Analytics Vidhya

The DataHour on

Build a LLM Application with Databricks DBRX

Sarbani Maiti (Specialist Solutions Architect ML & GenAI at Databricks)

Table of contents

- Introduction to the Databricks DBRX
- DBRX MoE architecture
- Get started with DBRX
- Build a LLM application with DBRX using RAG
  - RAG architecture
  - Demo

- Conclusion
- Q & A

Brief about me..

- My name is Sarbani Maiti
- Specialist Solutions Architect ML & GenAI @ Databricks
- ~23 years of work experience
- My work areas are ML, GenAI, Scalable Machine Learning with Spark, MLOPS, LLMOPS
- Hobbies : Painting, Travelling

Analytics
Vidhya

DBRX - The State of the Art Open Source LLM developed by Databricks

Introduction to the Databricks DBRX

- DBRX is a transformer-based decoder-only large language model (LLM) with 132B total parameters, utilizing a fine-grained mixture-of-experts (MoE) architecture.
- Pre-trained on 12T tokens of text and code data, DBRX employs next-token prediction and a maximum context length of 32k tokens.
- Unique features include rotary position encodings (RoPE), gated linear units (GLU), and grouped query attention (GQA).

Introduction to the Databricks DBRX

- DBRX uses the GPT-4 tokenizer and benefits from a curated dataset, estimated to be at least 2x better token-for-token than previous data.
- Pretraining involves curriculum learning, adjusting the data mix during training for improved model quality.
- DBRX's fine-grained MoE architecture provides 65x more possible combinations of experts, enhancing model performance.

Databricks DBRX - How it is built

- Explored data using Lilac AI, then processed and cleaned it with Apache Spark™ and Databricks notebooks.
- Trained DBRX using optimized versions of open-source training libraries: MegaBlocks, LLM Foundry, Composer, and Streaming.
- Managed large scale model training and fine-tuning with Mosaic AI Training service across thousands of GPUs.
- Logged results using Unity Catalog, MLflow, collected human feedback for quality and safety improvements via Mosaic AI Model Serving and Inference Tables.
- Manually experimented with the model using the Databricks Playground
- Found Databricks tools to be best-in-class for their purposes, benefiting from a unified product experience.

DBRX MoE architecture

MoE architecture

- DBRX uses the GPT-4 tokenizer and benefits from a curated dataset, estimated to be at least 2x better token-for-token than previous data
- DBRX utilizes a fine-grained mixture-of-experts (MoE) architecture, which allows the model to scale to 132B total parameters.
- The MoE architecture consists of 16 experts, with 4 active experts selected per input.
- This fine-grained approach provides 65x more possible combinations of experts, enhancing model performance.

MoE architecture

- The MoE architecture enables efficient scaling by activating only a subset of the total parameters for each input.
- This results in a more computationally efficient model compared to traditional transformer architectures.
- The MoE architecture in DBRX improves the model's ability to specialize and handle a diverse range of tasks and inputs.
- DBRX's MoE architecture is designed to optimize resource usage and provide better performance for large language models.
-

Get Started with

Get started with DBRX

Github : https://github.com/databricks/dbrx/tree/main

Hugging Face Databricks Space

- https://huggingface.co/spaces/databricks/dbrx-instruct
- https://huggingface.co/databricks/dbrx-base
- https://huggingface.co/databricks/dbrx-instruct

Databricks AI Playground or Foundation Model API

- https://$instance$.databricks.com/ml/playground

You.com: https://you.com/

Perplexity.ai : https://labs.perplexity.ai/

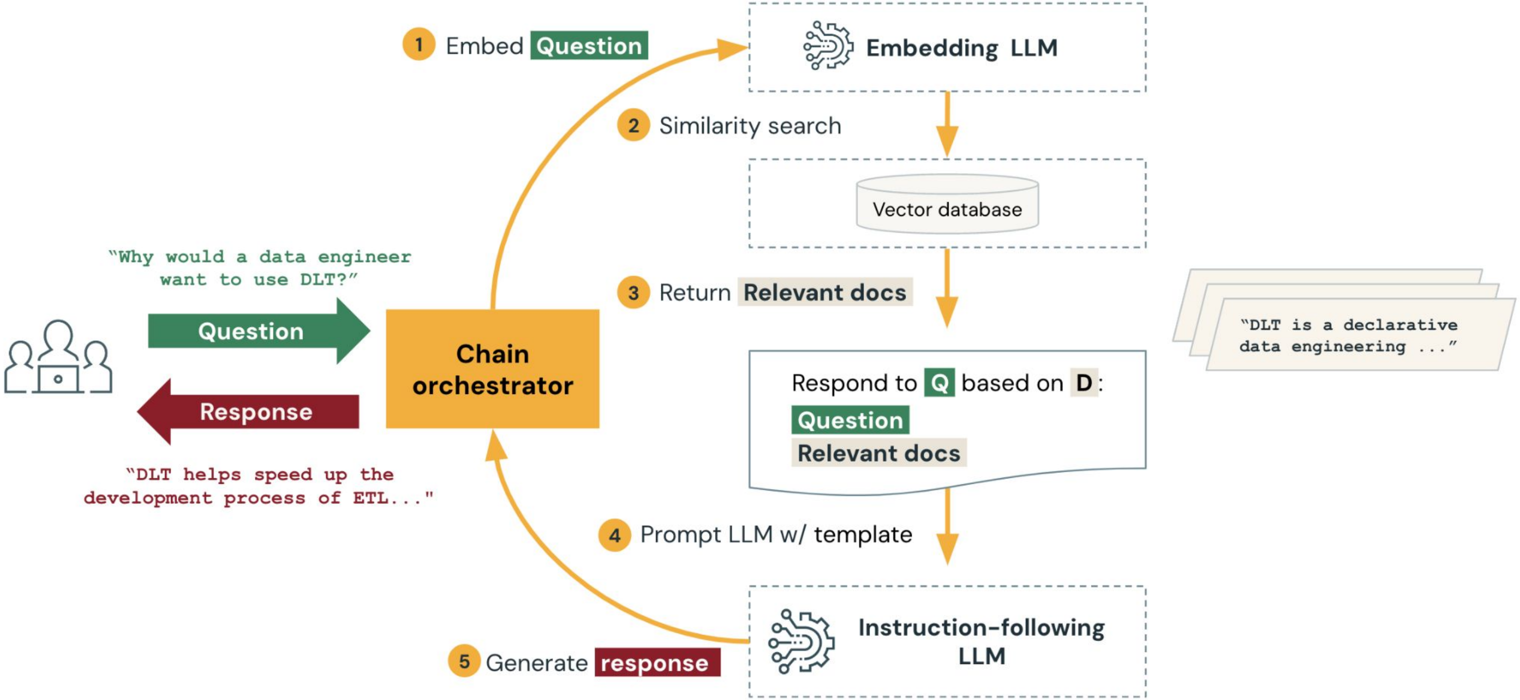
Build a LLM application with DBRX

How to build LLM application

- Prompt Engineering
- RAG
- Fine Tune
- Pre training

Today we will talk about RAG architecture and see how can we build RAG Chatbot using DBRX.

What is RAG


1 Embed Question
Embedding LLM
2 Similarity search
Vector database
3 Return Relevant docs
"DLT is a declarative data engineering ..."
"Why would a data engineer want to use DLT?"
Question
Response
Chain orchestrator
Respond to Q based on D:
Question
Relevant docs
"DLT helps speed up the development process of ETL..."
4 Prompt LLM w/ template
Instruction-following LLM
5 Generate response

Build a LLM application with DBRX & RAG - DEMO

LLM application with DBRX

DBRX: https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm

LLM application with DBRX with Databricks Vector Search

Production ready LLM application using DBRX with Databricks Vector Search for embeddings & Vector Index
And Custom model using Unity Catalog, MLFLOW & Databricks Model serving Endpoint

https://huggingface.co/spaces/databricks-demos/chatbot

Notebooks :
https://www.databricks.com/resources/demos/tutorials/data-science-and-ai/lakehouse-ai-deploy-your-llm-chatbot

```
 %pip install dbdemo

import dbdemos
  dbdemos.install('llm-rag-chatbot', catalog='main', schema='rag_chatbot')
```

LLM application with DBRX - Chroma db

Production ready LLM application using Chroma DB (LLM of you choice)

https://www.databricks.com/solutions/accelerators/biomedical-literature-qa-large-language-models-llms

Today’s Demo : https://github.com/sarbaniAi/databricks-dbrx

DataHour : Q&A

Analytics Vidhya

Analytics Vidhya

Thank you!