

L2.2 Intro to Neural Networks Quiz ANS

1. Why can a single perceptron not compute the XOR function?

- A) XOR needs more than two inputs, but a perceptron has only two
- B) The two classes are not linearly separable, so no single line $w \cdot x + b = 0$ separates them
- C) The step activation function is not differentiable at 0, so training fails
- D) A perceptron can only output the values 0.5 and 1

ANS: B) The two classes are not linearly separable, so no single line $w \cdot x + b = 0$ separates them

2. What would happen to a multi-layer perceptron if all non-linear activation functions were removed?

- A) The network would become non-differentiable and could no longer be trained
- B) The network would still be able to represent any non-linear function
- C) The number of hidden neurons would double automatically
- D) The stack of affine layers would collapse into a single linear transformation of the input

ANS: D) The stack of affine layers would collapse into a single linear transformation of the input

3. What does the Universal Approximation Theorem state, as presented in the lecture?

- A) A one-hidden-layer MLP with enough neurons can approximate any function we want; deeper networks are more effective in practice
- B) Every MLP with a ReLU activation converges to the global minimum of the loss
- C) Any dataset can be classified with 100 % accuracy by a perceptron
- D) A network needs at least ten hidden layers to approximate a non-linear function

ANS: A) A one-hidden-layer MLP with enough neurons can approximate any function we want; deeper networks are more effective in practice

4. What is the value of the sigmoid derivative $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ at $z = 0$?

- A) 0
- B) 0.5
- C) 0.25, which is also its maximum
- D) 1

ANS: C) 0.25, which is also its maximum

5. Why do deep networks with many sigmoid layers suffer from vanishing gradients?

- A) The sigmoid output is always negative, so the weights cannot be updated
- B) The sigmoid derivative is at most 0.25, so the product of many such factors during back-propagation shrinks toward zero
- C) The sigmoid function is not defined for large inputs
- D) The learning rate is automatically reduced to zero after a few layers

ANS: B) The sigmoid derivative is at most 0.25, so the product of many such factors during back-propagation shrinks toward zero

6. What is the 'dying ReLU' problem?

- A) A neuron whose input is always negative outputs 0 with zero gradient, so it stops learning
- B) The ReLU output grows without bound and the loss becomes infinite
- C) The ReLU derivative is exactly 0.25 for all positive inputs
- D) ReLU can only be used in the output layer of a network

ANS: A) A neuron whose input is always negative outputs 0 with zero gradient, so it stops learning

7. According to Clevert et al. (2016), as cited in the lecture, which activation function converged fastest and gave a lower test error than ReLU on CIFAR-10?

- A) Sigmoid
- B) Binary step
- C) tanh
- D) ELU

ANS: D) ELU

8. What do L1 and L2 weight regularization encourage, respectively?

- A) L1 keeps the weights small; L2 makes the weights sparse
- B) Both set all weights exactly to the same value
- C) L1 makes the weights sparse; L2 keeps the weights small in magnitude
- D) L1 increases the learning rate; L2 decreases it

ANS: C) L1 makes the weights sparse; L2 keeps the weights small in magnitude

9. A SoftMax layer receives four equal logits $z = (2, 2, 2, 2)$. What does it output?

- A) (0.25, 0.25, 0.25, 0.25)
- B) (1, 0, 0, 0)
- C) (0.5, 0.5, 0, 0)
- D) (2, 2, 2, 2)

ANS: A) (0.25, 0.25, 0.25, 0.25)

10. The true class has one-hot label y , and the network assigns probability 0.5 to the true class. What is the cross-entropy loss?

- A) 0.5
- B) $\ln 2 \approx 0.69$
- C) 0
- D) $-\ln 2 \approx -0.69$

ANS: B) $\ln 2 \approx 0.69$

11. For a SoftMax output p and a cross-entropy loss with one-hot label y , what is the gradient of the loss with respect to the logits z ?

- A) $y - p^2$
- B) $p \cdot (1 - p)$
- C) $\ln p - \ln y$
- D) $p - y$

ANS: D) $p - y$

12. A weight $w = 2$ has gradient $dL/dw = 4$. What is w after one gradient-descent step with learning rate $\eta = 0.1$?

- A) 2.4
- B) 0.4
- C) 1.6
- D) 3.6

ANS: C) 1.6

13. What typically happens if the learning rate of gradient descent is set far too large?

- A) The updates overshoot the minimum, so the loss oscillates or diverges
- B) The network converges faster to the global minimum without any risk
- C) The gradient becomes exactly zero after the first step
- D) The training loss always decreases but the validation loss stays constant

ANS: A) The updates overshoot the minimum, so the loss oscillates or diverges

14. Why is back-propagation efficient compared to computing every derivative separately?

- A) It avoids computing any derivatives by estimating them randomly
- B) It trains each layer independently of the others at the same time
- C) It applies the chain rule from output to input and reuses the local gradients, so all gradients cost about one backward pass
- D) It needs only the weights of the last layer to update the whole network

ANS: C) It applies the chain rule from output to input and reuses the local gradients, so all gradients cost about one backward pass

15. In $f = (x + y) \cdot z$ with $x = -2$, $y = 5$ and $z = -4$, what is $\partial f / \partial x$?

- A) 3
- B) -4

C) -12

D) 5

ANS: B) -4

16. In a computational graph, what does a multiply gate do during the backward pass?

A) It copies the upstream gradient unchanged to both inputs

B) It passes the gradient only to the larger of the two inputs

C) It sets both input gradients to zero

D) It gives each input the other input's value multiplied by the upstream gradient

ANS: D) It gives each input the other input's value multiplied by the upstream gradient

17. How does the lecture describe NVIDIA PilotNet (2016)?

A) A reinforcement-learning agent that learns from rewards in a simulator

B) A graph neural network that works on the road network of an HD map

C) A CNN that directly predicts steering commands from camera images as a regression task

D) A rule-based planner that follows hand-written lane-keeping rules

ANS: C) A CNN that directly predicts steering commands from camera images as a regression task

18. How many parameters (weights and biases) does an MLP with 784 inputs, one hidden layer of 128 neurons and 10 outputs have?

A) 101,770

B) 100,352

C) 100,480

D) 102,400

ANS: A) 101,770

19. Why are CNNs preferred over fully connected MLPs for camera images?

A) CNNs do not need any training data because the kernels are fixed

B) Convolution layers share a small kernel across the image, so there are far fewer parameters and some translation invariance, and the input size is not fixed

C) CNNs use no non-linear activation functions, which makes them faster

D) MLPs cannot be trained with back-propagation

ANS: B) Convolution layers share a small kernel across the image, so there are far fewer parameters and some translation invariance, and the input size is not fixed

20. During training, the training loss keeps decreasing but the validation loss starts to rise. What does this indicate, and what is a sensible reaction?

A) Underfitting; train the same model with a larger learning rate

B) A bug in the loss function; the validation loss must never rise

C) Perfect generalization; no further action is needed

D) Overfitting; use early stopping, dropout, weight regularization or more training data

ANS: D) Overfitting; use early stopping, dropout, weight regularization or more training data