Information Processing and the Brain 2020/2021

Course work: Analysing deep neural networks as brain models

For this course work you are asked to implement and explore the behaviour of backprop using supervised learning and a deep neural network as a potential model for information processing in the brain. For implementation we suggest that you use lab1 as a starting point and Python or Julia in particular.

1. **Implement the backpropagation algorithm in the supervised context**
   The backpropagation algorithm is often used in machine learning to solve the credit assignment problem. Here you are going to contrast its different elements (e.g. error backpropagation, or symmetric weights) to how the brain is organised. You should use a simple feedforward neural network with one (or two) hidden layers and sigmoidal units to teach the network to classify the widely used handwritten digit recognition dataset (MNIST) . Do not use autograd tools as in pytorch for this simple task. You should use a mean squared cost (as discussed in the lectures) and a one-hot output encoding to represent the MNIST classes. Include brief snippets of your code in the report. *Note*: Although here you are asked to implement the backprop algorithm in a supervised setting, backprop is also used in unsupervised (e.g. autoencoders) and reinforcement learning (e.g. deep RL).
   You should use the lectures and you may also use these (or related) papers as *references* when contrasting backprop with the brain:
   - Blake and Lillicrap, Dendritic solutions to the credit assignment problem, Current Opinion in Neurobiology 2018 (link)
   - Sacramento et al., Neural Information Processing Systems 2018 (link)

2. **Analyse the deep network using information theory**
   Recently information theory has been used to study the dynamics of learning in deep learning networks; here you will examine the mutual information between the image label and the network activity: you can imagine a communication channel

   label -> image -> hidden layer -> output

   in which the image label is encoded in the image itself and that in turn is encoded in the activity of the hidden layer and then in the output layer. The aim is to quantify information flow in this channel, at least in a discrete approximation in which neurons are either on or off. Thus a random variable representing a layer will take binary-sequence values with ones and zeros corresponding to neurons being “on” or “off” based on their activation. The probability for each binary sequence can be estimated by counting the number of images that produce that sequence; note that for the hidden layer many will have zero estimated probability.
   - *Opening the black box of deep neural networks via information.* Ravid Shwartz-Ziv and Naftali Tishby*. arXiv:1703.00810.*
   - *Adaptive estimators show information compression in deep neural networks.* Ivan Chelombiev, Conor Houghton and Cian O'Donnell. arXiv:1902.09037.

Questions

1. **Biological relevance of backprop**
    1. How does the algorithm work? Explain the algorithm and plot the different components over learning to help you explain its behaviour (e.g. what are the weight changes given by backprop and the different terms that make up the gradient). You should also plot the performance (i.e. cost) of the algorithm over learning iterations (i.e. the learning curve). You should include a snippet of the key component of your code in your report. [3/10 marks, max 500 words]

    2. *How does the algorithm relate to the brain?* Discuss how backprop relates to the brain. Can neural networks optimised with backprop explain neuroscience data? Here you are going to explore the biological plausibility of three of the key issues with backprop (you should discuss their implications and how important they are using simulations and the respective plots to support your arguments) [4/20 marks, max 500 words]:
        a. Weight transport problem
        b. The need for derivative of the activation function
        c. The need for a target (in the supervised setting)

    3. Discuss what are the key advantages of using supervised learning over the other two learning paradigms (unsupervised and reinforcement learning) both in terms of performance/data needs and biological plausibility. **Note** that the backpropagation algorithm can also be used in unsupervised and reinforcement learning networks. [1.5/20 marks, max 250 words]

    4. What about disadvantages in terms of performance/data needs and biological plausibility? And how could this be improved upon? [1.5/20 marks, max 250 words]

2. **Information theory analysis**

    1. If *X* is a random variable representing the input, *H(X)* will be log(10) since there are ten labels. If *Y* represents the discretized hidden layer activity then what is H(Y|X) and how does this change with learning? What about *I(X,Y)*? How are these dynamics changed if the number of hidden layer neurons is changed? If Z is the output layer, how do *H(Z|X)* and *I(X,Z)* change? [7/20 marks, maximum 250 words and four graphs]
    2. What happens if you consider some property of the image, for example, if the image is divided in four and W is the random variable giving which quadrant has the largest amount of white. How do *I(W,Y)* and *I(W,Z)* change with learning? [3/20 marks, maximum 140 words and two graphs]

Submit your report on Blackboard (IPB > Assessment, submission and feedback). No need to submit your full code (only add a snippet in the report), only the report in pdf or similar. Please add the necessary references at the end.

Note 1: Where possible cite papers and/or use simulations/plots to support your claims.
Note 2: Collaborative work is encouraged (e.g. for coding and understanding of the algorithm), but every submission should be individual / unique.

Support provided

We (Lecturers) will be available to provide some support, email either of us to make an appointment.

Deadline

The deadline for submission of all optional unit assignments is **16:00 on Monday 16th of August**. Students should submit all required materials to the "Assessment, submission and feedback" section of Blackboard - it is essential that this is done on the Blackboard page related to the "With Coursework" variant of the unit.

Time commitment

The expectation is that students will spend 3 full working weeks on their two assignments. The effort spent on the assignment for each unit should be approximately equal, being roughly equivalent to 1.5 working weeks each.