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]

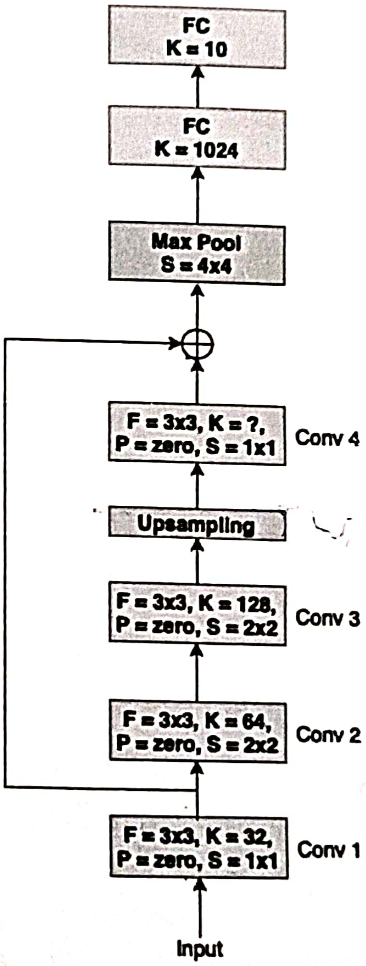
Consider a CNN shown on the right where Conv and FC represent convolutional and fully connected layers, respectively. F, K, and S represent filter size, number of filters or neurons, and stride, respectively. Each Conv layer uses zero padding (P). Answer the following for two input sizes (W×H×C), i)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

3. [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.



128X128X3

f = filter size

k = No. of filters

S = stride.

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

April 30, 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-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 co-variance.

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.

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

May 07, 2024

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]

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

February 9, 2024

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. $a(z) = \min(1, \frac{z+|z|}{2})$
   4. $a(z) = \ln(1 + e^z)$

$q = wx \quad [N \times D][D \times D] = [N$

$k \quad wx \quad [N \times 1D]$

$v \quad wx \quad [N \times D]$

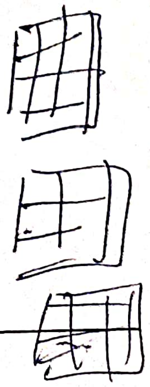
4. Explain the role of positional embeddings in transformers. [2]

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 updated once. [2]

7. Why is it critical to reduce learning rate in SGD? [2]

w^0



2 3 −4 −5   stochastic

1 −2 3 4   w   wx

~~7 −5 2 −3~~   ~~wx~~

1 2 2 −3

$$\frac{wx + b}{N \times D \times (D \times D)} = N \times D$$