Convergent Learning: Do different neural networks learn the same representations?

Statistics Student Seminar
22 March 2016

Yixuan ("e-shwen") Li, Jason Yosinski, Jeff Clune, Hod Lipson and John Hopcroft
http://www.cs.cornell.edu/~yli/
YES!

NOW THE DEEP LEARNING BEGIN!
MY DEEP LEARNING MODEL DOESN'T WORK, I HAVE NO IDEA WHY

MY MODEL WORKS, I HAVE NO IDEA WHY
Deep representations
Are representations local or distributed?
Invariant visual representation by single neurons in the human brain

R. Quiroga 1,2, L. Reddy 1, G. Kreiman 3, C. Koch 1 & I. Fried 2,4

It takes a fraction of a second to recognize a person or an object even when seen under strikingly different conditions. How such a robust, high-level representation is achieved by neurons in the human brain is still unclear 5, 6. In monkeys, neurons in the upper stages of the ventral visual pathway respond to complex images such as faces and objects and show some degree of invariance to metric properties such as the stimulus size, position and viewing angle 7, 8. We have previously shown that neurons in the human medial temporal lobe (MTL) fire selectively to images of faces, animals, objects or scenes 9, 10. Here we report on a remarkable subset of MTL neurons that are selectively activated by strikingly different pictures of given individuals, landmarks or objects and in some cases even by letter strings with their names. These results suggest an invariant, sparse and explicit code, which might be important in the transformation of complex visual percepts into long-term and more abstract memories.

The subjects were eight patients with pharmacologically intractable epilepsy who had been implanted with depth electrodes to localize the focus of seizure onset. For each patient, the placement of the depth electrodes, in combination with micro-wires, was determined exclusively by clinical criteria 11. We analysed responses of neurons from the hippocampus, amygdala, entorhinal cortex and parahippocampal gyrus to images shown on a laptop computer in 21 recording sessions. Stimuli were different pictures of individuals, animals, objects and landmark buildings presented for 1 s in pseudo-random order, six times each. An unpublished observation in our previous recordings was the sometimes surprising degree of invariance inherent in the neuron’s (that is, unit’s) firing behaviour. For example, in one case, a unit responded only to three completely different images of the ex-president Bill Clinton. Another unit (from a different patient) responded only to images of The Beatles, another one to cartoons from The Simpson’s television series and another one patient. The mean number of images in the screening session was 93.9 (range 71–114). The data were quickly analysed offline to determine the stimuli that elicited responses in at least one unit (see definition of response below). Subsequently, in later sessions (testing sessions) between three and eight variants of all the stimuli that had previously elicited a response were shown. If not enough stimuli elicited significant responses in the screening session, we chose those stimuli with the strongest responses. On average, 88.6 (range 70–110) different images showing distinct views of 14 individuals or objects (range 7–23) were used in the testing sessions. Single views of random stimuli (for example, famous and non-famous faces, houses, animals, etc.) were also included. The total number of stimuli was determined by the time available with the patient (about 30 min on average). Because of our clinical set-up, the recording conditions can sometimes change within a few hours, so we always tried to perform the testing sessions shortly after the screening sessions in order to maximize the probability of recording from the same units. Unless explicitly stated otherwise, all the data reported in this study are from the testing sessions. To hold their attention, patients had to perform a simple task during all sessions (indicating with a key press whether a human face was present in the image). Performance was close to 100%.

We recorded from a total of 993 units (343 single units and 650 multi-units), with an average of 47.3 units per session (16.3 single units and 31.0 multi-units). Of these, 132 (14% 64 single units and 68 multi-units) showed a statistically significant response to at least one picture. A response was considered significant if it was larger than the mean plus 5 standard deviations (s.d.) of the baseline and had at least two spikes in the post-stimulus time interval considered (300–1,000 ms). All these responses were highly selective: for the responsive units, an average of only 2.8% of the presented pictures (range: 0.9–22.8%) showed significant activations according to this
Invariant visual representation the human brain

R. Quiroga, L. Reddy, G. Kreiman, C. Koch & I. Fried

It takes a fraction of a second to recognize a person or an object even when seen under strikingly different conditions. How such a robust, high-level representation is achieved by neurons in the human brain is still unclear. In monkeys, neurons in the upper stages of the ventral visual pathway respond to complex images such as faces and objects and show some degree of invariance to metric properties such as the stimulus size, position and viewing angle. We have previously shown that neurons in the human medial temporal lobe (MTL) fire selectively to images of faces, animals, objects or scenes. Here we report on a remarkable subset of MTL neurons that are selectively activated by strikingly different pictures of given individuals, landmarks or objects and in some cases even by letter strings with their names. These results suggest an invariant, sparse and explicit code, which might be important in the transformation of complex visual percepts into long-term and more abstract memories.

The subjects were eight patients with pharmacologically intractable epilepsy who had been implanted with depth electrodes to localize the focus of seizure onset. For each patient, the placement of the depth electrodes, in combination with micro-wires, was determined exclusively by clinical criteria. We analysed responses of neurons from the hippocampus, amygdala, entorhinal cortex and parahippocampal gyrus to images shown on a laptop computer in 21 recording sessions. Stimuli were different pictures of individuals, animals, objects and landmark buildings presented for 1 s in pseudo-random order, six times each. An unpublished observation in our previous recordings was the sometimes surprising degree of invariance inherent in the neuron’s (that is, unit’s) firing behaviour. For example, in one case, a unit responded only to three completely different images of the ex-president Bill Clinton. Another unit (from a different patient) responded only to images of The Beatles, another one to cartoons from The Simpsons television series and another one to patient. 

39.9 (range 30–55) neurons in a single unit were responsive (range: 0.9–100% of total). We found that the firing rates of these neurons were more than twice that of other neurons that fired less frequently. This finding suggests that these neurons are more likely to be involved in the encoding of long-term memories. We conclude that neurons in the MTL are able to encode a vast range of visual stimuli in a way that is both invariant and specific.
Are representations local or distributed?

Local

Partially-Distributed
(spanning low-dimensional subspaces)

Distributed


Key idea:
Probe representation types by comparing multiple networks (with different random initializations)
Key idea:
Probes representation types by comparing multiple networks (with different random initializations)
Key idea:
Probe representation types by comparing multiple networks (with different random initializations)
Key idea:
Probe representation types by comparing multiple networks (with different random initializations)
Local
Local
Local

Partially-Distributed
(spanning low-dimensional subspaces)
Local

Partially-Distributed
(spanning low-dimensional subspaces)
Local

Partially-Distributed
(spanning low-dimensional subspaces)

Distributed
Local

Partially-Distributed
(spanning low-dimensional subspaces)

Distributed
Same architecture

Different random initialization

AlexNet trained on ImageNet (Krizhevsky, NIPS 2012)
Key idea:
Probe representation types by comparing multiple networks (with different random initializations)
Neuron Activation Statistics: Correlation
Neuron Activation Statistics: Correlation

high correlation
Neuron Activation Statistics: Correlation

**High Correlation**

**Low Correlation**
Neuron Activation Statistics: Correlation

Within-net correlation (Net1)
Within-net correlation (Net2)
Between-net correlation (Net1)

(Example for conv1: 96 x 96)
Aside: is correlation a powerful enough measure?

Between-net correlation:

\[ C_{l,i,j}^{(n)} = \mathbb{E}[(X_{l,i}^{(n)} - \mu_{l,i}^{(n)})(X_{l,j}^{(n)} - \mu_{l,j}^{(n)})] \sigma_{l,i}^{(n)} \sigma_{l,j}^{(n)} \]

Between-net mutual information:

\[ I_{l,i,j}^{(n)} = \sum_{a \in X_{l,i}^{(n)}} \sum_{b \in X_{l,j}^{(m)}} p(a, b) \log \frac{p(a, b)}{p(a)p(b)} \]

Short answer: yes. See Supplementary Information.

(Example for conv1: 96 x 96)
One-to-one alignment between features

Between-net correlation

(Example for conv1: 96 x 96)

Greedy matching:
Pick the max along each row
One-to-one alignment between features

Top 9 patches

Deconv (Zeiler & Fergus, 2014)
One-to-one alignment between features

best match

conv1
One-to-one alignment between features

best match

conv1
One-to-one alignment between features

best match

conv1

Net1

Net2
One-to-one alignment between features

conv1

best match

worst match

Net1

Net2
One-to-one alignment between features

conv2

best match

Net1

Net2

worst match
One-to-one alignment between features

best match
One-to-one alignment between features

Net1

Net2

conv3
One-to-one alignment between features

best match
One-to-one alignment between features

Greedy matching:
Pick the max along each row

(Example for conv1: 96 x 96)
One-to-one alignment between features

Find max weighted bipartite matching (Hopcroft & Karp, ‘73)

(Example for conv1: 96 x 96)
Find max weighted bipartite matching (Hopcroft & Karp, '73)

Net2

conv1 within-net c (Net1, Net2)

conv1 between-net natural (unaligned) order permuted (aligned) order c (Net1)

One-to-one alignment between features Net2
Find max weighted bipartite matching (Hopcroft & Karp, '73)

Conv1 within-net \(c\) (Net1, Net2)

Conv1 between-net natural (unaligned) order permuted (aligned) order \(c\) (Net1)

Net1

Net2

One-to-one alignment between features
One-to-one alignment between features

Find max weighted bipartite matching (Hopcroft & Karp, ‘73)

unique, highly correlated unit
Take aways:

- For many units, a one-to-one alignment is possible.
- Some units are network-specific and have no high-correlation pairing in the alternate network.
- The two networks learn different numbers of units to span certain subspaces.
One-to-one alignment between features

Take aways:
- For many units, a one-to-one alignment is possible.
- Some units are network-specific and have no high-correlation pairing in the alternate network.
- The two networks learn different numbers of units to span certain subspaces.
Take aways:

• For many units, a one-to-one alignment is possible.
• Some units are network-specific and have no high-correlation pairing in the alternate network.
• The two networks learn different numbers of units to span certain subspaces.
Few-to-one alignment between features

Net1 feature map

Net2 feature map

Sparse representations

Net2 feature map

Net1 feature map
### Few-to-one alignment between features

<table>
<thead>
<tr>
<th></th>
<th>Sparse Prediction Loss (after 4,500 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>decay 0</td>
</tr>
<tr>
<td>conv1</td>
<td></td>
</tr>
<tr>
<td>conv2</td>
<td></td>
</tr>
<tr>
<td>conv3</td>
<td></td>
</tr>
<tr>
<td>conv4</td>
<td></td>
</tr>
<tr>
<td>conv5</td>
<td></td>
</tr>
</tbody>
</table>

---

Both representations (input and target) are taken after the relu is applied. Before training, each channel is normalized to have mean zero and standard deviation 1. The average squared prediction errors for each layer are shown in Table 1 for a variety of L1 penalty weights (i.e. different decay values). For the conv1 layer, the prediction errors do not rise much compared to the dense (decay = 0) case with an average of 0.514. A sparsity penalty as high as $10^{-1}$ in the conv2 layer produced a model with an average of 2.5 non-zero connections per predicted unit. The mapping layers for higher layers (i.e. conv3, conv4, conv5) showed slightly larger errors. For the conv5 layer, the errors did not rise with the imposition of a sparsity penalty until a decay parameter of over $10^{-2}$. The results on those layers are not included here. Future investigation with different hyperparameters or codes will be necessary to shed light on the open, long-standing debate about the extent to which learned representations are local vs. distributed: The units that match well one-to-one suggest the presence of a local code, while the units that do not match well one-to-one, such extreme sparsity does not hurt performance implies that each neuron in one network can be predicted by only one or a few neurons in another network. For the conv2 layer, the prediction errors do not rise with the imposition of a sparsity penalty until a decay parameter of over $10^{-2}$. The results on those layers are not included here. Future investigation with different hyperparameters or codes will be necessary to shed light on the open, long-standing debate about the extent to which learned representations are local vs. distributed: The units that match well one-to-one suggest the presence of a local code, while the units that do not match well one-to-one, such extreme sparsity does not hurt performance.
**Few-to-one alignment between features**

<table>
<thead>
<tr>
<th></th>
<th>Sparse Prediction Loss (after 4,500 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>decay 0</td>
</tr>
<tr>
<td>conv1</td>
<td>0.170</td>
</tr>
<tr>
<td>conv2</td>
<td></td>
</tr>
<tr>
<td>conv3</td>
<td></td>
</tr>
<tr>
<td>conv4</td>
<td></td>
</tr>
<tr>
<td>conv5</td>
<td></td>
</tr>
</tbody>
</table>
Few-to-one alignment between features

<table>
<thead>
<tr>
<th></th>
<th>Sparse Prediction Loss (after 4,500 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>decay 0</td>
</tr>
<tr>
<td>conv1</td>
<td>0.170</td>
</tr>
<tr>
<td>conv2</td>
<td></td>
</tr>
<tr>
<td>conv3</td>
<td></td>
</tr>
<tr>
<td>conv4</td>
<td></td>
</tr>
<tr>
<td>conv5</td>
<td></td>
</tr>
</tbody>
</table>
Few-to-one alignment between features

<table>
<thead>
<tr>
<th>Layer (Conv)</th>
<th>Sparse Prediction Loss (after 4,500 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>decay 0</td>
</tr>
<tr>
<td>conv1</td>
<td>0.170</td>
</tr>
<tr>
<td>conv2</td>
<td></td>
</tr>
<tr>
<td>conv3</td>
<td></td>
</tr>
<tr>
<td>conv4</td>
<td></td>
</tr>
<tr>
<td>conv5</td>
<td></td>
</tr>
</tbody>
</table>
Few-to-one alignment between features

<table>
<thead>
<tr>
<th></th>
<th>Sparse Prediction Loss (after 4,500 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>decay 0</td>
</tr>
<tr>
<td>conv1</td>
<td>0.170</td>
</tr>
<tr>
<td>conv2</td>
<td>0.372</td>
</tr>
<tr>
<td>conv3</td>
<td>0.434</td>
</tr>
</tbody>
</table>
Few-to-one alignment between features

<table>
<thead>
<tr>
<th></th>
<th>Sparse Prediction Loss (after 4,500 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>decay 0</td>
</tr>
<tr>
<td>conv1</td>
<td>0.170</td>
</tr>
<tr>
<td>conv2</td>
<td>0.372</td>
</tr>
<tr>
<td>conv3</td>
<td>0.434</td>
</tr>
<tr>
<td>conv4</td>
<td>0.478</td>
</tr>
<tr>
<td>conv5</td>
<td>0.484</td>
</tr>
</tbody>
</table>

- For the conv1 and conv2 layers, the prediction errors do not rise with the imposition of a sparsity penalty until a penalty greater than $10e^{-3}$.

- That such extreme sparsity does not hurt performance implies that each neuron in one network can be predicted by only one or a few neurons in another network.
Few-to-one alignment between features

(Example for conv1: 96 x 96)
Few-to-one alignment between features

(Example for conv1: 96 x 96)
Finding the low-dimensional subspaces
Finding the low-dimensional subspaces
Finding the low-dimensional subspaces
Finding the low-dimensional subspaces
Conclusions

1. Some features are learned reliably in multiple networks, yet other features are not consistently learned.

2. The representation codes are a mix between a local code and partially distributed codes.

3. Units learn to span low-dimensional subspaces and, while these subspaces are common to multiple networks, the specific basis vectors learned are not.
Future work:

- Different architectures
- Different data
http://arxiv.org/abs/1511.07543

Thank you :-)