Faculty of Computer Science and Information Technology

A Framework for Parameter Optimization and Transfer Learning on Quartznet for Iban Automatic Speech Recognition

Steve Olsen Maikol @ Michael

Master of Science
2024

A Framework for Parameter Optimization and Transfer Learning on Quartznet for Iban Automatic Speech Recognition

Steve Olsen Maikol @ Michael

A thesis submitted

In fulfillment of the requirements for the degree of Master of Science

(Computer Science)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK

2024

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

………………………………

Signature

Name: Steve Olsen Maikol @ Michael

Matric No.: 21020006

Faculty of Computer Science and Information Technology

Universiti Malaysia Sarawak

Date : 25/3/2024

ACKNOWLEDGEMENT

First of all, I extend my sincere gratitude to Dr Sarah Flora Samson Juan, my supervisor, for her invaluable support and guidance throughout my master's journey. Her expert direction in conducting experimental tasks, insightful advice in identifying potential solutions, and, above all, her unwavering assistance in completing the writing of this research thesis have been instrumental in my success.

In addition, I would like to express my heartfelt appreciation to Dr Edwin Mit, my co-supervisor, for his invaluable assistance and guidance in the publication of my research paper.

Furthermore, I am deeply grateful to my family and friends for their unwavering support, both morally and financially, throughout my academic journey. Their constant encouragement and belief in my abilities have been a tremendous source of strength and motivation. I am truly fortunate to have such loving and supportive people by my side.

My sincere gratitude also to the Centre for Graduate Studies, the management of the Universiti Malaysia Sarawak, and my sponsor, Biasiswa Kerajaan Negeri Sabah, for making it possible for me to complete my study here in Sarawak.

Last but not least, I wish to express my profound gratitude to the almighty God for bestowing upon me the energy, resilience, and unwavering spirit needed to overcome the challenges encountered during my master's journey. I am humbled by the divine guidance and blessings that have sustained me throughout this endeavour, and I am deeply thankful for the strength and inspiration received from above.

ABSTRACT

The development of automatic speech recognition (ASR) systems for under-resourced languages poses challenges due to the lack of written resources required to train such systems. Traditionally, researchers have used language models to improve ASR model accuracy, some also resorts to the integration of pronunciation dictionaries, but these methods require abundance of written resources, which under-resourced languages often lack. The Iban language, spoken by the majority people of Sarawak in Malaysia, is an example of an under-resourced language for which previous attempts at developing an ASR system involved building a pronunciation dictionary and language model, transfer learning, and using DNN-HMM acoustic models. However, these methods proved challenging and costly. In this research, we propose a framework that uses a convolutional neural network (CNN) as an acoustic model to build an end-to-end ASR model for the Iban language. Three techniques are proposed to optimize the model without requiring additional data resources, including hyperparameter optimization, data augmentation and transfer learning. We report a significant reduction in word error rate (WER) in our experiments, demonstrating the effectiveness of our techniques. Overall, the proposed framework offers a promising approach for developing ASR systems for under-resourced languages that lack the necessary written resources for traditional methods.

Keywords: End-to-end, speech recognition, low-resource language, convolutional neural network, parameter optimization

Rangka Kerja untuk Pengoptimuman Parameter dan Pembelajaran Pemindahan pada Quartznet untuk Pengecaman Pertuturan Automatik Iban

ABSTRAK

Pembangunan sistem pengecaman pertuturan automatik (ASR) untuk bahasa bersumber rendah menimbulkan cabaran kerana kekurangan sumber bertulis yang diperlukan untuk melatih sistem tersebut. Secara tradisinya, penyelidik telah menggunakan kamus sebutan atau model bahasa untuk meningkatkan ketepatan model ASR, tetapi kaedah ini memerlukan banyak sumber bertulis yang sering bahasa sumber rendah tidak miliki. Bahasa Iban, yang digunakan oleh penduduk Sarawak di Malaysia adalah contoh bahasa sumber rendah yang mana percubaan sebelumnya untuk membangunkan sistem ASR melibatkan pembinaan kamus sebutan dan model bahasa, pemindahan pembelajaran, dan menggunakan model akustik DNN-HMM . Walau bagaimanapun, kaedah ini terbukti mencabar dan mahal. Dalam penyelidikan ini, kami mencadangkan rangka kerja yang menggunakan rangkaian neural konvolusi (CNN) sebagai model akustik untuk membina model ASR hujung ke hujung untuk bahasa Iban. Tiga teknik dicadangkan untuk mengoptimumkan model tanpa memerlukan sumber data tambahan, termasuk pengoptimuman hiperparameter, penambahan data dan pembelajaran pemindahan. Kami melaporkan pengurangan ketara dalam kadar ralat perkataan (WER) dalam eksperimen kami, menunjukkan keberkesanan teknik kami. Secara keseluruhan, rangka kerja yang dicadangkan menawarkan pendekatan yang menjanjikan untuk membangunkan sistem ASR untuk bahasa bersumber rendah yang tidak mempunyai sumber bertulis yang diperlukan untuk kaedah tradisional.

Kata kunci: *Hujung ke hujung, pengecaman pertuturan, bahasa sumber rendah, rangkaian neural konvolusi, pengoptimuman parameter*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AM	Acoustic Model
ANN	Artificial Neural Network
API	Application Programming Interface
ASR	Automatic Speech Recognition
CEL	Convolutional Embedding Layer
CER	Character Error Rate
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DNN	Deep Neural Network
DTW	Dynamic Time Warping
FedAvg	Federated Averaging Algorithm
fMLLR	Feature-Space Maximum Likelihood Linear Regression
G2P	Grapheme to Phoneme
GMM	Gaussian Mixture Model
GPU	Graphical Processing Unit
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
LDA	Linear Discriminant Analysis
LM	Language Model
LSTM	Long Short-Term Memory
LVCSR	Large Vocabulary Continuous Speech Recognition
MFCC	Mel-frequency cepstral coefficients

MLLT	Maximum Likelihood Linear Transform
MLT	Multi-task Learning
NGO	Northern Goshawk Optimization
NLP	Natural Language Processing
PD	Pronunciation Dictionary
PED	Pronunciation Error Detection (PED)
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SA	Spectrogram Augmentation
SCTK	Speech Recognition Scoring Toolkit
STT	Speech-to-Text
STTQznet	Speech-to-Text Quartznet
TDNN	Time Delay Neural Network
TF	Time-Frequency
TL	Transfer Learning
TTS	Text-to-Speech
WER	Word-Error-Rate
WRS	Weighted Random Search
WSJ	World Street Journal dataset

CHAPTER 1

INTRODUCTION

1.1 Background

The development of Automatic Speech Recognition (ASR) system has been trending in these recent years and it has been implemented in many software and applications such as Google Assistant and Amazon Siri, whereby this system receives our audio speech and translate it into text data for the system to recognize as input before handling out its programmed tasks. However, building an ASR for under-resourced language is still a challenge as under-resourced languages suffers the issue of data scarcity, causing the ASR developed to have low prediction accuracy due to insufficient training data. The Iban language is an under-resourced language spoken mainly in Sarawak, Malaysia and West Kalimantan, Indonesia (Aman et al., 2019). The local people of Sarawak use Iban a lot in terms of daily communication, however, written data on the language are lacking. To this day, the most prominent work in the development of Iban ASR was done by Juan (2015), in which the author initiated the development of the very first Iban ASR using Deep Neural Network (DNN). Aside from the works done by Juan (2015), no other committed effort involving the Iban language in ASR development was done. Meanwhile, research gaps regarding the studies of Iban ASR development are still many.

1.2 Convolutional Neural Network in Speech Recognition

Typically, three components are required to build a statistical ASR, these are Acoustic Model, Pronunciation Dictionary and Language Model. Each of these three crucial components is required to be developed using and trained with abundance speech data to help an ASR model to achieve excellent prediction accuracy. This requirement, however, is an issue for under-resourced language such as Iban as they do not possess enough language resources. CNN is a neural network that recently has present itself as a solution to overcome the issue of data scarcity faced by under-resourced languages in building ASR models by being its acoustic model (Arnel Fajardo, 2020; Lekshmi & Sherly, 2021; Thai et al., 2020). With end-to-end architecture and CNN excellent feature extraction capabilities, it helps researcher to exclude the necessity for the integration of pronunciation dictionary and language model into ASR system while still producing high accuracy predictions (Alsayadi et al., 2021; Parry et al., 2019; Yu et al., 2019; Zhang et al., 2021). However, no studies using CNN in the development of Iban ASR system has ever been conducted previously and no data discussing about its performance as an acoustic model for Iban ASR model, whether it is able to overcome Iban language's data scarcity, has ever been recorded yet. With that said, it serves as our motivation to conduct research on the CNN for Iban ASR to analyse its capability as an acoustic model for under-resourced language and to fill this study gap.

1.3 Problem Statement

As mentioned previously, developing an ASR model using the end-to-end CNN architecture for the Iban language has never been conducted yet and it is known that under-resourced language suffers a lack of language resource in building a statistical ASR. The proper steps to build an Iban ASR using CNN as acoustic model while excluding the integration of pronunciation dictionary and language model in the system and still achieving excellent prediction accuracy has never been documented previously and no framework describing its process has ever been proposed. Furthermore, it was required that for a CNN ASR model to perform well, its network structure has to be optimized (Aszemi & Dominic, 2019; Xie & Yuille, 2017). Currently, there is no known systematic way of building an optimized CNN ASR model for the under-resourced Iban language. Investigating this research gap would help us to identify a proper method to develop an end-to-end CNN Iban ASR with optimized model structure and propose its systematic framework which will act as a document for future references.

1.4 Research Questions and Objectives

In response to the problem statement described earlier, we have listed out several research questions and objectives as a guideline for the research to investigate the previously mentioned study gap. The details are as follows:

Research Questions:

1. How to obtain a CNN acoustic model for Iban ASR?
2. How to determine the parameters that can influence the performance of CNN acoustic modelling in Iban ASR?

Research Objectives:

I. To study the general architecture of CNN acoustic modelling in ASR and its benefits for under-resourced language.

II. To propose a CNN acoustic modelling framework for investigating the WER in Iban ASR.

III. To investigate the WERs obtained by the CNN-based Iban ASR model through hyperparameter optimization, spectrogram augmentation, and transfer learning and analyse its performance.

1.5 Scope of Research

Two constraints have been defined to help us focus on our scope when carrying out experiments that covers a wide field of knowledge. Although this research aims to study the ways on how to improve ASR for under-resourced language, only the Iban language that will be used as the target language in this research. The main corpus that is going to be used for the experiment will be the Iban corpus that was previously collected by (Juan, 2015). Other corpus may be imported but only for the sole purpose of improving the performance of the Iban ASR model (e.g., for transfer learning). Secondly, as the title of the research implies, studies on Neural Network model will be conducted only on the CNN. The scope of our research experiment will focus on the development of ASR model using end-to-end CNN model only and without the integration of pronunciation dictionary and the language model. Despite having to do comparison between CNN and other ANN models in the evaluation stage, thorough analysis and investigation will be done only for the CNN model. The focus of this research will follow these two rules; Iban language being the targeted under-resourced language and CNN being the only neural network model to be investigated, as its core to prevent straying away from the main purpose and the objectives of this research.

1.6 Significance of Study

Through the conduct of this research, we will be able to contribute an improvement towards ASR advancement specifically, for the Iban language, generally, for under-resourced language. First of all, we would be able to document our setup on the CNN architecture that will be used and be presented generally as a reference for future under-resourced language research that wants to implement the same architecture. In addition, this research would be beneficial to researcher as its baseline results produced during the research experiments can be taken for the conduct of comparison between different variant of Iban CNN models in the future. By doing this research, we would be able to identify which hyperparameters in CNN that may affect the performance of an ASR in training the Iban language, beneficially and detrimentally. We would also be able to identify what may be the weakness of CNN as an acoustic model through the conduct of this research experiment.

Moreover, this research would also help us analyse the potential of CNN in overcoming data scarcity issue of under resourced language in the development of ASR without the integration of pronunciation dictionary and language model. Furthermore, it is in our expectation that the very first framework to develop an end-to-end CNN Iban-ASR system will be proposed at the end of this research, thus, it will serve as a reference for future CNN Iban ASR model development and fine-tuning. The result and protocol obtained from this research experiment would also be prove useful for future analysis and reference for identifying effective fine-tuning techniques on the CNN architecture. Finally, this research will help us to set a new benchmark for Iban ASR model performance as we attempt to further improve the accuracy of speech recognition models while implementing other various method of algorithm in predicting Iban words.

1.7 Research Outlines

The thesis is organized into five chapters. Chapter 1 introduces the research work which includes our problem statement, research questions and objectives, as well as scope and significance of research. Chapter 2 discusses the literature review and previous existing works that has been done in improving under-resourced ASR model performance, as well as exploring currently trending CNN techniques while promoting the relevancy of conducting this research. Chapter 3 introduces our methodology and the description of our proposed framework to develop the first optimized Iban end-to-end CNN ASR model. The chapter explains our experimental steps and procedure in developing and optimizing our CNN model which includes implementing different model improvement techniques. Chapter 4 presents the results obtained from the experiments conducted in Chapter 3 as well as its analysis and discussion. Chapter 5 concludes the research thesis with a summary of the work that has