README.md

CUTIEPI

Overview

This module exposes an API using the TensorFlow model server. It will take the CUTIE model input in request and return model output in response.

Tech Stack

- `TensorFlow ModelServer` : It serves TensorFlow models over gRPC/HTTP and does things like model versioning or model server maintenance easily.
- `gRPC` : Uses HTTP/2 to support highly performant and scalable API's and makes use of binary data rather than just text which makes the communication more compact and more efficient.

How to run

- Run `pip install -r requrements.txt` to install required python libraries.
- Export Environment Variables using `source .env`
- Change latest model checkpoint to saved model using `ckpt_to_pb.py` .
- Start the model server using `tensorflow_model_server --port=8500 --rest_api_port=8501 --model_name=CUTIE --model_base_path="${MODEL_DIR}"`

Challenges Faced

- Converting ckpt file to saved model
- Calling api using gRPC client, had to provide model input and output
- Understanding protobuf