ADARLAB AI Training Course

Lec10 Homework Report

110511118 陳孟頡

Part I. Please explain the training methods in ChatGPT.

- **Training Process:**

  1. Pre-training:

  以大量的網路文章、資訊對 GPT 模型進行非監督式學習訓練，學習一般語言模式、文法、事實知識，習得基本推理能力以進行文字接龍任務。

  2. Supervised fine-tuning:

  藉由提供人類的對話、語句及偏好的答案，讓模型學習出較佳的文字接龍回答。

  3. Reinforcement Learning from Human Feedback:

  人類對模型所提出的答案進行好壞的評比，以此讓模型的行為與人類價值和偏好保持一致，以提高安全性和實用性。

Part II. Please explain why do we mostly use fine-tuning instead of training from scratch on language models? And explain what is instruction fine-tuning?

- **Benefits of Fine-Tuning Model:**

  1. Resource efficiency:

  訓練模型需耗費大量資源，且經過 pre-train 的模型已具備語言架構與一般知識，與其在訓練新模型時 train from scratch 消耗大量時間、金錢，不如對 pre-train model 進行 fine-tune，來減少資源消耗。

  2. Faster convergence:

  Fine-tune 模型通常比 train from scratch 模型收斂的快，以更少的時間達到最佳效能。

  3. Generalization:

  經過 pre-train 的模型擁有大量一般知識，因此在執行新熱霧上通常有著更佳的表現。

- **Instruction Fine-Tuning:**

  Instruction fine-tuning 對模型的 instructional prompt 及對應的 dataset 進行微調，不僅可以提高模型在特定任務上的表現，增強其對特定應用程式的實用性，還能在保持一般語言理解能力下，提高模型執行一般指令的效能。

- **KV Cache Calculation:**

  KV cache 在 LLaMA 模型中被用以減少計算量，以下計算所需 KV cache 大小:

  假設條件:

  1. Number of layers: 32
  2. Hidden size (d_model): 4096
  3. Number of attention heads: 32
  4. Max sequence length: 2048

(1) KV head size = hidden_size / num_heads = 4096 / 32 = 128

(2) Size per token = 2 * num_layers * num_heads * size_per_head

= 2 * 32 * 32 * 128 = 262,144 elements

(3) Number of elements = elements * max_sequence_length

= 262,144 * 2048 = 536,870,912

(4) Total cache size for max sequence = 536,870,912 * 2(float16)

= 1,073,741,824 bytes ≈ 1 GB