

Module 2

Monday, June 17, 2024 10:03 PM

Pretraining for Domain adaptation:

Pre-training for domain adaptation

Legal language

The prosecutor had difficulty proving mens rea, as the defendant seemed unaware that his actions were illegal.

The judge dismissed the case, citing the principle of res judicata as the issue had already been decided in a previous trial.

Despite the signed agreement, the contract was invalid as there was no consideration exchanged between the parties.

Medical language

After a strenuous workout, the patient experienced severe myalgia that lasted for several days.

After the biopsy, the doctor confirmed that the tumor was malignant and recommended immediate treatment.

Sig: 1 tab po qid pc & hs

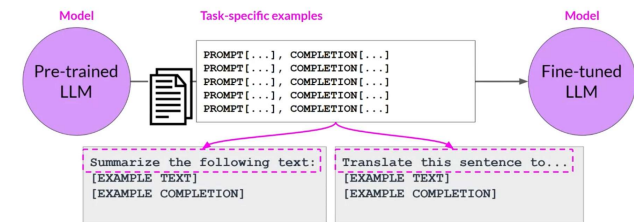
Why domain specific model?
→ Because language in that domain is different. Example

← When scaling data, we need to increase both data and number of parameters of model.

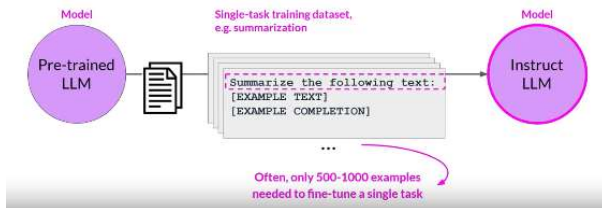
Fine-Tuning is supervised method to change LLM behavior

Using prompts to fine-tune LLMs with instruction

LLM fine-tuning



Fine-tuning on a single task



- Single-task training dataset
- Often only 500-1000 examples are needed to fine-tune a single task.

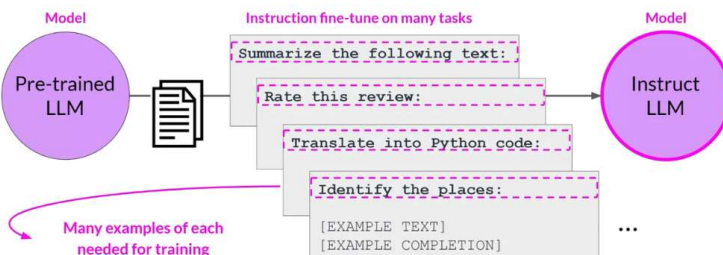
Problem: Catastrophic forgetting

Solution: Fine-tune on multiple tasks
(or) Parameter Efficient Fine-tuning (PEFT)

Multi-task, instruction fine-tuning:

→ contains example of different instruction fine-tuning on many tasks

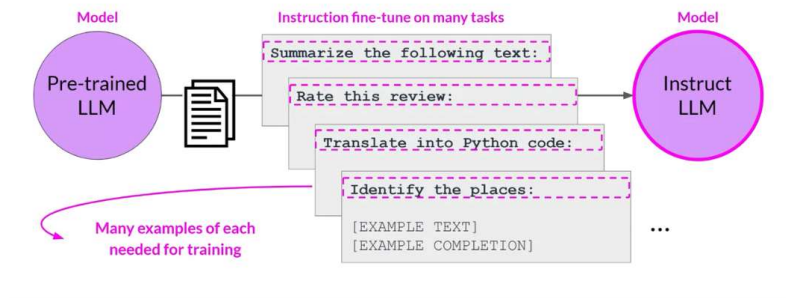
Multi-task, instruction fine-tuning



• Major Problem is that it requires lot of data, but it is worth it!

Multi-task, instruction fine-tuning

• Many people think it's too much data, but it is worth it!



Instruction fine-tuning with FLAN

→ Fine Tuned Language Net: It is a specific set of instruction used to perform instruction fine-tuning.

Model Evaluation:

ROUGE

Recall-Oriented Understudy for Gisting Evaluation

- Used for text summarization
- Compare a summary to one or more reference summaries

LLM Evaluation - Metrics - ROUGE-L

Reference (human): It is cold outside.	ROUGE-L Recall: = LCS(Gen, Ref) / unigrams in reference = $\frac{2}{4} = 0.5$
Generated output: It is very cold outside.	ROUGE-L Precision: = LCS(Gen, Ref) / unigrams in output = $\frac{2}{5} = 0.4$
LCS: Longest common subsequence	ROUGE-L F1: = $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \times \frac{0.2}{0.9} = 0.44$

BLEU Score (Bilingual evaluation understudy)

- Used for text translation
- Compare to human-generated translation

Benchmarks:

GLUE: General language understanding tasks
→ contains sentiment analysis and question answering

Super GLUE: Multi-sentence reasoning and reading comprehension

Big-Bench: Contains different dataset sizes to test against

HELM: Includes metrics like Fairness, Bias, Toxicity

MMPU: Massive Multi task Language Understanding
→ Designed specifically for modern LLMs.

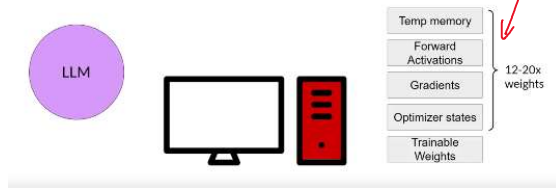
Parametric Efficient Fine-Tuning (PEFT)

- While fine tuning models, it involves storing model variables that are not usually loaded for inferencing. For fine tuning, we need atleast

Full fine-tuning of large LLMs is challenging



Full fine-tuning of large LLMs is challenging

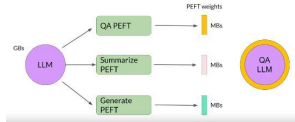


PEFT → fine-tuning a small subset of parameters.



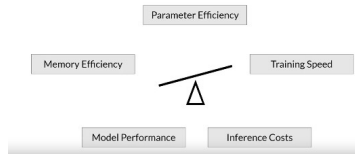
- Either freeze the model and train subset of layers.
- or freeze the whole model and add new layers.

PEFT fine-tuning saves space and is flexible



we can then add new layers on top of original layer for our usecase.

PEFT Trade-offs



PEFT methods

Selective

Select subset of initial LLM parameters to fine-tune

→ Both these methods use same models parameters.

Reparameterization

Reparameterize model weights using a low-rank representation

LoRA

Additive

Add trainable layers or parameters to model

Adapters

Soft Prompts

Ex → Prompt Tuning

Add new trainable layers to encoder/decoder architecture

Freezes architecture, and changes the input type to improve performance

By adding trainable parameters to the prompt embeddings (or) keeping the input fixed and retraining the embedding weights.

Source: Lialin et al. 2023, "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning".

Low-Rank Adaptation of Large Language Models

- Parametric Efficient Fine-Tuning on single GPU.
- Choosing rank of lower matrices is still Active Research (4-32 is good enough)

Widely used

Prompt Tuning

Prompt Tuning is not prompt engineering

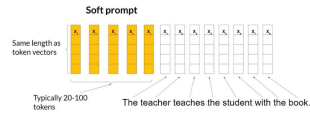
Limitations of Prompt Engineering:

- Requires lot of effort to write and try different prompts.
- Also limited by the length of context window.

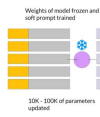
Prompt tuning adds trainable "soft prompt" to inputs



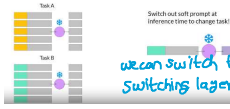
↳ Limitations of Prompt Engineering:
 → Requires lot of effort to write and try different prompts.
 → Also limited by the length of context window.
 With prompt tuning → we add additional trainable tokens to the prompt.



Full Fine-tuning vs prompt tuning



Prompt tuning for multiple tasks



we can switch tasks by switching layers.