SRN □□□□□□□□□□□□

PES University, Bengaluru-560085

(Estd. Under Karnataka Act 10 of 2013)

UE16CS414

DEC 2019: END SEMESTER ASSESSMENT (ESA) B.TECH. VII SEMESTER

UE16CS414 – TOPICS IN DEEP LEARNING

Time: 3 Hrs | Answer All Questions. | Max Marks: 100

Instructions: ANSWER TO THE POINT

1. a) Using appropriate equations, prove that the Adam optimizer usually converges faster than the RMSProp. — 6

b) Explain the Q-Learning algorithm for non-deterministic rewards and actions. Thence derive the Q-learning formula for the deterministic case. — 3+2

c) List the key features of Deep Learning. For the following use cases, suggest an effective deep learning technique/architecture with a strong one-line justification for your choice. — 3+6

i. Face Recognition (rotation invariant)
ii. Text translation from Hindi to English, each input sentence around 50 words long.
iii. Antarctic Ice formation pattern recognition, training data set size = 1000 instances.

2. a) Starting with an expression for the margin in terms of the weight vector, derive the "BOX" constraint for the soft-margin SVM and bring out the significance of the cost factor 'C' and the slack variable. — 2+3 +3

b) Prove mathematically that the Gaussian kernel is extremely effective in accomplishing linear separability. — 4

c) Prateek has 2D numerical data with him. There are 5 classes possible. Briefly explain to him the o-v-o and the o-v-r techniques for classification using SVM. If each classifier takes 2 seconds to train, find the **total** time needed for training, if he tests out **all of** the below techniques. — 4+4

a) o-v-o b) o-v-r and c) DAGSVM.

3. a) With a neat schematic diagram, explain the structure of a GRU cell and hence show how GRU overcomes the "memory" problem of the vanilla RNN. Also argue that a GRU-network generally trains faster than an LSTM-network and uses lesser memory. — 2+4 +2

b) Read the following keras code carefully and answer the questions that follow. (N is a variable) — 4+3

```
1   model = Sequential()
2   model.add(GRU(N, activation='relu',
                  input_shape=(None, 4),return_sequences=True))
3   model.add(GRU(N, activation='relu'))
4   model.add(Dense(1))
5   model.compile(optimizer='adam', loss='mse')
6   model.summary()
```

If LSTMs were used instead of GRUs, then the total parameters to be learnt would be 240 more (all biases included).

i) Solve for N
ii) If line 4 was changed to `model.add(Dense(1), activation='sigmoid'),` what would be the ideal change to line 5? Justify clearly.

SRN

<table>
<tr><td></td><td>c)</td><td>What do you understand by KL-Divergence? Explain its significance in the training of a Variational Autoencoder.</td><td>2+3</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>4.</td><td>a)</td><td>In a paper at ICLR 2016, Huszar argues that GANs perform better than VAEs because GANs try to reduce the Jensen-Shannon divergence instead of the KL-Divergence. Starting with the Gain function for a GAN, prove that the Generator indeed tries to minimize the JS-Drivergence between the real distribution and the generator's distribution. Hence prove that the Nash equilibrium is -2log(2) and that the corresponding Discriminator output is 0.5</td><td>1+4
+3</td></tr>
<tr><td></td><td>b)</td><td>Read the following keras code and answer the questions that follow.
<pre>
1   model = Sequential()
2   model.add(Conv2D(8, kernel_size=(2,2),
              padding="same", input_shape=(32,32,1)))
3   model.add(LeakyReLU(alpha=0.1))

4   model.add(Conv2D(16, kernel_size=(4,4),
              padding="valid", strides=2))
5   model.add(LeakyReLU(alpha=0.1))
6   model.add(BatchNormalization())
7   model.add(Dropout(0.2))

8   model.add(Conv2D(32, kernel_size=(2,2), padding="same"))
9   model.add(LeakyReLU(alpha=0.1))
10 model.add(MaxPooling2D(pool_size=(2,2)))
11 model.add(Flatten())
</pre>
i. Explain line 6 clearly.
ii. Explain line 7 briefly
iii. How many parameters are learnt at the 2nd Conv layer
iv. How many parameters are learnt at the 3rd Conv layer
v. What is the size of the activation map at the 1st conv layer?
vi. What is the size of the activation map at the 2nd conv layer?</td><td>3+2
+4</td></tr>
<tr><td></td><td>c)</td><td>List the three main advantages of Transfer Learning compared to normal training of Deep Networks.</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>a)</td><td>With a schematic diagram, briefly explain the basic encoder-decoder architecture for Seq2Seq models. With respect to this, what do you understand by "Teacher force"?</td><td>2+3
+2</td></tr>
<tr><td></td><td>b)</td><td>Justify the need for "Attention". Clearly bring out the working of Bahdanau Attention with the necessary math.</td><td>2+5</td></tr>
<tr><td></td><td>c)</td><td>Explain the intuition behind Capsule Networks. Clearly explain the CapsuleNets loss function as elucidated in the "Dynamic Routing between Capsules" paper by Hinton.</td><td>3+3</td></tr>
</table>