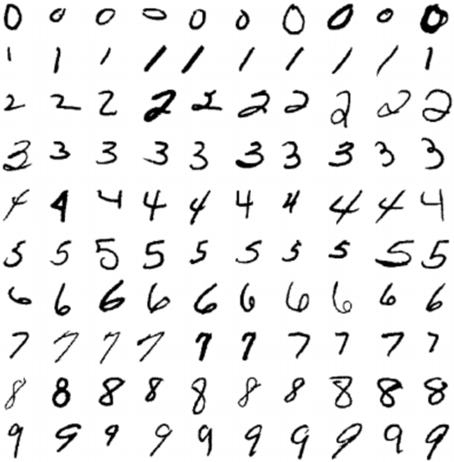


coal black color	0.79
hairpiece (hair)	0.71
dress	0.71
maroon color	0.71
person	0.58
toupee (hairpiece)	0.58
woman	0.56
Earrings	0.55
female	0.50

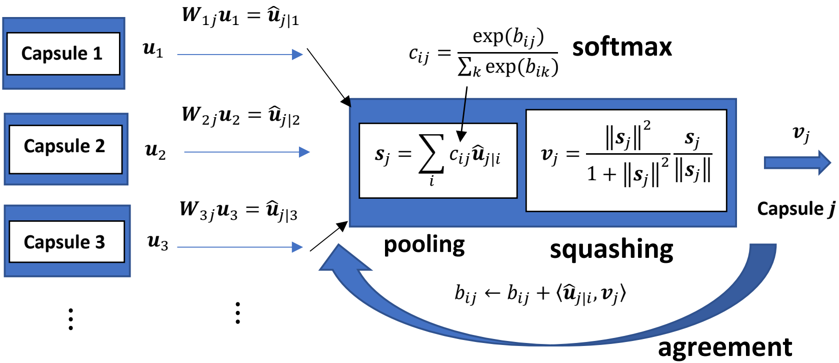
Capsules and Routing

1. Capsule is a group of neurons encoding properties (instantiation parameters) of entities in an image (e.g., circle, 1)
2. Parameters: existence, pose (position, size, orientation), deformation, hue, texture etc.
3. So unlike CNN (scalars), outputs and inputs are (high-dimensional) vectors
4. Norm of vectors encode probability of existence
5. Draws analogy from computer graphics, points on linear manifold and transformation
6. Max pooling is too crude, how to route vectors of instantiation parameters from one capsule layer to another?

MNIST



Dynamic Routing by Agreement



Capsule 1

Capsule 2

Capsule 3

⋮

$\boldsymbol{u}_1$

$\boldsymbol{u}_2$

$\boldsymbol{u}_3$

⋮

$\boldsymbol{W}_{1j}\boldsymbol{u}_1 = \hat{\boldsymbol{u}}_{j|1}$

$\boldsymbol{W}_{2j}\boldsymbol{u}_2 = \hat{\boldsymbol{u}}_{j|2}$

$\boldsymbol{W}_{3j}\boldsymbol{u}_3 = \hat{\boldsymbol{u}}_{j|3}$

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}$$

softmax

$$\boldsymbol{s}_j = \sum_i c_{ij}\hat{\boldsymbol{u}}_{j|i}$$

pooling

$$\boldsymbol{v}_j = \frac{\|\boldsymbol{s}_j\|^2}{1+\|\boldsymbol{s}_j\|^2}\frac{\boldsymbol{s}_j}{\|\boldsymbol{s}_j\|}$$

squashing

$\boldsymbol{v}_j$

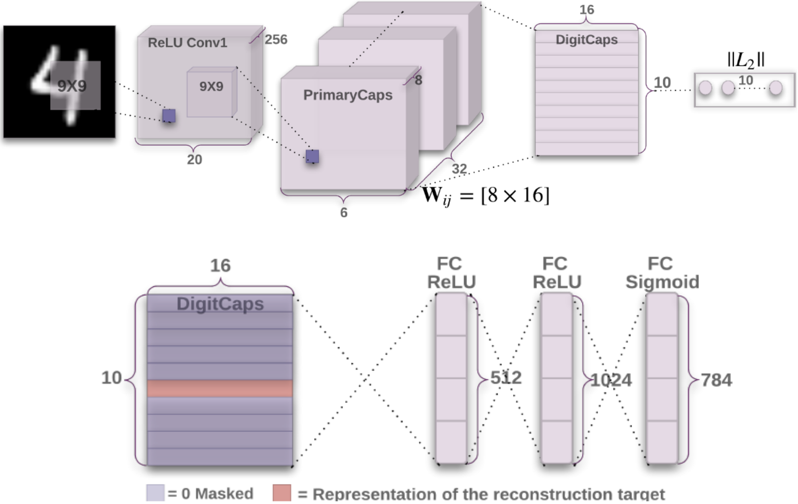
Capsule *j*

$$b_{ij} \leftarrow b_{ij} + \langle \hat{\boldsymbol{u}}_{j|i}, \boldsymbol{v}_j \rangle$$

agreement

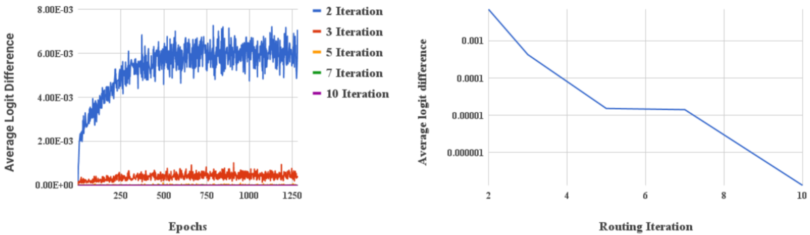
CapsNet Architecture

1 Conv. layer + 2 Capsule layers + 3 Fully connected



How many routing iterations?

Average change of b_{ij}(routing logit) vs. routing iterations:



Stabilizes after 500 epochs of training.
Negligible change (1×10^{-5}) by 5 iterations.

Margin Loss for Digit Existence

L_k: Margin loss for digit capsule k
$\boldsymbol{v}_k$: Output from digit capsule k

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

$T_k = 1$ iff digit k present
$m^+ = 0.9$ and $m^- = 0.1$
$\lambda = 0.5$

▷ Train using backpropagation (Rumelhart, 1986) with Adam optimizer (Kingma and Ba 2014) to minimize $\sum_{k=1}^{10} L_k$
Implemented using Tensor Flow (Abadi et al. 2016)

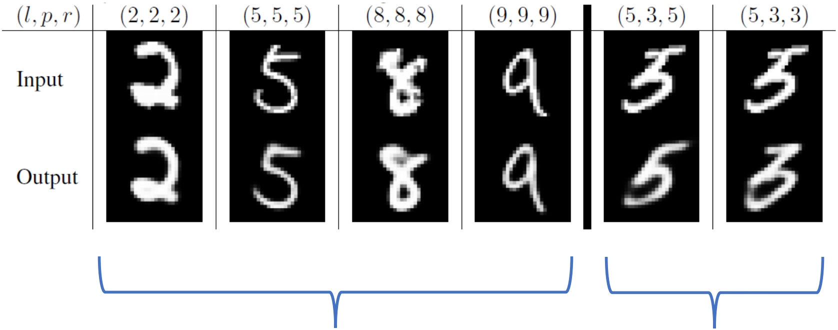
Total loss $=$ Margin loss $+0.0005 \times$(Reconstruction loss)

MNIST Reconstructions

l: label

p: prediction

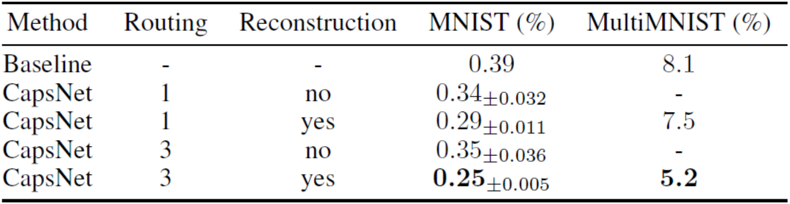
r: target reconstruction



Correct with noise smoothing

Wrong and confused

MNIST Classification Results



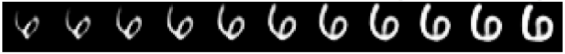
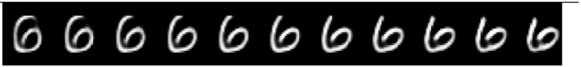
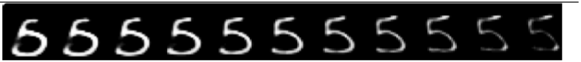
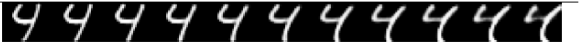
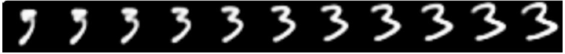
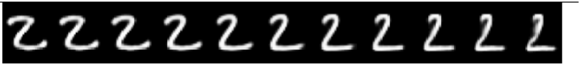
Method	Routing	Reconstruction	MNIST (%)	MultiMNIST (%)
Baseline	-	-	0.39	8.1
CapsNet	1	no	$0.34_{\pm 0.032}$	-
CapsNet	1	yes	$0.29_{\pm 0.011}$	7.5
CapsNet	3	no	$0.35_{\pm 0.036}$	-
CapsNet	3	yes	$\mathbf{0.25}_{\pm 0.005}$	**5.2**

- Baseline is 3 Conv. layer CNN + 2 FC (dropout) + softmax
- Designed so that comp. cost ≈ CapsNet
- Parameters: Baseline 35.4M vs. CapsNet 8.2M

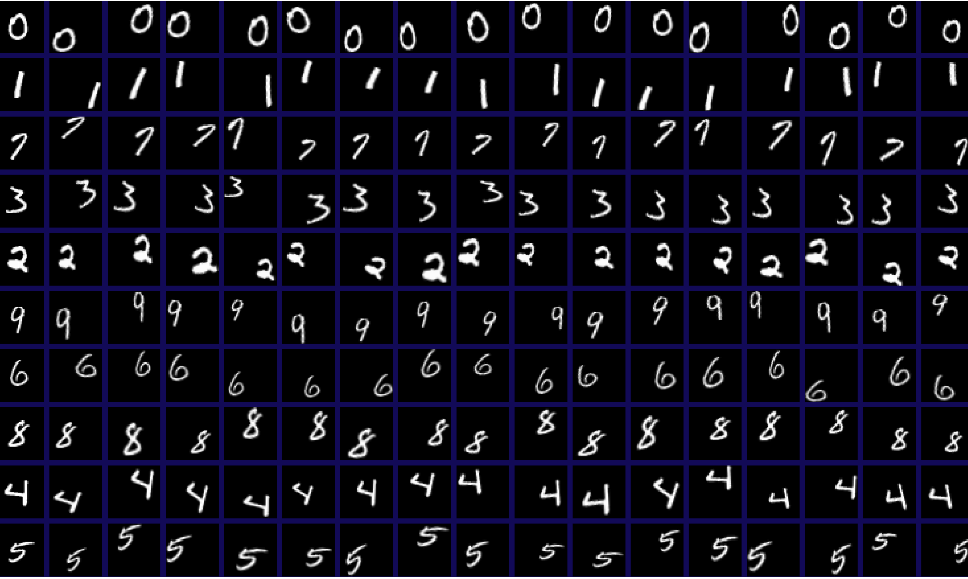
⇒ CapsNet can achieve state-of-art performance with relatively shallow (3) network

Dimension Perturbations

dth DigitCaps $(16 \times 1) + [-0.25, -0.2, \ldots, 0, \ldots, 0.2, 0.25]$ for $d = 1, \ldots, 6$:

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

AffNIST



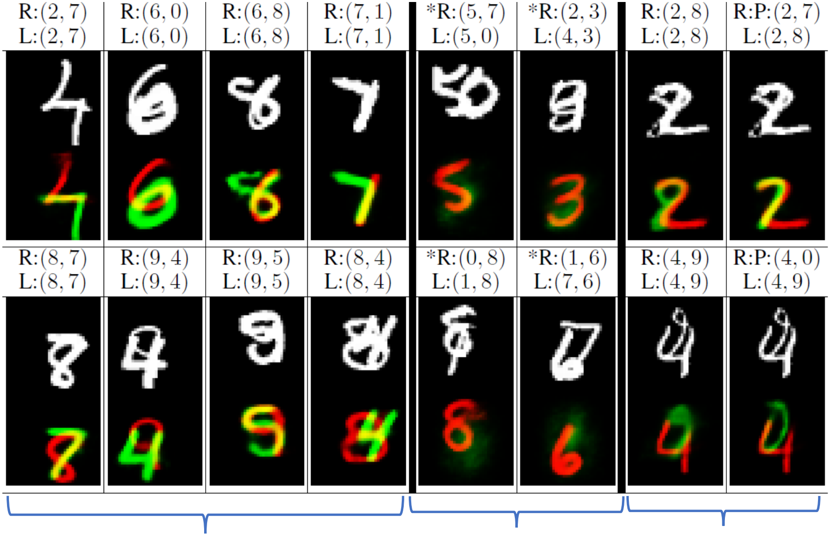
Robust to Affine Transformations

expanded MNIST: digits placed randomly on 40×40 pixels
affNIST: MNIST + random small affine transformation

	expanded MNIST	**affNIST**
CapsNet	99.23%	79%
CNN	99.22%	66%

Table: Test errors

MultiMNIST



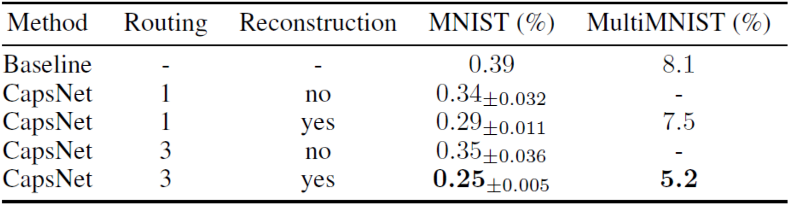
Correct

Reconstruction
not from label
or prediction

Wrong

MultiMNIST Classification Results

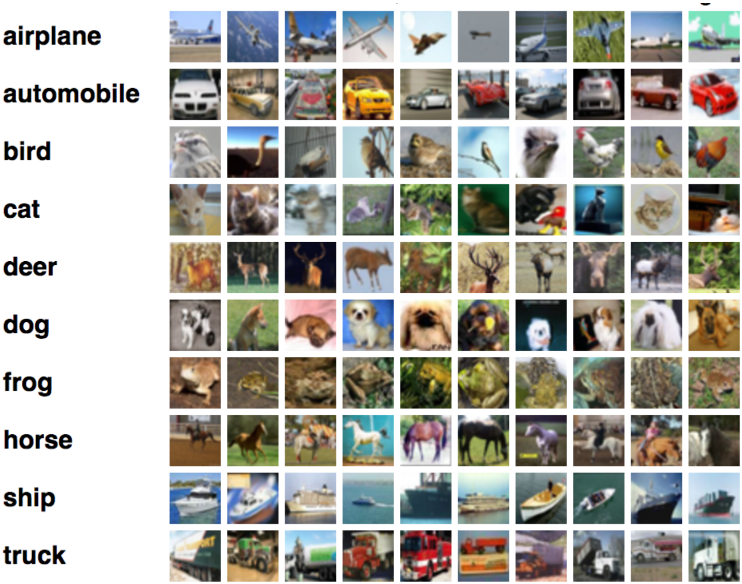
Take learning rate $10\times$ larger because training dataset is larger



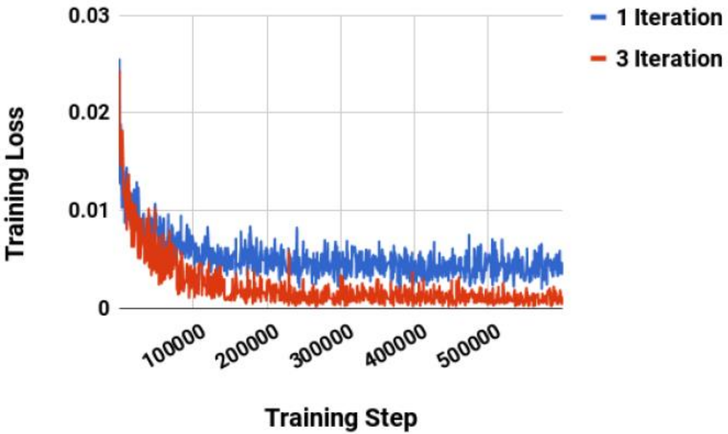
Method	Routing	Reconstruction	MNIST (%)	MultiMNIST (%)
Baseline	-	-	0.39	8.1
CapsNet	1	no	$0.34_{\pm 0.032}$	-
CapsNet	1	yes	$0.29_{\pm 0.011}$	7.5
CapsNet	3	no	$0.35_{\pm 0.036}$	-
CapsNet	3	yes	$\mathbf{0.25}_{\pm 0.005}$	**5.2**

- Baseline CNN: 2Conv. layer+2 FC (ReLU and pooling in between)
- Parameters: 24.56 M for CNN vs. 11.36 M CapsNet
- State-of-the-art performance on segmentation

CIFAR10: $10.6\% \approx$ CNN (Zeiler and Fergus 2013)

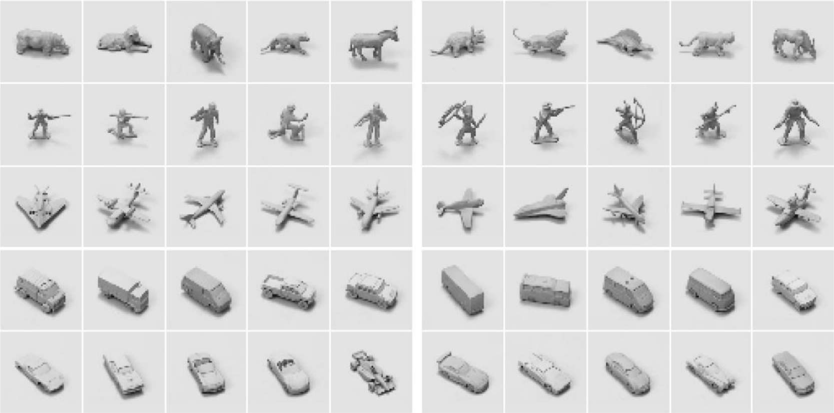


CIFAR10 Iterations (3 recommended)



smallNORB

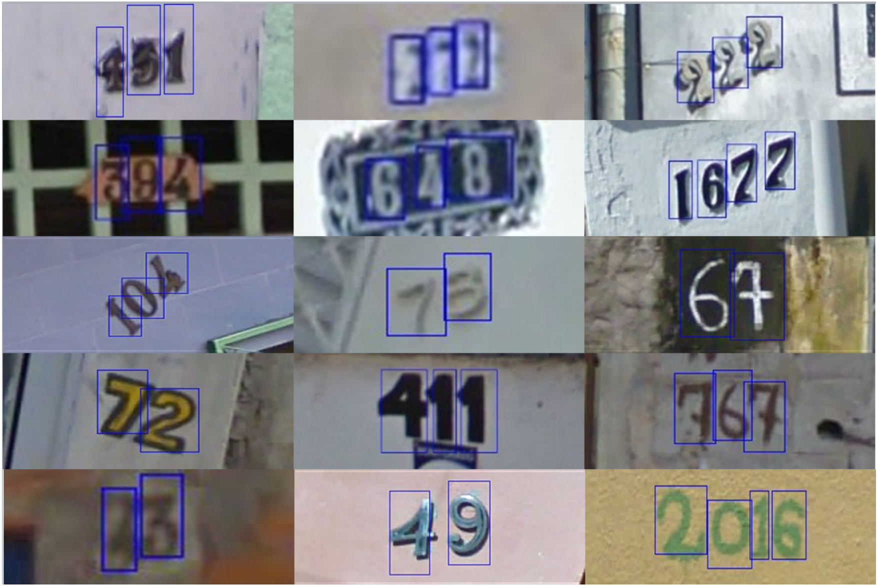
2.7% error ≈ state-of-the-art Cireșan et al. 2011



Training instances

Test instances

SVHN: 4.3% error



Tensor Flow implementation in Sara Sabour's GitHub:
https://github.com/Sarasra/models/tree/master/research/capsules

New (follow-up) paper for ICLR 2018 by the same authors:
Matrix Capsules with EM Routing