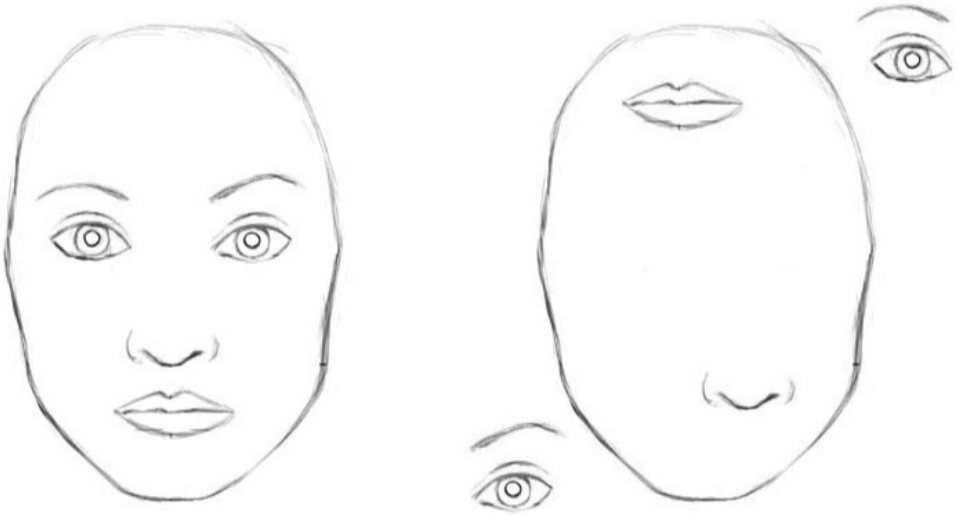
DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Cápsulas [CapsNets]

Fernando Berzal, berzal@acm.org

Limitaciones de las redes convolutivas

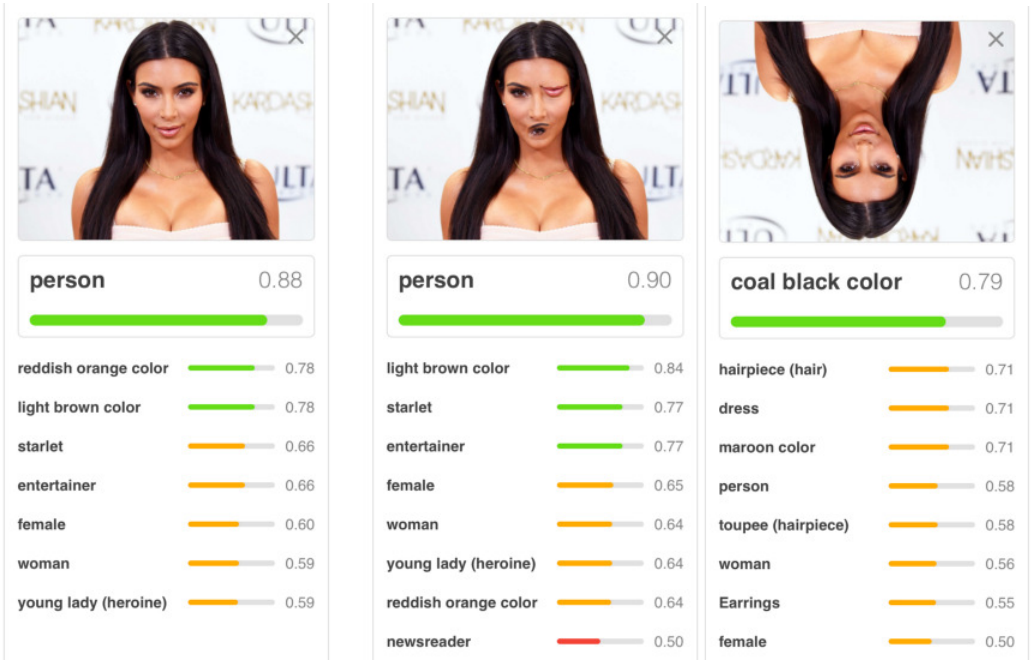
- Las redes convolutivas [CNNs] funcionan muy bien en la práctica, pero…



- … para una CNN, ambas imágenes son similares ☹

Limitaciones de las redes convolutivas

"Convolutional neural networks are doomed"
—Geoffrey Hinton

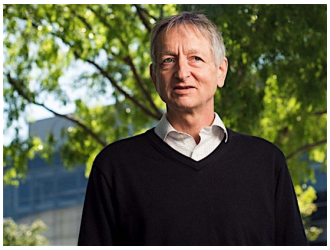


Limitaciones de las redes convolutivas

- Las redes convolutivas detectan características, pero no su colocación relativa (traslación & rotación).

- Las redes convolutivas ignoran las posiciones relativas utilizando "pooling", un apaño que funciona sorprendentemente bien en la práctica:

"The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster." – Geoffrey Hinton



Limitaciones de las redes convolutivas

Problema clave

La representación interna de una red convolutiva no tiene en cuenta las relaciones espaciales entre objetos, ni la jerarquía existente entre objetos simples y los objetos compuestos de los que forman parte.

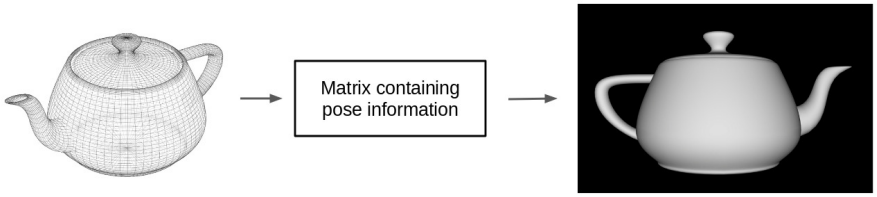
Idea de cápsula

Una red convolutiva aquí no funcionaría…

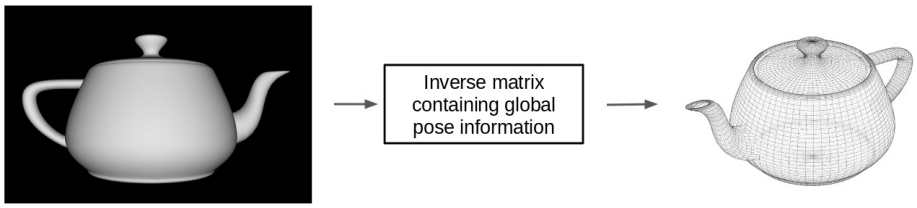
Idea de cápsula

You can think of (computer) vision as "Inverse Graphics"
- Geoffrey Hinton

Informática gráfica [rendering]



Visión artificial



Idea de cápsula

HIPÓTESIS:

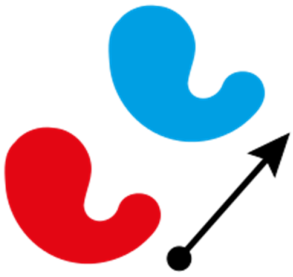
La representación de los objetos en el cerebro no depende de la perspectiva con la que recibimos información visual.

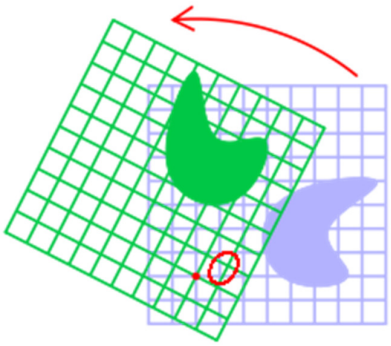
El cerebro extrae una representación jerárquica de nuestro entorno e intenta hacerla corresponder con patrones y relaciones previamente aprendidos (ya almacenados en el cerebro).

Idea de cápsula

TESIS

En Informática Gráfica, las relaciones entre objetos 3D se representan mediante transformaciones afines (su pose, básicamente traslación y rotación).



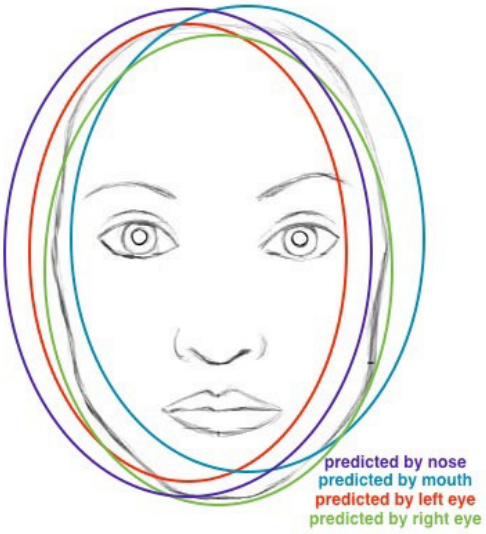


Idea de cápsula

Si las relaciones espaciales las modelamos en la representación interna de la red neuronal, resultaría muy sencillo para un modelo comprender que está viendo una vista diferente de algo que ya ha visto anteriormente…

CapsNet

Idea de cápsula

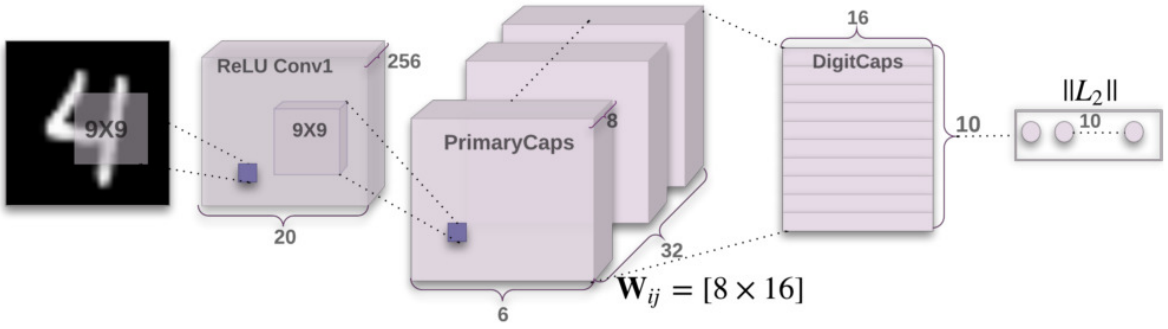


Arquitectura

ABSTRACT

"A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part. We use the length of the activity vector to represent the probability that the entity exists and its orientation to represent the instantiation parameters. Active capsules at one level make predictions, via transformation matrices, for the instantiation parameters of higher-level capsules. When multiple predictions agree, a higher level capsule becomes active."

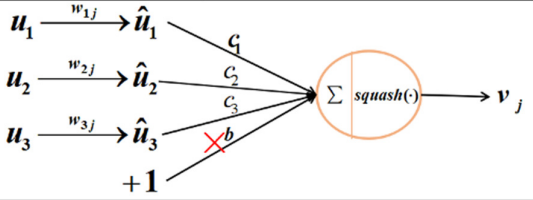
Arquitectura

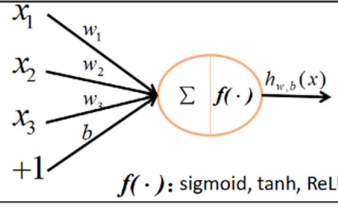


Arquitectura de una red de cápsulas [CapsNet].

Cápsula

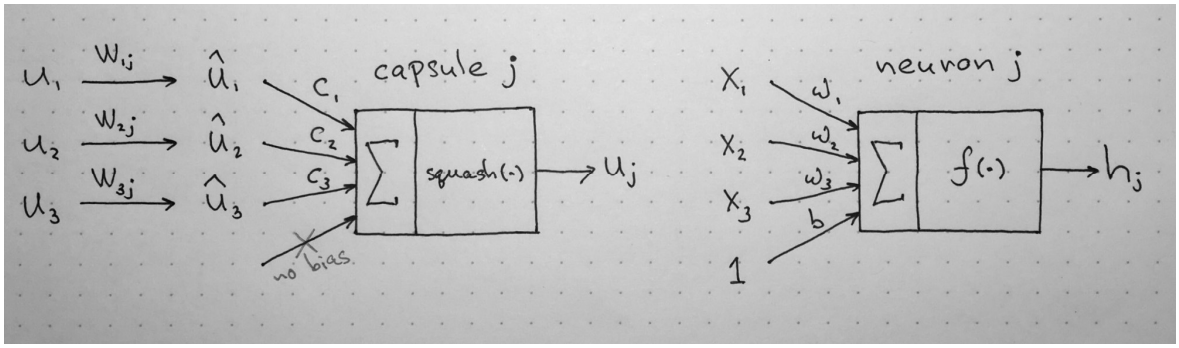
		capsule	VS.	traditional neuron
Input from low-level neuron/capsule		vector($\boldsymbol{u}_i$)		scalar(x_i)
Operation	Affine Transformation	$\hat{u}_{j\|i} = W_{ij} u_i$ (Eq. 2)		–
	Weighting	$s_j = \sum_i c_{ij} \hat{u}_{j\|i}$ (Eq. 2)		$a_j = \sum_{i=1}^{3} W_i x_i + b$
	Sum			
	Non-linearity activation fun	$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} \frac{s_j}{\|s_j\|}$ (Eq. 1)		$h_{w,b}(x) = f(a_j)$
output		vector($\boldsymbol{v}_j$)		scalar(h)





Capsule = New Version Neuron!
vector in, vector out VS. scalar in, scalar out

Cápsula



- Multiplicación matricial del vector de entrada
- Ponderación escalar de los vectores de entrada
- Suma de los vectores de entrada ponderados
- "Squash" (función vectorial no lineal: salida vectorial)

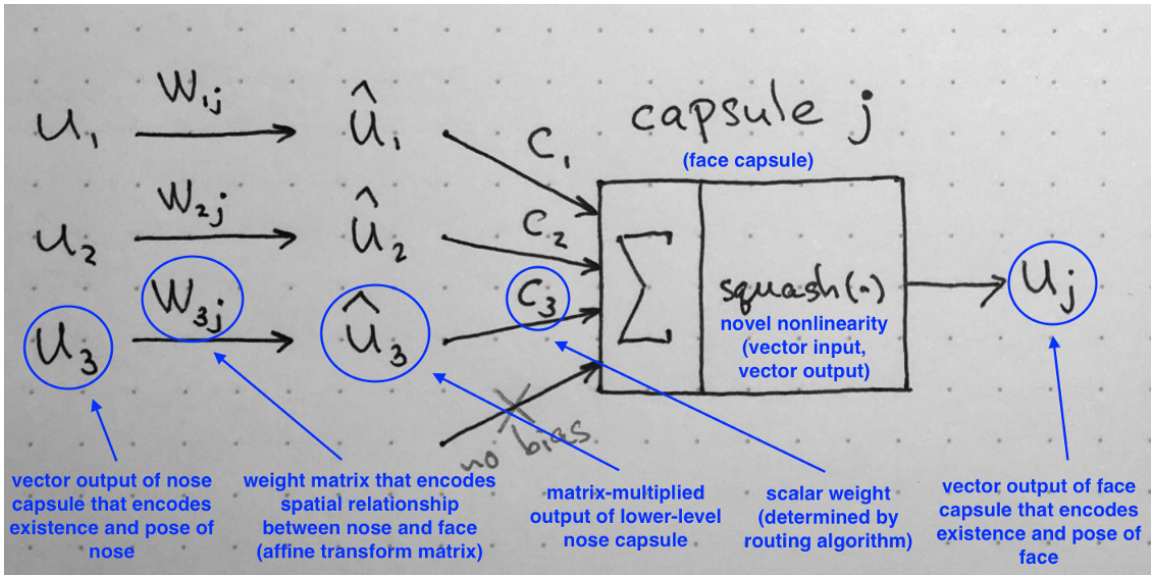
Cápsula

"Squashing"

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

additional "squashing" | unit scaling

Cápsula

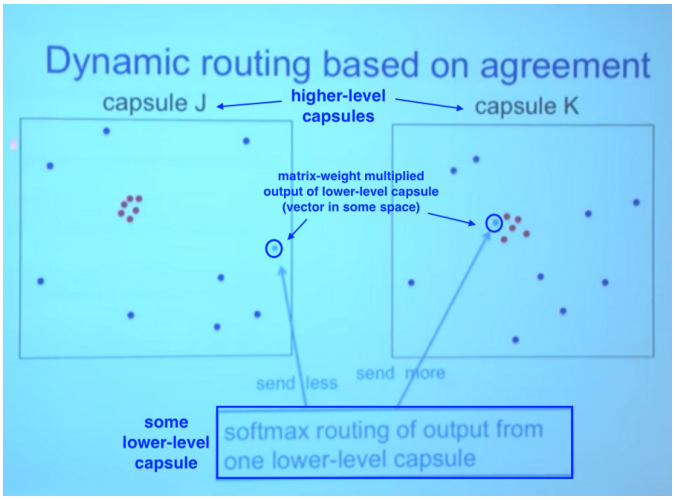


https://medium.com/ai³-theory-practice-business/understanding-hintons-capsule-networks-part-ii-how-capsules-work-153b6ade9f66

Algoritmo de entrenamiento

Enrutamiento dinámico [dynamic routing]

Lower level capsule will send its input to the higher level capsule that "agrees" with its input.



Algoritmo de entrenamiento

Enrutamiento dinámico [dynamic routing]

Lower level capsule will send its input to the higher level capsule that "agrees" with its input.

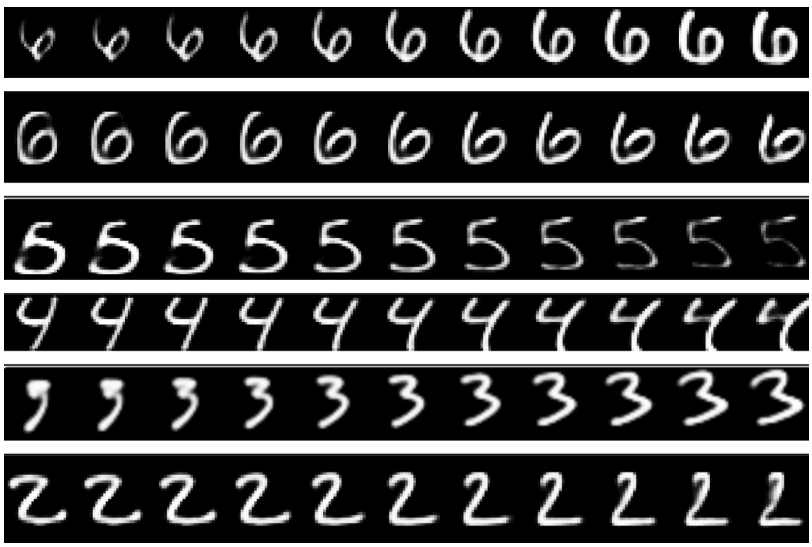
Procedure 1 Routing algorithm.

1: **procedure** ROUTING($\hat{\boldsymbol{u}}_{j|i}, r, l$)
2: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$.
3: **for** r iterations **do**
4: for all capsule i in layer l: $\mathbf{c}_i \leftarrow \texttt{softmax}(\mathbf{b}_i)$ ▷ `softmax` computes Eq. 3
5: for all capsule j in layer $(l+1)$: $\mathbf{s}_j \leftarrow \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}$
6: for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow \texttt{squash}(\mathbf{s}_j)$ ▷ `squash` computes Eq. 1
7: 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 $\mathbf{v}_j$

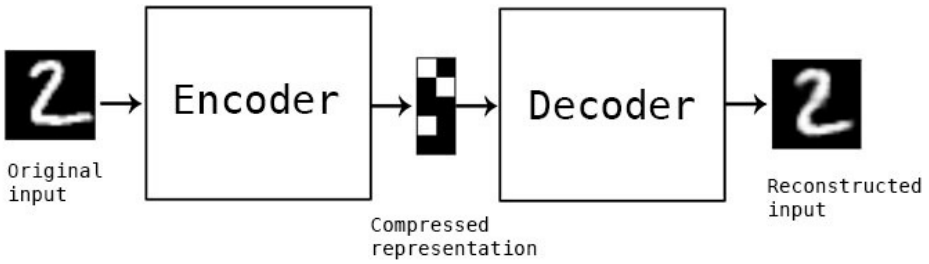
- More iterations tends to overfit the data.
- Recommendation: 3 routing iterations.

Modelo generativo

Una red de cápsulas [CapsNet] puede reconocer dígitos, pero también generarlos…



Modelo generativo



4.1 Reconstruction as a regularization method

We use an additional reconstruction loss to encourage the digit capsules to encode the instantiation parameters of the input digit. During training, we mask out all but the activity vector of the correct digit capsule. Then we use this activity vector to reconstruct the input image. The output of the digit capsule is fed into a decoder consisting of 3 fully connected layers that model the pixel intensities as described in Fig. 2. We minimize the sum of squared differences between the outputs of the logistic units and the pixel intensities. We scale down this reconstruction loss by 0.0005 so that it does not dominate the margin loss during training. As illustrated in Fig. 3 the reconstructions from the 16D output of the CapsNet are robust while keeping only important details.

Modelo generativo

Procedure 1 Routing algorithm[1] returns **activation** and **pose** of the capsules in layer $L+1$ given the **activations** and **votes** of capsules in layer L. V_{ich} is an H dimensional vote from capsule i with activation a_i in layer L to capsule c in layer $L+1$. β_a, β_v are learned discriminatively and the inverse temperature λ increases at each iteration with a fixed schedule.

procedure EM ROUTING($\boldsymbol{a}, V$)
  $\forall i, c$: $R_{ic} \leftarrow 1/size(L+1)$
  for t iterations **do**
    $\forall c$: $M_{c:}, S_{c:}, \boldsymbol{a}'_c \leftarrow$ M-STEP($R_{:c}, \boldsymbol{a}, V_{:c:}$)
    $\forall i$: $R_{i:} \leftarrow$ E-STEP($M, S, \boldsymbol{a}', V_{i::}$)
  return $\boldsymbol{a}', M$

procedure M-STEP($\boldsymbol{r}, \boldsymbol{a}, V'$) ▷ for one higher-level capsule
  $\forall i$: $\boldsymbol{r}'_i \leftarrow \boldsymbol{r}_i * \boldsymbol{a}_i$
  $\forall h$: $\boldsymbol{\mu}_h \leftarrow \frac{\sum_i \boldsymbol{r}'_i V'_{ih}}{\sum_i \boldsymbol{r}'_i}$
  $\forall h$: $\boldsymbol{\sigma}_h^2 \leftarrow \frac{\sum_i \boldsymbol{r}'_i (V'_{ih} - \boldsymbol{\mu}_h)^2}{\sum_i \boldsymbol{r}'_i}$
  $cost_h \leftarrow (\beta_v + log(\boldsymbol{\sigma}_h)) \sum_i \boldsymbol{r}'_i$
  $a' \leftarrow sigmoid(\lambda(\beta_a - \sum_h cost_h))$
  return $\boldsymbol{\mu}, \boldsymbol{\sigma}, a'$

procedure E-STEP($\boldsymbol{a}', S, M, V''$) ▷ for one lower-level capsule
  $\forall c$: $\boldsymbol{p}_c \leftarrow \frac{1}{\sqrt{\prod_h^H 2\pi \boldsymbol{S}_{ch}^2}} e^{-\sum_h^H \frac{(V''_{ch} - \boldsymbol{M}_{ch})^2}{2\boldsymbol{S}_{ch}^2}}$
  $\forall c$: $\boldsymbol{r}_c \leftarrow \frac{\boldsymbol{a}'_c p_c}{\sum_j \boldsymbol{a}'_j p_j}$
  return $\boldsymbol{r}$

Resultados: MNIST

0.25%

Resultados: smallNORB

Figure A.1: Sample smallNORB images at different viewpoints. All images in first row are at azimuth 0 and elevation 0. The second row shows a set of images at a higher elevation and different azimuth.

- Mejora del 45% con respecto al estado del arte.

Resumen

- Modelo bioinspirado simple y elegante.
- Resultados experimentales prometedores.
- Requiere menos datos de entrenamiento que las CNN.

¿Por qué no se hizo antes?

- Coste computacional elevado

Referencias

- Sara Sabour, Nicholas Frosst & Geoffrey E. Hinton:
  "Dynamic Routing Between Capsules"
  NIPS'2017 & arXiv, October 2017
  https://arxiv.org/abs/1710.09829
  http://papers.nips.cc/paper/6975-dynamic-routing-between-capsules
- ??? & Geoffrey E. Hinton:
  "Matrix Capsules with EM Routing"
  Submitted to ICLR'2018
  https://openreview.net/pdf?id=HJWLfGWRb
- Geoffrey E. Hinton, Alex Krizhevsky, Sida D. Wang:
  "Transforming Auto-Encoders"
  ICANN (1) 2011: 44-51
  https://doi.org/10.1007/978-3-642-21735-7_6
- Geoffrey E. Hinton:
  "What is wrong with convolutional neural nets?"
  MIT, 2014
  https://youtu.be/rTawFwUvnLE