Nptel Online Certification Course
Indian Institute of Technology Kharagpur
Computer Vision
Assignment - Week 12

Number of questions: 10 **Total marks: 10x2=20**

QUESTION 1: **Type: Numeric**

The softmax predictions of four different classes are [0.2 0.3 0.4 0.1]. Calculate the cross entropy loss for this sample (use logarithm with base 2) if the true labels are [0 1 0 0]. Round off the answer to 2 places of decimals.

Correct Answer: 1.74

Detailed Solution:

$L = -\sum y.log_2(p)$

where, y is the true label and p is the predicted probability. In this case we calculate it as

$L = -[1.log_2(0.3) = 1.74$

QUESTION 2: **Type: Numeric**

Up-convolve the following input x using the filter h. What is the value at (2,2) in the up-convolved feature map? We assume index of first cell of any 2D-matrix is (0,0).

x

1	2
3	1

h

0	1	0
2	2	1
0	1	0

Correct Answer: 7

Detailed Solution:

The up-convolution is done as

x1h1	x1h2+x2h1	x1h3+x2h2	x2h3
x1h4+x3h1	x1h5+x2h4+x3h2+x4h1	x1h6+x2h5+x3h3+x4h2	x2h6+x4h3
x1h7+x3h4	x1h8+x2h7+x3h5+x4h4	x1h9+x2h8+x3h6+x4h5	x2h9+x4h6
x3h7	x3h8+x4h7	x3h9+x4h8	x4h9

=

0	1+0	0+2	0+0
2+0	2+4+3+0	1+4+0+1	2+0
0+6	1+0+6+2	0+2+3+2	1
0	3+0	0+1	0+0

This will give the following result:

0	1	2	0
2	9	6	2
6	9	7	1
0	3	1	0

QUESTION 3: **Type: MSQ**

Which of the following statements are true for Pooling?

a) Creates a pool of data in order to improve the accuracy of the algorithm predicting images.

b) It assists in the detection of features, even if they are distorted.

c) Decreases the attribute size, in order to decrease the computational power.

d) It extracts recessive features.

Correct Answer: b), c)

Detailed Solution:

Pooling filter responses help gain robustness to the exact spatial location of features. As a result, even if the picture is a little tilted, the largest number in a certain region of the feature map would be recorded and hence, the feature would be preserved. It reduces the number of computations and the number of parameters in a pooling layer is 0. It progressively reduces the spatial size of the feature maps. It is useful in extracting dominant features.

QUESTION 4: **Type: MSQ**

Which of the following statements are true for Batch normalisation?

a) It returns the normalised mean and standard deviation of the weights.

b) It normalizes (changes) all the input before sending it to the next layer.

c) It is a very efficient backpropagation technique.

d) It allows using higher learning rate.

Correct Answer: b), d)

Detailed Solution:

Batch normalization is a layer that allows every layer of the network to do learning more independently. It is used to normalize the output of the previous layers. The activation scale the input layer in normalization. Using batch normalization learning becomes efficient also it can be used as regularization to avoid overfitting of the model and avoid vanishing gradient.

FOR QUESTIONS 5, 6 AND 7:

Consider an input image of dimension $8 \times 8 \times 5$. It passes through a convolutional layer with 16 kernels of dimension 3×3. The image has been zero padded on both sides with 1×1 padding and convolved using 1×1 stride. Based on the given data solve the following questions 5, 6 and 7:

QUESTION 5: **Type:Comprehensive**

What is the dimension of the activation map?

a) $16 \times 16 \times 5$

b) $8 \times 8 \times 5$

c) $16 \times 16 \times 16$

d) $8 \times 8 \times 16$

Correct Answer: d)

Detailed Solution:

The formula is given by:

$width\_of\_activation\_map = \frac{width\_of\_input\_layer - width\_of\_kernel + 2 \times padding}{stride} + 1$

$height\_of\_activation\_map = \frac{height\_of\_input\_layer - height\_of\_kernel + 2 \times padding}{stride} + 1$

$output\_channel = number\_of\_kernels$

Consider the dimension of the activation map to be $W \times H \times D$. Then each of the following is given by:

$W = \frac{8-3+2\times1}{1} + 1$

$H = \frac{8-3+2\times1}{1} + 1$

$D = 16$

QUESTION 6: **Type:Comprehensive**

Calculate the number of parameters in the convolutional layer.

Correct Answer: 736

Detailed Solution:

Number of parameters in convolution layer = (input channel $\times$ kernel height $\times$ kernel width $\times$ no. of kernels) + no. of kernels(bias). Thus, it will be $5 \times 3 \times 3 \times 16 + 16 = 736$

QUESTION 7: **Type:Comprehensive**

Assume that the convolutional layer is followed by a MaxPool layer with kernel size 2×2 and stride 2×2. Calculate the dimension of the feature map.

a) $4 \times 4 \times 5$

b) $2 \times 2 \times 5$

c) $4 \times 4 \times 16$

d) $2 \times 2 \times 16$

Correct Answer: c

Detailed Solution:

Consider the dimension of the activation map to be $W \times H \times D$. Then each of the following is given by:

$W = \frac{8-2}{2} + 1$

$H = \frac{8-2}{2} + 1$

$D = 16$

QUESTION 8: **Type: Numeric**

You have been given an encoder block consisting of a convolutional layer followed by a pooling layer. The input to the block is $16 \times 16 \times 8$. 16 kernels of size 5×5 are used to generate feature maps. The pooling layer uses kernel size of 2×2 with stride 2×2. Calculate the total number of parameters required in the encoder block.

Correct Answer: 3216

Detailed Solution:

Number of parameters in convolution layer = (input channel $\times$ kernel height $\times$ kernel width $\times$ no. of kernels) + no. of kernels(bias). Thus, it will be $8 \times 5 \times 5 \times 16 + 16 = 3216$

Number of parameters in pooling layer = 0

Parameters in encoder block = 3216 + 0 = 3216

QUESTION 9: **Type: MCQ**

You have been given four tasks: object detection, image segmentation, image classification and image captioning. Choose the most probable deep neural architectures you will use to complete these tasks.

a) object detection - RCNN.
image segmentation - Vgg-16.
image classification - U-net.
image captioning - RNN.

b) object detection - RCNN.
image segmentation - U-net.
image classification - Vgg-16.
image captioning - RNN.

c) object detection - RNN.
image segmentation - Vgg-16.
image classification - U-net.
image captioning - RCNN.

d) object detection - U-net.
image segmentation - Vgg-16.
image classification - RCNN.
image captioning - RNN.

Correct Answer: b)

Detailed Solution:

Vgg-16 is a supervised learning architecture used for image classification. U-net is mostly used for image segmentation problem. It uses down-sampling and up-sampling method for segmentation. RCNN is used for object detection where the feature extraction is done using a convolutional network on the region of proposals. RNN allows sequential input/output. Thus, it can be used for image captioning.

QUESTION 10: **Type:Numeric**

In an object detection problem, the time taken to generate all proposals is 0.75. There are 4 number of proposals generated. What is the inference time of a Faster RCNN if the convolution time and the time taken by the fully connected layer are 2.25 and 1.75, respectively? Round off the answer to 2 places of decimals.

Correct Answer: 9.25

Detailed Solution: Inference time of Faster RCNN = $1\times$ ConvTime + NumProp $\times$ fcTime = $2.25 + 4 \times 1.75 = 9.25$,

where,

NumProp = number of proposals generated

ConvTime = convolution time

fcTime = time taken by the fully connected layer.
