Semester End Examination
Master of Science
MCSE 304: Deep Learning
Unique Paper Code:223412304
Semester III
December 2022
Year of admission: 2021

Time: Three hours **Max. Marks: 70**

Note: ML denotes machine learning throughout the question paper.

1. For each of the following questions, answer in 3-4 lines: **(7×2 = 14)**

   a. You have 500 examples in a dataset, of which only 175 were examples of preterm births (positive examples, label = 1). To compensate for this class imbalance, an ML engineer decides to duplicate all of the positive examples and then splits the data into train, validation, and test sets. The domain experts tell you that the model should not miss preterm births, but false positives are okay. Using the above-mentioned split, an ML model achieves 100% recall. Will you recommend it for deployment? Or you see a problem with this approach. If so, how do you fix the issue? Justify your answer.

   b. Distinguish between bias and variance of an ML classifier.

   c. Identify a reason that an ML algorithm may be stuck with the problem of
      i. exploding gradients
      ii. vanishing gradients

   d. You are training a single-layer, feedforward neural network with a softmax activation function in the final layer to classify among two classes, with a cross-entropy loss as the training objective. An engineer You decide to independently sample your initial weights from a Gaussian distribution of mean 0, standard deviation 0.0001. You can assume perfect class balance in the dataset. Do you expect the neural network to work efficiently? Justify your answer.

   e. You would like to train a dog/cat image classifier using mini-batch gradient descent. You have already split your dataset into the train, dev, and test sets. The classes are balanced. You realize that within the training set, the images are ordered so that all the dog images come first and all the cat images come after. A friend tells you:" you need to shuffle your training set before the training procedure." Do you agree with your friend? Justify your answer.

   f. You are solving a biometric authentication task (modeled as binary classification) that uses fingerprint data to help users log into their devices. You train a classification model for user A until it achieves > 95%

accuracy on a validation set (for the same user). However, upon testing, you get complaints the model fails to correctly authenticate user A about half the time (50% misclassification rate). Mention one factor you think could have contributed to the mismatch in misclassification rates between the validation set and test set and how you would go about fixing this issue.

g. An ML engineer designs a fully connected neural network architecture where all activations are sigmoids and initialize the weights with large positive numbers. Is this a good idea? Justify your answer.

2. a. Consider the convolutional neural network defined by the layers in the left column below. Fill in the shape of the output volume and the number of parameters at each layer. You should write the shapes in the NumPy format, for example, (128,128,3). **(10 + 4)**

Notation:

- CONV-K-N denotes a convolutional layer with N filters, each of size K×K, Padding, and stride parameters are always 0 and 1, respectively.
- POOL-K indicates a K × K pooling layer with stride K and padding 0.
- FC-N stands for a fully-connected layer with N neurons.

Layer	Activation Volume Dimensions	Number of Parameters
INPUT	$128 \times 128 \times 3$	0
CONV-9-32		
POOL-2		
CONV-5-64		
POOL-2		
CONV-5-64		
POOL-2		
FC-3		

b. Design a convolutional neural network for the task of digit prediction: given a 32×32 image (with 3 channels i.e., RGB), you want to predict the lowercase alphabet shown in the image. Therefore, this is a 26-class classification problem. Your network will have four layers, given in order: a convolutional layer, a max-pooling layer, a flatten layer, and a fully-connected layer. Write a code snippet to design the neural network to solve this problem.

3\. a. What are the problems associated with the basic beam search algorithm? How do we solve them? **(2+4+5+3 = 14)**

b. Give the architecture of an LSTM network, clearly stating the semantics of each component.

c. Suppose you are training an LSTM unit for speech recognition application. You have a vocabulary of 100,000 words. Input(X) is 1000-dimensional, and all activations are 100-dimensional. Determine the dimension of the following at each time-step:

- $c^{<t>}$
- $\tilde{c}^{<t>}$
- Γ_u
- Γ_f
- Γ_o

d. What are the pros and cons of character-level language models compared to word-level language models?

4\. Answer the questions based on the different explainable AI methods: **(3 + 4 + 4 + 3 = 14)**

a. Consider the following explanation generated by LIME for classifying an image as a wolf. Interpret LIME's explanation.

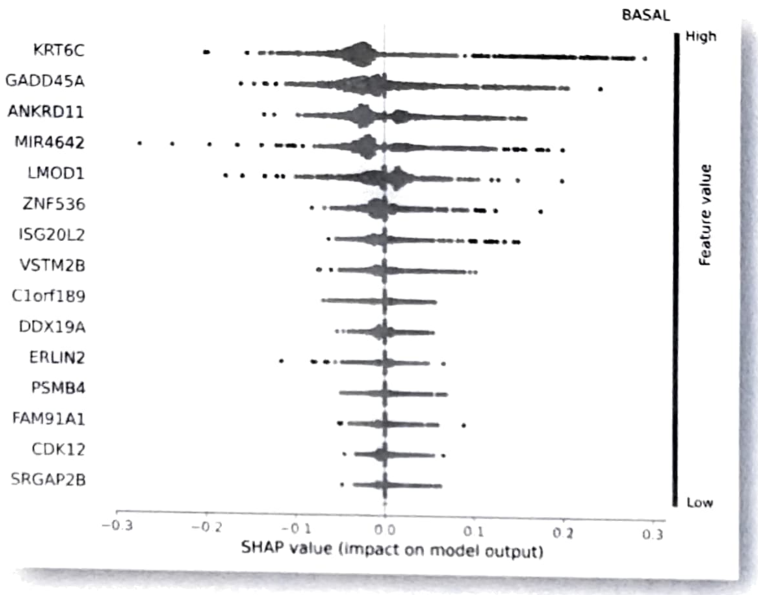
(a) Husky classified as wolf (b) Explanation

b. Describe the SHAP method. Consider a dataset with F as the Feature set. Let S be a subset of the feature S, and f_S be the output of the model using feature subset S. Consider the formula for determining the shapley value of feature i:

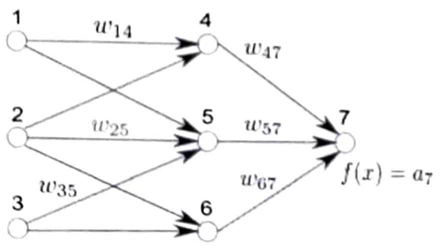
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \right].$$

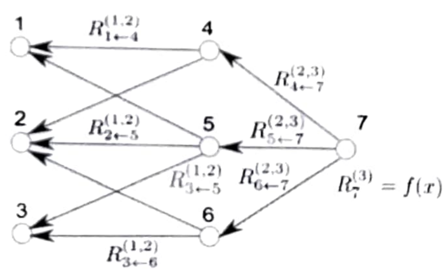
Describe how the above formula captures the role of i^{th} feature.

a. Consider the following figure depicting SHAP values of features determined by the SHAP method. Determine which feature is contributing most towards the prediction. Justify your answer.



b. Consider the following network with one input layer, followed by a hidden layer and an output layer:





Using the Layerwise Relevance Propagation Method, determine the contribution of node 5 towards R2 (the relevance of node 2).

5. a. Why does the MobileNet architecture prefer the depth-wise separable convolution over standard convolution? 14

b. Using the image that follows, describe the following terms in the context of YOLO detection system

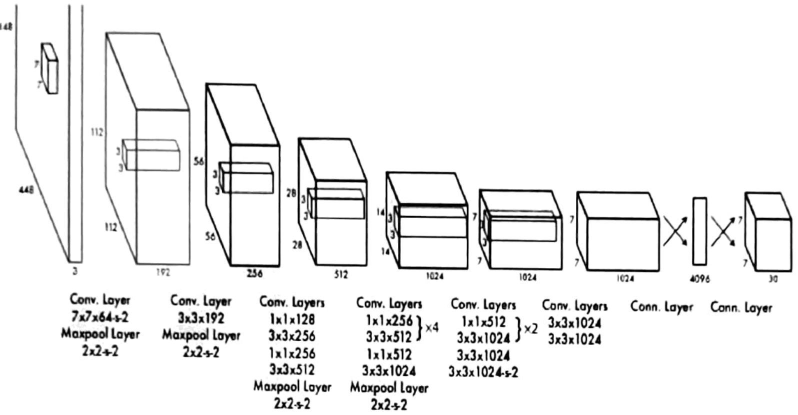
(i) Bounding Box

(ii) Intersection over Union

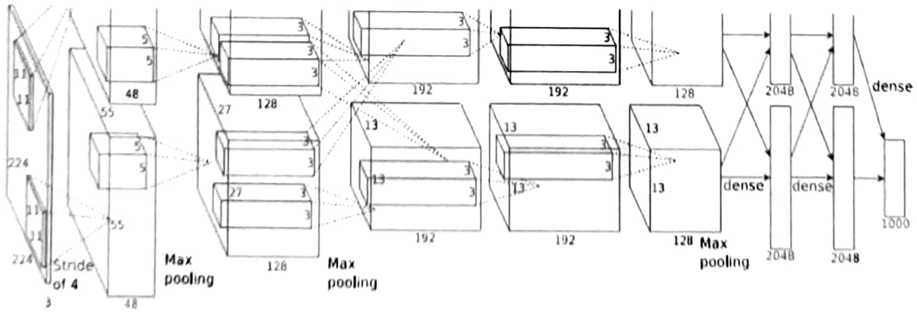
(iii) Non-max suppression

c. Which of the following neural network architectures corresponds to YOLO ~~ImageNet~~. What are its hyperparameters? What is the basis of the loss function? What are the merits of the proposed architecture?

I.



II.



III.

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$