1. What is the core drawback of using a plain MLP directly on images?

- A) It cannot use nonlinear activations
- B) It loses spatial structure and explodes parameter count after flattening
- C) It does not support backpropagation
- D) It cannot handle RGB images
  Answer: B

2. Why do CNNs use local receptive fields and weight sharing?

- A) To drastically increase parameters
- B) To capture spatial locality while reducing parameters
- C) To eliminate the need for pooling
- D) To remove nonlinear activations
  Answer: B

3. The 2D convolution output size follows $W_{out} = \frac{W-K+2P}{S} + 1$. What do K, P, and S denote?

- A) K kernel size, P channels, S number of filters
- B) K kernel size, P padding, S stride
- C) K number of filters, P padding, S channels
- D) K channels, P padding, S image depth
  Answer: B

4. For a Conv2D with kernel 3×3, input channels 3, and 64 filters, how many trainable parameters?

- A) 576
- B) 1,728
- C) 1,792
- D) 2,048
  Answer: C (3×3×3×64 + 64)

5. With “same” padding and stride 1, how do spatial dimensions typically change after Conv2D?

- A) They always decrease
- B) They remain equal to the input
- C) They double
- D) They are unrelated to padding
  Answer: B

6. What is the principal role of pooling in CNNs?

- A) Increase spatial dimensions
- B) Downsample while retaining salient information
- C) Replace convolution entirely
- D) Remove the need for activations
  Answer: B

7. Which innovations were central to AlexNet's success?

- A) Tanh only
- B) ReLU, GPU acceleration, dropout, and data augmentation
- C) No pooling
- D) Only 1×1 kernels
  Answer: B

8. What is the key idea behind ResNet that eases training very deep networks?

- A) Only fully connected layers
- B) 1×1 convolutions without shortcuts
- C) Residual skip connections that bypass blocks
- D) Removing activations
  Answer: C

9. What pattern characterizes VGG16 blocks?

- A) 7×7 kernels with stride 4 only
- B) 3×3 kernels, 2×2 max pooling, gradually increasing depth
- C) 5×5 kernels with no pooling

- D) 1×1 kernels only
  Answer: B

10. In transfer learning, which layers are commonly frozen first?

- A) Task-specific deeper layers
- B) Output layer only
- C) Shallow layers capturing edges/textures
- D) No layers are frozen
  Answer: C

11. Which loss fits multi-class softmax classification?

- A) MSE
- B) MAE
- C) Categorical Cross-Entropy
- D) Hinge
  Answer: C

12. What's a key difference between MSE and MAE?

- A) MSE is less sensitive to outliers than MAE
- B) MAE is less sensitive to outliers than MSE
- C) They are identical
- D) Both are non-differentiable at zero
  Answer: B

13. Binary Cross-Entropy is best described as:

- A) A regression loss
- B) A loss for binary classification measuring dissimilarity between predicted probabilities and labels
- C) A loss that does not use probabilities
- D) A loss that cannot be backpropagated
  Answer: B

14. Compared to cross-entropy, what does hinge loss emphasize?

- A) Only maximizing correct-class probability
- B) Enforcing a margin (max-margin) like SVM
- C) Handling continuous outputs
- D) Reducing sensitivity to outliers
  Answer: B

15. What is the goal of Triplet Loss?

- A) Bring all samples equally close
- B) Pull anchor toward positive and push away from negative by a margin
- C) Binary classification
- D) L2 weight decay
  Answer: B

16. What benefit does Momentum provide in optimization?

- A) Forces learning rate to decay too fast
- B) Accumulates a velocity to move faster along shallow valleys
- C) Stops updates at zero
- D) Deletes large gradients
  Answer: B

17. What common drawback of AdaGrad is addressed by RMSProp?

- A) Exploding learning rate
- B) Learning rate decays continually and becomes too small
- C) Inability to handle periodic data
- D) Requirement for very high momentum
  Answer: B

18. Adam combines which ideas (with bias correction)?

- A) Momentum and RMSProp
- B) SGD and Dropout
- C) AdaGrad and Cross-Entropy

- D) BatchNorm and pooling
  Answer: A

19. Which learning-rate schedule often benefits deep nets?

- A) Constant
- B) Fixed step decay only
- C) Cosine annealing or cyclical LR
- D) Linear increase
  Answer: C

20. Which regularization technique reduces overfitting by randomly deactivating units at train time?

- A) L1 only
- B) Batch Normalization only
- C) Dropout
- D) Data Augmentation only
  Answer: C