Semester End Examination
Master of Computer Applications
MCAE 505: Neural Networks
Unique Paper Code: 223402505
Semester V
Nov-Dec 2022
Year of admission: 2020

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

1. For each of the following questions, answer in 3-4 lines: **(7×2 = 14)**
   A. Weight sharing allows CNNs to deal with image data without using too many parameters. Does weight sharing increase the bias or the variance of a model?
   B. Which of the techniques can be used to reduce model overfitting?
   C. Why is the sigmoid activation function susceptible to the vanishing gradient problem?
   D. Why is it necessary to include non-linearities in a neural network?
   E. Say you are trying to solve a binary classification problem where the positive class is very underrepresented (e.g. 9 negatives for every positive). Describe precisely a technique which you can use during training which helps alleviate the class imbalance problem. Would you apply this technique at test time? Why or why not?
   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 same user). However, upon testing, you get complaints the model fails to correctly authenticate user A about half the time (50% misclassification rate). List one factor you think could have contributed to the mismatch in Misclassification rates between the validation set and test set, and how you'd go about fixing this issue.
   G. You are training a large feedforward neural network (100 layers) on a binary classification task, using a sigmoid activation in the final layer, and a mixture of tanh and ReLU activations for all other layers. You notice your weights to a subset of your layers stop updating after the first epoch of training, even though your network has not yet converged. Deeper analysis reveals the gradients to these layers completely, or almost completely, go to zero very early on in training. How will you resolve the problem?

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 can write the shapes in the numpy format (e.g. (128,128,3)). **(10 + 4)**
   Notation:
   - CONV5-N denotes a convolutional layer with N filters with height and width equal to 5, padding is 2, and stride is 1.
   - POOL2 denotes a 2x2 max-pooling layer with stride of 2 and 0 padding.

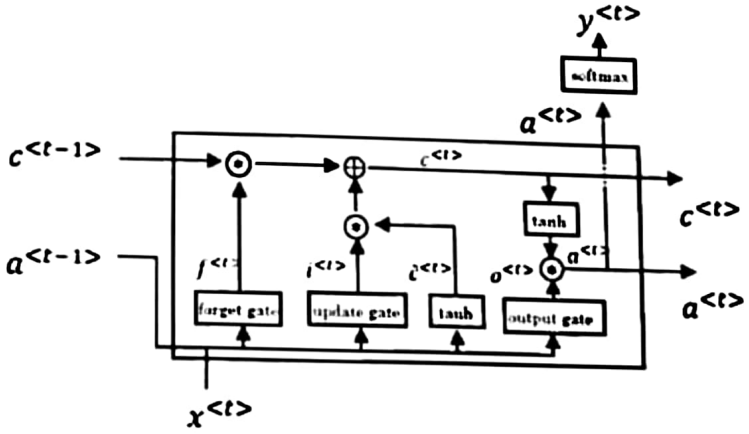
- FC-N denotes a fully-connected layer with N neurons

Layer	Activation Volume Dimensions	Number of Parameters
Input	$32 \times 32 \times 1$	
CONV5-10	$32 \times 32 \times 10$	$(5 \times 5 \times 1)10$
POOL2	$16 \times 16 \times 1$	$(2 \times 1 \times 10)$
CONV5-10	$16 \times 16 \times 10$	$(5 \times 5 \times 1)10$
POOL2	8×8	$(2 \times 1 \times 10)$
FC10	10×64	$10 \times 64 + 10$

B. Design a convolutional neural network for the task of digit prediction: given a 16 x16 image (with 3 channels i.e., RGB), you want to predict the digit shown in the image. Therefore, this is a 10-class classification problem. Your network will have 4 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. How does Beam Search Algorithm work? Why does it perform better than the Greedy Search Algorithm? How does length normalization improve its performance? How can we identify if our beam search algorithm or the underlying recurrent neural network model at fault? Illustrate with the help of an example in the context of language translation. **(4+2+5+3 14)**

B. Consider the following architecture of LSTM:



What is the role of the Forget gate? Give the formula to compute $c^{<t>}$?

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:

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

D. Illustrate the functioning of the skip-gram model with an example. What is the major disadvantage of the skip-gram model which makes it impractical? Give a method to overcome the disadvantage of skip-gram model.

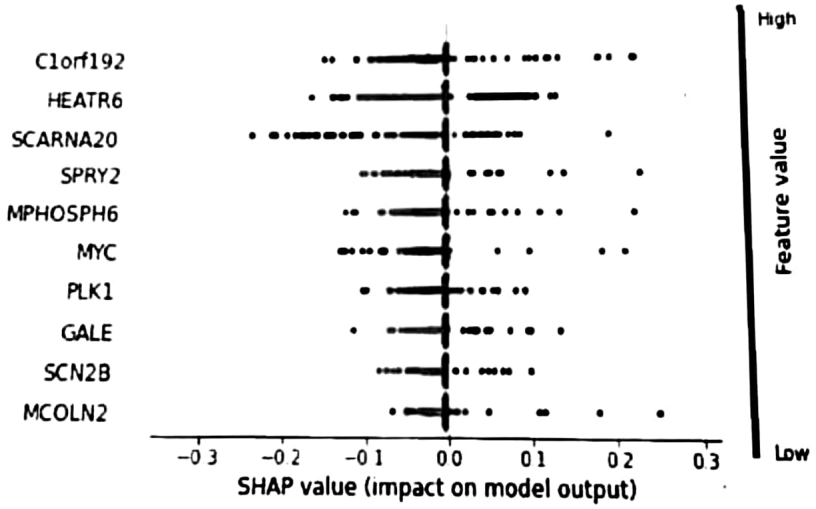
4. Answer the questions based on the different explainable AI methods: **(6 + 4 + 8 + 2 = 14)**

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

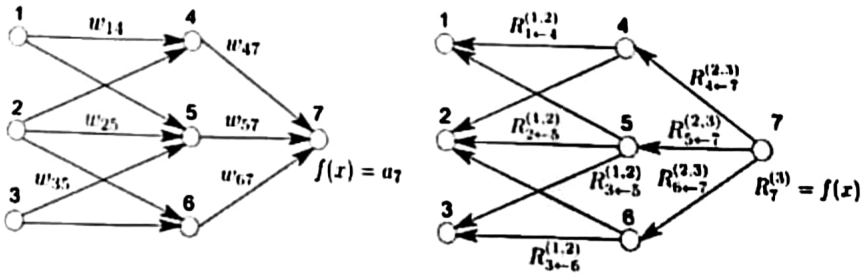
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].$$

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

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



C. 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 4 towards R2 (relevance of node 2). Further, determine the total relevance of node 2 (R2).

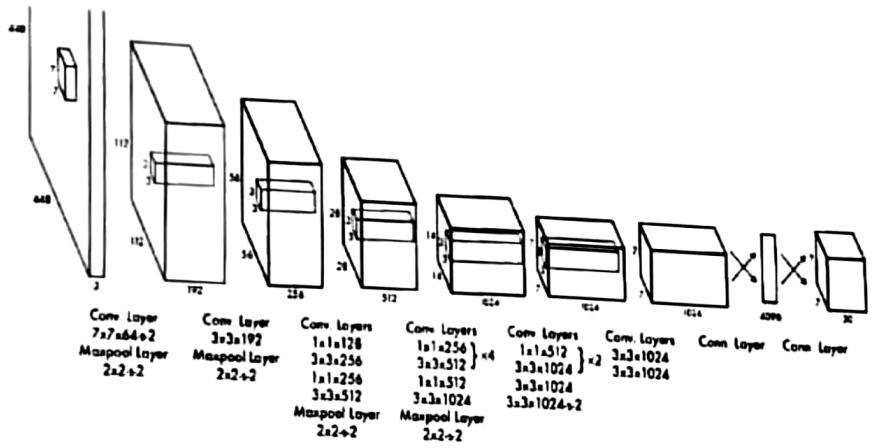
D. Consider the following attribute values which led the model to predict "STROKE":

Age = 59	Gender = M	Body Mass Index =30

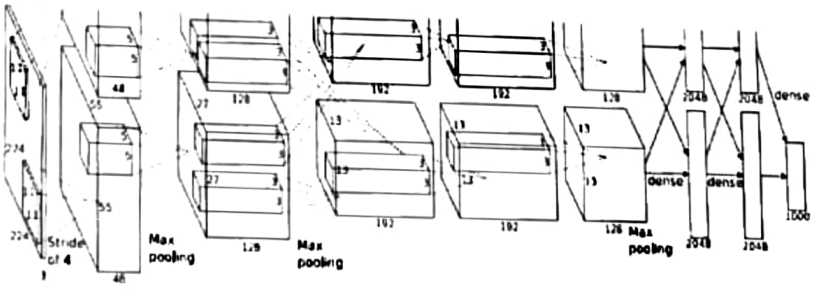
Suggest some counterfactuals for this record that might reverse this decision. Justify your answer.

5. Match the following neural network architectures to their names (i) YOLO detection system, (ii) ImageNet, and (iii) MobileNet (which do not necessarily appear in the correct order). For each of these networks, list the dataset used for training, its hyperparameters, loss function, and merits of the proposed architecture. 14

A.



B.



C.

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