A Technical Approach to the Preservation and Translation of Jeju Language: Application of Transformer Architecture and LoRA.

Yun Ji-yeong

Email: yunjiyeong0106@gmail.com

***Abstract*—This paper proposes an approach utilizing the Transformer architecture and Low-Rank Adaptation (LoRA) technique for the translation and preservation of the Jeju language. The Jeju language, retaining vocabulary from Middle Korean and the original forms of Hangul from the creation of Hunminjeongeum, possesses significant academic and cultural value. However, it is classified as one of the 'endangered languages' by UNESCO, making the technological approach to its preservation and translation a crucial challenge.**

In this study, we aimed to enhance both the efficiency and translation performance of the model based on the Transformer architecture by employing the LoRA technique, utilizing the JIT Dataset.

The model applying LoRA was experimented with under various settings (rank and alpha values), and further experiments were conducted with the settings that showed the best performance. Through this approach, this research aims to develop an effective translation model that considers the unique vocabulary, grammar, and cultural background of the Jeju language, thereby promoting the preservation and utilization of the language.

This paper introduces the linguistic and cultural characteristics of the Jeju language, discusses the application of the Transformer architecture and LoRA technique for translating Jeju language, and conducts a comparative analysis between the basic Transformer architecture and the model with LoRA applied. Furthermore, it explores the limitations of the research and the potential of technological approaches.

Keywords: transformer, LoRA, Machine translation

I. Introduction

The Jeju language, unique to the Jeju region, holds high academic and cultural value as it preserves vocabulary from Middle Korean and the original forms of Hangul created during the development of Hunminjeongeum. Classified by UNESCO into one of the five stages of 'endangered languages', the preservation of Jeju language, which embodies the unique culture and history of Jeju, is of utmost importance. However, the specificity of the Jeju language and limited datasets pose significant challenges to its translation and preservation using modern technology.

This study aims to develop a model utilizing the latest technology, the Transformer architecture, and applying the Low-Rank Adaptation (LoRA) technique for the preservation and efficient translation of the Jeju language. The Transformer has shown excellent performance in the field of natural language processing and is widely used in various language translation tasks. However, for languages with limited datasets like Jeju language, achieving sufficient performance with existing Transformer models alone is challenging. Therefore, this research introduces the LoRA technique to simultaneously enhance the model's efficiency and translation performance.

The task of translating the Jeju language seeks to overcome the limitations exhibited by existing large-scale models. Most models are optimized for standard languages or major languages, with insufficient consideration for minority languages like the Jeju language. For instance, existing models' attempts to translate Jeju expressions like "폭삭 속암수다" into "Talking loudly and secretly" fail to fully capture the subtle nuances and cultural context of the Jeju language. Recognizing these limitations, this study aims to develop an effective translation model that considers the unique vocabulary, grammar, and cultural background of the Jeju language, thereby promoting its preservation and utilization.

The paper is organized as follows: It begins by introducing the linguistic and cultural characteristics of the Jeju language and explaining why its preservation is important. Then, it discusses the concepts of the Transformer architecture and the LoRA technique, and how these will be applied to translate the Jeju language. Finally, the proposed model's experimental results are presented to demonstrate performance improvements over existing models, exploring the potential of a technical approach for the translation and preservation of the Jeju language.

II. Related Work

This section reviews existing research on Jeju dialect translation and studies on LoRA networks.

A. Dialect Translation Research

Amid growing research interest in low-resource languages (Zoph et al., 2016; Gu et al., 2018)[8], new challenges are being posed to existing deep learning approaches. These challenges are unique due to the limited availability of data and the various grammatical and semantic issues that low-resource languages present, which are not addressed by systems optimized for mainstream languages such as English, German, and Arabic (Bender, 2019). Research related to the Jeju language tackles these challenges, motivated by the unique characteristics of the Jeju language and the lack of computational resources for it. Preliminary work has focused on constructing

a machine-readable Jeju-Korean parallel corpus and a clean Jeju language single speaker voice dataset. This is mentioned in the paper "Jejueo Datasets for Machine Translation and Speech Synthesis," emphasizing the necessity and progress of such research in the field.

B. LoRA Networks

The concept of LoRA networks was first introduced in the paper "LoRA: Low-Rank Adaptation for Neural Language Models," which describes a methodology for improving model performance by applying low-rank adaptation matrices to the attention layers of Transformer models [1].

Specifically, Low-Rank Adaptation (LoRA) is proposed to enhance the adaptability of the model without directly modifying the parameters of the pre-trained model. When fine-tuning pre-trained models, LoRA helps to improve model performance without additional computational burden. LoRA increases the flexibility of the model during the fine-tuning process by introducing small adapters to specific parts of the model (e.g., attention modules or feedforward networks) and introducing additional training parameters. In our task, it operates by adding low-rank adapters to the feedforward networks. These adapters assist the model in better adapting its representational capacity to new data or tasks. In other words, using LoRA allows the model to learn a broader range of functions within the existing parameter structure. Although our task may not involve a pre-trained model, the application of LoRA is still beneficial, particularly when the dataset is limited. It can prevent overfitting of the model and contribute to maintaining or improving the model's expressiveness while keeping the number of training parameters limited.

III. Methodology

A. SentencePiece

SentencePiece is one of the processing techniques used for tokenizing text data. It is designed to learn subword units from raw text data without any prior knowledge of the language. SentencePiece is based on Byte Pair Encoding (BPE) and the Unigram language model, and it tokenizes by sequentially merging the most frequently occurring pairs of characters in the text data or optimizing the set of subwords based on the probability of tokens[2].

B. Transformer

The Transformer model is a deep learning architecture primarily used in the field of natural language processing (NLP). Its key components include the self-attention mechanism and positional encoding, which allow each word to learn its relationship with all other words in the sentence.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

The self-attention mechanism in the Transformer model allows each word in a sentence to attend to all other words in the same sentence. The attention function computes a weighted sum of values V, where the weight assigned to each value is computed by a compatibility function of the query Q with the corresponding key K. Self-attention mechanism enables the model to dynamically weigh the influence of each word in a sentence when constructing the contextual representation of every other word.[8] This is particularly powerful as it allows the model to capture a wide range of dependencies between words, irrespective of their positional distance from each other within the sentence.

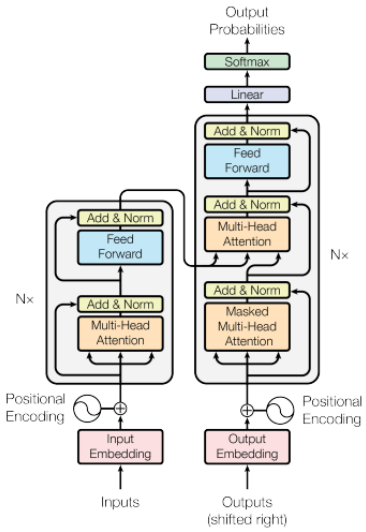

Fig. 1: Figure 1.Transformer architecture[8]

C. LoRA

Low-Rank Adaptation (LoRA) is applied within the Transformer architecture to improve the model's adaptability and expressiveness without significantly increasing the number of parameters. LoRA achieves this by adding low-rank adapters to the weight matrix of specific layers.

$$\tilde{W} = W + BA,$$

$\tilde{W}$ is the adapted weight matrix after applying the LoRA modification. It is the new weight matrix that will be used in the Transformer layer.

W represents the original weight matrix before adaptation. It contains the pre-trained weights of the model, which are learned from the data during the training phase[3]. By adding BA to W, LoRA achieves an adapted weight matrix $\tilde{W}$ that incorporates new knowledge or adaptations while keeping the increase in the number of parameters minimal. This allows the model to adapt to new tasks or fine-tune on specific datasets without the need to retrain all the parameters, thus maintaining the original model's strengths while extending its capabilities. The primary role of LoRA is to enhance the model's adaptability and expressiveness, enabling it to learn new tasks or adapt to new data more effectively without a significant increase in computational resources or parameters.

IV. Experiment

A. Dataset

In this study, we utilized the JIT Dataset, which consists of pairs of Standard Korean and Jeju dialect sentences. The dataset contains a total of 113519 sentence pairs[9].

B. Data Preprocessing

The data preprocessing process includes the following steps:

- **Duplicate Removal:** Removing duplicate sentence pairs from the dataset to ensure data diversity.
- **Dictionary Construction:** Creating a dictionary of 10,000 words using Jeju Special Self-Governing Province dialect information[6].
- **Sentence Transformation:** Transforming Jeju dialect sentences to words included in the dictionary to maintain consistency.
- **Special Character Removal:** Removing unnecessary special characters from sentences to refine the data.
- **Tokenization:** Using SentencePiece to tokenize with a vocab size of 20,000.
- **Sentence Length Filtering:** Removing sentences with a length of 50 or less to improve data quality.
- **Padding:** Padding all sentences to a consistent length of maxlen=40.sentence pairs.

C. Model

The experiment used a basic Transformer model, with the following configuration:

Configuration	Value
Vocabulary Size	20000
Maximum Length of Positional Encoding	40
Optimizer	Adam
Batch Size	64
Number of Epochs	10
Learning Rate	LearningRateScheduler(512)

TABLE I: Transformer model configuration.

When LoRA was applied to the feedforward network, the experiment was conducted under the following conditions:

Rank	Alpha
16	0.1
32	0.1
64	0.1
128	0.1

TABLE II: LoRA configurations for feedforward network experiments.

Further experiments were conducted based on the setting that showed the best results (rank = 128, alpha = 0.1):

Rank	Alpha
64	0.1
64	0.01
64	0.001

TABLE III: Further LoRA configurations based on best results.

Experiments were performed under equal conditions for each setting, and the impact of changes in LoRA's rank and alpha values on model performance was observed. This allowed for an evaluation of the effect of LoRA on the expressiveness and adaptability of the Transformer model.

V. Result

A. Comparison of LoRA and baseline

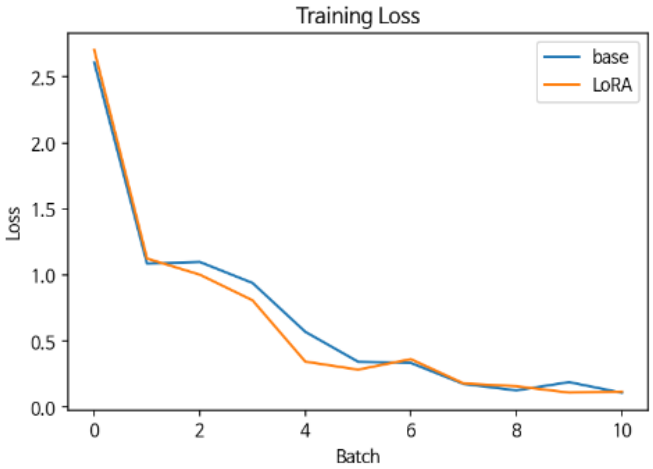

Fig. 2: Training loss for different ranks of LoRA.

The provided graph compares the training loss of a standard Transformer model (base) and a Transformer model with Low-Rank Adaptation (LoRA) applied to its feedforward layers over several batches of training. Initially, both models start with a comparably high loss, which is typical at the beginning of training as the models have not yet learned to fit the data.As the batches progress, we see a rapid decline in training loss for both models. This indicates that both models are learning and effectively reducing the error in their predictions. Throughout the training process, the LoRA model tends to have a slightly lower loss compared to the base model, suggesting that the LoRA model might be capturing and learning the patterns in the data more efficiently. The convergence of the loss for both models appears to be quite similar towards the end of the plotted batches, indicating that both models are reaching a point of optimization where improvements become incremental.When considering the actual translated outputs provided:

- Base Model Translation: "무사 경헷수과 속"
- LoRA Model Translation: "무사 경 속상헤붑서"

Although the graph shows only a slight difference in training loss, which might not indicate a substantial improvement in isolation, the predicted translations reveal a more nuanced improvement. The LoRA model produces a translation that seems to be more natural or correct from a linguistic perspective. This suggests that LoRA, despite the small numerical difference in training loss, is providing qualitative benefits to the model's translation capabilities.

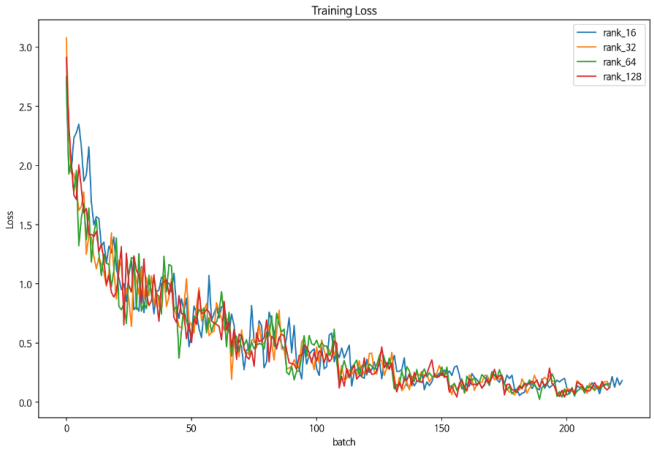

Fig. 3: Training loss for different ranks of LoRA.

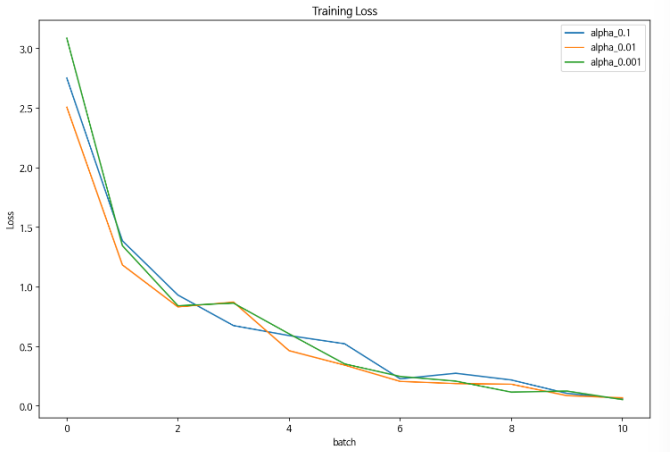

Fig. 4: Training loss for different alpha of LoRA.

B. Comparison of LoRA ranks

This graph displays the training loss according to the rank hyperparameter of Low-Rank Adaptation (LoRA). Here, the rank refers to the rank of the low-rank approximation applied to the weight matrix of the neural network during training. A higher rank value allows for more parameters to be adjusted, thereby increasing the model's capacity. The 'rank-16', 'rank-32', 'rank-64', 'rank-128' depicted on the graph represent different rank values, and the trend of training loss for each can be compared.

From the graph, it is evident that all rank settings show a tendency for the training loss to decrease, with higher rank values showing a more rapid decrease in loss initially. However, as training progresses, the benefit of loss reduction becomes less clear, even for higher rank values. Rank 128 shows the fastest initial loss reduction, but over time, it converges to similar loss values as the other ranks. Eventually, the rank-64 configuration exhibits the lowest loss, albeit marginally.

These results suggest that while a higher rank value may lead to faster initial learning speeds, ultimately all settings reach similar levels of training loss. Therefore, selecting the optimal rank value is crucial considering the model's complexity and computational costs. Additionally, in this experiment, the alpha value was fixed at 0.1 for all rank settings, so the impact of varying alpha values on the outcomes was not evaluated.

C. Comparison of LoRA alpha

In the provided graph, we observe the training loss trajectories for different alpha values in a Low-Rank Adaptation (LoRA) setting. Alpha is a hyperparameter in the LoRA technique that determines the learning rate of the low-rank adapters. A smaller alpha value leads to more conservative updates during training. - The model with alpha set to 0.001 (green line) shows the steepest decline in training loss initially, suggesting that a smaller alpha may allow for rapid improvements early in the training process. - However, as training progresses, the model with alpha set to 0.01 (orange line) maintains a lower loss compared to the others across most batches. This indicates that while an even smaller alpha can initiate a quick learning response, a slightly larger alpha (0.01 in this case) may provide a better balance between rapid convergence and stability. Towards the later stages of training, all models, irrespective of the alpha value, tend to converge to a similar loss value. This convergence implies that while different alpha settings may affect the speed and trajectory of learning, they may not significantly impact the final performance of the model once it has had sufficient training.As the training progressed, all models appeared to converge towards similar loss values, highlighting that the choice of alpha impacts the learning trajectory rather than the ultimate performance of the model. These findings underscore the importance of tuning the alpha parameter to achieve efficient training without compromising the final model quality.

VI. Conclusion

Throughout this study, we have pursued a technical approach to the preservation of the Jeju language, specifically by undertaking a Standard Korean-to-Jeju translation task. We applied the recently developed powerful method, Low-Rank Adaptation (LoRA), to the commonly used Transformer in machine translation, aiming to ensure the task could be well generalized.

In conclusion, the investigation into the application of LoRA to Transformer models for the translation and preservation of the Jeju language has yielded insightful outcomes. Meticulous experimentation with various LoRA rank configurations has demonstrated that LoRA can enhance the adaptability and expressiveness of Transformer models without a significant increase in the number of parameters. According to the research, despite initial differences in learning rates and convergence patterns, all tested ranks of LoRA adaptation led to similar levels of performance by the end of the training process. Ultimately, the results of this study not only advocate for the technical feasibility of preserving linguistic heritage through AI and machine learning but also lay the groundwork

for future research in this domain. The implications of this research are far-reaching and may potentially influence the way we approach language preservation and the development of adaptive language models in the future.

VII. Discussion

This study has clearly identified the limitations in processing the Jeju language. One of the primary limitations revealed is the need for specialized approaches and further research due to its unique language structure, which differs from the standard language. This challenge is particularly pronounced for minority languages like Jeju, which have limited datasets and are gradually disappearing.

The lack of data acts as a fundamental obstacle to problem-solving. While linguistic analysis and understanding are important, the potential for generalization can be greatly enhanced with the availability of sufficient data. To address this issue, exploring the possibilities of augmenting existing datasets could be a promising approach. The limitations in Jeju language processing presented through this study emphasize the necessity of a technical approach in Jeju language research, which holds the potential to contribute to the creation of datasets and the advancement of research in the future.

VIII. References

1) Efficient Fine-Tuning of Large Language Models with LoRA Technology. (2024). Retrieved from https://news.hada.io/topic?id=13004
2) SentencePiece. (2024). Retrieved from https://github.com/google/sentencepiece
3) LoRA: Low-Rank Adaptation of Large Language Models. [Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)].
4) Hossam, R. (2023). Comprehensive Guide to Adapters, LoRA, QLoRA, LongLoRA with implementation. Retrieved from https://medium.com/@raniahossam/comprehensive-guide-to-adapters-lora-qlora-longlora-with-implementation-de003be30352
5) Investing 16 billion won over the next 5 years for the preservation and cultivation of the Jeju language. (2023). Retrieved from http://www.jejudomin.co.kr/news/articleView.html?idxno=
6) Jeju Special Self-Governing Province Dialect Dictionary Information. (2022). Retrieved from https://www.data.go.kr/data/15058412/openapi.do
7) Park, K., Choe, Y. J., & Ham, J. (2021). Jejueo Datasets for Machine Translation and Speech Synthesis. [Submitted on 17 Jun 2021 (v1), last revised 16 Oct 2021 (this version, v2)].
8) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Retrieved from https://arxiv.org/abs/1706.03762
9) Park, B. (2014). Why is it essential to preserve the Jeju language? Retrieved from https://www.kaggle.com/datasets/bryanpark/jit-dataset/data

References