Scribes: Daniel Gomez
Hyun Dong Lee

COMS 6998-06: Computation and the Brain

Christos Papadimitriou

10/03/2018

Contents

Preface

TA begins class by scaring everyone into thinking there is a quiz when there is not.

1 Introduction and Recap

1.1 Types of random graphs

- Erdos-Renyi graph: random probability of connecting any two nodes
- Internet graph: probability of connection correlated with degree of each node

- Small world graph: Few steps necessary to reach any other node

Remark 1.1. Different areas of the brain seem to resemble different types of these graphs.

1.2 Growth of Neurons

The growth cone is given split and movement directions by the DNA and is more likely to connect to neurons that are neighbors of its connections. After birth, a great sparsification happens, where a lot of these created connections go away. Oja's rule is a way to approximate how synapses work, taking homeostasis into account. At a slower time scale, the sum of all presynaptic weights is renormalized to 1.

Remark 1.2. At this point, Papadimitriou went on a brief tangent to respond to a student's question from piazza/homework about what they were meant to gain from advanced talks, eventually giving students extra time to understand them Thalamus talk and read up on the background.

2 Kiran's Talk

2.1 Background and introduction to fMRI

fMRIs measure bloodflow to determine which areas of the brain are more active given certain stimuli. The precision is measured by voxel size, which is about $(3mm)^3$, holding about 10^6 neurons. The number of voxels is 10^5 and time precision is low.

Prior work in this space tried to create a model for what an fMRI would look like, given an object. The previous work often used 1 person at a time and tried to learn representations for common things like dogs, using LDA to generate words.

Some examples include:

- Mitchell et al 08 predicts fMRI responses induced by pictures of concrete nouns.
- Naselaris et al 11 tries to reconstruct movie images from fMRI signals measured while subjects watched movies.
- Pereira et al 16 decodes fMRI responses to word clouds and short sentences.

2.2 Kiran's Project

2.2.1 Structure

16 people watch an episode of Sherlock (BBC TV series) and every 1.5 seconds of the TV show has a description of exactly what's happening in the scene. The intention of the project is to map between the language in the description and the fMRI that gets measured from the subject at that time (adding a 3 sec delay for fMRI). The words were embedded as vectors using word2vec from a 1-hot encoding. Word2vec keeps a notion of similarity and distance between words. The word vectors were then combined by a weighted sum to create the sentences. The fMRI data was also dimensionally reduced.

2.2.2 Model

They checked four different brain regions:

- Default mode - generally thought of as the section of the brain that is used to think about oneself or narrative functions. A total of 2000 voxels.
- Language Areas - both ventral and dorsal, consisting of 2000 voxels
- "Whole brain" - the areas, scattered all over the brain, that have high intersubject connectivity. A total of 26000 voxels.
- Occipital lobe - consists of 6000 voxels

They used a shared response model which assumes that all 16 subjects must have some sort of similarities among them, hoping to find what they all have in common. The SRM is functionally very similar to PCA, but with added orthogonality constraints. Went from 26,000 dimensions to 20.

2.2.3 Linear Maps Between fMRI and Text

The basic model is:

$$WX = Y,\ W \in R^{m \times n} \tag{1}$$

where X represents the fMRI data matrix and y represents the semantic annotation data matrix. We learn the map through procrustes or ridge regression. With procrustes, we restrict the map to rotations of the data and impose strong constraint on map. With ridge regression, we restrict map weights to be uniformly small (not sparse). As a result, we get multiplicative improvements.

2.2.4 Training and Evaluation

To train the model, they segmented into groups of 5 timepoints and took the first half for training and the second half for testing, then measured using the cosine similarity between predicted and observed semantic space. The data was then chunked by time into 25 pieces and the fMRI data was used to guess which of the 25 word vectors was most likely. The accuracy depended on the brain region used, but tended to be around 85% for first choice.

3 Jacob's Talk

3.1 Fundamental Axioms of Neurobiology

The fundamental axioms of neurobiology, greatly influenced by Cajal, are the neuron doctrine, the law of dynamic polarization, canonical circuits, and stereotyped cell types. They The neuron doctrine states that the neuron is the basic unit of the brain, and information travels from one neuron to another. The law of dynamic polarization states that nerve cells in the brain are polarized; neurons receive information with their dendrites and cell bodies, and conduct information to remote cells via axons. In addition, there are many kinds of neurons with different shapes and functions. Lastly, although the brain of one person will differ from that of another person, the basic layout, called the canonical circuits, will be very similar to each other. These axioms by Cajal inspired many scientists to find all neural connections in the brain, which may lead to new insights.

3.2 Why the Fly?

Flies are pretty simple neurally, with only 100,000 neurons, 60% of which are dedicated to optical processing. They also have orders of magnitude fewer types of neurons, making them easier to understand. Also, we have lots of genetic tools to work with flies, unlike possibly more interesting brains like honeybees. There is work where people turn on certain areas of fly brains using genetic modification to create mess with their aggression, escape maneuvers, and courtship behaviors. There are three levels of fly neuronal activity to theorize about:

- Why? - What computation causes the response?
- What? - What is the algorithm that decides the actions?
- How? - What is the process by which this is implemented?

Jacob focuses on the algorithms.

3.3 Mushroom Body

It seems that most memory and learning happens in the mushroom body of the fly. The olfactory sense gets projected into a high dimensional space and is directly connected to memory.

Additionally, path integration, which seems to be very primitive in flies (compared to say honeybees or cats, who have quite good path integration), may be found here. The theorized circuitry for this is a “ring attractor”, a circle of neurons that light up when the animal is moving in a certain direction. For a ring attractor to work, global inhibitory neurons are necessary. It is theorized that these could be the precursors for grid cells, if stacked correctly.

Remark 3.1. We have a full connectome for the fly visual cortex and mushroom body, but not the rest of its brain.

3.4 Vision

3.4.1 Motion Detection

Though there is a lot of data on fly vision, motion detection is still considered ”unsolved”. There are many models, but no consensus:

- HR Correlator - Uses temporal delays on the two different sides to detect motion
- Barlow Levick - similar to the HR Correlator, but instead uses inhibition on the opposite direction.
- Motion Energy - primate model, but found to not work on flies.

3.4.2 Natural Images and Movement

Remark 3.2. Flies often have their legs ripped off to do the following experiments.

When an LSTM was trained on natural videos based on the fly connectome data, they were able to create a T4 area that evolved correctly, but that was the only one. It seems that different parts of the fly brains do very similar things and it is still unknown whether or not flies recognize objects.

Remark 3.3. At this point, Jacob listed a variety of project ideas, all centered on flies.

4 STDP

When the neuronal connections are developing, the response is highly based on timing. Synapses that are slightly late are heavily discouraged, while synapses that are slightly early are heavily encouraged. This is called Spike-timing-dependent plasticity. At this point, the backpropagation paper that we read was discussed, talking about whether or not it was possible to train a network with backpropagation without exact signals of the a posteriori neurons. The "Biologically Plausible Deep Learning" paper relates how gradient descent and STDP work using differential equations and the Dirac-Delta function.

For differential equations:

$$\Delta x^t = \alpha \delta(t) \nabla f(x^t) \tag{2}$$

where t is the update time. For STDP:

$$\Delta w^t = \beta \delta(t) \nabla V(w^t) \tag{3}$$

where t is the spike time.

5 Information Theory

It all began with the Shannon paper. Shannon wants the "surprisingness" of a sequence to be able to sum, so he uses a logarithm.

5.1 Entropy

The entropy of a generative process is maximized when the probability distribution governing it is the uniform distribution. The more ordered something is, the less entropy it generally has. As an example, the English language has an entropy of about 1.6, whereas random English letters has an entropy of about 4.8.

The Kolmogorov complexity of a source is equivalent to its entropy for large n. To calculate the KC, one can use the length of the smallest program that would generate the string. Given this, it is clear that the KC of the random string is n.

5.2 Shannon's First Theorem

Theorem 1: Suppose the channel has capacity C bits/sec, the source has entropy H, and there is no noise. Then the maximum rate (symbols/sec) is C/H.

- No rate greater than C/H can be achieved.
- For any $\epsilon > 0$, $rate \geq C/H - \epsilon$ can be achieved by coding.

5.3 Shannon's Second Theorem

Theorem 2: If the channel has capacity C and noise $p < \frac{1}{2}$ then

- Any rate R¡C(1-h(p)) can be achieved by coding
- No rate greater than C(1-h(p)) can be achieved
- If equivocation is allowed, the attainable region is above the line with slope 1 and intercept -C.

5.4 Properties

The chain rule for entropy states that:

$$H(x, y) = H(x) + H(y|x) = H(y) + H(x|y) \tag{4}$$

The mutual information then is:

$$I(x, y) = \sum_{i,j} P(i, j) log_2 \left(\frac{P(i, j)}{P(i)P(j)} \right) \tag{5}$$

Note that when the two are independent, this is 0. The Mutual Information can therefore be thought of as the distance of a distribution from independence, where the distance used is the KL divergence:

$$KL(P, Q) = \sum_{i} log_2 \left(\frac{P[i]}{Q[i]} \right) \tag{6}$$

6 Distribution of synaptic responses

Is the brain a good machine for relaying information? We look at the way different things are encoded in mind signals.

Ex: We look at what contrast a fly sees in its general world and create a distribution. We then separate the distribution into chunks of equal probability instead of equal distance and we see that the fly's responses actually correspond to this, firing more rapidly at rates equivalent to those seen on the CDF model.

Remark 6.1. A student asked a very good question that Papadimitriou didn't have an answer for based on the energy usage of such a system.