DLP - Week-4
Graded Assignment - 4

Assume that we want to continue pertaining a model containing 7 billion parameters with a context length of 2048 and the embedding dimension (model dimension) of 2048. The dataset contains 1 million samples (the size of each sample is exactly equal to the context length of the model). Suppose we set "per device batch size" to 16 and the gradient accumulation steps to 4.

1) How many iterations does it take to complete 1 epoch in a single GPU setting?

batch size = 16, gradient = 4

samples processed per iteration = 16 × 4 = 64

$$\text{Iteration} = \frac{\text{dataset}}{\text{samples processed per iteration}} = \frac{1,000,000}{64} = 15,625$$

2) The statement that the weight update happens **only** at the end of one epoch is *1 point*

- [ ] True
- [x] False

the wt. updates occur in neural network training after processing 1 batch of data.

3) How much memory (GB) is required to load the model with fp32? Use 1GB = 1e9.

1 param = 4 bytes (how? → fp 32 → $\frac{32}{8} = 4$)

$$\text{memory req. (in GB)} = \frac{\text{no. of param} \times \text{bytes per param}}{\text{bytes per GB}}$$

7 billion $= \frac{7 \times 10^9 \times 4}{1 \times 10^9} = 28$ (1×10^9)

4) Suppose the context length is increased from 2048 to 8000. Then, how much memory (GB) is required to load the model with fp32? Use 1GB = 1e9.

There is no role of context len in the memory req., the context len is more or less req. for creating the vector store.

5) Suppose we fine-tune all the parameters of the model for Named Entity Recognition (NER) with a suitable dataset in a supervised manner.

There are 8 labels (entities). Suppose we prefer to tune only the classification head (one linear layer with appropriate size). Enter the number of parameters in the classification head (exclude the bias).

no. of params = context_len (also known as embedding_dim) × no. of labels (this includes the bias)

= 2048 × 8 = 16384 − 2048 (1 each bias)

= 14336

6) Task-specific (selective/partial) fine-tuning with a classification head takes far less memory than task-specific full-fine-tuning. The statement is *1 point*

- [x] True
- [ ] False

true as it changes only the output layer of classifica" head, which is the linear layer. This is always better than task-specific fine-tuning (which messes up all the layers of 12 self attention & FFN)

7) The training objective for both continued pre-training and Instruction tuning is the same. The statement is 1 point

- True
- False

it might be the same because both want to achieve that model is fine-tuned.

8) Suppose we apply LoRA adapters to the model for parameter-efficient fine-tuning. Increasing the rank r will increase the number of parameters to be tuned. The statement is 1 point

- True
- False

$r \uparrow$, no. of params $\uparrow$

$A = (d_{in} \times r)$
$B = (d_{out} \times r)$
} A & B matrices are used in LORA optimizⁿ

Total no. of trainable params $= r \times (d_{in} + d_{out})$

$\therefore$, $r \uparrow$, params $\uparrow$