Deep Learning final exam

## What is the difference between Bias and Variance and why the trade off is important in Machine Learning?

Bias is the simplifying assumptions that a model makes in order to fit the target function. On the other hand, variance means that the amount that the performance of the model will change on the test data. So a high amount of variance leads to overfitting on the training data. A high amount of bias leads to a bad performance on the training data itself. Therefore, bias and variance are topics which are inversely related. A model cannot have both a low bias and a low variance. Bias and variance are inversely connected. The tradeoff is important in machine learning because of this inverse relation. The data scientist will have to choose a spot where the bias and variance values are both acceptable.

## Describe Stochastic Gradient Descent.

In normal gradient descent what happens is that the gradient of the objective function is used in order to find the minima. This is usually a long process and it takes a lot of steps for the minima to be reached. In gradient descent, when calculating the new values for the coefficients, all the data points in the dataset are taken into consideration. However, if the dataset is extremely huge, the computational complexity will go up by a huge amount. However, if only one data point is taken into consideration, the computational complexity is reduced by a huge factor. This is what is done in stochastic gradient descent. The data point is randomly chosen.

## Why normalization speeds up Deep Learning training?

If normalization is not done in deep learning training, some of the columns in the input values might have extreme values compared to the other columns. Because of this the initial effect of the columns with extreme values is very high compared to the other columns. The weights of the connections of the neural network will have to adjust a lot in order for the model to work effectively. This will take a lot of time. On the other hand, if normalization is done, there are no columns with extreme values which will dominate the columns which have less extreme values. Therefore, the model will take less time to converge.

## What is softmax and when do we use it?

Softmax is a function that is used in the field of multiclass classification. It is usually used in the last layer of a neural network. It converts a vector of numbers into a vector of probabilities with the sum of the probabilities coming up to 1. This can be calculated by calculating the exponent of each value in the list and dividing it by the sum of the exponent values.

## 5- Name a few activation functions, and go into details on Sigmoid activation function.

The activation functions include binary step function, linear function, relu, sigmoid, tanh, leaky relu, parameterized relu.

The sigmoid function has a characteristic s shaped curve or “sigmoid curve”. It exists between 0 and 1. Therefore it is commonly used where the output is a probability. It is a continuous function and is differentiable everywhere.

## 6- What does the convolution layer do in CNN?

In a convolutional layer there exists a kernel or a filter. This filter/kernel is much smaller dimension-wise compared to the input matrix. Convolution is conducted by doing a dot product between the input matrix and the filter/kernel all the while sliding the filter after each dot product operation. The amount by which the filter slides is called as the stride. This process decreases the size of the input image and therefore sometimes padding is done. Padding involved adding fake pixels to the borders of the image.

## 7- Describe LSTM and mention why LSTM is beneficial for timeseries forecasting?

RNNs very often suffer from a problem called as the vanishing or exploding gradient problem. This is because of accumulated multiplicative terms in the rnn. This causes the values to become very huge or very small. LSTMs solve this by introducing new gates, such as input, output and forget gates, which allow for a better control over the gradient flow and enable better preservation of long-range dependencies.

LSTMs are extremely good for time series forecasting because there are lags between important events in a time series. LSTMs are able to capture patterns of both long term seasonalities and short term seasonalities. The different gates inside a LSTM boost its capability for capturing non linear relationships for forecasting.

## 8- How do we split a timeseries data into train and test?

In time series data, we can use a small part of the rear end of the dataset for testing and the rest of the dataset for training. For example if we have 22 years of time series data, one way we could split it into train and test would be to use the first 20 years for training and the last 2 years for testing.