Quiz 3 - Results

Attempt 1 of 1

Written Mar 10, 2024 7:31 PM - Mar 10, 2024 8:16 PM

Released Mar 17, 2024 1:00 PM

Attempt Score **23.35 / 25 - A-**

Overall Grade (Highest Attempt) **23.35 / 25 - A-**

Question 1 — 1.5 / 1.5 points

Which of the following statements on Inception Networks are true? (Check all that apply.)

- ✓ A single inception block allows the network to use a combination of 1x1, 3x3, 5x5 convolutions and pooling.
- ✓ Inception blocks usually use 1x1 convolutions to reduce the input data volume's size before applying 3x3 and 5x5 convolutions.
- ✓ Inception networks incorporate a variety of network architectures (similar to dropout, which randomly chooses a network architecture on each step) and thus has a similar regularizing effect as dropout.

Question 2 — 1.5 / 1.5 points

Mark all that is true about activation functions.

- ✓ *Leaky ReLU* can be used to mitigate the dying *ReLU* problem.
- ✓ Derivative of *tanh(z)* is maximum at z=0 and its value is 1.
- ✓ The derivative of *softmax* with z as input is maximum at z=0.5, while its value is maximum at z=0.25.
- ✓ *tanh*, *sigmoid*, and *ReLU* all squash the input to a value in the positive quadrant.

Question 3 **1 / 1 point**

Suppose your input is a 200 by 200 color (RGB) image, and you are **not** using a convolutional network. If the first hidden layer has 200 neurons, each one fully connected to the input, how many parameters does this hidden layer have (including the bias parameters)?

- ○ 8,000,200
- ✓ ○ 24,000,200
- ○ 24,000,000
- ○ 24,002,000
- ○ 8,000,000

Question 4 **0.75 / 0.75 points**

In Stochastic Gradient Descent, loss is calculated over all the training points at each weight update.

- ○ True
- ✓ ○ False

Question 5 **1.5 / 1.5 points**

One benefit of using convolutional networks is "parameter sharing". Which of the following statements about parameter sharing in ConvNets are true? (Check all that apply.)

- ✓ ☐ It allows parameters learned for one task to be shared even for a different task (transfer learning).
- ✓ ☐ It allows a feature detector to be used in multiple locations throughout the whole input image/input volume.
- ✓ ☐ It allows gradient descent to set many of the parameters to zero, thus making the connections sparse.
- ✓ ☐ It reduces the total number of parameters, thus reducing overfitting.

Question 6 **1 / 1 point**

Convolutional layers in deep neural networks exhibit "sparsity of connections". What does this mean?

- ○ Each layer in a convolutional network is connected only to two other layers.
- ○ Each filter is connected to every channel in the previous layer.

- ✓ Each activation in the next layer depends on only a small number of activations from the previous layer.
- Regularization causes gradient descent to set many of the parameters to zero.

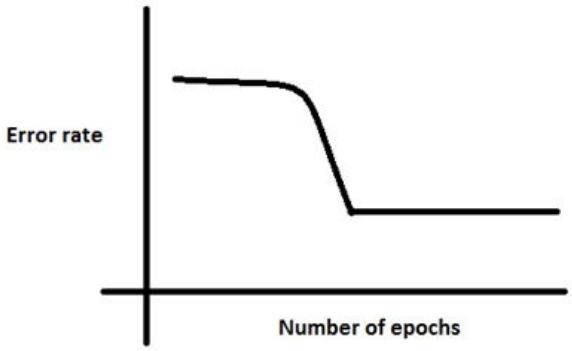
Question 7 **1 / 1 point**

Which of the following methods DOES NOT prevent a model from overfitting to the training set?

- Dropout
- Data augmentation
- Early stopping
- ✓ Pooling

Question 8 **1 / 1 point**

In training a neural network, you notice that the loss does not decrease in the few starting epochs.



What might be the reason for this?

1. The learning is rate is low
2. Regularization parameter is high
3. Stuck at local minima

- 1 and 3
- 2 and 3
- ✓ Any of them

◯ 1 and 2

Question 9 **0.75 / 0.75 points**

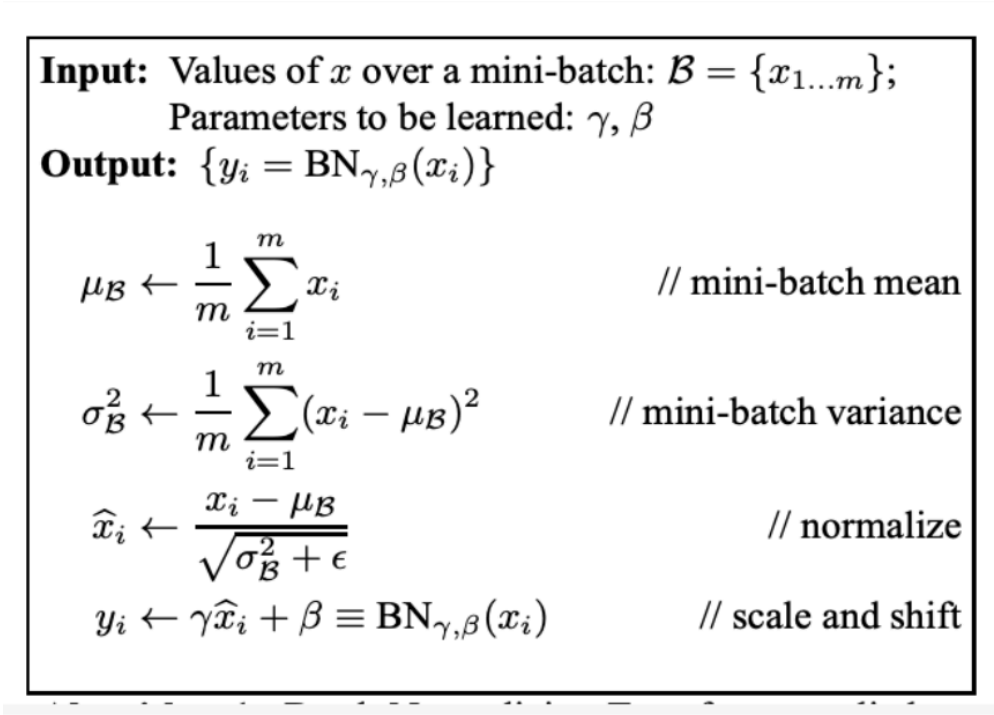
In order to normalize our data (i.e. subtract mean and divide by standard deviation), we typically compute the mean and standard deviation across the entire dataset before splitting the data into train/val/test splits.

◯ True

✔ ◯ False

Question 10 **1.5 / 1.5 points**

Consider the Batch Normalizing Transform, applied to activation x over a mini-batch:



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Select all options that apply.

✔ ☐ Batch norm can reduce internal covariance shift.

✔ ☐ Batch norm can enable faster convergence as it allows working with larger learning rates.

✔ ☐ Batch norm can be applied with any activation function.

✔ ☐ Batch norm can be effective with any batch size, including 1.

✔ ☐ The trainable parameters are γ (scale) and β (shift) .

✔ ☐ It may be possible that learned values of γ and β give back the original activation value (before applying batch norm transformation), i.e., y_i will be

same as x_i.

Question 11 **0.75 / 0.75 points**

Because pooling layers do not have parameters, they do not affect the backpropagation (derivatives) calculation.

- ○ True
- ✓ ○ False

Question 12 **0.75 / 1.5 points**

Which ones of the following statements on Residual Networks are true? (Check all that apply.)

- ✓ ☐ A ResNet with L layers would have on the order of L^2 skip connections in total.
- ➡ ✓ ☐ Using a skip-connection helps the gradient to back propagate and thus helps you to train deeper networks.
- ➡ ✓ ☐ The skip-connection makes it easy for the network to learn an identity mapping between the input and the output within the ResNet block.
- ✗ ☐ The skip-connections compute a complex non-linear function of the input to pass to a deeper layer in the network.

Question 13 **1.05 / 1.05 points**

Which of these is NOT a good learning rate decay scheme? Here, *t* is the epoch number.

- ✓ ☐ $\alpha = e^t \alpha_0$
- ✓ ☐ $\alpha = \frac{1}{1+2t}\alpha_0$
- ✓ ☐ $\alpha = 0.95^t \alpha_0$

✔ ☐

$$\alpha = \frac{1}{\sqrt{t}}\alpha_0$$

Question 14 **1.8 / 2.7 points**

Write your answers in the form: H*W*C , where H is height,W is width and C is the number of channels. For example, 5*5*3)

You have an input volume that is 63*63*16, find the output shape when you:

Convolve it with 32 filters that are each 7x7, using a stride of 2 and no padding.
___29*29*32___ ✔(33.33 %)

Convolve it with 32 filters that are each 7x7, using stride of 1 and padding of 3
___63*63*32___ ✔(33.33 %)

Apply max pooling with a stride of 2 and a filter size of 3 and no padding
___31*31*32___ ✖ **(31*31*16)**

Question 15 **1 / 1 point**

Which of the following weight initializers work better with ReLU?

- ◯ Random Uniform
- ◯ Xavier
- ✔ ◯ He
- ◯ Glorot

Question 16 **1.5 / 1.5 points**

Select all that is true about weight/bias initialization

- ✔ ☐ Typically all biases are initialized to random values.
- ✔ ☐ Typically all biases are initialized to the same value.
- ✔ ☐ Typically all weights are initialized to random values.
- ✔ ☐ Typically all weights are initialized to the same value.

Question 17 **1.5 / 1.5 points**

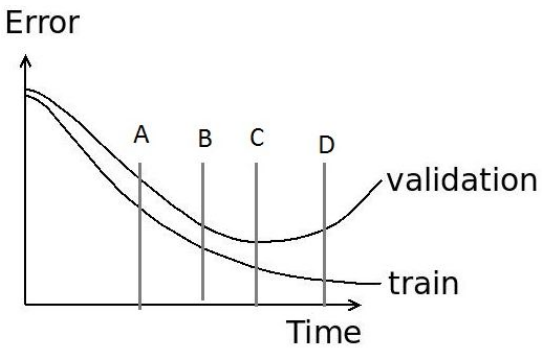
Suppose you have an input volume of dimension nH x nW x nC, where nH is height, nW is width, and nC is the number of channels. Which of the following statements do you agree with?

(Assume that "1x1 convolutional layer" below always uses a stride of 1 and no padding.)

- [x] You can use a 1x1 convolutional layer to reduce nC but not nH, nW.
- [x] You can use a pooling layer to reduce nH, nW, and nC.
- [x] You can use a 1x1 convolutional layer to reduce nH, nW, and nC.
- [x] You can use a pooling layer to reduce nH, nW, but not nC.

Question 18 **1 / 1 point**

While training a neural network for an image recognition task, we plot the graph of training error and validation error.



What is the best place in the graph for early stopping?

- [ ] D
- [ ] B
- [ ] A
- [x] C

Question 19 **1.5 / 1.5 points**

Which of the following do you typically see in a ConvNet?

(Check all that apply.)

- [x] FC layers in the first few layers
- [x] CONV layer followed by a POOL layer
- [x] Multiple POOL layers followed by a CONV layer

- [x] FC layers in the last few layers

Question 20 **1 / 1 point**

Suppose your input is a 200 by 200 color (RGB) image, and you use a convolutional layer with 100 filters that are each 3x3. How many parameters does this hidden layer have (including the bias parameters)?

- [ ] 2700
- [x] 2800
- [ ] 901
- [ ] 1000

Done