Introduction to Deep Learning for Computer Vision

Adhyayan '23 - ACA Summer School
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur

Lecture 10

Quiz 1 Solutions

Question 1

- Which of the following statements is true for a Multi-Layered Perceptron (MLP)?
  a. The output of all the nodes of a layer is input to all the nodes of the next layer.
  b. The output of all the nodes of a layer is input to all the nodes of the same layer.
  c. The output of all the nodes of a layer is input to all the nodes of the previous layer.
  d. The Output of all the nodes of a layer is input to all the nodes of the output layer.
- Ans: a.

Question 2

- Which of the following statements is true?
  a. In batch gradient descent, we update the weights and biases of the neural network after forward passing over each training example.
  b. In batch gradient descent, we update our neural network weights and biases after forwarding all the training examples.
  c. Each step of stochastic gradient descent takes more time than batch gradient descent.
  d. None of these three options is correct.
- Ans: b

Question 3

- While training a batch neural network, you notice that the loss does not decrease after running the first few epochs. The reasons for this could be
  - 1. The learning rate is low.
  - 2. The neural net is stuck in local minima or a plateau.
  - 3. The neural net has too many units in the hidden layer.
- Ans: 1 or 2

Question 4

- Which of the following statements is true about Gradient Descent using Backpropagation?
  a. It always converges to the global minimum.
  b. Convergence does not depend on the initialisation of the weights.
  c. It may converge to local minima.
  d. The speed of convergence is directly proportional to the number of hidden layers.
- Ans: c

Question 5

- Pooling layers accomplish which of the following?
  a. To progressively reduce the spatial size of the representation.
  b. To reduce the number of parameters and computations in the network.
  c. To select maximum value over pooling region always.
  d. The pooling layer operates on each feature map independently.
- Ans: a, b, d.

Question 6

- Which of the following statements about parameter sharing in CNNs are true?
  a. It reduces the total number of parameters, thus reducing overfitting.
  b. It allows a feature detector to be used in multiple locations throughout the whole input image/input volume.
  c. It allows parameters learned for one task to be shared even for a different task (transfer learning).
  d. It allows gradient descent to set many parameters to zero, thus making the connections sparse.
- Ans: a,b

Question 7

- Which of the following statement is true?
  - 1. In CNN, having max-pooling can decrease the number of trainable parameters.
  - 2. In CNN, having max-pooling can keep the number of trainable parameters the same.
  - 3. In CNN, having max-pooling can increase the number of trainable parameters.
- Ans: 1 or 2

Question 8

- Consider the following statements regarding a Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN):
  - 1. A CNN has sparse connections between inputs and outputs between two consecutive layers.
  - 2. CNNs can be used only for image data
  - 3. Parameters are shared between output neurons in a CNN layer.
  - 4. Both CNNs and MLPs can take image data as input

  Which of the above statements are TRUE?

- Ans: 1, 3 and 4

Question 9

- Given a convolutional layer with an input size of 32x32, a filter size of 5x5, and a stride of 1, how many unique convolutional operations are performed on the input image?
  - 625
  - 900
  - 784
  - 225
- Ans: 784

Question 9: Calculation

- To calculate the number of unique convolutional operations performed on the input image, we need to consider the spatial dimensions of the input image, the filter size, and the stride
  - **Given**:
  - Input size: 32x32
  - Filter size: 5x5
  - Stride: 1
- To calculate the number of unique convolutional operations, we need to determine how many times the filter can be applied to the input image while staying within its boundaries.
- For each row and column, we can perform the convolution operation by sliding the filter across the input image with a stride of 1. Since the filter size is 5x5, we can perform the convolution operation within a 28x28 window (32 - 5 + 1 = 28).
- Therefore, the number of unique convolutional operations can be calculated as:
- Number of operations = (28 * 28) = 784
- So, there are 784 unique convolutional operations performed on the input image.

Question 10

- In a CNN, a pooling layer with a pooling window size of 2x2 and a stride of 2 is applied to a convolutional layer output with dimensions 16x16. What will be the output size after applying the pooling operation?
  a. 8x8
  b. 4x4
  c. 6x6
  d. 12x12
- Ans: a. Formula: floot((h - k)/s) + 1 [h: image size, k: pooling size, s: stride]

Question 11

- In a CNN, a fully connected layer receives input from a preceding convolutional layer with dimensions 8x8x64. If the fully connected layer has 512 neurons, how many parameters (weights and biases) are there in this layer?
  - 32768
  - 65536
  - 2097664
  - 524288
- Ans: None of them are correct. Correct Answer is: 2097664.
- 8*8*64 = 4096 inputs + 1 bias. And next layer has 512 nodes. So, total parameters are 4097*512 = 2097664.

Question 12

- In a CNN, a 3x3 filter is applied to a grayscale image with pixel values ranging from 0 to 255. If the filter weights are set as [1, 0, -1; 2, 0, -2; 1, 0, -1], and the centre pixel of the filter overlaps the pixel with a value of 128, what will be the output value after the convolution operation?
  - 130
  - 128
  - 132
  - 136
- Ans: Cannot say. Insufficient information.

Question 13

- Which weight initialization technique is commonly used for training deep convolutional neural networks (CNNs) and involves initializing the weights with small random values drawn from a normal distribution with zero mean and a standard deviation proportional to the square root of the number of inputs to the neuron?
    a. He initialization
    b. Glorot initialization
    c. LeCun initialization
    d. Orthogonal initialization
- Ans: b

Question 14

- Weight decay is a regularization technique used to prevent overfitting in neural networks. Which of the following statements about weight decay is correct?
  a. Weight decay increases the learning rate during training to reduce model complexity.
  b. Weight decay adds an additional term to the loss function that penalizes large weights.
  c. Weight decay randomly drops weights during training to improve model generalization.
  d. Weight decay decreases the number of neurons in the neural network to improve model performance.
- Ans: b

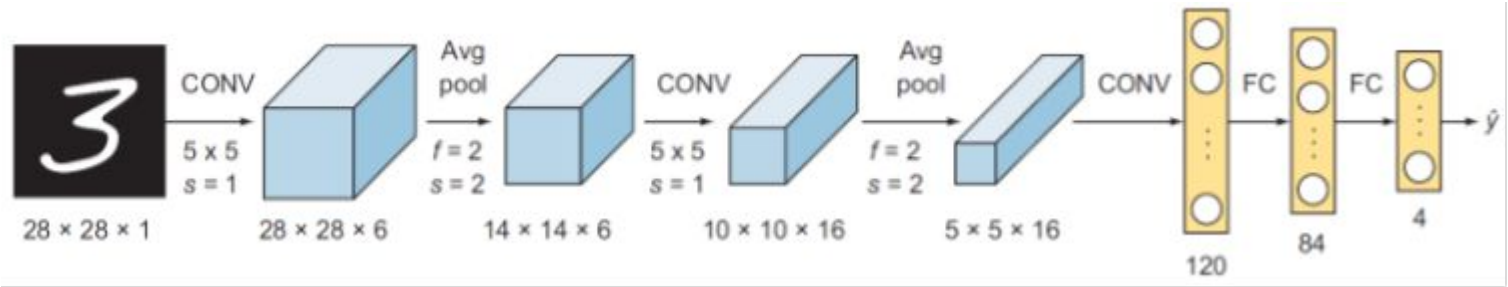
Question 15

- Dropout is a regularization technique commonly used in neural networks. Which of the following statements about dropout is correct?
  a. Dropout assigns random weights to neurons during training to reduce model complexity.
  b. Dropout randomly drops input features during training to improve model generalization.
  c. Dropout adjusts the learning rate dynamically based on the model's performance.
  d. Dropout combines multiple models with different architectures to improve model performance.
- Ans: b

Quiz 2 Solutions

Question 1

- Consider the CNN architecture shown in the figure. You will be using this for learning a 4-way classification problem. Which activation must be applied following the last FC layer?



- ❖ ReLU
- ❖ Hyperbolic Tangent
- ❖ Softmax (Ans.)
- ❖ Leaky ReLU

Question 2

- AlexNet network architecture has 60 million parameters. It is not straightforward to learn so many parameters without considerable overfitting. What are the tricks AlexNet used to combat overfitting?

- ❖ Data Augmentation (Ans.)
- ❖ Data Normalization
- ❖ Dropout (Ans.)
- ❖ Early Stopping

Question 3

- Which of the following statements are **FALSE**?

❖ Transfer learning migrates the knowledge learned from the source dataset to the target dataset, to save training time and computational cost.
❖ Early layers in the network learn low-level features like lines, blobs, and edges.
❖ Transfer learning transfers knowledge from one network to another and requires the model architecture of both the networks to be the same. (Ans.)
❖ Deep neural networks are immensely data-efficient and require very few labelled data to achieve high performance. (Ans.)

Question 4

- Which of the following statements are **TRUE**?

❖ In a typical transfer learning scenario, data for the source task is more abundantly available than the target task. (Ans.)
❖ Features from earlier layers of a trained CNN are more domain-specific than the deeper layers.
❖ In a scenario where the target dataset is small and similar to the source dataset, the best approach for transfer learning is to fine-tune the entire network.
❖ While fine-tuning a pre-trained network with a new classifier, we should use a smaller learning rate for the weights that are being fine-tuned in comparison to the randomly initialized weights for the new classifier. (Ans.)

Question 5

- Suppose you want to replace a 11x11 convolutional layer (stride=1) in a CNN with a stack of 3x3 convolutional layers (stride=1) but keep the effective receptive field same as original. How many 3x3 conv layers would you use? You can ignore the bias parameters in your calculation.

- 7
- 4
- 5 (Ans.)
- 8

Question 5: Calculation

- Let f= input shape (considering same height and width), k= kernel size, s= stride and p= padding.
- So, output shape formula = floor((f-k+2p)/s)+1
- Let's assume image size is 22x22. So, f=22. According to the question, k=11, s=1, p=0 (not given). So, output shape is floor((22-11+0)/1)+1 = 12x12
- for k=3, output shape is floor((22-3+0)/1)+1 = 20 after 1 3x3 filter layer.
- Similarly, output = 18 after another 3x3 filter layer.
- ...
- ...
- Similarly after 5 3x3 filter layers, output becomes 12x2. Hence, 5 3x3 conv layers are needed.

Question 6

- Suppose you want to replace a 11x11 convolutional layer (stride=1) in a CNN with a stack of 3x3 convolutional layers (stride=1) but keep the effective receptive field same as original. Assuming that the number of channels is same in each layer, what would be the ratio of number of parameters between the two configurations (#parameters(original)/#parameters(new))? You can ignore the bias parameters in your calculation.

- 2.689 (Ans.)
- 4.367
- 3.156
- 1.119

Question 6: Calculation

- Suppose you want to replace a 11x11 convolutional layer (stride=1) in a CNN with a stack of 3x3 convolutional layers (stride=1) but keep the effective receptive field same as original. Assuming that the number of channels is same in each layer, what would be the ratio of number of parameters between the two configurations (#parameters(original)/#parameters(new))? You can ignore the bias parameters in your calculation.

- Number of weights in 11x11 filter is 11*11=121 (#parameters(original))
- Number of weights in 5 3x3 filters = 5*3*3 = 45 #parameters(new)
- Ratio = 121/45 = 2.689

Question 7

- Which one of the following statements are true for tanh function T(x) and sigmoid function S(x)?

- S(x) = 1 - S(-x)
- T(x) = 2S(2x) - 1
- Derivative of T(x), T'(x) = 1 - (T(x))^2
- All of the above (Ans.)

Question 8

- Which one of the following activation functions is not a continuous function?

- Heaviside step function (Ans.)
- ReLU
- Sigmoid function
- Leaky ReLU

Question 9

- Given a model with prediction y' = S(sum(w_i*x_i) + b) and loss E = 0.5*(y' - y)^2. Here, S(.) is the sigmoid activation function, y is the target value, and x_i and w_i are the i-th input and i-th weight respectively. The gradient of loss E w.r.t. weight w_i is:

- (y'- y)y'(1-y')w_i
- (y' - y)y(1-y')x_i
- (y' - y)y'(1-y')x_i (Ans.)
- (y' - y)(1-y')x_i

Question 9: Derivation

$$\hat{y} = S(\textstyle\sum_i w_i x_i + b)$$

$$E = \tfrac{1}{2}(\hat{y} - y)^2$$

We know: $S'(x) = S(x)(1 - S(x))$

$$\begin{aligned}
\frac{\partial E(W)}{\partial w_i} &= (\hat{y} - y)\frac{\partial}{\partial w_i}(\hat{y} - y) \\
&= (\hat{y} - y)\frac{\partial}{\partial w_i}(\hat{y}) \\
&= (\hat{y} - y)\frac{\partial}{\partial w_i}(S(\textstyle\sum_i w_i x_i + b)) \\
&= (\hat{y} - y)S(\textstyle\sum_i w_i x_i + b)(1 - S(\textstyle\sum_i w_i x_i + b))\frac{\partial}{\partial w_i}(\textstyle\sum_i w_i x_i + b) \\
&= (\hat{y} - y)\hat{y}(1 - \hat{y})\frac{\partial}{\partial w_i}\textstyle\sum_i w_i x_i \\
&= (\hat{y} - y)\hat{y}(1 - \hat{y})\frac{\partial}{\partial w_i} w_i x_i \\
&= (\hat{y} - y)\hat{y}(1 - \hat{y})x_i
\end{aligned}$$

Question 10

- Consider the image transformation G(x, y) = 255 * (F(x, y)/255)^a (a > 0). Which of the following statements are true for this transformation?

- ❖ G(x, y) is always brighter than F(x, y)
- ❖ G(x, y) is always darker than F(x, y)
- ❖ G(x, y) is brighter than F(x, y) if a > 1
- ❖ G(x, y) is darker than F(x, y) if a > 1 (Ans.)

Question 11

- Which of the following object detection methods use a separate non-convolutional method for generating region proposals?

- R-CNN (Ans.)
- Fast R-CNN (Ans.)
- Faster R-CNN
- SSD

Question 12

- Complete the following statement: In ____ R-CNN, selective search is replaced by a convolutional network called _____.

- ❖ Fast, Spatial Pyramid Pooling
- ❖ Faster, Spatial Pyramid Pooling
- ❖ Faster, Region Proposal Network (RPN) (Ans.)
- ❖ Fast, Region Proposal Network (RPN)

Question 13

- Consider a regression problem where each feature-map is an m-dimensional vector. Following this, if we want to train a feed-forward neural network to solve the problem, then how many nodes will be in the output layer of the neural network?

❖ m nodes
❖ One node (Ans.)
❖ Require more information
❖ None of the above

Question 14

- For a regression problem, what is the error function that should be used for optimization? (a) Cross-entropy, (b) Mean Square Error, (c) Mean Absolute Error, (d) b or c.

- (a)
- (b)
- (c)
- (d) (Ans.)

Question 15

- Consider the following the feed-forward neural network design for a classification problem:

  model = Sequential(Linear(in_features=1280, out_features=512, bias=True), Linear(in_features=512, out_features=512, bias=True), Linear(in_features=512, out_features=10, bias=False))

  What kind of non-linear function has it learned?

- Sigmoid
- ReLU
- Softmax
- None of the above (Ans.)

Thank You!