ADVANCED AI

Quiz 4

Friday 9 October 2026

ANSWER KEY

10 questions, one mark each. Circle one answer per question.

1. Several networks reach zero training loss on the same data, but behave differently between the training points. What is regularisation for in this situation?

(A) Helping the optimiser reach the same solution in fewer steps
(B) Removing the noise from the training targets before fitting
(C) Favouring the fits that are more likely to generalise to new data
(D) Making the fits generalise equally well, so the choice stops mattering
(E) Lowering the training loss further, so the fit generalises better

2. Compared with a dense layer, which two assumptions about images does a convolutional layer build in?

(A) Each pixel can be classified on its own, without looking at its neighbours.
(B) Objects look the same when rotated, and the same when rescaled.
(C) Distant pixels matter most, and patterns mean different things by position.
(D) Pixel order carries no information, and each position needs its own detector.
(E) Nearby pixels are related, and a useful pattern can appear at any position.

3. The strength λ of an L2 penalty is increased step by step from zero. What typically happens?

(A) Training error stays the same, because the penalty ignores the data.
(B) The weights grow larger, so the fitted function can bend more sharply.
(C) Training and test error both keep falling, however large λ becomes.
(D) Training error rises and the fit gets smoother, until the model underfits.
(E) Only the biases shrink, so the overall shape of the fit is unchanged.

4. An object moves across an image. Which pairing of task and desired behaviour is correct?

(A) Both want invariance: the output ignores the move.
(B) Classification wants invariance, and segmentation wants equivariance.
(C) Classification wants equivariance, and segmentation wants invariance.
(D) Neither: the move should change both outputs.
(E) Both want equivariance: the output follows the move.

5. Dropout zeroes a random subset of hidden units at each training step, drawing a new subset every time. What does this encourage that training a fixed, smaller network would not?

(A) A different random subnetwork for every prediction in standard inference
(B) Units that each specialise on one fixed subset of the examples
(C) Units that stay useful whichever other units happen to be active
(D) Faster training, because fewer units are updated at each step
(E) Units that all learn the same feature, as copies of one another

6. In a stack of convolutional layers with 3×3 kernels, how does the receptive field of a unit change as you go deeper?

(A) It stays the size of one kernel, however deep the unit is.

(B) It shrinks, because each layer loses positions at the edges.

(C) It already covers the whole input from the first layer onwards.

(D) It grows only if the layers use pooling or a stride above one.

(E) It grows, so deeper units respond to larger regions of the input.

7. Training loss keeps falling, while validation loss falls and then rises. Which parameters does early stopping keep?

(A) Those from the checkpoint with the lowest validation loss

(B) Those from the checkpoint with the lowest training loss

(C) Those from where the training and validation losses are equal

(D) Those at the end of training, once training loss stops falling

(E) Those from the checkpoint with the lowest test loss

8. Max pooling and strided convolution both shrink a feature map's height and width. What do both give up?

(A) Mixing of information across channels

(B) Receptive field size in later layers

(C) Channels, in proportion to the shrinkage

(D) Precise location and fine spatial detail

(E) Weight sharing across positions

9. A model is trained to classify handwritten digits. Which data augmentation is most likely to hurt accuracy rather than help it?

(A) Moving the digit elsewhere in the frame

(B) Rotating images by 180 degrees

(C) Scaling images up or down by 20%

(D) Doubling or halving stroke thickness

(E) Rotating images by up to 10 degrees

10. On MNIST-1D, a CNN with about 2,000 parameters and a dense network with about 59,000 both fit the training set perfectly. The dense network could represent the same function as the CNN, yet the CNN's test error is far lower. Which interpretation of this result is most appropriate?

(A) The CNN's locality and weight sharing match how the data are structured.

(B) Convolution generalises better than a dense layer on any kind of data, not just this.

(C) The CNN has too few parameters to memorise the training set, so it can't overfit.

(D) The CNN's output ignores where a pattern occurs, since convolution is invariant.

(E) Weight sharing acts like an L2 penalty, keeping the CNN's weights small.