An English-to-Spanish Translator - Shan Ali Shah Sayed

1. Project Overview
This project presents a Transformer-based neural machine translation (NMT) system for English-to-Spanish translation. The objective was to build a fully functional model capable of real-time inference using an attention-based encoder-decoder architecture.

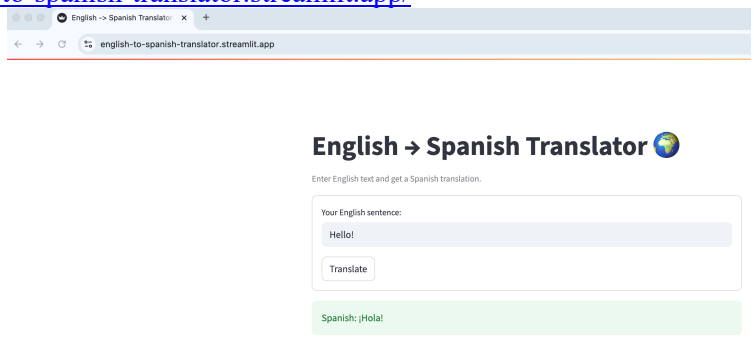
2. Model Architecture and Training
The model employs a Transformer structure with 4 layers, 8 attention heads, 128-dimensional embeddings, and a 512-dimensional feed-forward network. It supports sequences of 20 tokens and a vocabulary size of 15,000. The model was trained over 10 epochs on a parallel English-Spanish dataset. The training objective used masked sparse categorical cross-entropy, optimized with the Adam optimizer and a custom learning rate schedule. The final model achieved a validation masked accuracy of 69.89%, demonstrating effective learning of source-target alignments across varied sentence structures.

3. Preprocessing and Vectorization
Preprocessing was implemented using Keras's TextVectorization layer, alongside a custom standardization function registered with @register_keras_serializable() to ensure consistent token formatting. Source and target vectorizers were saved as .keras files to preserve vocabulary and sequence transformations for reproducibility during inference.

4. Inference and Decoding
Transformer model is reconstructed at inference using real vectorized inputs, ensuring that all layers and attention mechanisms are properly initialized. Translations are generated using greedy decoding, predicting one token at a time in an autoregressive manner. A dedicated Python script is provided for direct command-line translation, enabling real-time usage without requiring model retraining.

5. Deployment and External Integration
The system supports two modes of interaction. Translations can be executed programmatically via the included Python script using the saved model and vectorizers. Additionally, a web-based Streamlit application has been deployed to allow interactive usage through a browser interface. Both modes operate using the same pre-trained model, with weights and vectorizer files dynamically loaded from Google Drive using the gdown library.

Link: https://english-to-spanish-translator.streamlit.app/



6. Conclusion
This project demonstrates a complete application of Transformer architecture to neural machine translation. The system combines attention-based sequence modeling, preprocessing robustness, and reproducible deployment, achieving a validation accuracy of ~70% while supporting real-time translation access.