PASS QUESTION

QUESTION & ANSWER

HIGHER QUALITY, BETTER SERVICE

Provide One Year Free Update!

https://www.passquestion.com

Exam : **NCA-GENL**

Title : NVIDIA Generative AI LLMs

Version : DEMO

1.Why do we need positional encoding in transformer-based models?
A. To represent the order of elements in a sequence.
B. To prevent overfitting of the model.
C. To reduce the dimensionality of the input data.
D. To increase the throughput of the model.
Answer: A
Explanation:
Positional encoding is a critical component in transformer-based models because, unlike recurrent neural networks (RNNs), transformers process input sequences in parallel and lack an inherent sense of word order. Positional encoding addresses this by embedding information about the position of each token in the sequence, enabling the model to understand the sequential relationships between tokens. According to the original transformer paper ("Attention is All You Need" by Vaswani et al., 2017), positional encodings are added to the input embeddings to provide the model with information about the relative or absolute position of tokens. NVIDIA's documentation on transformer-based models, such as those supported by the NeMo framework, emphasizes that positional encodings are typically implemented using sinusoidal functions or learned embeddings to preserve sequence order, which is essential for tasks like natural language processing (NLP).
Options B, C, and D are incorrect because positional encoding does not address overfitting, dimensionality reduction, or throughput directly; these are handled by other techniques like regularization, dimensionality reduction methods, or hardware optimization.
Reference: Vaswani, A., et al. (2017). "Attention is All You Need."
NVIDIA NeMo Documentation: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html

2.What is Retrieval Augmented Generation (RAG)?
A. RAG is an architecture used to optimize the output of an LLM by retraining the model with domain-specific data.
B. RAG is a methodology that combines an information retrieval component with a response generator.
C. RAG is a method for manipulating and generating text-based data using Transformer-based LLMs.
D. RAG is a technique used to fine-tune pre-trained LLMs for improved performance.
Answer: B
Explanation:
Retrieval-Augmented Generation (RAG) is a methodology that enhances the performance of large language models (LLMs) by integrating an information retrieval component with a generative model. As described in the seminal paper by Lewis et al. (2020), RAG retrieves relevant documents from an external knowledge base (e.g., using dense vector representations) and uses them to inform the generative process, enabling more accurate and contextually relevant responses. NVIDIA’s documentation on generative AI workflows, particularly in the context of NeMo and Triton Inference Server, highlights RAG as a technique to improve LLM outputs by grounding them in external data, especially for tasks requiring factual accuracy or domain-specific knowledge.
Option A is incorrect because RAG does not involve retraining the model but rather augments it with retrieved data.
Option C is too vague and does not capture the retrieval aspect, while Option D refers to fine-tuning, which is a separate process.

Reference: Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." NVIDIA NeMo Documentation: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html

3.In the context of fine-tuning LLMs, which of the following metrics is most commonly used to assess the performance of a fine-tuned model?
A. Model size
B. Accuracy on a validation set
C. Training duration
D. Number of layers
Answer: B
Explanation:
When fine-tuning large language models (LLMs), the primary goal is to improve the model's performance on a specific task. The most common metric for assessing this performance is accuracy on a validation set, as it directly measures how well the model generalizes to unseen data. NVIDIA's NeMo framework documentation for fine-tuning LLMs emphasizes the use of validation metrics such as accuracy, F1 score, or task-specific metrics (e.g., BLEU for translation) to evaluate model performance during and after fine-tuning. These metrics provide a quantitative measure of the model's effectiveness on the target task.
Options A, C, and D (model size, training duration, and number of layers) are not performance metrics; they are either architectural characteristics or training parameters that do not directly reflect the model's effectiveness.
Reference: NVIDIA NeMo Documentation: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/model_finetuning.html

4.Which of the following claims is correct about quantization in the context of Deep Learning? (Pick the 2 correct responses)
A. Quantization might help in saving power and reducing heat production.
B. It consists of removing a quantity of weights whose values are zero.
C. It leads to a substantial loss of model accuracy.
D. Helps reduce memory requirements and achieve better cache utilization.
E. It only involves reducing the number of bits of the parameters.
Answer: A, D
Explanation:
Quantization in deep learning involves reducing the precision of model weights and activations (e.g., from 32-bit floating-point to 8-bit integers) to optimize performance. According to NVIDIA's documentation on model optimization and deployment (e.g., TensorRT and Triton Inference Server), quantization offers several benefits:
Option A: Quantization reduces power consumption and heat production by lowering the computational intensity of operations, making it ideal for edge devices.
Option D: By reducing the memory footprint of models, quantization decreases memory requirements and improves cache utilization, leading to faster inference.
Option B is incorrect because removing zero-valued weights is pruning, not quantization.
Option C is misleading, as modern quantization techniques (e.g., post-training quantization or

quantization-aware training) minimize accuracy loss.
Option E is overly restrictive, as quantization involves more than just reducing bit precision (e.g., it may include scaling and calibration).
Reference: NVIDIA TensorRT Documentation: https://docs.nvidia.com/deeplearning/tensorrt/developer-guide/index.html
NVIDIA Triton Inference Server Documentation: https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/index.html

5.What is the primary purpose of applying various image transformation techniques (e.g., flipping, rotation, zooming) to a dataset?
A. To simplify the model's architecture, making it easier to interpret the results.
B. To artificially expand the dataset's size and improve the model's ability to generalize.
C. To ensure perfect alignment and uniformity across all images in the dataset.
D. To reduce the computational resources required for training deep learning models.
Answer: B
Explanation:
Image transformation techniques such as flipping, rotation, and zooming are forms of data augmentation used to artificially increase the size and diversity of a dataset. NVIDIA's Deep Learning AI documentation, particularly for computer vision tasks using frameworks like DALI (Data Loading Library), explains that data augmentation improves a model's ability to generalize by exposing it to varied versions of the training data, thus reducing overfitting. For example, flipping an image horizontally creates a new training sample that helps the model learn invariance to certain transformations.
Option A is incorrect because transformations do not simplify the model architecture.
Option C is wrong, as augmentation introduces variability, not uniformity.
Option D is also incorrect, as augmentation typically increases computational requirements due to additional data processing.
Reference: NVIDIA DALI Documentation: https://docs.nvidia.com/deeplearning/dali/user-guide/docs/index.html