Here is a draft README file summarizing my approach for the English to French translation model fine-tuning and demo project:

English to French Translation Model

Model Details:

For this project, I fine-tuned the Hugging Face "Helsinki-NLP/opus-mt-en-fr" translation model. This is a MarianMT model originally trained on a large English-French parallel corpus.

To further adapt it to my use case, I fine-tuned the model on the opus_books English-French sentence pairs dataset.

Training Process

The training was performed using a Python script and leveraged a GPU runtime for faster training. The model was trained for 3 epochs, with a batch size of 8 sentence pairs.

The AdamW optimizer was used along with a linear learning rate warmup and decay schedule. I tracked validation loss during training to monitor for overfitting. This was done by tracking BLEU score to monitor translation quality

After training, the updated model state dict containing the tuned weights was saved locally to be used in the demo application.

Demo Application

I built an interactive web-based demo with Gradio that allows a user to input English text and see the French translation output from my fine-tuned model. The interface also includes example translations and the ability to customize parameters like temperature.

The app is designed to showcase the model's translation capabilities in a simple, user-friendly way.

Challenges

The main challenges I faced were around managing memory and compute limitations during the fine-tuning process, and optimizing the model size for deployment in the web app.

I overcame these by experimenting with batch sizes, checkpointing, and quantization techniques to reduce model size.

Overall, through this project I gained valuable hands-on experience in training and deploying NLP models to create an end-to-end machine learning pipeline tailored to a real-world application.