Gemma LLM Guide

Introduction to Gemma

- Overview of Gemma as a family of open-weight Large Language Models (LLMs) developed by Google DeepMind, based on Gemini research and technology.

Installation

- Step-by-step instructions for installing Gemma:
    1. Install JAX for CPU, GPU, or TPU as per the instructions on the JAX website.
    2. Run `pip install gemma` to install the Gemma package.

Model Checkpoints

- Guidance on downloading and loading pre-trained Gemma model checkpoints from KaggleHub, including instructions for manual and programmatic downloads.

Tokenizer Usage

- Explanation of the Gemma tokenizer, including encoding and decoding methods, and the significance of control tokens like `<bos>`, `<eos>`, `<start_of_turn>`, and `<end_of_turn>`.

Sampling Methods

- Examples of how to perform multi-turn conversations using `gm.text.ChatSampler`, and other sampling techniques.

Fine-Tuning Gemma

- Instructions on fine-tuning Gemma models using the Kauldron library, including setting up the data pipeline and training loop.

Parameter-Efficient Fine-Tuning (PEFT) with LoRA

- Guidelines on implementing LoRA (Low-Rank Adaptation) for efficient fine-tuning, including both training and inference processes.

API Reference

- Detailed descriptions of key modules and classes in the Gemma library, such as `gm.nn.Gemma3_4B`, `gm.text.ChatSampler`, and `gm.nn.LoRA`.