Birla Institute of Technology & Science - Pilani, Hyderabad Campus
First Semester 2020-21
CS F441: Selected Topics from Computer Science (Deep Learning)
Comprehensive Examination

Type: Open (Online) **Time: 120 mins** **Max Marks: 68** **Date: 21.12.2020**

All parts of the same question should be answered together.

1. Suppose you are asked to solve the binary classification task of classifying images as dog vs. non-dog. Suppose you design a CNN with a single output neuron. Let the output of this neuron be z. The final output of your network, y is given by:

y = sigmoid (ReLU(z))

You would like classify all inputs with a final value y ≥ 0.5 as dog images. What major problem are you going to encounter? [8 Marks]

Sol:

we show that the network always outputs the image as 'dog' irrespective of the value z. Equivalently we show that $y \geq 0.5$ irrespective of the input image if the final output of the network is given by $y = \text{sigmoid}(\text{RELU}(z))$.

Case (i): $z \geq 0$

$z \geq 0 \Rightarrow e^z \geq e^0 \Rightarrow e^z \geq 1 \Rightarrow 2e^z \geq 1+e^z$

$\Rightarrow \frac{e^z}{1+e^z} \geq \frac{1}{2} \Rightarrow \frac{1}{1+e^{-z}} \geq \frac{1}{2}$

$\Rightarrow \text{sigmoid}(\text{RELU}(z)) \geq 0.5 \Rightarrow y \geq 0.5$

Case(ii) $z < 0$

$z < 0 \Rightarrow \text{RELU}(z) = 0$

$\text{Sigmoid}(\text{RELU}(z)) = \text{sigmoid}(0)$

$\Rightarrow y = \frac{1}{1+e^{-0}} \Rightarrow y = \frac{1}{1+1}$

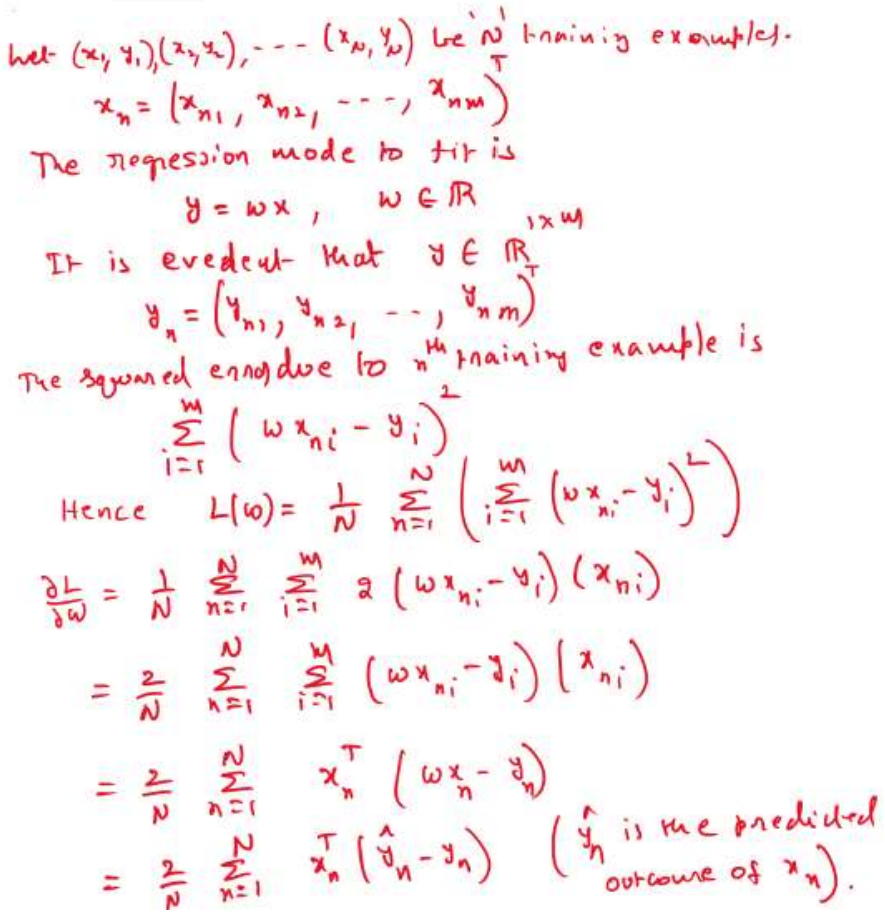
$\Rightarrow y = 0.5$

Thus we proved that irrespective of value of z, $y \geq 0.5$. And hence the output of the network is always a 'dog'.

2. Suppose you are given 'N' training examples. Consider a univariate regression y = wx where w € R, and x € R^{1Xm}. Let the classical error function i.e., L(w) = sum of squares of errors / m. Find the first derivative of L with respect to w, i.e., dL/dw. The final expression should be as compact as possible.

[4 Marks]

Sol:



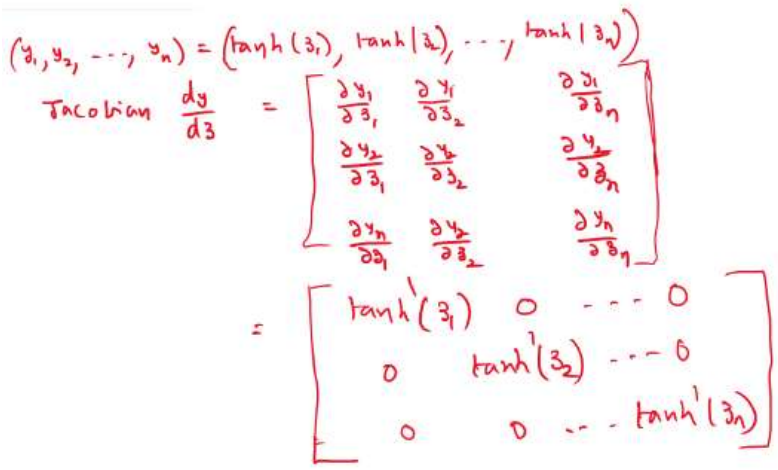
3. Explain the significance of tensors in deep learning and provide a case study in which its usage is maximally exploited? [4 Marks]

Sol: Tensors are multidimensional arrays in Deep Learning that are used to represent data. They represent the data with higher dimensions. Due to the high-level nature of the programming languages, the syntax of tensors are easily understood and broadly used.

Images data – (i, j, (R,G,B))

4. Suppose y = (y1, y2, . . , yn) be the outcome of a layer after tanh is applied on (z1, z2, . . , zn). i.e., (y1, y2, . . , yn) = (tanh z1, tanh z2, . . , tanh zn). Write down the Jacobian matrix dy/dz. You may give your answer in terms of $\tanh^{l}(z)$, the univariate derivative of the tanh function. [6 Marks]

Sol:



5. One of the difficulties with the logistic activation function is that of saturated units. Briefly explain the problem, and whether switching to tanh fixes the problem. (You may refer to your answer from the above question or sketch the activation functions.) [6 Marks]

Sol: No, switching to tanh does not fix the problem. The derivative of sigmoid(z) is small for large negative or positive z. The same problem persists in tanh(z). Both function has a sigmoidal shape. We can see tanh is effectively a scaled and translated sigmoid function: tanh(z) = 2 sigmoid(2z) – 1.

6. Give an example of a data augmentation technique that would be useful for classifying images of cats vs. dogs, but not for classifying handwritten digits. Briefly explain your answer. [4 Marks]

Sol: Flipping the image horizontally; doing this to a dog image would result in a plausible dog image, but not so for an image of a digit.

7. Suppose that you have a model that provides around 80% accuracy on the training as well as on the out-of-sample data test data. Would you recommend increasing the amount of data or adjusting the model to improve accuracy? Please provide suitable reasoning in support of your answer. [4 Marks]

Sol: As the error in the training data and testing data (unseen data) is more or less the same the chosen model is probably not overfitting. So to improve the accuracy, adjusting the model (for example, in the case of polynomial regression, changing the degree of the polynomial etc.) might improve the accuracy.

8. We have seen that multilayer perceptrons are universal for the set of functions mapping binary-valued input vectors to binary valued outputs. [4 Marks]

(a). What do we mean by universal?

Sol: For any function in the set, there is some network that computes it.

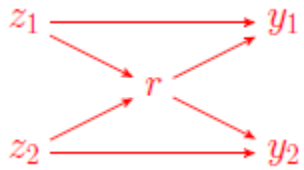
(b). If multilayer perceptrons are universal, why do we still consider other architectures?

Sol: The function might not be compactly representable, i.e. it might require exponentially many units. This means it may be prohibitively expensive to compute, and training such a large architecture wouldn't generalize to new data.

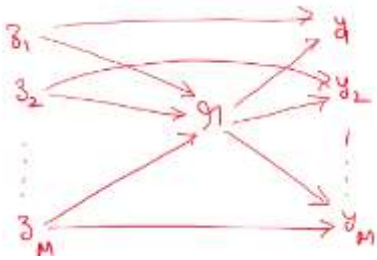
9. We know that the softmax function takes a vector (z1, z2, zM) and returns a vector (y1, y2, . . ,yM). We can express it in the following term:

$r = (e^{z1} + e^{z2} + - - + e^{zM})$ and $yi = e^{zi} / r$

Suppose M = 2, i.e., just two inputs and outputs to the softmax. Draw the computation graph relating z1, z2, r, y1, y2.
For any generic M, draw the computation graph relating z1, z2, r, y1, y2. [6 Marks]
Sol: Consider M = 2, i.e. just two inputs and outputs to the softmax. Draw the computation graph relating z1, z2, r, y1, and y2.



For a generic M, the computational graph is as shown below:



10. Can you think of top two hurdles that computer vision algorithms face in dealing with images? i.e. two characteristics of image formation that make it difficult to recover the image content. [4 Marks]
Sol: There are several examples, such as: (1) they can be from various poses and perspectives, (2) they may vary in color and brightness, (3) there might be occlusion or ambiguity of objects

11. Consider a 1-diemnsional time-series with values 2, 1, 3, 4, 7. Perform a convolution with a 1-dimensional filter 1, 0, 1. Also, for a one-dimensional time series of length L and a filter of size F, what is the length of the output? [4 Marks]
Sol: If a 1–dimensional filter 1, 0, 1 is applied on the 1-diemnsional time-series with values 2, 1, 3, 4, 7 the the output is 5, 5, 10.
If a filter of size F is applied on a one-dimensional time series of length L then the length of the output is L – F + 1.

12. If a data block in a convolutional network has dimension $H \times W \times D = 300 \times 300 \times 256$, and we apply a convolutional filter (or kernel) to it of dimensions $HF \times WF \times D = 11 \times 11 \times 256$, what is the dimension of the output data block? How many weight parameters are there in this simple operation with and without bias? [6 points]
Sol: The size of the output data block is 290 X 290 X 1.
The weight parameters (11 X 11 X 265) without bias parameter and (11 X 11 X 265) + 1 with bias parameter.

13. Can you mention couple of reasons to support the argument that convolutional layers more commonly used than fully-connected layers for image processing? [4 Marks]
Sol: Convolutional layers are better suited for processing images because: (a) they assume that nearby pixels share similar structure, (b) the parameters for a single filter are shared across the entire input image in the sense that a single filter "slides" across an image and produces a new

output with one image channel. This allows the parameters to be invariant to the location of objects in images.

14. Can you identify one option amongst the below mentioned four option to answer the question - When should multi-task learning be used? You should provide appropriate reasoning to support the selected option. [4 Marks]
(i) When your problem involves more than one class label.
(ii) When two tasks have the same dataset.
(iii) When you have a small amount of data for a particular task that would benefit from the large dataset of another task.
(iv) When the tasks have datasets of different formats (text and images).

Sol: Option (ii) is correct. The prior belief is “among the factors that explain the variations observed in the data associated with the different tasks, some are shared across two or more tasks”.
Option (iii) typically refers to refers to transfer learning problem. However with decent explanation, a partial credit can be given.