

PRACTICE QUESTIONS

IABAC

Certified Deep Learning Expert

SHARED BY [u/noble_pelican_90](#) on aiexamcerts

EDITION	Free Set 2
---------	------------

QUESTIONS	10
-----------	----

SHARED	October 15, 2026
--------	------------------

EXAM	IABAC Certified Deep Learning Expert
------	--------------------------------------

Shared by [u/noble_pelican_90](#) 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. What is the vanishing gradient problem most closely associated with?
 - A Weight matrices initialized with values much too large for the layer width, which causes overflow in the forward pass of the network during training
 - B Training runs in which the batch size is so large that the loss stops decreasing during the epoch, even with a well tuned learning rate
 - C Models that store so many parameters that the optimizer cannot hold the gradients in memory at once, so each update has to be skipped
 - D Deep networks with saturating activations such as sigmoid, where gradients shrink layer by layer during backpropagation

Answer: D

When the derivative of each activation is small, the chain rule multiplies many small factors, so gradients reaching early layers become tiny. Oversized initial weights are the tempting distractor, and they more often cause exploding values rather than vanishing ones. Batch size and memory limits describe different problems.

2. Adding more hidden layers always reduces the training loss to zero on any dataset.
 - A True
 - B False

Answer: B (False)

Extra layers add capacity, but optimization may fail to converge, and some datasets have label noise that keeps the loss above zero. The idea that depth guarantees zero loss is a tempting error. Capacity increases the chance of fitting training data, not a guarantee.

3. What is the role of the learning rate in gradient descent?
 - A It scales the size of each parameter update relative to the negative gradient
 - B It sets the number of layers that are updated in each training iteration of the loop
 - C It determines how many examples are loaded into each mini-batch from the data loader
 - D It fixes the maximum value that any single activation in the network can reach

Answer: A

The update rule subtracts the learning rate times the gradient from each weight, so the rate controls the step size. Layer selection is the tempting distractor, but freezing layers is a separate decision from the step size. Batch loading and activation bounds are determined by other settings.

4. The optimizer that adds a decoupled weight decay term to Adam's update is named _____.

Answer: AdamW (also accepted: AdamW optimizer)

AdamW keeps Adam's adaptive updates and applies weight decay directly to the weights rather than through the gradient. Adagrad is a tempting alternative, since it is another adaptive optimizer, but it has no decoupled decay. The answer is AdamW.

5. Which two techniques reduce overfitting in deep networks? (Choose two.)

- A Adding weight decay to the training objective
- B Removing the validation set to make more data available for fitting
- C Evaluating on a larger test set inside the training loop
- D Applying dropout to hidden layers during training
- E Increasing the model width until the training loss reaches zero

Answer: A, D

Dropout and weight decay both constrain the model, which reduces its ability to memorize training noise. Growing the model until training loss hits zero is the tempting distractor, since it looks like a fit, but it usually increases overfitting. Removing validation and checking the test set during training both hurt generalization.

6. Batch normalization is commonly credited with stabilizing training because it does what?

- A Guarantees that every activation in the network is exactly zero-centered at inference time, which removes the need for running averages entirely
- B Replaces the validation split, since the batch statistics act as a built-in test of how well the model generalizes to unseen examples
- C Removes the bias terms from all layers, so that fewer parameters need to be learned by the optimizer during each update of the training run
- D Stabilizes the statistics of activations across layers, which permits larger learning rates and smoother optimization

Answer: D

Normalizing each layer's inputs with batch statistics reduces sensitivity to scale changes, which is why larger learning rates become workable. The zero-centered claim is the tempting distractor, since the layer does center activations during training, but at inference it uses running averages, so the output is not guaranteed to be zero-centered. The other options are incorrect.

7. What is the core operation of the transformer encoder that lets each token weigh every other token in the sequence?
- A Multi-head self-attention, which computes weighted sums over all positions using query, key and value projections
 - B Max pooling across the sequence, which keeps only the largest activation for each feature channel and discards all other positions entirely
 - C A one-dimensional convolution with stride two, which halves the sequence length at each layer and reduces the cost of attention
 - D Recurrent state passing, where each token receives only the hidden state of the previous token in the sequence, one step at a time

Answer: A

Self-attention forms attention weights from query-key similarity and mixes value vectors across all positions, and multiple heads do this in parallel subspaces. Max pooling is the tempting distractor, since it also combines positions, but it is not a learned all-pairs mixing step. The convolution and recurrence options are different architectures.

8. Transfer learning can reduce the labeled data needed for a new task when the source and target tasks are related.
- A True
 - B False

Answer: A (True)

Pretrained features give a head start, so fine-tuning on a smaller labeled set can reach good accuracy when the tasks share useful structure. Relatedness matters because unrelated features transfer poorly. The statement is true.

9. Focal loss was introduced mainly to address which problem?
- A Exploding gradients in very deep recurrent networks during backpropagation through time, which is usually fixed with gradient clipping and careful initialization
 - B Extreme class imbalance in dense detection, by down-weighting easy, well-classified examples so that training focuses on hard ones
 - C Overfitting caused by a training set that contains many duplicate images across the classes, which is usually fixed with more augmentation
 - D Slow inference caused by a softmax layer computed over a very large output vocabulary, which is usually fixed with sampled softmax methods

Answer: B

Focal loss scales cross-entropy by a factor that shrinks for confident correct predictions, which keeps the many easy background examples from dominating training. Exploding gradients in recurrent networks are the tempting distractor, and they are a different problem with different fixes. Duplicate images and vocabulary-size softmax costs are unrelated to the design of focal loss.

10. In a language model, what does next-token cross-entropy reward?
- A Producing the longest possible continuation for each prompt during the training runs
 - B Matching the exact hidden state vector that was observed in the previous training step
 - C Assigning high probability to the token that actually follows in the training text
 - D Keeping every token probability equal across the full vocabulary at all positions in the text

Answer: C

Next-token cross-entropy is the negative log probability of the observed next token, so it rewards placing mass on that token. Length is the tempting distractor, since generation length matters at inference, but the training loss does not reward it directly. Matching hidden states and uniform distributions are not what the loss measures.