STATS 507: Final Project Report

Zhijie Qiao
Department of Civil and Environmental Engineering
University of Michigan
Ann Arbor, USA, 48109
zhijieq@umich.edu

***Abstract*—The transformer architecture has revolutionized generative AI and reshaped numerous domains within machine learning. Drawing inspiration from the groundbreaking successes of transformer-powered tools like OpenAI's GPT and Google's Gemini, this project focuses on developing a transformer-based English-to-Chinese translator. The model was trained on the English-Chinese section of the OPUS-100 dataset on Hugging Face, utilizing a single GPU over approximately 12 hours. Evaluation using BERT scoring yielded a Precision score of 0.795, a Recall score of 0.766, and an F1 score of 0.779. These results demonstrate the model's effectiveness and highlight its potential for further advancements in multilingual translation and AI-driven language processing. The source code for the project is available at https://github.com/QZJGeorge/STATS507-Final-Project.**

***Index Terms*—natural language processing, transformer architecture, open dataset**

I. Introduction

Generative AI has transformed diverse domains, enabling remarkable applications like high-quality image creation, coherent text generation, natural language translation, and interactive conversations. At the heart of this revolution lies the transformer architecture, first introduced in the seminal paper ”Attention Is All You Need” [1]. By replacing traditional recurrence neural networks [2], [3] and convolutional layers [4], [5] with attention mechanisms, transformers redefined natural language processing (NLP), delivering unprecedented efficiency and scalability. Building on this foundation, BERT (Bidirectional Encoder Representations from Transformers) [6] showcased the effectiveness of bidirectional pre-training for contextual understanding, achieving state-of-the-art performance in tasks such as question answering and language inference. T5 (Text-to-Text Transfer Transformer) [7] introduced a unified text-to-text framework, enabling a flexible and scalable approach to NLP tasks. Further innovations, such as Sparse Transformers [8], which mitigated the quadratic cost of self-attention through sparse patterns, and Reformer [9], which utilized locality-sensitive hashing to efficiently handle longer sequences, significantly expanded the transformative potential of these models.

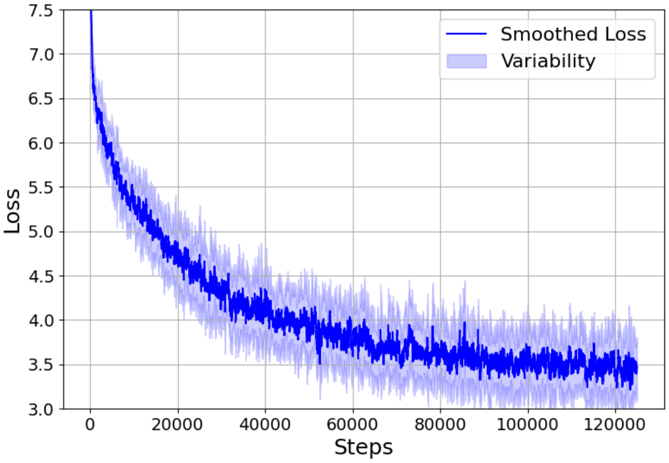

Fig. 1: Training Loss with Respect to Epochs

Transformers have also made significant strides beyond text. For example, the Vision Transformer (ViT) [10] applied attention mechanisms to image classification, achieving state-of-the-art results. These breakthroughs paved the way for large-scale models such as OpenAI's GPT and Google's Gemini, which harness vast datasets and advanced computational capabilities to push the boundaries of AI, shaping the future of innovation across industries.

Inspired by these advancements, this project implements a seminal transformer architecture, for an English-to-Chinese translation task to gain a deeper understanding of its principles and applications. We utilize the open-source OPUS-100 dataset from Hugging Face [11], [12], which provides one million sentence pairs for training, 2,000 for validation, and 2,000 for testing.

II. Method

A. Model Design

The overall structure of the transformer model follows the design in [1]. The main coding implementation, including the encoder and decoder block design, residual connections, and projection layers, was adapted from a tutorial at [13], which implements a transformer for an English-to-Italian translation task.

B. Tokenization

Building on that, several modifications were made to the tokenization process. For the English portion, tokenization adheres to the conventional word-level approach, where words are separated by whitespace. For Chinese text, we explored a more advanced tokenization method.

Source Text	**Target Text**	**Prediction**	**Prediction Translation**
The report deals extensively with the specific difficulties confronted by landlocked developing countries.	该报告对内陆发展中国家所面临的各种具体困难做了广泛的分析。	报告深入地处理内陆发展中国家面临的具体困难。	The report deals in depth with the specific difficulties faced by landlocked developing countries.
Women working in the informal and agricultural sectors were particularly badly off.	在非正规部门和农业部门工作的妇女尤其贫困。	在非正规和农业部门工作的妇女特别被。	Women working in the informal and agricultural sectors are particularly.
Tell you what. I can give you a ride to Belle's, a diner five miles down.	这样吧，我载你们到5英里外的贝拉餐厅	告诉你我可以给你5英里	Tell you I can give you 5 miles.
Some paternal wisdom. How about that?	来点父辈的智慧，怎么样?	的，怎么了?	Yes, what's wrong?

TABLE I: Sample Translations from the Model

Instead of using a character-level approach that predicts one character at a time, we employed a pre-segmentation strategy using the Python library *Jieba*, which segments Chinese sentences into words. These segmented words are then tokenized and used during both training and testing. We believe this approach improves model accuracy by capturing meaningful linguistic units more effectively. For both languages, the minimum word frequency required to create a token is set to 10, reducing noise and minimizing the impact of typos.

More advanced pre-trained tokenization datasets can improve model performance with optimized vocabularies and strategies fine-tuned on diverse corpora. However, in this project, we stick to the process described above to gain some hands-on experience on tokenization and better understand the model implementation details.

III. Results

A. Training

The training dataset comprises 1 million samples. A batch size of 16 is used, resulting in 62,500 steps per epoch. The maximum sequence length is set to 256, with sentences longer than this being padded and excluded from the training process. The model's embedding dimension is set to 512, and the multi-head attention mechanism uses 8 heads. For simplicity, the learning rate is fixed at 10^{-4} and remains constant throughout the training. Loss function is cross entropy loss based on the softmax probability of the token predicted against the target token, with label smoothing equal to 0.1.

The training was conducted on an NVIDIA 3070 GPU with 8GB of RAM for 2 epochs. Each epoch took approximately 6 hours, amounting to a total training time of 12 hours. After 2 epoches, the training loss converged and stabilized as shown in Fig. 1.

B. Evaluation

Due to the characteristics of the Chinese language, popular evaluation metrics such as BLEU may not fit well. Instead, we adopt the BERTScore metric in [14] to evaluate our model's performance. This metric compares model predictions with ground truth sentences, summing and averaging the scores over all samples. For simplicity, we combine the validation and test datasets, resulting in a total of 4,000 samples. Our model achieved a Precision score of 0.795, Recall score of 0.766, and F1 score of 0.779, indicating that the model yields reasonably good translation performance.

Table I lists some outputs from the model to provide a qualitative representation. Four examples are selected from the test dataset to represent varying levels of translation accuracy. The first column in the table shows the original English sentence, the second column contains the target Chinese sentence, the third column presents the model's predicted Chinese sentence, and the fourth column includes the Google-translated version of the predicted Chinese sentence for reference. Row 1 demonstrates a high-quality translation that captures all key details accurately. Row 2 shows a translation that is mostly correct but incomplete or missing details. Row 3 contains a translation that is somewhat similar to the original sentence but conveys an incorrect meaning. Row 4 depicts a translation that is entirely off-target. From visual inspection, most translations fall into the second category, representing translations that are neither entirely accurate nor completely incorrect, although more rigorous analysis could be used. This suggests that while the model performs adequately, there is room for improvement, particularly in achieving higher completeness and accuracy. Due to time and resource constraints, this project did not explore different parameter settings for the transformer model, which is a potential area for future improvement.

IV. Conclusion

In this project, we developed and trained a transformer-based model for English-to-Chinese translation using the OPUS-100 dataset from Hugging Face for training and evaluation. The model achieved promising performance, though there is still room for enhancement. Future efforts will focus on optimizing the tokenization process, refining dataset preprocessing to improve data quality, and investigating alternative transformer architecture variants to further improve performance.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[3] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, “Recurrent neural network based language model,” in *Interspeech*, 2010, pp. 1045–1048.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 34, no. 3, pp. 770–778, 2016.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[8] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating long sequences with sparse transformers,” *CoRR*, vol. abs/1904.10509, 2019.

[9] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” *CoRR*, vol. abs/2001.04451, 2020.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.

[11] J. Tiedemann, “Parallel data, tools and interfaces in OPUS,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218.

[12] B. Zhang, P. Williams, I. Titov, and R. Sennrich, “Improving massively multilingual neural machine translation and zero-shot translation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 1628–1639.

[13] U. Jamil, “Coding a transformer from scratch on pytorch, with full explanation, training and inference,” 2023.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” *CoRR*, vol. abs/1904.09675, 2019.