

PRACTICE QUESTIONS

IABAC

Certified Deep Learning Expert

SHARED BY [u/curious_panda_99](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM IABAC Certified Deep Learning Expert

Shared by [u/curious_panda_99](#) on aiexamcerts

10 practice questions for the IABAC Certified Deep Learning Expert exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/iabac/certified-deep-learning-expert. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Which activation function outputs zero for negative inputs and returns the input unchanged for positive inputs?
 - A Sigmoid, which squashes every input into the open interval from zero to one
 - B Tanh, which squashes every input into the open interval from minus one to one
 - C ReLU, a piecewise linear function that is cheap to compute and widely used in hidden layers
 - D Softmax, which turns a vector of scores into a probability distribution across classes

Answer: C

ReLU computes $\max(0, x)$, which gives zero for negative inputs and the identity for positive ones. Sigmoid is the tempting distractor, since it is the other common activation that is often shown as a simple curve, but it saturates at both ends rather than clipping at zero. Tanh and softmax have different output ranges and roles.

2. The activation function $f(x) = \max(0, x)$ is called the _____ linear unit.

Answer: rectified

The function is the rectified linear unit, abbreviated ReLU, which clips negative inputs to zero. Exponential linear unit is the tempting alternative, since ELU is a related activation with a smooth negative part. The blank is rectified.

3. In stochastic gradient descent, what does a mini-batch provide?
 - A An exact gradient of the loss over the entire dataset, computed in one pass through memory without sampling any examples at all
 - B A fixed set of weights that the optimizer is forbidden to update during that step, which keeps the training run deterministic
 - C A complete copy of the validation set, used to score each update before it is applied, so that bad steps are rejected
 - D A noisy but cheap estimate of the full-dataset gradient, computed from a small subset of the examples

Answer: D

A mini-batch gradient is an unbiased but noisy estimate of the gradient over the full training set, and its cost per step is much lower. The exact full-dataset gradient is the tempting distractor, since that is what batch gradient descent computes, but it is far more expensive per update. The other options describe things a mini-batch does not provide.

4. Gradient descent is guaranteed to reach the global minimum of a non-convex neural network loss.
- A True
 - B False

Answer: B (False)

Neural network losses are non-convex, so gradient methods typically find a local minimum or a saddle point, and many good solutions are used in practice. The guarantee of a global optimum belongs to convex problems. The statement is false.

5. Which two statements about the Adam optimizer are true? (Choose two.)
- A It uses one global learning rate that never adapts for each parameter separately
 - B Its bias correction matters most in the first training steps, while the averages warm up
 - C It always converges to a better minimum than SGD with momentum on every task
 - D It cannot be combined with a learning rate schedule during training
 - E It keeps running averages of the first and second moments of the gradients

Answer: B, E

Adam maintains exponential moving averages of gradients and squared gradients and corrects their initial bias, which makes early steps more reliable. The claim that it always beats SGD with momentum is the tempting distractor, but no optimizer is universally better. A global fixed rate contradicts the per-parameter scaling, and Adam works with schedules.

6. Weight decay as a regularizer is equivalent, for plain SGD, to adding which term to the loss?
- A An L1 penalty that drives most of the weights to exactly zero on every single update
 - B An L2 penalty on the weights, scaled by the decay coefficient and added to the training loss
 - C A term proportional to the number of training steps that have been completed so far
 - D A penalty on the variance of the activations computed in each hidden layer of the network

Answer: B

Adding an L2 penalty with coefficient λ to the loss shrinks each weight in proportion to its value, which matches the weight decay update for plain SGD. The L1 penalty is the tempting distractor, since it is also a common regularizer, but it produces sparsity rather than uniform shrinkage. The other options do not correspond to weight decay.

7. What is early stopping?

- A Stopping training after the first epoch so that the model does not overfit the initial batch of examples the optimizer happens to see
- B Ending training as soon as the training loss reaches exactly zero on every single example in the training set, without any validation check
- C Halting training when the validation metric stops improving, and keeping the weights from the best checkpoint
- D Halting the optimizer whenever the gradient norm exceeds a fixed threshold for a single step, which resets the training run each time

Answer: C

Early stopping monitors a validation metric, stops when it plateaus or worsens for a set patience, and restores the best checkpoint. Stopping after one epoch is the tempting distractor, but that ignores whether validation performance is still improving. Stopping at zero training loss and gradient-norm halts are not the method.

8. The normalization that computes statistics across the features of each example, rather than across the batch, is _____ normalization.

Answer: layer

Layer normalization normalizes each example over its feature dimension, which is why it is common in transformers. Batch normalization is the tempting distractor, since it uses batch statistics instead. The blank is layer.

9. What problem do the gates in an LSTM cell address that a plain recurrent network struggles with?

- A Processing high-resolution images, by sliding a fixed kernel across every frame in the sequence and pooling the results at each stage
- B Computing attention scores between every pair of tokens in parallel across the whole input, which removes the need for any recurrence
- C Reducing the parameter count to one per time step, so that the model fits on small devices with very little memory available
- D Learning long-range dependencies, by controlling what information is kept in the cell state across many time steps

Answer: D

The forget, input and output gates let an LSTM preserve gradient flow over long spans and forget information when it is no longer useful. Convolution over frames is the tempting distractor, since sequences of images can use it, but it does not address recurrence. Attention and parameter counts are different mechanisms.

10. What is the usual output activation and loss pairing for binary classification with a single logit?
- A Sigmoid output with binary cross-entropy loss, so the single logit becomes a probability that is scored against the label
 - B Linear output with no activation function, paired with mean squared error against the 0 or 1 label
 - C ReLU output, paired with hinge loss so that the output stays non-negative for every example
 - D Tanh output, paired with categorical cross-entropy computed over a one-hot target vector

Answer: A

A sigmoid maps the logit to a probability, and binary cross-entropy is the matching negative log-likelihood. A linear output with mean squared error is the tempting distractor, and it is sometimes used, but it is not the standard pairing. ReLU and tanh outputs and categorical loss do not fit a single-logit binary task.