Machine Learning
Ecosystem

Lecture 16 for 14-763/18-763

Guannan Qu

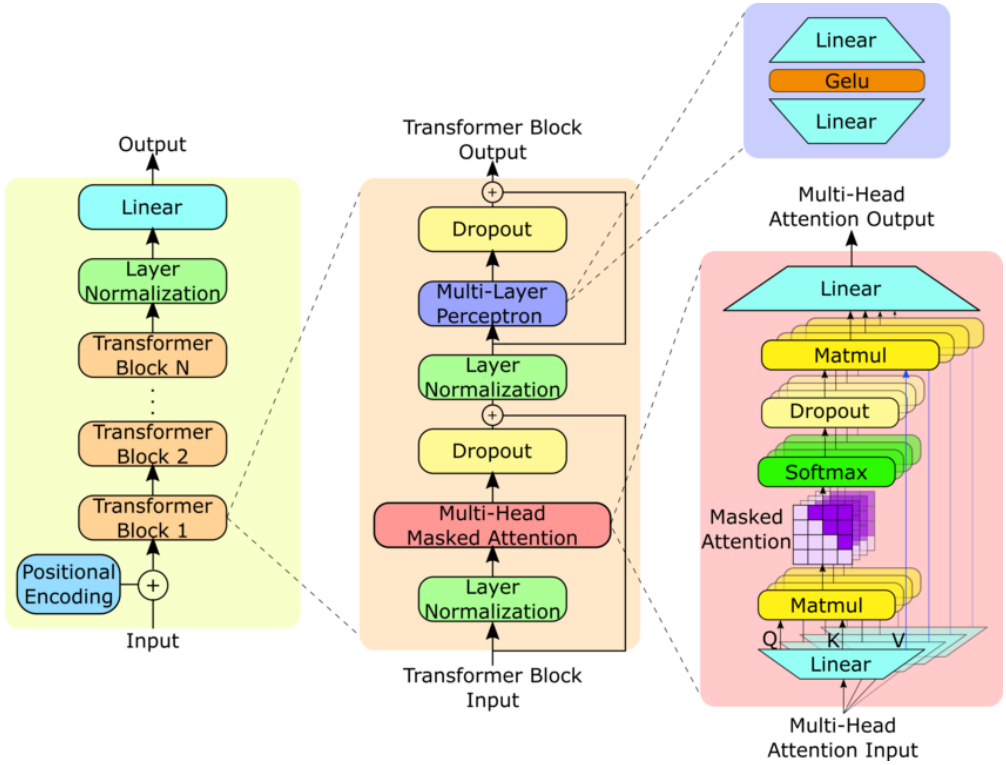
Oct 30, 2024

Today’s Agenda

- Large foundational models
- HuggingFace Hub
- Model size, memory and GPU
- Lab: fine-tuning a language model

What is Large Foundational Model?

You must have heard about GPT, Llama, BERT, etc. Foundational models are deep learning models that are **Large**:



Mode Architecture of GPT-2

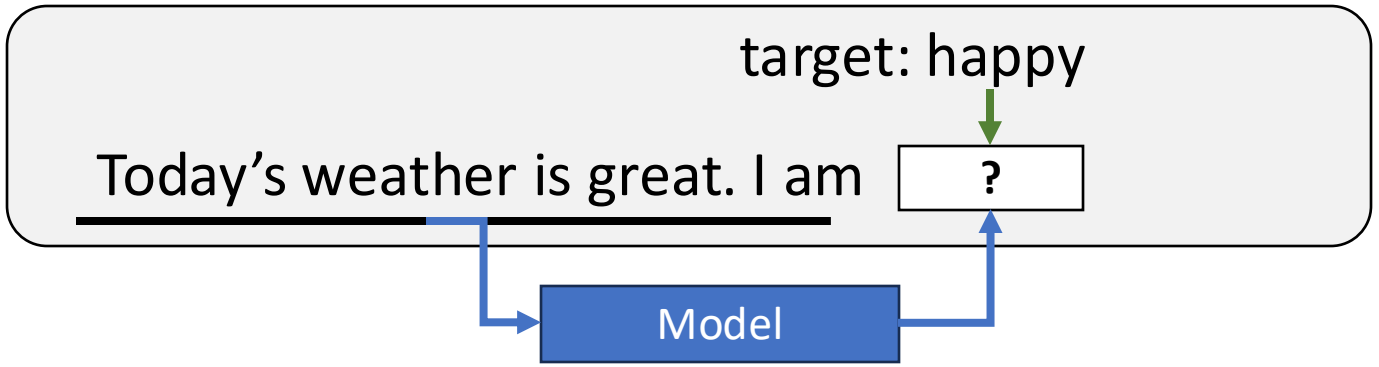
Can contain many layers, totaling billions of parameters

What is Large Foundational Model?

You must have heard about GPT, Llama, BERT, etc. Foundational models are deep learning models that are **Large, Pretrained**:

- Data source is **large** and **unlabeled**. Including Books, Websites, Wikipedia, Scientific Papers, Code Repositories (GPT-3 used 45 TB of data)

- Pretrained with **self-supervised learning**: construct targets from unlabeled data



Model is trained to predict the next word

What is Large Foundational Model?

You must have heard about GPT, Llama, BERT, etc. Foundational models are deep learning models that are **Large, Pretrained, Finetuned**:

- The pre-trained model “understands” language, but is not designed for specific tasks.
- Finetuning means using small amount of **labeled** data to train the pretrained model to perform tasks like sentiment classification, question answering, translation, …

Common Large Language Models

Increasing Model Size

ChatGPT

Year	Model Name	Company	Model Size
2018	BERT	Google	110M - 340 M
2018	GPT	OpenAI	117 M
2019	GPT-2	OpenAI	117M - 1.5 B
2019	RoBERTa	Facebook	125M - 355 M
2019	T5	Google	220M - 11 B
2020	GPT-3	OpenAI	175 B
2022	GPT-3.5	OpenAI	UNKNOWN, Similar to GPT-3
2023	GPT-4	OpenAI	UNKNOWN
2023	Llama 1 & 2	Meta	7 B – 65B
2023	Gemini	Google	UNKNOWN

Many more since 2023-2024, including Mistrial, Claude, …

Ecosystem

- HuggingFace is an AI company that hosts Model Hub, a platform where users can discover, share, and deploy pre-trained models and datasets.
- It hosts thousands of models contributed by both Hugging Face and the community, covering a wide range of languages and tasks.
- It also hosts many datasets
- Provide libraries “**transformers**”, ”**datasets**” to facilitate easy loading and sharing of models and datasets

Ecosystem

Hugging Face | Search mo | Models | Datasets | Spaces | Posts | Docs | Pricing | Log In | Sign Up

Tasks 1 | Libraries | Datasets | Languages | Licenses | Other

Filter Tasks by name | Reset Tasks

Multimodal

- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering
- Video-Text-to-Text

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video

Models 135,715 | Filter by name | Full-text search | Sort: Most likes

meta-llama/Meta-Llama-3-8B
Text Generation • Updated May 13 • 2.06M • 5.59k

bigscience/bloom
Text Generation • Updated Jul 28, 2023 • 8.42k • 4.7k

mistralai/Mixtral-8x7B-Instruct-v0.1
Text Generation • Updated 19 days ago • 406k • • 4.09k

meta-llama/Llama-2-7b
Text Generation • Updated Apr 17 • 4.04k

meta-llama/Llama-2-7b-chat-hf
Text Generation • Updated Apr 17 • 649k • • 3.82k

meta-llama/Meta-Llama-3-8B-Instruct

Ecosystem

Hugging Face | Search mo | Models | Datasets | Spaces | Posts | Docs | Pricing | Log In | Sign Up

Main | Tasks | Libraries | Languages | Licenses | Other

Modalities | Reset Modalities

3D | Audio | Geospatial | Image | Tabular | Text × | Time-series | Video

Size (rows)

< 1K — > 1T

Format

json | csv | parquet | imagefolder | soundfolder | webdataset | text | arrow

Datasets 126,778 | Filter by na | Full-text search | Sort: Most downloads

lighteval/mmlu
Viewer • Updated Jun 9, 2023 • 5.82M • 19.2M • 29

argilla/databricks-dolly-15k-curated-en
Viewer • Updated Oct 2, 2023 • 15k • 2.67M • 41

lavita/medical-qa-shared-task-v1-toy
Viewer • Updated Jul 19, 2023 • 64 • 1.55M • 14

ceval/ceval-exam
Viewer • Updated Aug 31, 2023 • 13.9k • 1.2M • 230

allenai/ai2_arc
Viewer • Updated Dec 21, 2023 • 7.79k • 883k • 122

lukaemon/bbh

Lab: download a language model & dataset

See notebook for more details

Understanding Model Size

What does model size mean? e.g. Llama 1 has 7B parameters

- 1 parameter = 1 FloatingPoint32 (FP32) = 4 Bytes (1 Byte = 8bits)
- 1 Billion parameters = 4 Billion Bytes ≈ 4GB memory

It takes 28GB of memory to *store* the Llama 1 7B model!

For inference, a bit higher

For training/full-parameter-finetuning, >3x higher

} Highly dependent on input length, batch size, number of layers

Storing the gradient takes the same memory as model weights
Storing the Adam optimizer state also takes the same memory

Memory size of typical GPUs

Series	Target Users	Model Number	Memory
GeForce RTX	Desktops/Laptops, gaming	4070/4080/4090	12 GB – 24 GB
RTX Series (formerly Quadro)	Workstations (professional video creation/editing…)	A4000	20GB
		A5000	32 GB
		A6000	48 GB
Data Center (formerly Tesla)	Data center, for ML	V100 (2018)	16 GB/32 GB
		A100 (2020)	40 GB / 80 GB
		H100 (2022)	80 GB
		B100 (2024/2025)	192 GB

Memory need for storing Llama 2 in FP32

28GB Llama 2 7B

52GB Llama 2 13B

280GB Llama 2 70B

Reducing Memory Need

Use a lower precision float point. e.g. FP16 takes half memory as FP32.

- 1B parameters in FP16 takes 2GB, half that of FP32

Use further quantization methods to store model parameters in int8 (1/4 memory compared to FP32) or even less…

Reducing Memory Need

Machine Learning Research

Overview Research Events Work with us

Featured Highlight

Introducing Apple's On-Device and Server Foundation Models

June 10, 2024

For on-device inference, we use low-bit palletization, a critical optimization technique that achieves the necessary memory, power, and performance requirements. To maintain model quality, we developed a new framework using LoRA adapters that incorporates a mixed 2-bit and 4-bit configuration strategy — averaging 3.7 bits-per-weight — to achieve the same accuracy as the uncompressed models. More aggressively, the model can be compressed to 3.5 bits-per-weight without significant quality loss.

https://machinelearning.apple.com/research/introducing-apple-foundation-models

Reducing Memory Need - PEFT

Another popular method to reduce memory need for Fine-Tuning is **Parameter Efficient Fine Tuning**, more specifically the **LoRA** method

- It reduces the number of trainable parameters, often to less than 1% of the total parameter number
- Significantly reduces the memory need to store gradient and optimizer state
- Previously we said >3x memory is needed to fine-tune, but with LoRA, this number can be significantly reduced

LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* Yelong Shen* Phillip Wallis Zeyuan Allen-Zhu
Yuanzhi Li Shean Wang Lu Wang Weizhu Chen
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)

https://arxiv.org/pdf/2106.09685

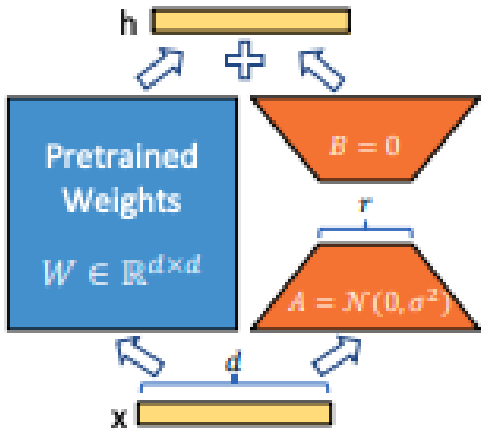


Figure 1: Our reparametrization. We only train A and B.

Lab: Fine-tuning a language model

See notebook for more details

Summary

- We are in the era of large foundation models
- Tools like pytorch is the foundation
- A huge ecosystem, particularly the HuggingFace community, is built to facilitate the research and development of large foundation models
  - transformers, datasets, peft
- Understand the memory need for large foundation models and ways to reduce it