Title: Fine-tuning (deep learning)

URL: https://en.wikipedia.org/wiki/Fine-tuning_(deep_learning)

PageID: 73250406

Categories: Category:Deep learning, Category:Machine learning

-----

In deep learning , fine-tuning is an approach to transfer learning in which the parameters of a pre-trained neural network model are trained on new data. Fine-tuning can be done on the entire neural network, or on only a subset of its layers , in which case the layers that are not being fine-tuned are "frozen" (i.e., not changed during backpropagation). A model may also be augmented with "adapters"—lightweight modules inserted into the model's architecture that nudge the embedding space for domain adaptation. These contain far fewer parameters than the original model and can be fine-tuned in a parameter-efficient way by tuning only their weights and leaving the rest of the model's weights frozen.

For some architectures, such as convolutional neural networks , it is common to keep the earlier layers (those closest to the input layer) frozen, as they capture lower-level features , while later layers often discern high-level features that can be more related to the task that the model is trained on.

Models that are pre-trained on large, general corpora are usually fine-tuned by reusing their parameters as a starting point and adding a task-specific layer trained from scratch. Fine-tuning the full model is also common and often yields better results, but is more computationally expensive.

Fine-tuning is typically accomplished via supervised learning , but there are also techniques to fine-tune a model using weak supervision . Fine-tuning can be combined with a reinforcement learning from human feedback -based objective to produce language models such as ChatGPT (a fine-tuned version of GPT models) and Sparrow .

Robustness

Fine-tuning can degrade a model's robustness to distribution shifts . One mitigation is to linearly interpolate a fine-tuned model's weights with the weights of the original model, which can greatly increase out-of-distribution performance while largely retaining the in-distribution performance of the fine-tuned model.

Variants

Low-rank adaptation

Low-rank adaptation (LoRA) is an adapter-based technique for efficiently fine-tuning models. The basic idea is to design a low- rank matrix that is then added to the original matrix. An adapter, in this context, is a collection of low-rank matrices which, when added to a base model, produces a fine-tuned model. It allows for performance that approaches full-model fine-tuning with lower space requirements. A language model with billions of parameters may be LoRA fine-tuned with only several millions of parameters.

LoRA-based fine-tuning has become popular in the Stable Diffusion community. Support for LoRA was integrated into the diffusers library from Hugging Face . Support for LoRA and similar techniques is also available for a wide range of other models through Hugging Face's parameter-efficient fine-tuning (PEFT) package.

Representation fine-tuning

Representation fine-tuning (ReFT) is a technique developed by researchers at Stanford University aimed at fine-tuning large language models (LLMs) by modifying less than 1% of their representations. Unlike parameter-efficient fine-tuning (PEFT) methods, which mainly focus on updating weights, ReFT targets representations, suggesting that modifying representations might be a more effective strategy than updating weights.

ReFT methods operate on a frozen base model and learn task-specific interventions on hidden representations and train interventions that manipulate a small fraction of model representations to steer model behaviors towards solving downstream tasks at inference time. One specific method within the ReFT family is low-rank linear subspace ReFT (LoReFT) , which intervenes on hidden representations in the linear subspace spanned by a low-rank projection matrix. LoReFT can be seen as the representation-based equivalent of low-rank adaptation (LoRA).

Applications

Natural language processing

Fine-tuning is common in natural language processing (NLP), especially in the domain of language modeling . Large language models like OpenAI 's series of GPT foundation models can be fine-tuned on data for specific downstream NLP tasks (tasks that use a pre-trained model) to improve performance over the unmodified pre-trained model.

Commercial models

Commercially-offered large language models can sometimes be fine-tuned if the provider offers a fine-tuning API. As of June 19, 2023, language model fine-tuning APIs are offered by OpenAI and Microsoft Azure 's Azure OpenAI Service for a subset of their models, as well as by Google Cloud Platform for some of their PaLM models, and by others.

See also

Catastrophic forgetting

Continual learning

Domain adaptation

Foundation model

Hyperparameter optimization

Overfitting

References

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder

Vibe coding

Vision transformer

Waluigi effect

Word embedding

Character.ai

ChatGPT

DeepSeek

Ernie

Gemini

Grok

Copilot

Claude

Gemini

Gemma

GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5

1

2

3

J

4

4o

4.5

4.1

OSS

5

Llama

o1

o3

o4-mini

Qwen

Base44

Claude Code

Cursor

Devstral

GitHub Copilot

Kimi-Dev

Qwen3-Coder

Replit

Xcode

Aurora

Firefly

Flux

GPT Image 1

Ideogram

Imagen

Midjourney

Qwen-Image

Recraft

Seedream

Stable Diffusion

Dream Machine

Hailuo AI

Kling

Midjourney Video

Runway Gen

Seedance

Sora

Veo

Wan

15.ai

Eleven

MiniMax Speech 2.5

WaveNet

Eleven Music

Endel

Lyria

Riffusion

Suno AI

Udio

Agentforce

AutoGLM

AutoGPT

ChatGPT Agent

Devin AI

Manus

OpenAI Codex

Operator

Replit Agent

01.AI

Aleph Alpha

Anthropic

Baichuan

Canva

Cognition AI

Cohere

Contextual AI

DeepSeek

ElevenLabs

Google DeepMind

HeyGen

Hugging Face

Inflection AI

Krikey AI

Kuaishou

Luma Labs

Meta AI

MiniMax

Mistral AI

Moonshot AI

OpenAI

Perplexity AI

Runway

Safe Superintelligence

Salesforce

Scale AI

SoundHound

Stability AI

Synthesia

Thinking Machines Lab

Upstage

xAI

Z.ai

Category