ANDRIY BURKOV's book

NEURAL NETS

- the dense network is a nested function
  - $y = fNN(x) = f3(f2(f1(x)))$.
  - where f1 is
    - $fl(z)def = gl(Wlz + bl)$
    - where Wis the weights, b is the bias, z is the input, g is activation function, l is the layer index.
    - g is a vector function, W is a matrix with same size as z, b is a vector.
    - *f2 and f3 have the same function with different variables*.
- The parameters W l (a matrix) and b l (a vector) for each layer are learned using the familiar gradient descent by optimizing, depending on the task, a particular cost function (such as MSE).

MULTI LAYER PERCEPTRON

- This FFNN can be a regression or a classification model, depending on the activation function used in the third, output layer.
- The output of each unit is the result of the mathematical operation written inside the rectangle.
- The following happens in each rectangle unit. Firstly, all inputs of the unit are joined together to form an input vector.
- Then the unit applies a linear transformation to the input vector, exactly like linear regression model does with its input feature vector. Finally, the unit applies an activation function g to the result of the linear transformation and obtains the output value, a real number.
- w l,u and b l,u , where u is the index of the unit, and l is the index of the layer.

Feed Forward network

- g last of the last unit is linear, then the neural network is a regression model. If the g last is a logistic function, the neural network is a binary classification model.
- g l,u , assuming it's differentiable 2 . The latter property is essential for gradient descent used to find the values of the parameters w l,u and b l,u for all l and u.
- the neural network to approximate nonlinear functions.
- Without nonlinearities, f NN would be linear, no matter how many layers it has.
- The dimensionality of the vector w l,u equals to the number of units in the layer l − 1.

Deep learning

- exploding gradient and vanishing gradient as gradient descent was used to train the network parameters.
- While the problem of exploding gradient was easier to deal with by applying simple techniques like gradient clipping and L1 or L2 regularization, the problem of vanishing

gradient remained intractable for decades..

- backpropagation using the chain rule.
- the term “deep learning” refers to training neural networks using the modern algorithmic and mathematical toolkit independently of how deep the neural network is.

CONVOLITIONAL NEURAL NETWORKS

- A convolutional neural network (CNN) is a special kind of FFNN that significantly reduces the number of parameters in a deep neural network with many units without losing too much in the quality of the model.
- CNNs have found applications in image and text processing where they beat many previously established benchmarks.
- Because CNNs were invented with image processing in mind.
- If we can train the neural network to recognize regions of the same information as well as the edges, this knowledge would allow the neural network to predict the object represented in the image.
- most important information in the image is local, we can split the image into square patches using a moving window approach.
- We can then train multiple smaller regression models at once, each small regression model receiving a square patch as input..
- The goal of each small regression model is to learn to detect a specific kind of pattern in the input patch.
- a small regression model has to learn the parameters of a matrix F (for “filter”) of size p × p, where p is the size of a patch.
- would need to learn a 3 by 3 parameter matrix F where parameters at positions corresponding to the 1s in the input patch would be positive.
- while the parameters in positions corresponding to 0s would be close to zero.
- convolution of matrices P and F, the value we obtain is higher the more similar F is to P.
- The filter matrix (one for each filter in each layer) and bias values are trainable parameters that are optimized using gradient descent with backpropagation..
- Typically, the ReLU activation function is used in all hidden layers. The activation function of the output layer depends on the task.
- If the CNN has one convolution layer following another convolution layer, then the subsequent layer l + 1 treats the output of the preceding layer l as a collection of size l image matrices. Such a collection is called a volume.
- pooling layer applies a fixed operator, usually either max or average.
- pooling has hyperparameters: the size of the filter and the stride.
- pooling layer follows a convolution layer.
- It also improves the speed of training by reducing the number of parameters of the neural network.
-