Dynamic Routing Between Capsules

Comment

Fixation

The maintaining of the visual gaze on a single location. Regular eye movement alternates between saccades and visual fixation. The term "fixation" can either be used to refer to the point in time and space of focus or the act of fixating. Fixation, in the act of fixating, is the point between any two saccades, during which the eyes are relatively stationary and virtually all visual input occurs. In the absence of retinal jitter, a laboratory condition known as retinal stabilization, perceptions tend to rapidly fade away. (Pritchard R.M. et al. (1960), Coppola, D. et al. (1996))

- The number of presenting classes were given and it is guaranteed to be exclusively selected.

Abstract

Capsule

A group of neurons whose activity vector represents the instantiation parameters of a specific type of entity

1. Introduction

Motivation

Human visions ignores irrelevant details by using a carefully determinded sequence of fixation points to ensure that only a tiny fraction of the optic array is ever processed at the highest resolution

Assumption

- A single fixation gives much more than just a single identified object and its properties
- Our multi-layer visual system creates a parse tree-like structure on each fixation
- For a single fixation, a parse tree is carved out of a fixed multilayer neural network like a sculpture is carved from a lock

Details

- Special property - the existence of the instantiated entity
- Overall length of the representing vector as the existence of the entity, orientation as the properties of the entity
- Routing-by-agreement - far more effective than max-pooling which ignores all but the most active feature detector in a local pool in the layer below

2. How the vector inputs and outputs of a capsule are computed

- *Squashing* - short vectors get shrunk to almost zero length and long vectors get shurnk to slightly below 1

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

- The first layer of capsules will be,

$$s_j = \sum_i c_{ij} \hat{\mathbf{u}_{j|i}}, \qquad \hat{\mathbf{u}_{j|i}} = \mathbf{W}_{ij} \mathbf{u}_i$$

  where c_{ij} are coupling coefficients that are determineded by the iterative dynamic *routing* process as following

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

- The agreement is simply, $a_{ij} = \mathbf{v}_j \cdot \hat{\mathbf{u}}_{j|i}$

Routing Algorithm

- 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: $c_i \leftarrow softmax(b_i)$ for all capsule j in layer $(l + 1)$: sj $\leftarrow \sum_i c_{ij} \hat{\mathbf{u}}_{j|i}$ for all capsule j in layer $(l + 1)$: vj $\leftarrow$ $squash(s_j)$ for all capsule i in layer l and capsule j in layer $(l + 1)$: $b_{ij} b_{ij} + \hat{\mathbf{u}}_{j|i} \cdot v_j$
- return v_j

3. Margin loss for digit existence

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

where $T_k = 1$ iff class k is present and empirically $m^+ = 0.9, m^- = 0.1, \lambda = 0.5$

4. CapsNet architecture

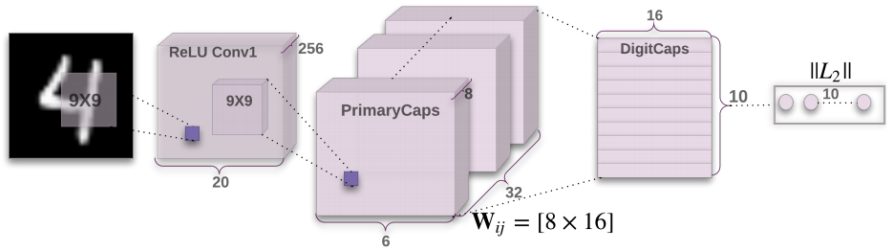

Figure 1: enter image description here

Decoder

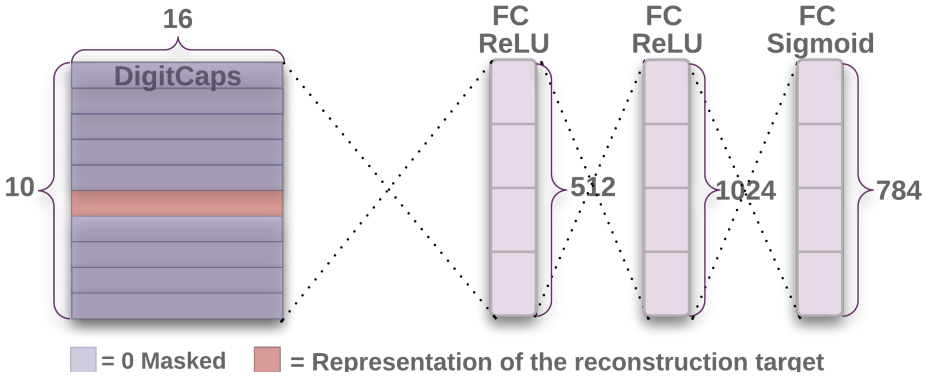

Figure 2: enter image description here

5. Capsules on MNIST

6. Segmenting highly overlapping digits

- The *Routing-by-agreement* obviate the need to make higher-level segmentation decisions in the domain of pixels.

7. Other datasets

- A drawback of Capsules is, because it shares with generative models, it prefers to account for everything in the image so it does better when it can

model the clutter than when it just uses an additional *orphan* category in the dynamic routing

8. Discussion and previous work

- We had to choose between replicating feature detectors on a grid that grows exponentially with the number of dimensions, or increasing the size of the labeled training set in a similarly exponential way.
- Capsules avoid these exponential inefficiencies by converting pixel intensities into ***vectors of instantiation parameters*** of recognized fragments and then applying transformation matrices to the fragments to predict the instantiation parameters of larger fragments
- This model made a very strong representational assumption to reduce the complexity by exponential degree – At each location in the image, there is at most one instance of the type of entity that a capsule represents. That's why they intentionally exclude the image generated with the same digits.
- Capsules use neural activities that vary as viewpoint varies rather than trying to eliminate viewpoint variation from the activities.
- The importance of dynamic routing procedure is also backed by biologically plausible models of invariant pattern recognition in the visual cortex