Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
CSL7590 - Deep Learning

February 6, 2024

Instructions:
Read the questions carefully. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

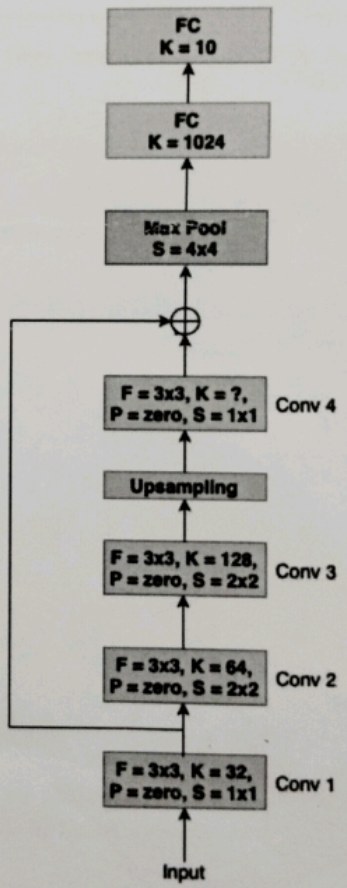
Time: 45 min | *Quiz-1* | Maximum Marks: 10

1. How do CNNs achieve invariance and equivariance? [2]

2. Explain the weight sharing characteristic of CNNs. [2]

3. Consider a CNN shown on the right where Conv and FC represent convolutional and fully connected layers, respectively. F, K, and S represent filter size, number of filters or neurons, and stride, respectively. Each Conv layer uses zero padding (P). Answer the following for two input sizes (W×H×C), i) 128×128×3, ii) 96×96×1 while showing calculation steps. [6]

   1. Calculate the size of feature (activation) map before and after the Upsampling layer.
   2. Calculate the value of K for Conv 4 layer.
   3. Calculate the total number of trainable parameters while assuming that there are no parameters in the Upsampling layer.



-Optimizers

128-3

126×126×32

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

February 9, 2024

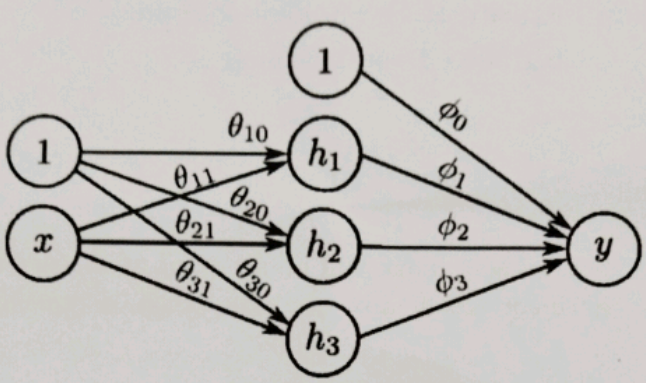
CSL7590 - Deep Learning

Instructions:
1. Read the questions carefully.
2. All questions are mandatory.
3. If a question requires justification, zero mark will be awarded in absence of the justification.
4. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 1 hour | *Minor-I* | Maximum Marks: 18

1. [True/False] Applying ReLU activation before or after average pooling has no difference. Justify your answer. [2]
2. [True/False] Residual connection in CNNs reduces the problem of vanishing gradient? Justify your answer. [2]
3. Which of the followings can or cannot be used as activation function for the forget gate in LSTM and why? [2]
   1. $a(z) = \frac{1}{1+e^{-z}}$
   2. $a(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$
   3. $a(z) = \min(1, \frac{z+\lvert z\rvert}{2})$
   4. $a(z) = \ln(1 + e^z)$
4. Explain the role of positional embeddings in transformers. [2]
5. Consider a self-attention mechanism that processes N inputs of length D to produce N outputs of the same size. How many weights (excluding biases) are used to compute the queries, keys, and values? How many attention weights will there be? [2]
6. Consider a simple function $f(x, y, z) = q_1(x, y) \times q_2(x, z)$, where $q_1(x, y) = x + y$ and $q_2(x, z) = x^z$. Now let us assume that we are evaluating this function at $x = -2, y = 5,$ and $z = -4$. In addition let the value of the upstream gradient (gradient of the loss with respect to our function, $\frac{\partial L}{\partial f}$) is equal to 1. We use gradient descent to update $x, y,$ and z with a learning rate of 0.1. Find out the values of the parameters after all of those are updated once. [2]
7. Why is it critical to reduce learning rate in SGD? [2]

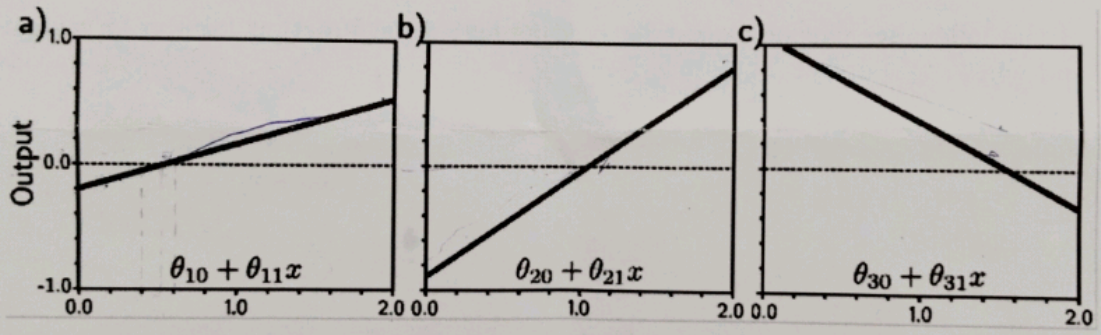
8. Consider the neural network shown below. This network realizes the following mapping [4]



between input (x) and output (y).

$$y = \phi_0 + \phi_1 a(\theta_{10} + \theta_{11}x) + \phi_2 a(\theta_{20} + \theta_{21}x) + \phi_3 a(\theta_{30} + \theta_{31}x)$$

where $a(z) = \text{ReLU}(z)$ represents ReLU activation. Furthermore, the preactivation outputs corresponding to the three hidden layer neurons, $h_1, h_2,$ and h_3, are shown below. If $\phi_0 =$



$-0.2, \phi_1 = -2, \phi_2 = 2, \phi_3 = 0.5$ draw the hidden layer outputs obtained after the activation and the corresponding output of the network (y). Please show the axis marks carefully.

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
CSL7590 - Deep Learning

March 15, 2024

Instructions:
Read the questions carefully. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 45 min | *Quiz-2* | Maximum Marks: 10

1. For a transfer learning task, which layers are generally transferred to another task? [1]

a) Higher layers (close to o/p) b) Lower layers (close to i/p)
c) Task Specific d) None

2. A sentiment predictor is trained on customer reviews of media content such as books, videos and music, it is then used to analyze comments about consumer electronics such as televisions or smartphones, this scenario describes? [1]

a) Concept drift b) Domain adaptation
c) Multi task learning d) None

3. Which of the following is an effective way of network compression? [1]

a) Knowledge distillation b) Low-precision arithmetic, like converting a float to an int.
c) Weight pruning d) Pruning of model activations

4. In SSL, what is the typical approach to create labels for pretext tasks? [1]

a) Manually label the data b) Use existing labels from another dataset
c) Automatically generate labels from the data d) Ignore labels and train without them

5. Which of the following can be chosen as a pretext image task for SSL? [1]

a) Rotation b) Jigsaw puzzles
c) Noise removal d) Reconstruction

6. How can self-supervised pre-training benefit downstream tasks like image classification? [1]

a) By making the model's predictions more uncertain b) By adding noise to the training data
c) By providing a better initialization point for the model d) By reducing the model's capacity

7. Consider the softmax prediction scores produced by a neural network for a given input at temperature (T = 1). [2]

   0.01, 0.03, 0.25, 0.01, 0.4, 0.1, 0.05, 0.15

   Find the prediction scores when the temperature is raised to T = 2.

8. How does Adam combine the advantages of RMSProp and momentum? [2]

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

March 22, 2024

CSL7590 - Deep Learning

Instructions:
1. Read the questions carefully.
2. All questions are mandatory.
3. If a question requires justification, zero mark will be awarded in absence of the justification.
4. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 1 hour | *Minor-II* | Maximum Marks: 15

1. [True/False] Co-occurrence matrix based word embedding is an example of distributed representation. (Justify your answer) [1]

2. What is the advantage of layer normalization over batch normalization? [1]

3. Recall the rotation based pretext task for self-supervised pretraining on image data. Suggest a way to realize a similar rotation based pretext task for audio data. [1]

4. Consider a convolutional layer with 64 convolution filters of size 5×3 applied on an input with 32 channels. What will be the amount of reduction (%) in number of trainable parameters if the flattened convolutions are used? [1]

5. Show that for a rectangular weight matrix $W \in R^{m \times n}$ such that $m >> n$, a compression rate of $O\left(\frac{n}{r}\right)$ can be achieved using singular value decomposition where r denotes the number of non-zero singular values preserved during low rank approximation of W. [2]

6. Consider an exhibition where different products are presented with never-seen-before shapes and designs. You have got a collection of several photographs of these products where few of them are labeled. You are asked to design a classifier using the collection. You decide to use a domain adaptation approach with the help of hand drawn sketches of the products seen in the labeled photographs. Among the different domain adaptation approaches we discussed in the class, which one is the best choice for the considered problem and why? [3]

7. Show that for the single task transfer learning, minimizing the loss on target data is equivalent of minimizing the same loss on source data with a weight factor which is calculated as the ratio of probability values of source samples under the target and source distribution. [3]

8. Knowledge distillation generally employs softmax with a temperature parameter (τ). Let us assume that there is a set of 20 possible discrete values of τ and we are interested in finding out the optimal one. Can we find it using gradient descent? Justify your answer. [3]

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

April 30, 2024

CSL7590 - Deep Learning

Instructions:
Read the questions carefully. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 45 min | *Quiz-3* | Maximum Marks: 10

1. Which of the following is TRUE for Autoencoders? [1]

a) Can be used for Dimensionality Reduction
b) Can Reconstruct masked image patches
c) Autoencoders can learn from labels rather than data
d) Can be used for Image Compression

2. Which of the following is/are FALSE about Autoencoders? [1]

a) It is an unsupervised deep learning algorithm
b) It is like a data compression algorithm which performs dimensionality reduction
c) More the number of code layers, more is the data compression
d) In it, output is nearly same as that of the input

3. Which of the following is TRUE with respect to a VAE? [1]

a) VAE learns an intractable posterior distribution in the presence of continuous latent variable.
b) Standard Stochastic Gradient Descent (SGD) cannot be used to optimize the variational lower-bound of a VAE, due to the presence of continuous latent variable z.
c) The prior for a latent variable z, is taken as a centered isotropic multivariate Gaussian, in a standard VAE.
d) In a standard VAE, the latent variable z, is sampled from a Gaussian with diagonal covariance.

4. Which of the following is incorrect regarding comparative study of GAN and VAE models? [1]

a) VAEs learn a given data distribution by comparing it's input to the output i.e. the reconstructed version.
b) GANS use a network to distinguishing the real data from the generated by returning a number between 0 and 1, where 0 meaning the data is fake and 1 meaning it is real.
c) Given data X, it's easy to find the corresponding latent z for GANs, but not for VAEs.
d) GANS are generally better than VAEs for generating sharp images.

5. In the context of VAE, consider that encoder outputs μ= [0.2, 0.3, 0.1], σ = [0.1, 0.2, 0.1] and sampled $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$ is equal to [0.5, 0.2, 0.8] then find out the latent vector. [2]

6. Which problem of GAN is solved by the non-saturating loss. Why does the problem occur and how does the non-saturating loss solve it? [4]

$[0.2, 0.3, 0.1] + [0.1, 0.2, 0.1] \odot [0.5, 0.2, 0.8]$

$[0.3, 0.5, 0.2] \odot [0.5, 0.2, 0.8]$

$[0.2, 0.3, 0.1] + [0.05, 0.04, 0.08]$

$[0.25, 0.34, 0.18]$

0.15
0.06
0.24
0.45

80

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

May 07, 2024

CSL7590 - Deep Learning

Instructions:
1. Read the questions carefully.
2. All questions are mandatory.
3. If a question requires justification, zero mark will be awarded in absence of the justification.
4. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 2 hour | *Major* | Maximum Marks: 30

1. Which among the following is computationally the most inefficient method for model compression? [1]

a) Knowledge Distillation
b) Weight Pruning
c) Neural Architecture Search
d) Quantization

2. Considering the traditional data-free knowledge distillation, which of the following is/are not true? [1]

a) Teacher is released with weights and metadata
b) Original data is available at the time of distillation
c) Synthetic data is available at the time of distillation
d) Teacher is trained on synthetic data

3. [True/False] Differential pooling in GNNs makes the graph coarser by assigning each node at the input to exactly a single cluster corresponding to a node at the output. [1]

4. [True/False] An autoregressive model predicts the next component in a sequence by taking measurements from previous components in the sequence. [1]

5. Recall our discussion on adding a virtual node for graph classification using GCNN. In general, what would be the degree of this node? [1]

6. Dynamic Network Surgery uses two thresholds to control pruning, a_k and $b_k = a_k + t$ where t is a pre-defined margin. Discuss the impact of large and small values of a_k and t. [2]

7. Explain the significance of cycle consistency loss in CycleGAN. [2]

8. What is the non-convergence problem of standard GANs? How can this problem be solved? [3]

9. Let us consider an autoencoder where $\boldsymbol{z}$, which represents the output of the encoder, is quantized into a corresponding vector ($\hat{\boldsymbol{z}}$) of 0's and 1's as [3]

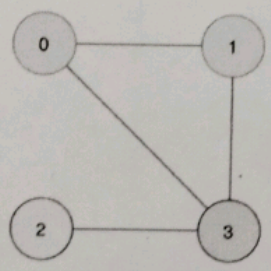
$$\hat{\boldsymbol{z}}(i) = \begin{cases} 1 & \text{if } \boldsymbol{z}(i) > T \\ 0 & \text{if } \boldsymbol{z}(i) \leq T \end{cases}$$

where $\boldsymbol{z}(i)$ is the i^{th} element of the vector $\boldsymbol{z}$ and T is a user-defined threshold. $\hat{\boldsymbol{z}}$ is fed into the decoder for reconstruction. Do you observe any challenge in training of this autoencoder? Suggest a way to overcome the challenge.

10. Show that, at the convergence of a standard GAN, the discriminator accuracy would be equal to 50%. [5]

11. In context of VAEs, explain the requirement and implementation of the reparameterization trick. [5]

12. Consider a graph with four nodes shown below. Consider a GCN layer which transforms the current node embeddings H as $\tilde{H} = \text{ReLU}(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}HW)$ where $\hat{A} = A + I$ with A representing adjacency matrix of the given graph and I representing the identity matrix. $\hat{D}$ is the degree matrix corresponding to $\hat{A}$. ReLU is the rectified linear unit activation. Find $\tilde{H}$ for the given W [5]



$$H = \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 2 & 0 \\ -0.5 & 1 \\ 5 & 10 \end{bmatrix} \qquad W = \begin{bmatrix} 0.1 & -1.5 & 0 \\ 2.5 & 0.2 & -3 \end{bmatrix}$$

[Note: Self-loop or self-edge is counted twice while calculating degree matrix.]