Quiz 5 - Results

Attempt 1 of 1

Written Dec 5, 2025 1:07 PM - Dec 5, 2025 1:18 PM

Attempt Score **11 / 12 - A-**

Overall Grade (Highest Attempt) **11 / 12 - A-**

Question 1 1 / 1 point

What are the contributions of Flash Attention:

- ○ Fused kernel implementation
- ○ Computing the SoftMax without realizing NxN Attention matrix A
- ○ Leveraging the recomputation of the attention matrix in the backward pass
- ○ Using the tiling technique to chunk the computations of the NxN softmax/scores matrix into blocks.
- ✓ ◉ All the above

Question 2 1 / 1 point

To compute the softmax in Flash attention, these statistics are being tracked:

- Block size and max score of each block
- ✓ The maximum score of each block and the sum of the exponent scores
- The sum of the exponent scores
- Block size, max score of each block, and the sum of the exponent scores

Question 3 — 0 / 1 point

What is the primary innovation introduced by FlashAttention, as described in the paper (https://arxiv.org/pdf/2205.14135)?

- Enhanced gradient descent optimization for faster convergence
- ✗ Improved memory utilization in neural networks
- ➡ Exact attention mechanism with reduced computational cost
- Introduction of a new activation function for better feature extraction

Question 4 — 1 / 1 point

Computing softmax in attention requires the entire input, posing a challenge when dealing with large attention matrices (nxn). To address this, FlashAttention uses tiling, where HBM sends each block to SRAM for attention computation, achieved through algorithm restructuring. The large softmax is then decomposed into smaller ones.

- ✓ True
- False

Question 5 — 1 / 1 point

Which type of knowledge in a neural network focuses on the final output layer of

the teacher model?

- [x] Response-based knowledge
- [ ] Feature-based knowledge
- [ ] Relation-based knowledge
- [ ] Procedural knowledge

Question 6 1 / 1 point

What is the primary difference between offline and online distillation methods?

- [ ] Offline distillation updates both teacher and student models simultaneously.
- [ ] Online distillation requires a pre-trained teacher model.
- [ ] Offline distillation updates the teacher model after the student model.
- [x] Online distillation updates both teacher and student models simultaneously.

Question 7 1 / 1 point

In knowledge distillation, what does the term "soft targets" refer to?

- [ ] The difficulty of training the student model
- [ ] The temperature parameter used in distillation loss
- [ ] The final output layer of the student model
- [x] The probability distribution over the output classes from the teacher model

Question 8

1 / 1 point

How does increasing the temperature parameter in the softmax function affect the probability distribution?

- It makes the distribution sharper.
- ✓ It makes the distribution smoother.
- It has no effect on the distribution.
- It increases the number of classes in the distribution.

Question 9

1 / 1 point

What is the purpose of using KL divergence as a loss function in knowledge distillation?

- ✓ To measure the difference between the student and teacher model outputs.
- To regularize the weights of the student model.
- To control the learning rate during training.
- To determine the architecture of the student model.

Question 10

1 / 1 point

How does minimizing the KL divergence loss contribute to the training of the student model?

- ✓ It encourages the student model to precisely match the teacher model's outputs.
- It increases the complexity of the student model.
-

It slows down the convergence of the training process.

It reduces the capacity of the student model.

Question 11 2 / 2 points

According to the Work-depth cost model, what is the work and depth for calculating the sum of 64 floating-point numbers?

16 and 63

6 and 64

16 and 64

16 and 31

✔ 6 and 63

Done