RAJ GARG

rajgarg021.github.io/blog
rajgarg021@gmail.com
(+91) 75670 25289

SKILLS

Languages: Python, R, SQL, C, LISP, Javascript, Mojo
Libraries/Frameworks: PyTorch, Transformers, Diffusers, LlamaIndex, LangChain, LangGraph, Jax

EXPERIENCE

Career Break **Dec'22 - Present**

- Took a planned sabbatical to focus on personal growth and skill development.
- Spent time reading research papers, learning the fundamentals and staying up-to-date with the latest advancements in the AI field.
- Implemented some research papers, experimented with open-source language models and created some cool projects mentioned below.

Monsoon CreditTech (Senior Data Scientist, Gurgaon) **Jul'22 - Nov'22**

- Took full ownership of the project from the start, built and deployed Application Scorecard (Credit underwriting risk model to assess likelihood of default) for loan applications at a leading NBFC.
- Automated the feature selection pipeline using Boruta, RFE and Null point importance and the Hyper-parameter tuning pipeline using Optuna.
- Managed and mentored junior data scientists in the team.

MateLabs (Data Scientist, Bangalore) **Mar'20 - Jun'22**

- Built an end-to-end AutoML based SaaS platform, developed the data preprocessing and training modules, which is now being used across industry in various domains.
- Worked on optimising supply chain for some of the biggest FMCG companies using machine learning. Created models for Demand planning and helped them understand the bottlenecks by model interpretability as a result saving up on their revenue.
- Developed a META learning algorithm that selects right model and hyper-parameters for a dataset, which was in future used in the AutoML platform.

SELECTED PROJECTS

Stable diffusion from scratch

https://github.com/rajgarg021/stable-diffusion-from-scratch

- PyTorch implementation of stable diffusion from scratch for both **text-to-image** and **image-to-image** generation.
- Implemented a **VAE** to map between pixel space and latent space, **CLIP Encoder** for generating embeddings from text prompt, **U-net** for the **reverse diffusion process** conditioned using **classifier-free guidance** and **DDPM Sampler** for removing noise.

Llama 3 from scratch

https://github.com/rajgarg021/llama3-from-scratch

- Clean implementation of **Llama 3** inference code from scratch with lots of comments explaining every step and every matrix shape change.
- Read research papers for all the building blocks - **Grouped-Query Attention with KV cache, Rotary Position Embeddings, RMSNorm, SwiGLU activation function** to gain a deeper understanding before implementing.

Receipt image to JSON

https://github.com/rajgarg021/image-to-json

- **Finetuned Google's PaliGemma** on a custom dataset where the goal for the model is to turn a receipt image into a JSON containing all fields.
- Used **Pytorch Lightning** for training the model and **Weights and Biases** for logging.
- Implemented **QLoRA (Quantization + Low Rank Adaption)** by freezing and quantizing the existing pre-trained weights and training only few adapter layers on top of the base model using **HuggingFace's PEFT** library and **BitsAndBytes**.

Agentic RAG Application with LangChain and LangGraph

https://github.com/rajgarg021/agentic-rag

- Engineered a cyclic, multi-node graph for **dynamic RAG**, integrating external APIs and LCEL for **efficient chain composition** and **async operations**.
- Implemented **state management** and **conditional routing** to enhance context retention and query processing in LLM applications.

Implemented various quantization techniques for neural networks

https://github.com/rajgarg021/quantization

- Implemented **affine and scale quantization** from scratch.
- Trained a simple neural network and implemented **post-training quantization** using PyTorch which resulted in ∼75% reduction of the model size.
- Improved the accuracy of the above model by using **quantization aware training**.

Video tracking and re-identification

https://github.com/rajgarg021/VideoReID

- Implemented a system for **detection, tracking** and **re-identifying** people across video frames, even if they temporarily leave and re-enter the frame or get occluded by other objects.
- Used **Yolov10** for object detection and created a **custom algorithm** for tracking and re-identifying.

Diffusion from scratch

https://github.com/rajgarg021/diffusion-from-scratch

- Implemented the **Denoising Diffusion Probabilistic Models** paper from scratch to obtain fundamental understanding about diffusion models.
- Added comprehensive notes about my learning in the repo readme file.

GPT from scratch

https://github.com/rajgarg021/gpt-from-scratch

- Implemented a **Decoder-Transformer network (10M parameters)** from scratch.
- Helped gain intuitive understanding of **self-attention**.
- Used optimization techniques like **residual connections** and **layer normaliztion**.

Experimented with various large language models

https://github.com/rajgarg021/LLM-experiments

- Experimented with various **open-source models** like **LLama2, Mistral, Mixtral 8x7B, Gemma**, etc to build a **RAG** application for generating answers given a textual context.
- Played around with **LangChain, OpenAI API (GPT-3.5-Turbo), Whisper** and **Pinecone** (to store embeddings) to transcribe a video and generate answers using that as context.

Teenytoken

https://github.com/rajgarg021/teenytoken

- Implemented **byte-level byte pair encoding** algorithm from scratch to create my own **tokenizer**.
- Working on improving it further to make it equivalent to the **GPT-4 tokenizer**.

Fine tuned BERT for sentiment analysis

https://github.com/rajgarg021/Fine-tuning-BERT-for-sentiment-analysis

- Used **HuggingFace's transformers** library to fine tune a **BERT** model for sentiment analysis task.

ACHIEVEMENTS

- Scored **98.89%** in CAT 2018 and converted several **IIM** interviews to admission offers.
- Received **scholarship** from **Bill & Melinda Gates Foundation** during college.