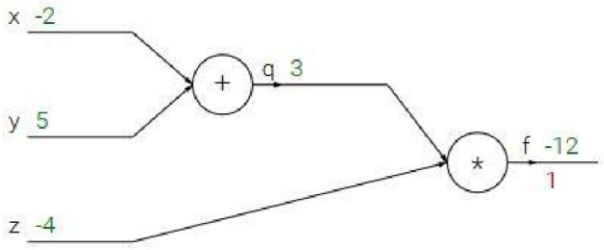
DEEP LEARNING – WORKSHEET 3

Q1 to Q8 are MCQs with only one correct answer. Choose the correct option.

1. Which of the following is true about model capacity (where model capacity means the ability of neural network to approximate complex functions)?
   A) As dropout ratio increases, model capacity increases
   ➡ B) As number of hidden layers increase, model capacity increases
   C) As learning rate increases, model capacity increases
   D) None of the above
2. Batch Normalization is helpful because?
   A) It is a very efficient backpropagation technique
   B) It returns back the normalized mean and standard deviation of weights
   ➡ C) It normalizes (changes) all the input before sending it to the next layer
   D) None of the above
3. What if we use a learning rate that's too large?
   A) Network will not converge
   ➡ B) Network will converge
   C) either A or B
   D) None of the above
4. What are the factors to select the depth of neural network?
   i) Type of neural network (e.g. MLP, CNN etc.)
   ii) Input data
   iii) Computation power, i.e. Hardware capabilities and software capabilities
   iv) Learning Rate
   v) The output function to map
   A) 1, 2, 4, 5
   ➡ B) 2, 3, 4, 5
   C) 1, 3, 4, 5
   D) All of these
5. Suppose you have inputs as x, y, and z with values -2, 5, and -4 respectively. You have a neuron 'q' and neuron 'f' with functions:
   q = x + y
   f = q * z
   Graphical representation of the functions is as follows:



   What is the gradient of F with respect to x, y, and z? (use chain rule of derivatives to find the solution)
   A) (3, -4, -4)
   B) (-3, 4, 4)
   ➡ C) (-4, -4, 3)
   D) (4, 4, 3)
6. Which of the following statement is the best description of early stopping?
   A) Train the network until a local minimum in the error function is reached
   ➡ B) Simulate the network on a test dataset after every epoch of training. Stop training when the generalization error starts to increase
   C) Add a momentum term to the weight update in the Generalized Delta Rule, so that training converges more quickly
   D) None of the above

7. Which gradient descent technique is more advantageous when the data is too big to handle in RAM simultaneously?
   A) Mini Batch Gradient Descent
   ➡ B) Stochastic Gradient Descent
   C) Full Batch Gradient Descent
   D) either A or B
8. Consider the scenario. The problem you are trying to solve has a small amount of data. Fortunately, you have a pre-trained neural network that was trained on a similar problem. Which of the following methodologieswould you choose to make use of this pre-trained network?
   ➡ A) Freeze all the layers except the last, re-train the last layer
   B) Assess on every layer how the model performs and only select a few of them
   C) Fine tune the last couple of layers only
   D) Re-train the model for the new dataset

Q9 and Q10 are MCQs with one or more correct answers. Choose all the correct options.

9. Which of the following neural network training challenge can be solved using batch normalization?
   A) Overfitting
   ➡ B) Training is too slow
   ➡ C) Restrict activations to become too high or low
   D) None of these
10. For a binary classification problem, which of the following activations may be used in output layer?
    ➡ A) ReLU
    B) sigmoid
    ➡ C) softmax
    D) Leaky ReLU

Q11 to Q15 are subjective answer type question. Answer them briefly.

11. What will happen if we do not use activation function in artificial neural networks?
12. How does forward propagation and backpropagation work in deep learning?
13. Explain briefly the following variant of Gradient Descent: Stochastic, Batch, and Mini-batch?
14. What are the main benefits of Mini-batch Gradient Descent?
15. What is transfer learning?

11) If we do not apply an activation function then the output signal would simply be a simple linear function. A linear function is just a polynomial of one degree. A linear equation is easy to solve but they are limited in their complexity and have less power to learn complex functional mappings from data. A Neural Network without Activation function would simply be a linear regression Model, which has limited power and does not perform good most of the times.

12) As the name suggests, the input data is fed in the forward direction through the network. Each hidden layer accepts the input data, processes it as per the activation function and passes to the successive layer.

13) Stochastic gradient descent updates the weight parameters after evaluation the cost function after each sample. That is, rather than summing up the cost function results for all the sample then taking the mean, stochastic gradient descent (or SGD) updates the weights after every training sample is analysed.

In Batch Gradient Descent, all the training data is taken into consideration to take a single step. We take the average of the gradients of all the training examples and then use that mean gradient to update our parameters. So that's just one step of gradient descent in one epoch.

Mini-batch gradient descent is a trade-off between stochastic gradient descent and batch gradient descent. In mini-batch gradient descent, the cost function (and therefore gradient) is averaged over a small number of samples, from around 10-500. This is opposed to the SGD batch size of *1* sample, and the BGD size of *all* the training samples.

14) The advantages of using mini-batch gradient descent is as follows:

1. Easily fits in the memory
2. It is computationally efficient
3. Benefit from vectorization
4. If stuck in local minimums, some noisy steps can lead the way out of them
5. Average of the training samples produces stable error gradients and convergence

15) Transfer learning (TL) is a research problem in machine learning (ML) that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks.