RakutenAI LLM

https://huggingface.co/Rakuten

Continued Pretraining

- Trained with Mistral 7b v1 w/ Japanese Data
- Evaluation: Language Model Evaluation Harness

Supervised fine-tuning

- Task-oriented fine-tuning

RLHF (Alignment)

- Small amount of chat data
- Positive and negative examples
- Evaluate w/ human or much larger LLM

Designing Architecture

- Lightweight vs Heavyweight
- Tokenization
  - Challenge: byte-fallback for non-English text
  - Designing more efficient tokenizers-more characters per token

Preparing Data

Format > Cleaning > Lang Detection > Quality Filter > Deduplication > Downsizing by Quality > Dataset

Application

- Assumes role of concierge replying to an R-travel review

PEFT with LoRA

Specific instruction datasets

- User: Extract sentiment from the following text: This movie is great!
- Assistant: The sentiment is positive

.

GPU Resources

- Inference:
  - Half precision: 14 GB
  - Full precision: 28 GB
- LoRA fine-tuning
  - Gradients: 0.1-0.5 GB
  - Optimizer States: 0.1-1 GB
- LoRA Parameters can be merged with the original model

LoRA Disadvantages

- Data scaling is **ineffective**

- LoRA fine-tuning is more aligned with the pre-trained model

Improving LLMs with Preference Optimisation

- Improving translation
    - E.g.: Variations in translation (such as from formal to informal)
- LLM Alignment
    - Making an LLM behave according to user expectations
    - Telling a LLM what to do and what **not** to do
- Human feedback:
    - Creating sentences: Generate parallel sentences
        - Pros: Highly customisable
        - Cons: Time consuming/expensive
    - Postedit LLM output
        - Pros: Less effort than creating sentences and fine-grained feedback
        - Cons: Time consuming/expensive
    - Scoring
        - Pros: Less expensive
        - Cons: Subjective scoring, how much should each aspect influence?
    - Deciding between alternatives
        - Pros: Least expensive
        - Cons: Does not say anything about output:
            - Why is one better than the other?
            - What is both are wrong?
- Direct Preference Optimisation (DPO)
    - Alternative to RLHF
    - No need for a separate reward model