Deep Learning (CS 590)
Test 1 (July - November 2020)
27th October, 2020

Total 20 questions, each having 1 points. Attempt all question. Full Marks: 20, Full Time: 30 minutes

* This form will record your name, please fill your name.

1

What steps can we take to prevent overfitting in a Neural Network?
(1 Point)

- [ ] Data Augmentation
- [ ] Weight Sharing
- [ ] Early Stopping
- [ ] Dropout

2

(1 Point)

Given a convolution layer ($\mathbf{L}$) with weights and biases as $\mathbf{W^L}$, $\mathbf{b^L}$, let input to the $\mathbf{L}$ be $\mathbf{x}$ while inference. The pre-activations of layer $\mathbf{L}$ can be defined as

$$f^L = B.N^L (\mathbf{W^L} * \mathbf{x} + \mathbf{b^L}),$$

where $B.N^L$ denotes the Batch Normalization. Assume μ^L, σ^L are moving averages of batch statistics during training, and γ^L, β^L are B.N's scale and shift parameters. Let $L+1^{st}$ layer consists of a convolution layer with weights and biases as $\mathbf{W^{L+1}}$, $\mathbf{b^{L+1}}$, followed by $B.N^{L+1}$ with parameters μ^{L+1}, σ^{L+1}, γ^{L+1}, β^{L+1}, followed by the ReLU6 ($\mathbf{R^{L+1}}$) activation function. Denoting the activations of layer $L+1^{st}$ with f^{L+1}, What is the distribution of f^{L+1} ? Also mention the range of the activations.

- [ ] Clipped Normal Distribution, (0,6)
- [ ] Clipped Uniform Distribution, (0,6)
- [ ] Clipped Normal Distribution, (6, inf)
- [ ] Clipped Uniform Distribution, (6, inf)

3

On training a predictive model using some learning algorithm, it is apprehended that the model is underfitting. Which of the following could be the nature of predictive errors ?
(1 Point)

- [ ] Low bias and High variance
- [ ] Low bias and low variance
- [ ] High bias and low variance
- [ ] High bias and high variance

4

Why ResNet works well for deeper architecture?
(1 Point)

- [ ] Deeper network without skip connections may suffer from vanishing gradient problems
- [ ] The residual part of ResNet receives the input as an amplifier to its output
- [ ] Both (a) and (b)
- [ ] None of the above

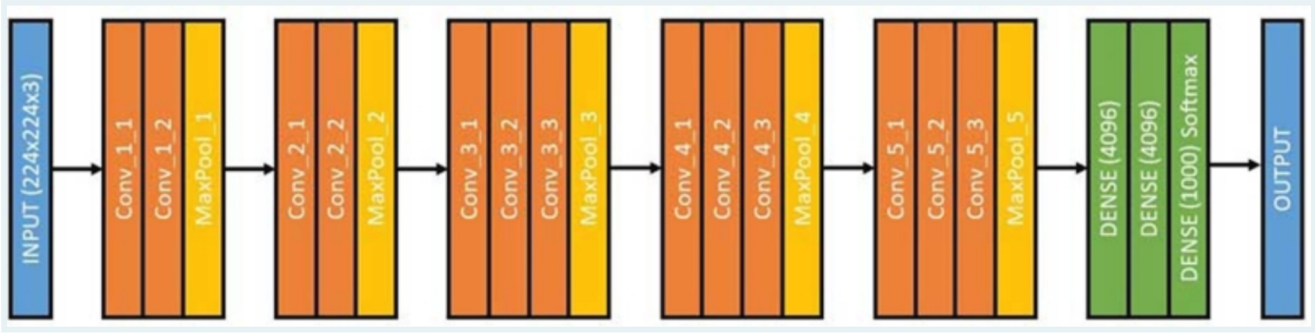
5

Pooling layers are used in CNNs because
(1 Point)

- [ ] It avoids underfitting
- [ ] Spatial downsampling of the incoming feature map makes backpropagation faster
- [ ] Number of learnable parameters are reduced
- [ ] It provides translation invariance to the internal representation of the feature maps

6

In the diagram below, a CNN is performing image classification. Which of the following statements about the inclusion or exclusion of a batch normalization layer (BNL) in the network holds true?
(1 Point)



- [ ] Inclusion of a BNL after an activation layer makes sense and total number of iterations to converge decreases.
- [ ] Inclusion of a BNL in a pre-trained setting of the network below, and then testing on a given image sample makes sense because it will help prevent internal covariance shift.
- [ ] Exclusion of a BNL generally allows us to explore a broader range of learning rates.
- [ ] Exclusion of a BNL will give us a significant reduction in training time.

7

Batch Normalization is helpful because
(1 Point)

- [ ] It normalizes (changes) all the input before sending it to the next layer
- [ ] It returns back the normalized mean and standard deviation of weights
- [ ] It helps gradient descent to reach local optima more steadily
- [ ] None of these

8

Assertion (A): An implementation of VGG-100 or VGG-256 would be disastrous.
Reason (R): During backpropagation, first the gradient of the loss is calculated with respect to the weights and then it is propagated backwards along the network. Hence, for so many layers, it would run into the vanishing gradient problem.
(1 Point)

- [ ] Both A and R are true and R is the correct explanation of A
- [ ] Both A and R are true but R is not the correct explanation of A
- [ ] Both A and R are false
- [ ] A is false but R is true
- [ ] A is true but R is false

9

What is the use of regularization parameter while performing a regularized linear regression?
(1 Point)

- ◯ It reduces the bias in the model and hence reduces overfitting
- ◯ Until some point, increasing it reduces the variance of the model significantly without significant addition of bias to the model
- ◯ Controls the trade-off between the need for the model to fit the training set well and also have a large number of model parameters
- ◯ All of the above

10

Which of the following gives non-linearity to a neural network?
(1 Point)

- [ ] Stochastic Gradient Descent
- [ ] Rectified Linear Unit
- [ ] Convolution operator
- [ ] None of the above

11

A stack of two small 3x3 convolutional filters, for example, has an effective receptive field of 5x5 filter. Then why does VGG-Net use a stack of small 3x3 filters instead of a single, high-dimensional kxk filter?
(1 Point)

- [ ] To reduce the number of parameters.
- [ ] To make the decision function more discriminative
- [ ] Neither A nor B
- [ ] Both A and B

12

Which of the following statements are True in general, in the context of training a deep CNN?
(1 Point)

- [ ] In the gradient ascent training, weights update follows the gradient downhill.
- [ ] Actual gradients may have fairly large values, scale it using a learning rate.
- [ ] No need to change the learning rate for first 100 epochs.
- [ ] In dropout, for each batch, different set of nodes are selected and their weights are not updated.
- [ ] Batch Normalization is a simple non-linear differentiable transformation during inference.

13

The main disadvantage of the RNN is
(1 Point)

- [ ] Sufferers from vanishing gradient problem
- [ ] Not capable to model long term dependencies well
- [ ] No. of parameters are very high
- [ ] Even a variant cannot model both way dependencies.

14

(1 Point)

Given a convolution layer ($\mathbf{L}$) with weights and biases as $\mathbf{W}^L$, $\mathbf{b}^L$, let input to the $\mathbf{L}$ be $\mathbf{x}$ while inference. The pre-activations of layer $\mathbf{L}$ can be defined as

$$f^L = \text{B.N}^L(\mathbf{W}^L * \mathbf{x} + \mathbf{b}^L),$$

where B.N^L denotes the Batch Normalization. Assume μ^L, σ^L are moving averages of batch statistics during training, and γ^L, β^L are B.N's scale and shift parameters. If the B.N^L is folded into the weights and biases of $\mathbf{L}$, what will be the new values of $\mathbf{W}^L$, $\mathbf{b}^L$?

a) $\mathbf{W}^{L,N} = (\beta^L.\mathbf{W}^L) / \mu^L$, $\mathbf{b}^{L,N} = \gamma^L - (\beta^L.\ \sigma^L) / \mu^L$
b) $\mathbf{W}^{L,N} = (\sigma^L.\mathbf{W}^L) / \gamma^L$, $\mathbf{b}^{L,N} = \mu^L - (\beta^L.\ \sigma^L) / \gamma^L$
c) $\mathbf{W}^{L,N} = (\mu^L.\mathbf{W}^L) / \beta^L$, $\mathbf{b}^{L,N} = \sigma^L - (\gamma^L.\ \mu^L) / \beta^L$
d) $\mathbf{W}^{L,N} = (\gamma^L.\mathbf{W}^L) / \sigma^L$, $\mathbf{b}^{L,N} = \beta^L - (\gamma^L.\ \mu^L) / \sigma^L$

- [ ] a
- [ ] b
- [ ] c
- [ ] d

15

In neural networks, nonlinear activation functions such as sigmoid, tanh, and ReLU
(1 Point)

- [ ] speed up the gradient calculation in backpropagation, as compared to linear units
- [ ] help to learn nonlinear decision boundaries
- [ ] are applied only to the output units
- [ ] always output values between 0 and 1

16

How the computation cost of the inception module is reduced
(1 Point)

- [ ] Use 1x1 convolution
- [ ] Using dropout
- [ ] Use Bottleneck layer
- [ ] Using Batch normalization
- [ ] Use regularization

17

Question
(1 Point)

Given a convolution layer ($\mathbf{L}$) with weights and biases as $\mathbf{W}^{L}$, $\mathbf{b}^{L}$, let input to the $\mathbf{L}$ be $\mathbf{x}$. The pre-activations of layer $\mathbf{L}$ can be defined as

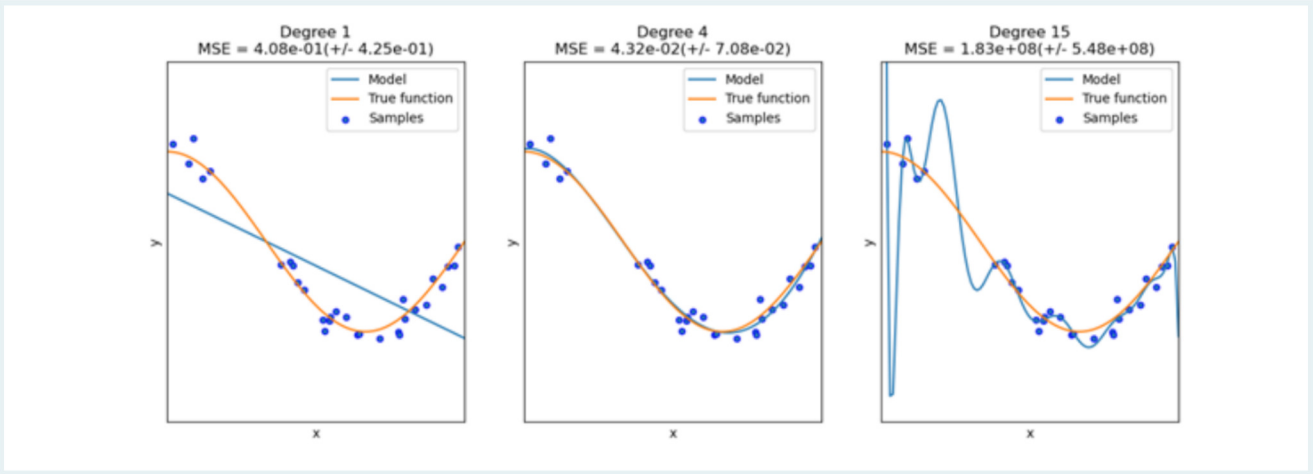
$$f^{L} = \mathrm{B.N}^{L}(\mathbf{W}^{L} * \mathbf{x} + \mathbf{b}^{L}),$$

where $\mathrm{B.N}^{L}$ denotes the Batch Normalization. Assume μ^{L}, σ^{L} are moving averages of batch statistics during training, and γ^{L}, β^{L} are B.N's scale and shift parameters. Let $L+1^{st}$ layer consists of a convolution layer with weights and biases as $\mathbf{W}^{L+1}$, $\mathbf{b}^{L+1}$, followed by $\mathrm{B.N}^{L+1}$ with parameters μ^{L+1}, σ^{L+1}, γ^{L+1}, β^{L+1}, followed by activation function ($\mathbf{R}^{L+1}$). Denoting the activations of layer $L+1^{st}$ with f^{L+1},, if $\mathbf{R}^{L+1}$ is hyperbolic tangent activation function, what can you say about the distribution of f^{L+1} in general?

- [ ] Still a Uniform Distribution
- [ ] Still a Gaussian Distribution
- [ ] Union of Gaussian and Uniform Distribution
- [ ] Not a Gaussian Distribution

18

Analyze the following figure. Assuming Co(P), Ca(M) denote the complexity of the problem and capacity of the model, respectively. Which of the following statements are True?
(1 Point)



- [ ] Ca(M) denotes the types of activations functions used in the model M.
- [ ] Ca(M) denotes the size of the model M in Megabytes (MBs).
- [ ] Ca(M) denotes the no. of trainable parameters in the model M.
- [ ] Ca(M) denotes the ratio of no. of convolution layers to the no. of fully-connected layers used in the model M.

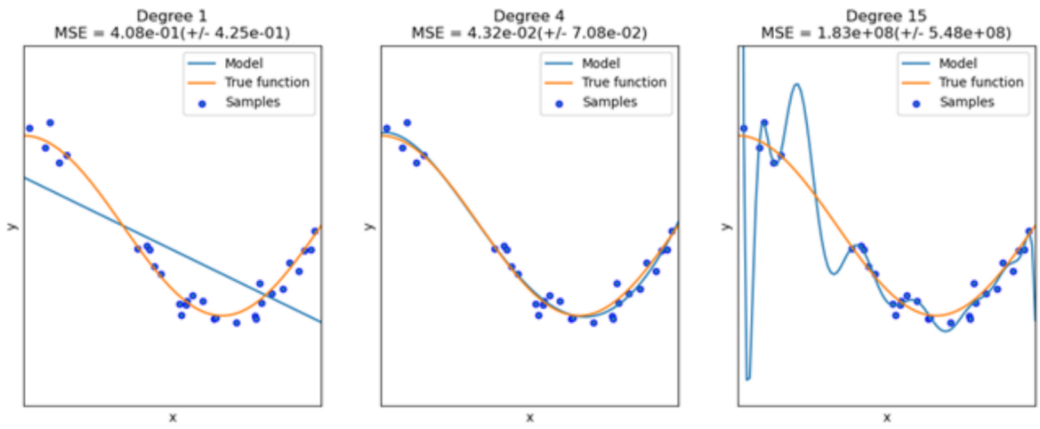
19

Let input volume: 32x32x3, Receptive field of convolution kernel: 5x5, Stride:3, number of filters: 5. Calculate output volume.
(1 Point)

- [ ] 32x32x5
- [ ] 32x32x3
- [ ] 10x10x5
- [ ] 5x5x10

20

Analyze the following figure. Assuming Co(P), Ca(M) denote the complexity of the problem and capacity of the model, respectively. Identify the closest behaviors of left, center and rightmost sub-figures
(1 Point)



- [ ] Perfect fit, Memorized, Underfitting
- [ ] Underfitting, Memorized, Perfect fit
- [ ] Underfitting, Perfect fit, Memorized
- [ ] Ca(M) > Co(P), Ca(M) ~ Co(P), Ca(M) < Co(P)
- [ ] Ca(M) < Co(P), Ca(M) ~ Co(P), Ca(M) > Co(P)
- [ ] Ca(M) > Co(P), Ca(M) ~ Co(P), β.Ca(M) < Co(P), where β denoted the no. of samples in training set.