Deep learning Assignment 1 (2022)

MPhil in Computational Biology

December 13, 2022

If there are errors found, I will update the assignment. **Due date: 2023-02-01 23:45**

Please submit your report to moodle, anonymised as before. Your report must be a maximum of twenty pages, excluding the appendix. Your appendix should contain only a copy of any code.

This assignment is worth 100% of your overall mark for this module.

All references to papers can be found in the Paperpile repository for the course: `https://paperpile.com/shared/pb4w0p`.

1 Self-supervised learning [25 marks]

Write your own multi-layer perceptron with the following specification: eight input units, three hidden units (sigmoidal activation) and eight output units (sigmoidal activation). There are eight -d input vectors **x**, for each of the eight cases where one input unit is "1" and the other seven units are "0". Your task is to solve the self-supervision problem of reproducing the input vector on the output vector, so **t = x**. Describe how your network attempts to solve this problem.

Show how an autoencoder can implement image compression (Cottrell et al. 1987). How do your results vary with the number of hidden units?

For this exercise, you should write your own code. you can adopt `https://github.com/sje30/dl2022/blob/main/code/xor/xor2.R` if you wish.

2 MNIST classification [20 marks]

Build a MNIST classifier using a convolutional framework (**Model 1**) in the same spirit of the LeCun et al (1996) paper. You should investigate the effect of varying the number of convolutional filters in the network. Show the typical convolutional filters that are learnt.

Compare its performance to a traditional feedforward network (**Model 2**) without convolutions that you should also build.

What are the pros and cons of each model?

You can Use Flux.jl, starting from this model zoo, or another framework, to create your networks. `https://github.com/FluxML/model-zoo/tree/master/vision/conv_mnist`

3 t-SNE and UMAP [30 marks]

1. Summarise the steps involved in the t-SNE algorithm (van der Maaten and Hinton, 2008).

2. Kobak and Berens (2019) highlight a protocol for the use of t-SNE for visualising high-dimensional data. Summarise and demonstrate their recommendations on a high-dimensional dataset of your choosing. How useful is their protocol?

3. Kobak and Berens (2019) state "... previous claims that UMAP vastly outperforms t-SNE [Becht et al 2019] might have been partially due to t-SNE being applied in a suboptimal way." How would you compare the pros and cons of UMAP and t-SNE? Present evidence to support your arguments.

Packages: `https://github.com/lejon/TSne.jl` `https://github.com/jkrijthe/Rtsne` `https://github.com/tkonopka/umap` `https://github.com/dillondaudert/UMAP.jl`

4   Alpha Fold2 [25 marks]

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. This news article was published in Nature, November 2020. `https://doi.org/10.1038/d41586-020-03348-4`. Write a three page (maximum essay) summarising what AlphaFold2 is, what has happened in the field since its release, and what outstanding challenges there are. Has it changed everything?

For this essay you should find relevant papers published about this topic, and reference them appropriately.