1) The universal approximation theorem states that a feedforward neural network (NN) with a single hidden layer can approximate any function over some compact set, provided that it has enough neurons on that layer. This suggests that the number of neurons is more important than the number of layers. But in practice deep learning is obviously very successful at various prediction tasks. Why is that? Shouldn’t all deep NNs be equivalent to single layered NNs with enough neurons? Why do we need depth when we could theoretically rewrite that neural network with a single layer?

2) In neural networks, what is the role of the activation function and why do we need it?

3) For a softmax of the values (3,4,1,7), without calculating the softmax, can you tell which one is a possible result?
   
   a) 0.0171, 0.0465, 0.0023, 0.9341
   b) 0.0011, 0.0085, 0.0003, 0.7362
   c) 0.0023, 0.0171, 0.0465, 0.9341
   d) 0.0311, 0.0785, 0.0103, 0.9362

4) Let’s say you have three problems which involve a linear regression, a logistic regression, and a small neural net. Which one of them is more likely to benefit from a newly discovered super-fast large matrix multiplication algorithm? Why?

5) Consider the dataset: \( D = \{ x_1, \ldots, x_{100} \} \) with \( x \) in \( \mathbb{R}^3 \). The problem is a 3-class classification problem. Consider a neural network architecture with 2 hidden layers of dimension 4 and 5 - using a sigmoid and a softmax respectively.
   
   a) How would you graphically represent this neural network?
   b) What are the feedforward equations for this neural network?
   c) Describe the relationship between the graphical representation and the feedforward equations. What do the nodes represent? What do the edges represent?
   d) What are the dimensions of \( W_1, b_1, W_2, \) and \( b_2 \), the parameters of the 1st and 2nd hidden layers of the neural network?
   e) What is the total number of parameters in this neural network?

6) Consider a large neural network that is trained to determine the species of bird present in an input image. Briefly explain what the first few layers of the network are likely trying to detect in the input image, and also what the last few layers are trying to detect.