LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR

Zheshu Song[1], *Jianheng Zhuo*[1], *Yifan Yang*[1], *Ziyang Ma*[1], *Shixiong Zhang*[2], *Xie Chen*[1,†]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, China
[2]Tencent AI Lab, USA

{songzheshu, chenxie95}@sjtu.edu.cn

Abstract

Recent years have witnessed significant progress in multilingual automatic speech recognition (ASR), driven by the emergence of end-to-end (E2E) models and the scaling of multilingual datasets. Despite that, two main challenges persist in multilingual ASR: language interference and the incorporation of new languages without degrading the performance of the existing ones. This paper proposes LoRA-Whisper, which incorporates LoRA matrix into Whisper for multilingual ASR, effectively mitigating language interference. Furthermore, by leveraging LoRA and the similarities between languages, we can achieve better performance on new languages while upholding consistent performance on original ones. Experiments on a real-world task across eight languages demonstrate that our proposed LoRA-Whisper yields a relative gain of 18.5% and 23.0% over the baseline system for multilingual ASR and language expansion respectively.

Index Terms: multilingual speech recognition, language expansion, Whisper, LoRA

1. Introduction

Automatic speech recognition (ASR) has traditionally concentrated on transcribing speech into written text for single languages [1–5]. Nevertheless, as the demand for cross-lingual communication grows and vast multilingual datasets [6–9] become more accessible, attention has recently turned towards the development of massively multilingual ASR models. With the emergence of large-scale multilingual speech recognition models such as Whisper [10], Google USM [11], and MMS [12], individuals now have the opportunity to construct **customized multilingual speech recognition models** tailored to specific languages based on these foundational models.

However, two significant challenges are still yet to be addressed in multilingual ASR. One is language interference, primarily stemming from language overlap, data imbalance, dialectal accents, etc. Another challenge involves incorporating new languages without compromising the performance of existing ones. To resolve the former problem, there are a series of previous works attempting to mitigate this issue by leveraging language ID information [13,14] or designing language-specific modules [15–21] such as languages-specific encoders to differentiate each language. Besides, some works [22–24] utilize a pruning strategy in multilingual ASR with a dedicated submodel for each language, while others propose new sampling methods [25] to address the data imbalance issue. Although the methods mentioned above alleviate language interference to some extent, they are somewhat cumbersome in design and fail to account for language expansion. When new languages need to be integrated into a multilingual ASR system, a naive approach is to fine-tune the ASR model using data from these new languages. Unfortunately, this often results in *catastrophic forgetting*, referring to the phenomenon that the recognition performance of base languages tends to decline. To solve the above problem, Li et al. [26] proposes lifelong learning [27] solution which remedies the language interference problem by mixing base language data and new language data. However, this approach is inefficient and time-consuming. Libera et al. [28] explores various continual learning methods [29–34] to address the issue of catastrophic forgetting. While these approaches have helped alleviate the problem, it still persists.

Towards this end, we introduce LoRA-Whisper, a parameter-efficient and extensible model for multilingual ASR. LoRA [35], originally introduced in natural language processing (NLP), effectively customizes large language models (LLMs) for specific domains. Drawing inspiration from this, it can also be used to tailor speech recognition models for specific languages. In practice, we assign a language-specific LoRA matrix for each language. This approach allows shared information across languages to be stored within the Whisper model, while language-specific information can be captured in the respective LoRA matrices. When incorporating a new language, a new LoRA matrix is assigned for it, ensuring no impact on the performance of existing languages. Furthermore, by capitalizing on the similarities between the new language and base languages, we can enhance performance on the new language through improved initialization of the new LoRA matrix or by employing mixture of experts (MoE) [36]. Note that the foundational model is not restricted to Whisper but can encompass other open-source speech recognition models as well, we are simply utilizing Whisper as an exemplar in this paper. In summary, the contributions of this paper are as follows:

- We propose LoRA-Whisper to mitigate language interference and avoid catastrophic forgetting when incorporating new languages by attaching language-specific LoRA modules to the Whisper model.
- By utilizing the similarity between languages, notable performance improvement can be achieved on new languages via better initialization of the new LoRA matrix or the employment of MoE.

2. Background

2.1. Whisper

Whisper [10] is an encoder-decoder Transformer model that is capable of multiple speech tasks, including multilingual speech recognition, speech translation, language identification,

† Corresponding author