GNR 638 QUIZ 1 FEB 7, 2025

Answer all questions. 17 x 1.5 = 25.5 (negative marks of 1 for every two wrong answers.)

1. Which of the following is not a primary benefit of using Batch Normalization in deep neural networks?
   (a) It helps in stabilizing and accelerating training by reducing internal covariate shifts.
   (b) It acts as a form of regularization, reducing the need for dropout.
   (c) It completely eliminates the need for careful weight initialization.
   (d) It allows the use of higher learning rates, improving convergence speed.

2. Which of the following is a key characteristic of Stochastic Gradient Descent (SGD) compared to Batch Gradient Descent?
   (a) SGD updates model parameters using the entire dataset at each iteration.
   (b) SGD tends to converge to the global minimum more efficiently than Batch Gradient Descent in all cases.
   (c) SGD introduces noise in gradient updates, which can help escape sharp local minima and find flatter minima.
   (d) SGD always requires a momentum term to work effectively.

3. What is the primary goal of Xavier Initialization in deep neural networks?
   (a) To ensure weights are initialized to small constant values close to zero.
   (b) To prevent activation outputs from exploding or vanishing by maintaining a balance in variance.
   (c) To assign random values to weights without considering the number of neurons in the layers.
   (d) To make training faster by skipping the need for backpropagation.

4. Which of the following activation functions is LEAST likely to be used in the hidden layers of a modern deep MLP, and why?
   (a) ReLU
   (b) Sigmoid
   (c) Tanh
   (d) Leaky ReLU

5. Which of the following regularization techniques is most likely to lead to some weights in the MLP being exactly zero?
   (a) L1 regularization
   (b) L2 regularization
   (c) Dropout
   (d) Early stopping

6. You are experimenting with different weight initialization strategies for your deep learning model. You notice that the training loss decreases very rapidly in the first few epochs but then plateaus and doesn't improve much further. What might this suggest about your initialization?
   (a) The initialization is likely too small
   (b) The initialization is likely too large
   (c) The initialization is probably fine
   (d) The initialization is perfect

7. VGGNet's primary contribution to the field of CNNs was:
   (a) Introducing the concept of residual connections
   (b) Demonstrating the effectiveness of very small (3x3) convolutional filters.
   (c) Proposing a novel pooling strategy
   (d) Introducing batch normalization

8. If the learning rate is too high, what is most likely to happen during backpropagation?
   (a) The model will converge faster to an optimal solution
   (b) The loss function may oscillate or diverge instead of converging
   (c) The gradients will become zero, leading to no learning
   (d) The model will memorize the training data perfectly

9. The input to a one-layer CNN model is a $6 \times 6 \times 3$ image, where 3 refers to the channels of the image. There are 32 learnable filters of the shape 2×2. With no zero padding and with stride $= 1$, what is the output shape and the number of parameters in the CNN model?
   (a) $6 \times 6 \times 32$, 128 parameters
   (b) $6 \times 6 \times 32$, 416 parameters
   (c) $5 \times 5 \times 32$, 416 parameters
   (d) $5 \times 5 \times 32$, 160 parameters

10. A two-layer feedforward Neural Network has been designed for the task of regression using the MSE loss function. The weights are $W_1 = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, $W_2 = \begin{bmatrix} 1 & 1 \end{bmatrix}$. Assume there is no bias term involved; what is the loss value for an input [1,1,1] and ground truth 50?
    (a) 6
    (b) 36
    (c) 44
    (d) 0

11. A deep neural network with L layers use the sigmoid activation function in all hidden layers. The weights are initialized from a normal distribution $W_l \sim \mathcal{N}(0, \sigma^2)$. During the training process, the gradients are observed to become extremely small, leading to slow learning. Which of the following is the most likely reason?
    (a) The variance (σ^2) of the weight initialization is too large, causing activations to explode.
    (b) The variance (σ^2) of the weight initialization is too small, causing activations to shrink exponentially.
    (c) This type of observation is impossible in the case of deep networks.
    (d) Using sigmoid activation inherently prevents gradient vanishing due to its smoother nature, so the activation function is not the issue.

12. Consider a deep MLP trained using backpropagation with Sigmoid activation. Which of the following best explains why training slows down in deeper layers?
    (a) Sigmoid activations saturate at extreme values, leading to near-zero gradients.
    (b) Weight initialization is responsible for reducing gradient magnitude.
    (c) The number of neurons increases, leading to more computational complexity.
    (d) The loss function does not affect backpropagation.

Cont.

13. Why does training a CNN with very large convolutional kernels (e.g., 11×11 or 9×9) often lead to poor generalization?

    (a) Large kernels increase the receptive field, which always improves accuracy.
    (b) Large kernels lead to excessive parameterization, causing overfitting.
    (c) Large kernels remove the need for deep networks.
    (d) Large kernels reduce the number of computations.

14. Consider a ResNet model where the identity skip connection is removed. What is the most likely effect?

    (a) The network will train significantly faster.
    (b) The network will suffer from performance degradation in deeper layers due to vanishing gradients.
    (c) The network will have a lower parameter count.
    (d) The network will generalize better.

15. Why do deep CNN architectures like ResNet perform better than shallow CNNs with similar parameter counts?

    (a) Deep networks can always memorize training data better.
    (b) Deep networks learn hierarchical representations efficiently.
    (c) Shallow networks always suffer from gradient explosion.
    (d) Shallow networks cannot use residual connections.

16. Consider a network where the first few layers are frozen, and only the last few layers are fine-tuned on a new dataset. What is the primary reason this works well in transfer learning?

    (a) The initial layers learn general features, while later layers specialize.
    (b) Freezing layers prevents overfitting entirely.
    (c) The last layers are computationally less expensive to train.
    (d) The optimization algorithm only updates shallow networks.

17. What is the primary reason why Dropout works effectively as a regularization technique in deep networks?

    (a) It reduces the number of parameters in the model.
    (b) It forces the network to learn redundant representations, improving generalization.
    (c) It replaces batch normalization.
    (d) It helps networks converge faster.

The End.