Dokumentacja do projektu

Repozytorium składa się z 3 katalogów. Każdy z katalogów rozwiązuje pojedynczy problem.

1. **Uruchom inferencję modelu Qwen3-0.6B przy użyciu Colab lub MLX** (dla użytkowników Maca). (Google Codelabs)
2. **Wytrenuj model LLaMA o 10 milionach parametrów na zbiorze 0,5M–1 miliarda tokenów** (laptop z kartą Nvidia, zoptymalizowany pod CUDA i GTX 1650).
3. **Przeprowadź nadzorowane dostrajanie (supervised finetuning) modelu Qwen2.5-0.5B Base z wykorzystaniem QLoRA**. (Google Codelabs)

Instrukcja.

Pobierz pliki i uruchom w swoim środowisku.

Środowisko z ćwiczeń 1 i 3 było uruchamiane na platformie GitHub z wykorzystaniem Codespace. Środowisko z ćwiczenia 2 realizowane lokalnie z wykorzystaniem 3 metod obliczeń:

a) z wykorzystaniem CPU – *BabyLLama_CPU*

b) z wykorzystaniem GPU (Nvidia GTX 1650) – *BabyLLama_GPU2*

c) z wykorzystaniem GPU Intela (i7 155H – ARC8) – *BabyLLama_XPU*

Dla procesora Intela wykorzystano biblioteki OneAPI, oraz bibliotek dedykowanych dla XPU Intela.

Wyniki dla prezentowanych zadań prezentowane są na stronie:

pawelzm/babyLlama · Hugging Face

pawelzm/lora_model · Hugging Face

Project Documentation

The repository consists of three directories. Each directory addresses a specific task:

1. **Run inference using the Qwen3-0.6B model with Colab or MLX** (for Mac users). (Google Codelabs)
2. **Train a LLaMA model with 10 million parameters on a dataset ranging from 0.5M to 1 billion tokens** (laptop with Nvidia GPU, CUDA optimized for GTX 1650).
3. **Perform supervised fine-tuning of the Qwen2.5-0.5B Base model using QLoRA.** (Google Codelabs)

Instructions:

Download the files and run them in your environment.

Exercises 1 and 3 were executed on the GitHub platform using Codespace.

Exercise 2 was carried out locally using three computing methods:

a) Using the CPU – *BabyLLama_CPU*

b) Using an Nvidia GPU (GTX 1650) – *BabyLLama_GPU2*

c) Using Intel GPU (i7 155H – ARC8) – *BabyLLama_XPU*

For the Intel processor, OneAPI libraries and dedicated Intel XPU libraries were used.

The results for the presented tasks are available at:

pawelzm/babyLlama · Hugging Face

pawelzm/lora_model · Hugging Face