INSTRUCTIONS:

1. This quiz is for **10min** and **10pts**. It is *open* book/notes/electronics.

2. Mark/write your answers using a dark pen or pencil. Ensure your writing is clear and legible.

**Q1. [2pts]** Suppose you triple the model size and halve the mini-batch size. By what factor does the communication cost of Horovod change?

(A) 1.5x  
(B) 2x  
(C) 3x  
(D) 6x

**Q2. [2pts]** For which data modality are deep CNNs most commonly used?

(A) Tabular  
(B) Images  
(C) Text  
(D) None of these

**Q3. [2pts]** Which compiler technique in TVM has the main benefit for reducing the number of memory stalls during DL inference?

(A) Operator fusion  
(B) Nested parallelism  
(C) Tensorization of ops  
(D) Pipelining

**Q4. [6pts]** Write brief pseudocode for a MapReduce job to scale the k-nearest neighbor algorithm for binary classification. It should be scalable along the number of rows. Make sure to explain how your input is split to begin with.

(Hint: You need to explain 3 things: sharding scheme; row-independent computations in Map; cross-worker aggregation in Reduce.)
INSTRUCTIONS:

1. This quiz is for 10min and 10pts. It is open book/notes/electronics.

2. Mark/write your answers using a dark pen or pencil. Ensure your writing is clear and legible.

Q1. [2pts] For which data modality are Transformers most commonly used?

(A) Tabular  (B) Images  (C) Text  (D) None of these

Q2. [2pts] Which compiler technique in TVM has the main benefit for reducing the number of memory stalls during DL inference?

(A) Operator fusion  (B) Nested parallelism  (C) Tensorization of ops  (D) Pipelining

Q3. [2pts] Suppose you double the model size and halve the mini-batch size. By what factor does the communication cost of Horovod change?

(A) No change  (B) 0.5x  (C) 2x  (D) 4x

Q4. [6pts] Write brief pseudocode for a MapReduce job to scale the k-nearest neighbor algorithm for binary classification. It should be scalable along the number of rows. Make sure to explain how your input is split to begin with.

(Hint: You need to explain 3 things: sharding scheme; row-independent computations in Map; cross-worker aggregation in Reduce.)
CSE 234 Fall 2021 Surprise Quiz 2

INSTRUCTIONS:

1. This quiz is for 10 min and 10 pts. It is open book/notes/electronics.

2. Mark/write your answers using a dark pen or pencil. Ensure your writing is clear and legible.

Q1. [2 pts] Which compiler technique in TVM has the main benefit for reducing the number of memory stalls during DL inference?

(A) Operator fusion  (B) Nested parallelism  (C) Tensorization of ops  (D) Pipelining

Q2. [2 pts] Suppose you triple the model size and double the mini-batch size. By what factor does the communication cost of Horovod change?

(A) 1.5x  (B) 2x  (C) 3x  (D) 6x

Q3. [2 pts] On which data modality is DL still not that popular?

(A) Tabular  (B) Images  (C) Text  (D) None of these

Q4. [6 pts] Write brief pseudocode for a MapReduce job to scale the k-nearest neighbor algorithm for binary classification. It should be scalable along the number of rows. Make sure to explain how your input is split to begin with.

(Hint: You need to explain 3 things: sharding scheme; row-independent computations in Map; cross-worker aggregation in Reduce.)