國立陽明交通大學 電機工程學系

Department of Electronics and Electrical Engineering

NYCU Llama for everyone

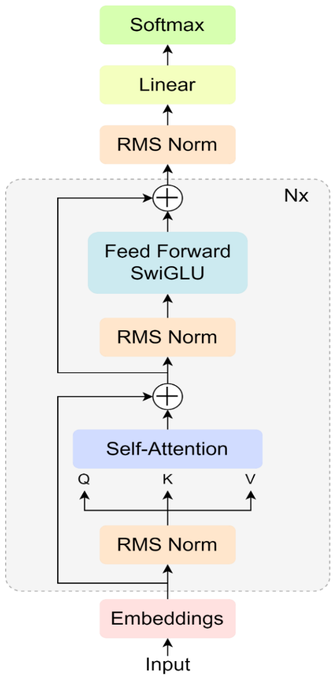
專題指導教授：黃俊達　　專題學生：張語楹、蔡雅婷、許睿洋

Abstract

Our goal is to establish a large language model (LLM) capable of answering questions related to NYCU. Our approach involves leveraging Taiwan Llama, a 7-billion-parameter model, then subsequently finetunes it using LoRA. To collect our data, we generate question-answer pairs using GPT-3.5. The resultant model is deployed on the web for seamless inference.

Llama 2

Llama2, the large language model launched by Meta, underscores the concept of smaller model trained on more data. The released model comes in three sizes (7B, 13B, and 70B), and they are accessible on Hugging Face with a license.



Demo website

GitHub repo

Dataset

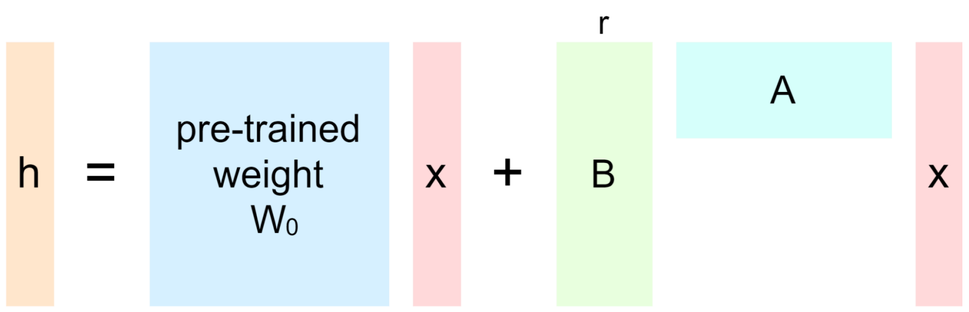
We collect data containing documents from NYCU and information of our department, then use GPT3.5 to generate QA pairs. Finally, the dataset includes 3000 pairs.

data from websites　　creating QA pairs　　integrate into json file

LoRA

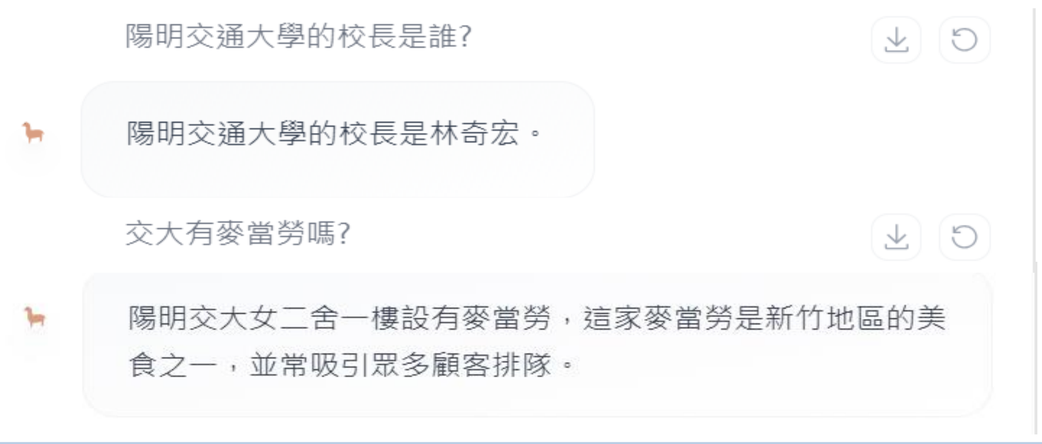
Low-Rank Adaptation enables updating dense layer ΔW indirectly by using rank decomposition $\Delta W = BA$, $B \in R^{d\times r}, A \in R^{r\times k}, r \ll \min(d,k)$, and all pre-trained weights $W_0 \in R^{d\times k}$ are kept frozen.

$$h = W_0 x + \Delta W x$$



Train B and A matrices only
Reduce trainable parameters by 10,000 times
Reduce GPU memory by 3 times

Inference

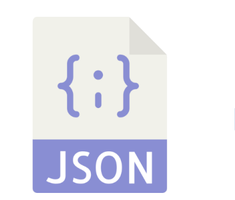


Experiment Flow

1. Collect data about NYCU and our department
2. Finetune pre-trained model, Taiwan Llama, using PEFT
3. Search suitable hyperparameters of LoRA config and use wandb to track training logs
4. Deploy on the web for easy inference

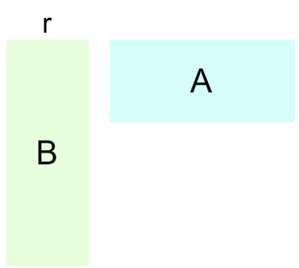
Future work

Our future work includes collecting more data based on NYCU and use RAG (retrieval augmented generation) for a more reliable LLM.

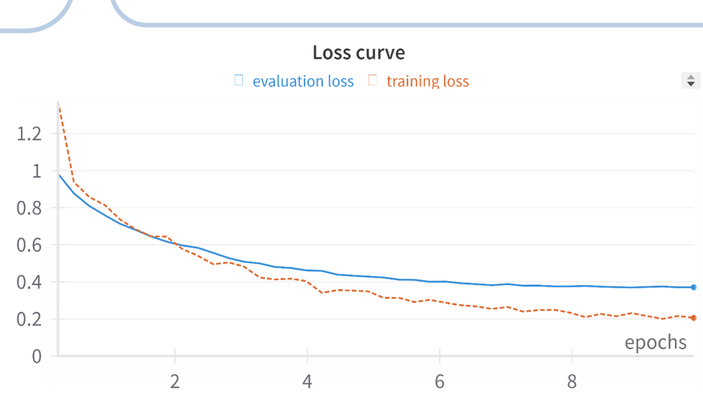


3000 QA pairs about NYCU

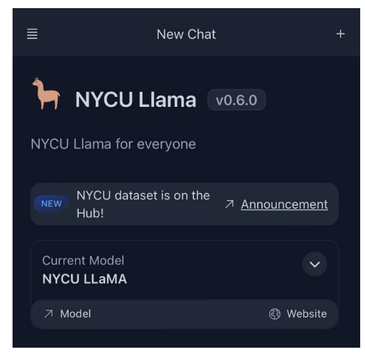
better performance on traditional Chinese



10 min/epoch
23 GB VRAM



LoRA hyperparameters:
$\alpha = 16, r = 16$ (α is the weight of ΔW with respect to W_0, r is intrinsic rank)



User interface