Medical Data Science, WS 2023/2024

Prof. Dr. Nico Pfeifer

17.01.2024

Oğuz Ata Çal 6661014

Edward Beach 5451904

Assignment 5

**Problem 1**

1. One improvement is the increase of computational power (deep learning algorithms are usually computationally demanding and older hardware would have been too inefficient). Another factor is the rise of big data, we previously did not have many large (labeled) data sets that are useful for deep learning methods. Another factor is the improvement of Neural Network architecture, e.g. introducing CNNs for images, developing new algorithms for optimization, etc.. Some uses of deep learning are computer vision, object detection, classification, or medical image analysis.
2. In sparse encoding, data is represented as a linear combination of a set of basis functions. Sparse encoding optimizes by trying to find a sparse set of coefficients to represent the data where only a few elements are non zero. This can be used for dimensionality reduction, compact feature representation or signal denoising
3. The encoder transforms input data into a compressed representation with the most important features of the data (i.e. dimensionality reduction). The decoder tries to reconstruct the original input data from the transformed representation that it received from the encoder. It is useful since we can extract only the most important features from our data ignoring unnecessary computation or distracting noise. We can avoid identity mapping by using regularization techniques or purposefully adding noise to the input data.
4. Activation function are usually non-linear functions that calculate the output of a given neuron in a neural network. An example would be the Rectified Linear Unit (ReLu). The function is defined as f(x) = max(0,x). Thus, if the input into the neuron is negative the result will always be 0 and the value x otherwise.
5. A 2-D contact map in the context of amino acids is a type of matrix that represents the spatial proximity of given amino acids. The rows and columns correspond to the amino acids and the matrix values indicate if those amino acids are spatially in contact with each other, allowing us to gain insights into the 3-D structure of the protein.
6. Amino acids can be represented by 25-dimensional vectors. 20 values represent the evolutionary information for each amino acid type, 3 binary values represent the predicted secondary structure the amino acid is in and 2 binary values represent the predicted accessibility of the amino acid in the 3-D structure (copied from lecture slide 43)
7. L is the length of the protein sequence and L/5 is the top-scored pairs of amino acids we would analyze for performance measure. The L/5 contract pairs are compared to the real contacts of the amino acids and the rate of false positives/false negatives can be calculated, which in turn allows us to evaluate the performance.
8. One method to measure dependence between two variables is by calculating their correlation. If the correlation is 0 they are independent, if they are between (-1,1)\ {0} they are dependent.
9. Max pooling is used to reduce the dimension of the spatial information in the input data. It selects the max value from neighboring values (e.g. pixels in images) in the input data. It is a procedure with information loss since we lose the non-max values and thus lose some dimensions of the input data.