Introduction

This report details the development of a machine learning model using the Transformer architecture on the Persian Wikipedia dataset. Transformers have significantly advanced the field of natural language processing (NLP) by providing a robust framework for understanding and generating human language. This project aims to utilize these capabilities for processing Persian text data.

Problem Statement

The primary objective of this project is to build and train a Transformer model on the Persian Wikipedia dataset. The goal is to develop a language model capable of understanding and generating text in Persian. This involves preprocessing the dataset, defining an appropriate model architecture, and training the model to achieve high performance on NLP tasks.

Data Loading and Preprocessing

The Persian Wikipedia dataset is sourced from Kaggle and includes extensive text data in Persian. The preprocessing steps include:

1. Downloading the Dataset: The dataset is fetched from Kaggle.
2. Extracting the Dataset: The dataset, typically in a compressed format, is extracted to make the text files accessible.
3. Reading the Data: The text data is read into the environment for further processing.
4. Text Preprocessing: This involves several steps:
   - Tokenization: Breaking down the text into individual tokens or subwords.
   - Normalization: Standardizing the text to ensure consistency.
   - Encoding: Converting the text into numerical representations suitable for model training.

Design Model Architecture

The model architecture is based on the Transformer framework, known for its encoder-decoder structure. The design includes:

1. Tokenization and Encoding:
   - The text is tokenized into subwords, ensuring that the model can handle rare and common words effectively.
   - Tokens are then converted into numerical encodings for model processing.

2. Model Definition:

- Embedding Layers: These layers transform token encodings into dense vectors of fixed size.
- Multi-Head Attention Mechanisms: These mechanisms allow the model to focus on different parts of the input sequence simultaneously, capturing various aspects of the text.
- Feed-Forward Neural Networks: Positioned after the attention layers, these networks process the attended information.
- Normalization Layers: These layers help stabilize and accelerate the training process by normalizing the inputs to each layer.
- Positional Encodings: These encodings provide information about the position of tokens within the sequence, crucial for the model to understand the order of words.

3. Training:
- The model is compiled with an appropriate optimizer and loss function.
- Training is conducted on the preprocessed dataset, with performance metrics such as loss and accuracy monitored throughout the process.

4. Evaluation:
- The model's performance is evaluated on a validation set to ensure it generalizes well to new, unseen data.
- Fine-tuning is performed if necessary, adjusting hyperparameters and employing techniques like regularization to improve performance.

Conclusion

This project successfully demonstrates the use of a Transformer model on the Persian Wikipedia dataset. Through systematic data preprocessing and the application of a robust Transformer architecture, the model effectively handles Persian text for language modeling tasks. This approach highlights the versatility and strength of Transformers in NLP, providing a foundation for further applications and extensions to other datasets and languages.