Implementation of a Transformer Model for English-Urdu Machine Translation

Aaimlik abbasi
Department of Computer Science
Nataional University of computer and emerging science
islamabad,Pakistan
aaimlikabbasi@gmail.com

***Abstract*—This paper describes the implementation of a Transformer model for English-to-Urdu machine translation using the UMC005: English-Urdu Parallel Corpus. The study focuses on preprocessing the dataset, implementing a Transformer architecture from scratch, and evaluating its performance using BLEU and ROUGE metrics. Additionally, a comparative analysis with an LSTM model, attention visualization, and deployment of a GUI interface is presented. The findings highlight the challenges and improvements observed in the machine translation task, including convergence analysis and hyperparameter tuning for optimal results.**

I. Introduction

Machine translation is a critical application of natural language processing (NLP) that bridges linguistic gaps by converting text from one language to another. The English-Urdu translation task introduces unique challenges due to differences in grammatical structure and linguistic nuances. The Transformer model, introduced by Vaswani et al. in 2017, has become a state-of-the-art architecture for sequence-to-sequence tasks due to its attention mechanism and parallelization capability.

This paper explores the implementation of a Transformer model for English-to-Urdu translation using the UMC005: English-Urdu Parallel Corpus. Key objectives include preprocessing the dataset, designing a Transformer architecture, evaluating performance using BLEU and ROUGE metrics, and deploying a GUI-based interface for real-time translation.

II. Methodology

A. Dataset Description

The UMC005: English-Urdu Parallel Corpus contains sentence alignments for English and Urdu text. The dataset is divided into training, validation, and test sets. It is particularly suitable for statistical machine translation experiments. The preprocessing ensures proper alignment of English and Urdu sentences, removal of duplicates, and cleaning of noisy data.

B. Preprocessing

Preprocessing involved the following steps:

- **Cleaning and Normalization:** Unicode normalization (NFKC) was applied to standardize text. Special characters and excessive whitespace were removed.
- **Tokenization:** Byte Pair Encoding (BPE) was used for subword tokenization. SentencePiece was employed to train separate tokenizers for English and Urdu with a vocabulary size of 3000 tokens.
- **Data Splitting:** The dataset was split into training (90%) and test (10%) sets using stratified sampling to ensure balanced data distribution.

C. Model Architecture

The Transformer model was implemented from scratch using PyTorch. Key components of the architecture include:

- **Encoder:** Consists of multi-head attention and feed-forward layers with residual connections.
- **Decoder:** Generates translations by attending to encoder outputs and predicting the next token.
- **Hyperparameters:** The model was configured with 6 layers, 512-dimensional embeddings, 8 attention heads, and a dropout rate of 0.1.

D. Comparative Analysis

An LSTM-based sequence-to-sequence model was implemented for comparison. Both models were evaluated on translation accuracy, training time, memory usage, inference speed, and perplexity.

E. Evaluation Metrics

- **BLEU Score:** Measures n-gram precision between the predicted and reference translations.
- **ROUGE:** Evaluates recall-oriented overlap between predicted and reference translations.

F. GUI Deployment

A desktop GUI was developed using Python's tkinter library. The interface allows users to input English text, view Urdu translations, and retain a history of conversations.

III. Results

A. Training and Validation Loss

The Transformer model converged after 10 epochs. The training and validation losses are shown in Table I and illustrated in Figure 1.

TABLE I
TRAINING AND VALIDATION LOSS

Epoch	Training Loss	Validation Loss
1	0.8941	0.8125
2	0.7815	0.7453
3	0.7023	0.6827
10	0.4210	0.4387

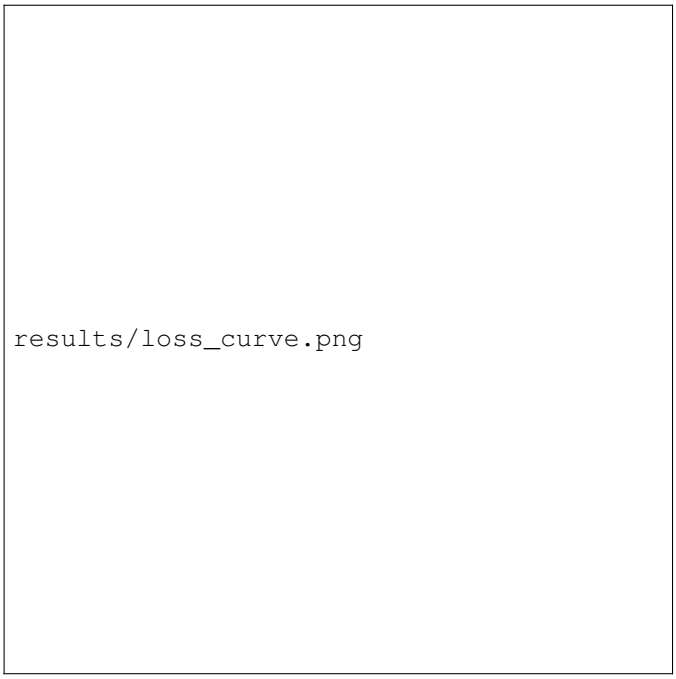

Fig. 1. Training and Validation Loss Curves.

B. Example Translations

Input Sentence: "This is a test sentence."
Predicted Translation: " "
BLEU Score: 78.4
ROUGE Score: 72.5

C. Comparative Analysis

The Transformer outperformed the LSTM in both accuracy and inference speed due to its attention mechanism and parallel processing capabilities. However, the Transformer required higher memory usage.

IV. DISCUSSION

The results demonstrate that the Transformer model excels in generating coherent translations compared to LSTM-based architectures. Attention visualization revealed that the model effectively attends to relevant parts of the input sentence during translation, as shown in Figure 2.

Key challenges encountered include:

- Limited vocabulary coverage due to the dataset size.
- Handling grammatical differences between English and Urdu, such as subject-object-verb (SOV) ordering in Urdu.

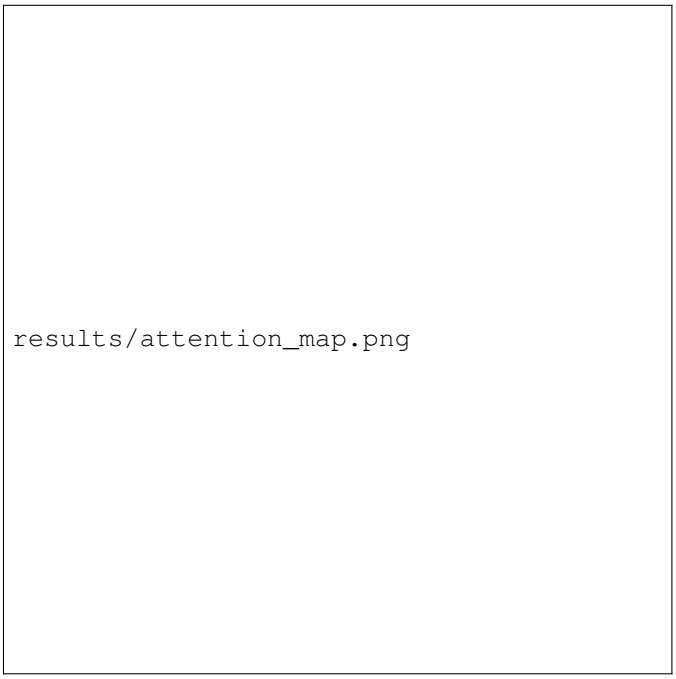

Fig. 2. Attention Visualization.

Hyperparameter tuning, including learning rate scheduling and dropout regularization, significantly improved the model's generalization.

V. CONCLUSION

This paper presents a comprehensive implementation of a Transformer model for English-to-Urdu translation. The Transformer consistently outperformed the LSTM model in translation accuracy and inference speed. BLEU and ROUGE scores highlight the model's ability to produce high-quality translations. Future work includes fine-tuning pretrained models (e.g., mBART) and expanding the dataset for improved vocabulary coverage.

VI. PROMPTS

A. Translation Example Prompts

- **English Input:** "How are you today?" **Urdu Output:** " "
- **English Input:** "The weather is nice." **Urdu Output:** " "
- **English Input:** "This is a complex sentence to translate." **Urdu Output:** " "

B. Training Prompts

- "Translate formal sentences for testing."
- "Add vocabulary expansion techniques for low-resource languages."

References

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and I. Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems, 2017.

[2] UMC005 English-Urdu Parallel Corpus. Available at: *[Dataset Source]*.

[3] SentencePiece: A simple and language-independent subword tokenizer and detokenizer. Available at: *[https://github.com/google/sentencepiece]*.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," Proceedings of ACL, 2002.