National University of Computer and Emerging Sciences

Generative AI (AI-4009)
Date: September 23, 2024
Course Instructor
Akhtar Jamil

Sessional-I
Total Time: 1 Hour
Total Marks: 50
Total Questions: 04

Semester: FALL-2024
Campus: Islamabad

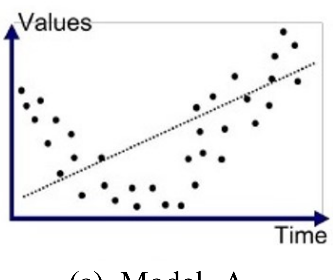
________________________________ ______________ ___________ ________________

Student Name Roll No Section Signature

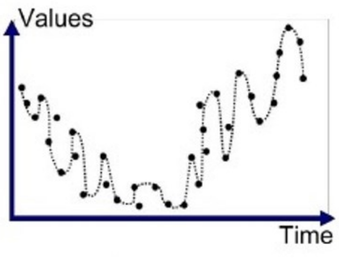
Question No 1. MCQ [20]

Answer the MCQs on the given answer sheet attached at the end of the question paper. Answers marked on the question paper will not be evaluated.

1. **Which of the following models has a triangular context?**
   a). PixelCNN
   b). Row LSTM
   c). Diagonal BiLSTM
   d). None of the above
2. **What does the receptive field of a neuron in a convolutional neural network (CNN) refer to?**
   a). The number of filters applied to an input.
   b). The number of parameters in a convolutional layer.
   c). The stride and padding used in convolution.
   d). **The area of the input image that influences the activation of a neuron.**
3. **Which of the following models shows that it is overfitting?**



(a) Model -A



(b) Model-B

   a). Model-A
   b). Model-B
   c). Both
   d). None
4. **In a BiLSTM, if we pass the original input through the network and then reverse the input and pass it again through the same network, is it equivalent to how a BiLSTM operates?**

   a). True
   b). False

5. **Consider a model that learns the following joint probability distribution for three sequential inputs (X_1), (X_2), and (X_3) along with the label (Y):**

$$P(X_1, X_2, X_3, Y) = P(Y) \cdot P(X_1 \mid Y) \cdot P(X_2 \mid X_1, Y) \cdot P(X_3 \mid X_2, X_1, Y)$$

Is this model generative or discriminative?

a). Generative
b). Discriminative
c). None

6. **Long-range dependencies can be introduced in autoregressive models using**

a). CNN
b). ANN
c). RNN
d). None of the above

7. **What is the primary function of the hidden layer(s) in an Artificial Neural Network (ANN)?**

a) To store training data for backpropagation.
b) To calculate the error for weight updates.
c) To match the number of input and output neurons.
d) To capture non-linear relationships in the data.

8. **Which activation function is most appropriate if your model needs to output probabilities for multiple classes?**

a) ReLU
b) Tanh
c) Softmax
d) None

9. **Which of the following is not a hyperparameter**

a) Learning rate
b) Batch size
c) Epochs
d) Bias

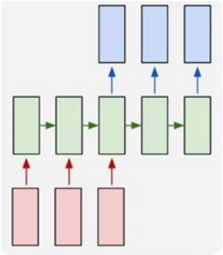
10. **When a model has high bias, this indicates that:**

a) The model is overfitting
b) The model is underfitting
c) None of the above

11. **Consider an input of size 64×64×10. To reduce the depth to 5 without changing the spatial dimensions, which of the following strategies would be the most suitable?**

a) Use a convolution with a kernel size of 3×3, and 5 filters.
b) Use a convolution with a kernel size of 1×1, and 5 filters.
c) Apply max pooling with a pool size of 2×2.
d) None of the above.

12. **Consider an RNN architecture shown below. For which of the following problems is this architecture most appropriate?**



a) Video Classification
b) Image Captioning
c) Text Summarization
d) Sentiment classification

13. What is one of the key advantages of using skip connections in deep neural networks?

a) They reduce the number of parameters.
b) **They help train deeper networks effectively.**
c) They prevent vanishing gradient.
d) They eliminate the need for activation functions.

14. What is the effect of setting the regularization parameter λ too large?

a) The model becomes more likely to underfit the data.
b) The model becomes more likely to overfit the data.
c) The model's capacity to generalize improves.
d) The optimization process becomes faster.

15. In a VAE, how does a high KL Divergence affect data generation?

a) The generated data will be more diverse but less accurate.
b) The generated data will lack diversity.
c) The generated data will be unrealistic and may include garbage.
d) The model will generate high-quality samples.

16. What is the role of the entropy term in a Variational Autoencoder (VAE)?

a) It helps generate specific outputs.
b) **It encourages diversity in generated samples.**
c) It minimizes the reconstruction error of the decoder.
d) It ensures that the latent variables follow a Gaussian distribution.

17. What is the primary purpose of the reparameterization trick in a Variational Autoencoder (VAE)?

a) To train the encoder and decoder.
b) **To make backpropagation possible through the network.**
c) To fix the latent variable distribution.
d) To reduce input dimensionality.

18. What is the primary purpose of transpose convolutions (also known as deconvolutions) in neural networks?

a) To increase the depth of the feature maps.
b) To reduce the spatial dimensions of the input.
c) To perform element-wise multiplication between feature maps.
d) **To increase the spatial dimensions of the input**.

19. For which of the following problems, RNN is not suitable:

a) Time series forecasting
b) Sentiment analysis on text data
c) Image classification
d) Speech recognition

20. When a model has high variance, this indicates that:

a) The model is overfitting
b) The model is underfitting
c) None of the above

Question No 2. [2 x 10=20]

Write short answers to the following questions (not more than five lines recommended)

1. **In the DiagonalBiLSTM, a 2x1 convolution kernel is used. Is it useful to increase its size to 3x1 or 5x1? Write reason.**

In the case of DiagonalBiLSTM, the receptive field is already designed to capture global dependencies across the data, increasing the kernel size beyond 2x1 does not significantly contribute to broadening the receptive field.

2. **What are Skip connections? Why are they useful? How can we select the number of skip connections?**

Skip connections are used in neural network where outputs from earlier layers are fed directly to later layers, bypassing one or more intermediate layers. They are useful for preventing the vanishing gradient problem, allowing for deeper networks, and preserving information across layers. The number of skip connections is usually determined based on network depth, task complexity, and empirical testing, balancing performance improvements with computational cost.

3. **What will be the impact if we do not normalize the input data, e.g., in the case of images.**

   If data is not normalized, the model may struggle to converge during training, leading to slower learning and potentially poorer performance.

4. **Why do we need activation functions in neural networks?**

- Non-linearity help models to learn complex patterns in the data
- Enable the network to model complex relationships between inputs and outputs

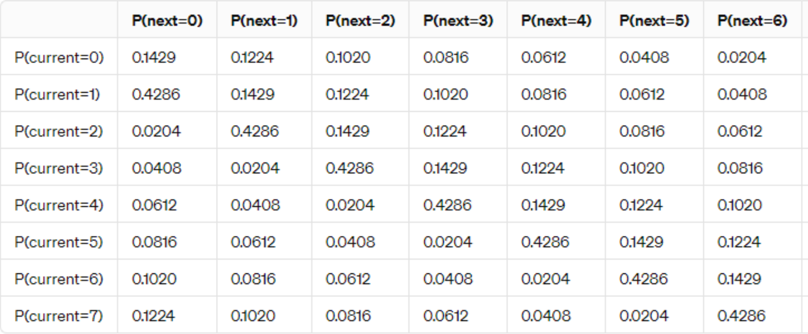
5. **What is the main disadvantage of finite memory autoregressive models?**
   Not able to capture long-term dependencies and may fail if relevant historical context is needed.

6. **In an LSTM, what is the impact on the performance of the model if the forget gate consistently outputs values close to 0 during training?**

If the forget gate outputs values close to 0, the LSTM will quickly forget previous information, preventing it from learning long-term dependencies and leading to poor performance on tasks requiring memory of earlier data.

7. **What is the role of regularization in training deep neural networks?**

   Regularization help to prevent overfitting, where the model performs well on training data but poorly on unseen data.



	P(next=0)	P(next=1)	P(next=2)	P(next=3)	P(next=4)	P(next=5)	P(next=6)
P(current=0)	0.1429	0.1224	0.1020	0.0816	0.0612	0.0408	0.0204
P(current=1)	0.4286	0.1429	0.1224	0.1020	0.0816	0.0612	0.0408
P(current=2)	0.0204	0.4286	0.1429	0.1224	0.1020	0.0816	0.0612
P(current=3)	0.0408	0.0204	0.4286	0.1429	0.1224	0.1020	0.0816
P(current=4)	0.0612	0.0408	0.0204	0.4286	0.1429	0.1224	0.1020
P(current=5)	0.0816	0.0612	0.0408	0.0204	0.4286	0.1429	0.1224
P(current=6)	0.1020	0.0816	0.0612	0.0408	0.0204	0.4286	0.1429
P(current=7)	0.1224	0.1020	0.0816	0.0612	0.0408	0.0204	0.4286

8. Consider the following transition probabilities derived for a 3-bit image shown above.

If the current pixel sequence is […,2,3,4]. Calculate the joint probability of the next two pixel values being 5 and 6: $P(\text{next} = 5, 6 \mid \text{current} = 4)$. Assume that you know the probability of the current sequence up to pixel value 4.

Here are the steps with their respective probabilities:

$$P(\text{next} = 5|\text{current} = 4) = 0.1224$$
$$P(\text{next} = 6|\text{current} = 5) = 0.1224$$

The joint probability is calculated as:

$$P(\text{next} = 5,6 \mid \text{current} = 4) = \text{P}(\text{next} = 5| \text{ current} = 4) \times P(\text{next} = 6|\text{current} = 5)$$
$$= 0.1224 \times 0.1224$$
$$= 0.01498$$

9. Perform a 1D convolution operation with the input vector $\boxed{3}\boxed{5}\boxed{2}\boxed{7}\boxed{1}$ using the kernel [−1,0,1] with no padding and a stride of 2. Show your work for each step of the convolution process and compute the final output.

$[-1,-1]$

10. Consider a gradient vector $\mathbf{g} = [4, 3, 2]$ and a threshold value $\boldsymbol{\theta} =$ 5. Perform gradient clipping. Show all the steps.

Calculate the norm of the gradient vector:

$$||\text{g}||= \sqrt{g_1^2 + g_2^2 + g_3^2}$$

$$||g|| = \sqrt{g_1^2 + g_2^2 + g_3^2]}$$

$$||g|| = \sqrt{4^2 + 3^2 + 2^2}$$

$$||g|| = 5.385$$

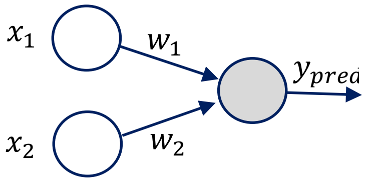
As ||g||> θ i.e. , 5.385 > 5 so, we clip the gradient as:

$$g = g \times \frac{\theta}{||g||}$$

$$g = 4\,x\,\frac{5}{5.385},\ 3\,x\,\frac{5}{5.385},\ 2\,x\,\frac{5}{5.385}$$

$$g = [3.71,2.79,1.86]$$

Question No 3. [10]



Consider an Artificial Neural Network (ANN) with two inputs ($x_1 = 2$) and ($x_2 = 3$) and a single output (y_{pred}). The network has a single layer where both weights ($w_1 = w_2 = 0.5$), no activation function is used, and there is no bias term. You are given the true target value ($y_{true} = 10$). The loss function is the Mean Squared Error (MSE), defined as:

$$L = \frac{1}{2}\left(y_{true} - y_{pred}\right)^2$$

1. Calculate the output (y_{pred}) for the given inputs and weights.
2. Perform one step of backpropagation to update the weights using gradient descent with a learning rate ($\eta = 0.01$). Calculate the new values for w_1 and w_2.

Note: You need to apply chain rule, e.g. $\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial w_1}$

Solution:

1. Forward Pass

$$y_{pred} = w_1 \cdot x_1 + w_2 \cdot x_2$$
$$y_{pred} = (0.5 \cdot 2) + (0.5 \cdot 3) = 1 + 1.5 = 2.5$$

2. Calculate Loss

$$L = \frac{1}{2}\left(y_{true} - y_{pred}\right)^2$$

$$L = \frac{1}{2}(10 - 2.5)^2$$
$$= \frac{1}{2} \cdot (7.5)^2$$
$$= \frac{1}{2} \cdot 56.25 = 28.125$$

3. Backward Pass (Weight Update)

$$L = \frac{1}{2}\left(y_{true} - y_{pred}\right)^2$$

$$\frac{\partial L}{\partial y_{pred}} = -\left(y_{true} - y_{pred}\right)$$
$$= -(10 - 2.5)$$
$$= -7.5$$

Let us calculate the gradients:

$$L = \frac{1}{2}\left(y_{true} - y_{pred}\right)^2$$
$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial w_1}$$

$$y_{pred} = w_1 \cdot x_1 + w_2 \cdot x_2$$

$$\frac{\partial y_{pred}}{\partial w_1} = x_1$$

Similarly,

$$\frac{y_{pred}}{\partial w_2} = x_2$$

Now, calculate the gradient of each weight:

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial w_1}$$
$$= -7.5 \cdot x_1$$
$$= -7.5 \cdot 2$$
$$= -15$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial y_{pred}} \cdot \frac{\partial y_{pred}}{\partial w_2}$$

$$= -7.5 \, . \, x_2$$
$$= -7.5 \, . \, 3$$
$$= -22.5$$

Weight Update using gradient descent:

$$w_1' = w_1 - \eta \cdot \frac{\partial L}{\partial w_1}$$
$$= 0.5 - 0.01 \cdot (-15)$$
$$= 0.5 + 0.15$$
$$= 0.65$$

$$w_2' = w_2 - \eta \cdot \frac{\partial L}{\partial w_2}$$
$$= 0.5 - 0.01 \cdot (-22.5)$$
$$= 0.5 + 0.225$$
$$= 0.725$$