សាកលវិទ្យាល័យភូមិន្ទភ្នំពេញ

ROYAL UNIVERSITY OF PHNOM PENH

ការស្រាវជ្រាវវិធីសាស្ត្រថ្មី សម្រាប់កំណត់សម្គាល់អត្ថបទអក្សរខ្មែរទូទៅ នឹង បានប្រើប្រាស់ស្ថាបត្យកម្ម Craft ជាមួយនឹង TrOCR

A novel End-to-End approach for General Khmer Text Recognition using Craft with TrOCR Architecture

Mr. Vitou Soy

A Thesis

In Partial Fulfilment of the Requirement for the Degree of
Bachelor of Engineering in Information-Technology-Engineering

Examination committee:	Mr. Sokchea Kor (Advisor)
	Mr. Chanpiseth Chap (committee)
	Mrs. Daly Chea (committee)
	Dr. ....

June 2025

មូលន័យសង្ខេប

ក្នុងសហគមន៍បច្ចេកវិទ្យាព័ត៌មានសម័យថ្មី ការចាប់យកអត្ថបទចេញពីរូបភាព – [OCR] (Optical Character Recognition) ក្លាយជាបច្ចេកវិទ្យាសំខាន់មួយដែលត្រូវបានប្រើប្រាស់ យ៉ាងទូលំទូលាយ សម្រាប់បំលែងឯកសារ សរសេរ ឬរូបភាពអក្សរឱ្យទៅជាអត្ថបទ អេឡិចត្រូនិច (digital text)។ ការអភិវឌ្ឍ OCR សម្រាប់ភាសាខ្មែរ តែងតែ ប្រឈមនឹងបញ្ហាជាច្រើន ដោយសារកង្វះនៃប្រភពទិន្នន័យ និងឯកសារសម្រាប់ train AI model។ ដើម្បីដោះស្រាយ បញ្ហានេះ យើងបានបង្កើតទិន្នន័យសិប្បនិម្មិត (Synthetic Dataset) ដោយប្រើវិធីសាស្ត្របច្ចេកទេសកម្រិតខ្ពស់។

ក្នុងដំណើរការបង្កើតទិន្នន័យសិប្បនិម្មិត (Synthetic Dataset) រួមមាន៖

- វិធីសាស្ត្រក្នុងការប្រមូលអត្ថបទចេញពីអ៊ីនធឺណិត មានដូចខាងក្រោម (Scrape data) ៖
  - ដំណាក់កាលទីមួយ៖ យើងបានប្រមូលអត្ថបទចេញពី khsearch.com, Chuon-Nath-Dictionary, Alpha-Word, Google-Word, និងចុងក្រោយគឺ Huggingface.com ។
  - ដំណាក់កាលទីពីរ៖ យើងបានសម្អាត ទិន្នន័យទាំងអស់នោះ ឆ្លងកាត់ដំណើរការ ដូចជា លុបចោលតួរ អក្សរណាដែលមិនសូវមាន វត្តមាននៅលើ រូបភាព ញឹកញាប់ និងបានលុបចោល តួរអក្សរណាដែល Fonts renders អត់ចេញ។
  - ដំណាក់កាលទីបី៖ ដំណាក់កាលមួយនេះ យើងបានធ្វើការ កាត់ប្រយោគទាំងអស់នោះ ជាពាក្យៗ ដោយ ប្រើប្រាស់ library ឈ្មោះ khmer-nltk
  - ដំណាក់កាលទីបួន៖ ចុងក្រោយ ក៏បានរៀបចំជា ប្រយោគដែល មានប្រវែង Random ពី ១ អក្សរ រហូតដល់ ១១០ អក្សរ ។
- បង្កើតរូបភាពដោយអនុវត្តតាមលក្ខខណ្ឌខាងក្រោម ៖
  - ផ្ទៃខាងក្រោយចៃដន្យ (Apply Different backgrounds)
  - បំពាក់ពុម្ពអក្សរផ្សេងៗគ្នា (Apply Different fonts)
  - Noise: `gaussian_noise`, `salt_pepper_noise`, `speckle_noise`, `blur`
  - បង្វិលអក្សរបន្តិច (random rotation text)
  - បញ្ចូល Margin Randomly (1, 5) pixels
- សរុបមកយើងបានបង្កើត Data ជាង ៤ លាន records សម្រាប់ train OCR model

Architecture OCR ត្រូវបានបែងចែកជា ២ ផ្នែក៖ Text Detection និង Text Recognition:

Text Detection: យើងប្រើម៉ូដែល CRAFT ដោយបានធ្វើការ Train ឡើងវិញដោយ បាន annotation ទៅលើ លើរូបភាពប្រហែល ៥០០ images និងសរុបចំនួន bounding box ជាង ១០,០០០ boxes។

Text Recognition: យើងប្រើ TrOCR base model ចេញពី Microsoft (មាននៅក្នុង Hugging Face) ហើយ បាន fine-tune ទៅលើ dataset ខ្មែរសិប្បនិម្មិត (Synthetic Dataset) ដើម្បីបង្កើនសមត្ថភាពក្នុងការសម្គាល់ អក្សរខ្មែរ។

លទ្ធផលសិក្សាបានបង្ហាញថា OCR របស់ពួកយើងអាចសម្គាល់អត្ថបទចេញពីរូបភាព បានដោយភាពត្រឹមត្រូវ លើសពី ៩០%។ ដូច្នេះ ការសិក្សានេះបង្ហាញអំពីសក្តានុពលនៃការបង្កើត dataset និងការប្រើម៉ូដែលជំនាន់ថ្មី ដើម្បីអភិវឌ្ឍន៍ OCR ភាសាខ្មែរឱ្យមានប្រសិទ្ធភាពកាន់តែខ្ពស់។ វាមានសមត្ថភាព អាចចាប់យកអត្ថបទមិនត្រឹមតែ ពាក្យខ្លីៗ ប៉ុណ្ណោះទេ តែវាក៏អាចធ្វើការចាប់យក ដូចជា មួយតួអក្សរដោយមួយតួអក្សរ, ពាក្យដោយពាក្យ, ប្រយោគ ដោយប្រយោគ រហូតដល់ មួយប្រយោគវែង ១១០ តួអក្សរថែមទៀតផង ។ ហើយលើសពីនោះទៀត វាក៏អាចធ្វើការ កំណត់សម្គាល់ទៅលើ ពីរ ភាសាចម្បង ទាំងភាសាខ្មែរ និងភាសាអង់គ្លេស ។

Abstract

In the modern era of information technology, Optical Character Recognition (OCR) has emerged as a crucial technology for converting printed or handwritten text from images into digital form. However, the development of OCR systems for the Khmer language presents significant challenges, primarily due to the lack of large-scale annotated datasets. To address this limitation, we constructed a high-quality synthetic dataset using an advanced data generation pipeline. Our Khmer OCR system consists of two core components:

- Text Collection: We gathered Khmer text data from various online sources, including khsearch.com, Chuon-Nath Dictionary, Alpha-Word, Google-Word, and Hugging Face.
- Data Cleaning: We processed and cleaned the collected text by removing uncommon characters, symbols that are rarely rendered correctly by fonts, and excessive whitespace.
- Text Segmentation: Sentences were tokenized into words using the khmer-nltk library, and then reconstructed into randomized sentence lengths ranging from 1 to 110 characters.
- Image Generation: We rendered text into synthetic images by:
  - Applying random backgrounds and a variety of Khmer fonts
  - Adding diverse noise types such as Gaussian noise, salt-and-pepper noise, speckle noise, and blur
  - Introducing slight random rotations and random margins (1–5 pixels)
- As a result, we generated over 4 million high-quality synthetic image-text pairs to train the OCR model.

Our Khmer OCR system consists of two core components:

- **Text Detection:** We fine-tuned the CRAFT (Character Region Awareness for Text Detection) model using 500 manually annotated images, totaling over 10,000 bounding boxes.
- **Text Recognition:** We fine-tuned Microsoft's TrOCR base model (available on Hugging Face) on our synthetic Khmer dataset to improve its ability to recognize Khmer text.

The evaluation results demonstrate that our system achieves a recognition accuracy exceeding 90%. These findings highlight the effectiveness of combining synthetic data generation with modern transformer-based architectures to significantly advance Khmer OCR capabilities. Notably, the system can accurately recognize a wide range of text—from single characters and individual words to full sentences of up to 110 characters—and supports both Khmer and English languages.

SUPERVISOR's RESEARCH SUPERVISION STATEMENT

Name of program: Khmer Studies
Name of candidate: Vitou Soy

Title of research report: A novel End-to-End approach for General Khmer Text Recognition using Craft with TrOCR Architecture

This is to certify that the research carried out for the above titled master's research report was completed by the above named candidate under my direct supervision. This thesis material has not been used for any other degree. The candidate has demonstrated strong research capabilities and independence in developing novel approaches for Khmer text recognition. The research methodology, implementation, and results are original contributions to the field of Khmer OCR technology. I have provided guidance and oversight throughout the research process while allowing the candidate to explore innovative solutions.

Supervisor's name: Sokchea Kor
Supervisor's signature:……………………
Date……………………………………………

CANDIDATE'S STATEMENT

TO WHOM IT MAY CONCERN

This is to certify that the dissertation that I, Vitou Soy, hereby present, entitled ”Advancing Khmer Optical Character Recognition: A Synthetic Data-Driven Approach,” for the degree of Bachelor of Engineering in Information Technology at the Royal University of Phnom Penh, is entirely my own work. Furthermore, it has not been used to fulfill the requirements of any other qualification, in the whole or in part, at this or any other University or equivalent institution. The research methodology, implementation, and findings represent original con- tributions to the field of Khmer OCR technology, particularly in developing novel approaches for synthetic data generation and transformer-based text recognition. Through this work, I have demonstrated strong research capabilities and independence in addressing the critical challenges of Khmer text digitization and recognition.

No reference to, or quotation from, this document may be made without the written approval of the author.

Name of Candidate: Vitou Soy

Signed by the candidate:

Date: . . . .

Name of Supervisor: Mr. Sokchea Kor

Countersigned by the Supervisor: . . . .

Date: . . . .

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Mr. Sokchea Kor, for his guidance and expertise. His feedback and support throughout the research process have been instrumental in shaping this work. This research was inspired by Dr. Rina Buoy's contributions to the field. I appreciate the Royal University of Phnom Penh management for establishing this program within the Faculty of Engineering. I would also like to acknowledge khsearch.com, Chuon-Nath Dictionary, Alpha-Word, Google-Word, and Hugging Face for providing essential datasets.

I am profoundly thankful to the entire Faculty of Engineering community for their exceptional support and contributions to my academic journey. The knowledge, resources, and supportive environment they provided have been crucial to my success. The faculty's unwavering commitment to excellence and their dedication to nurturing future engineers have created an atmosphere that truly fosters innovation and learning. The state-of-the-art facilities and cutting-edge technology available have enabled me to conduct advanced research in optical character recognition with unprecedented precision. The collaborative spirit among faculty members, researchers, and students has fostered an environment of intellectual growth and continuous innovation. Through numerous workshops, seminars, and technical discussions, I have gained deep insights into the field of computer vision and machine learning. The faculty's strong industry connections and emphasis on practical applications have ensured that my research remains relevant and impactful. Their guidance in implementing transformer-based architectures and synthetic data generation techniques has been particularly valuable. The mentorship provided by senior researchers and the opportunity to participate in various research projects have significantly enhanced my technical capabilities and research methodology.

Finally, I would like to thank my friends for their encouragement throughout my studies. Their support and belief in my capabilities have helped make this journey meaningful.

TABLE OF CONTENTS

Preliminary Pages

Appendices

List of Tables

List of Figures

LIST OF ABBREVIATIONS

OCR: Optical Character Recognition
CNN: Convolutional Neural Network
RNN: Recurrent Neural Network
LSTM: Long Short-Term Memory
GRU: Gated Recurrent Unit
Transformer: Transformer Model
BERT: Bidirectional Encoder Representations from Transformers
TrOCR: Transformer OCR
ViT: Vision Transformer
ViT-OCR: ViT OCR
ViT-OCR-S: ViT OCR Small
ViT-OCR-B: ViT OCR Base
ViT-OCR-L: ViT OCR Large
ViT-OCR-H: ViT OCR Huge

Chapter 1

Introduction

This is the introduction chapter.