An End-to-End Khmer Optical Character Recognition using Sequence-to-Sequence with Attention

Rina Buoy*, Nguonly Taing*, Sovisal Chenda* and Sokchea Kor**
*Techo Startup Center (TSC), **Royal University of Phnom Penh (RUPP)

Research Motivation

- To develop an end-to-end OCR pipeline for multi-font Khmer text recognition utilizing a deep learning-based sequence-to-sequence model with attention mechanism
- To achieve the state-of-art performance in Khmer text recognition

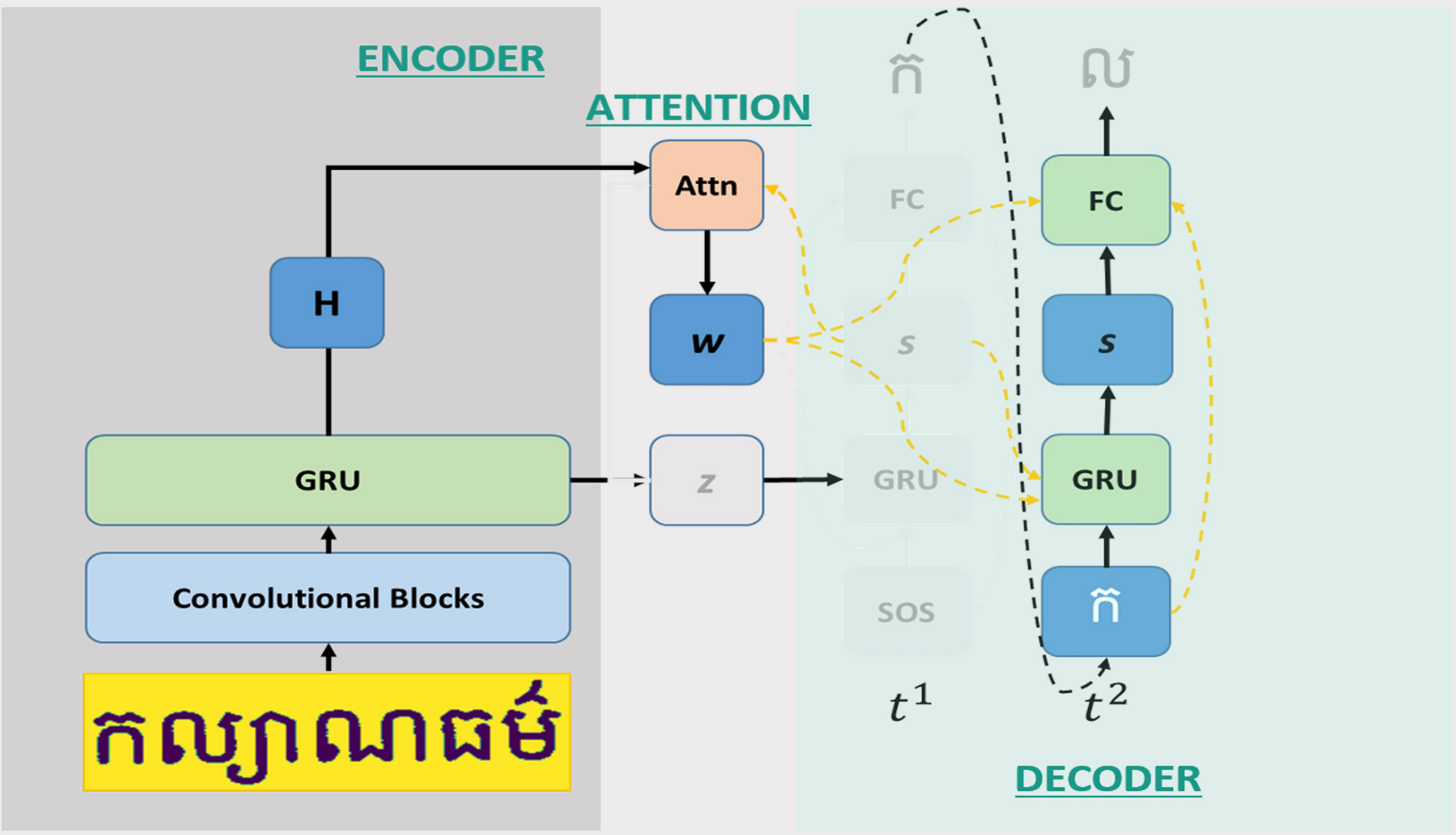
Problem in Khmer OCR

- Complex pre-processing or feature extraction and post-processing steps
- Being able to predict only a standalone character (ex. consonants only)
- Unable to handle noisy background

Proposed Solutions

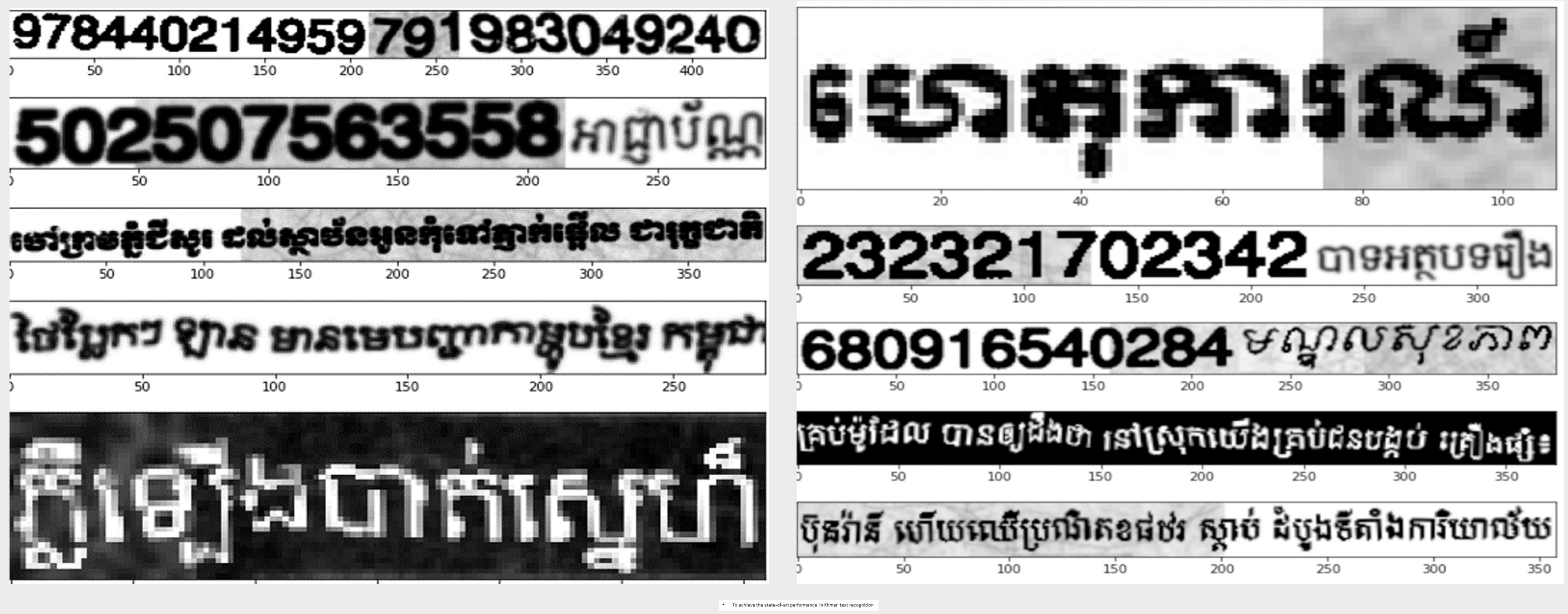
- We introduce one of the first end-to-end solutions to Khmer OCR
- The proposed solution outperforms the current state-of-art Tesseract engine for Khmer language.

Model Architecture



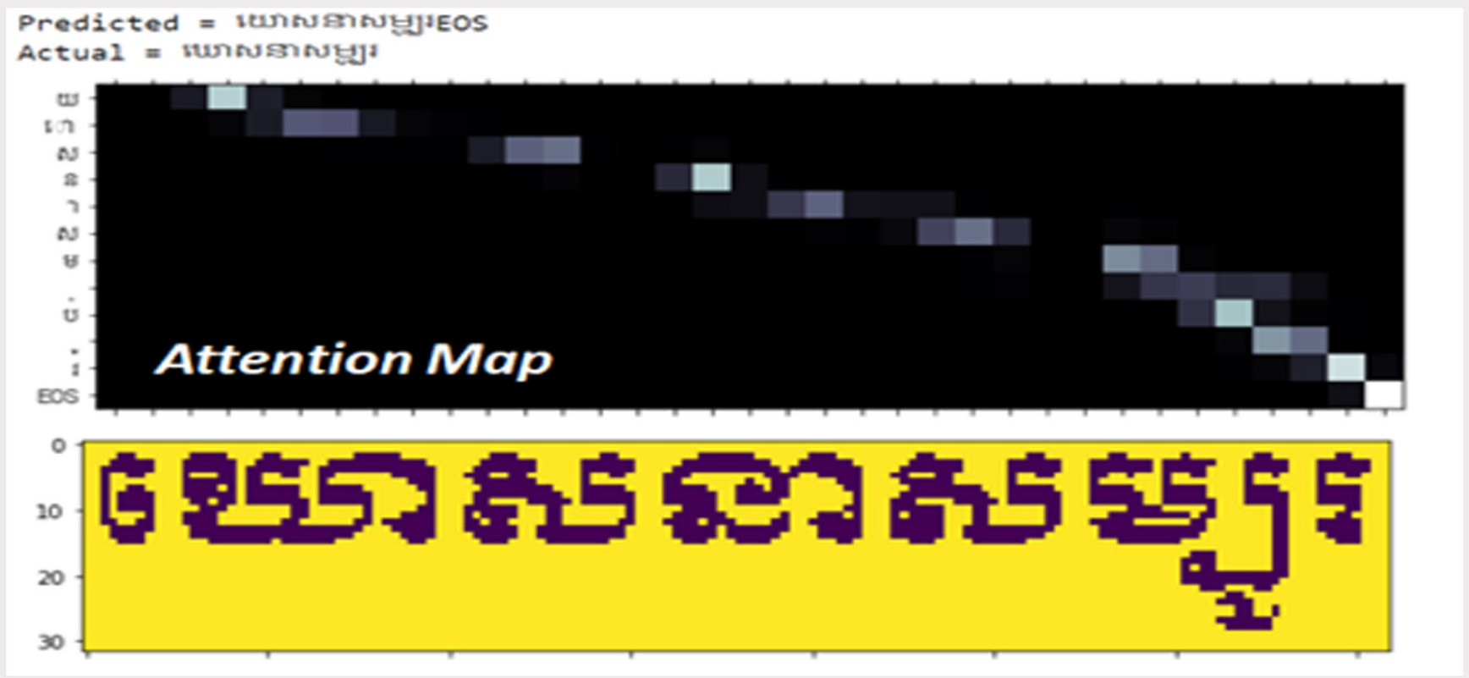
The end-to-end, attention-based Seq2Seq network for Khmer OCR

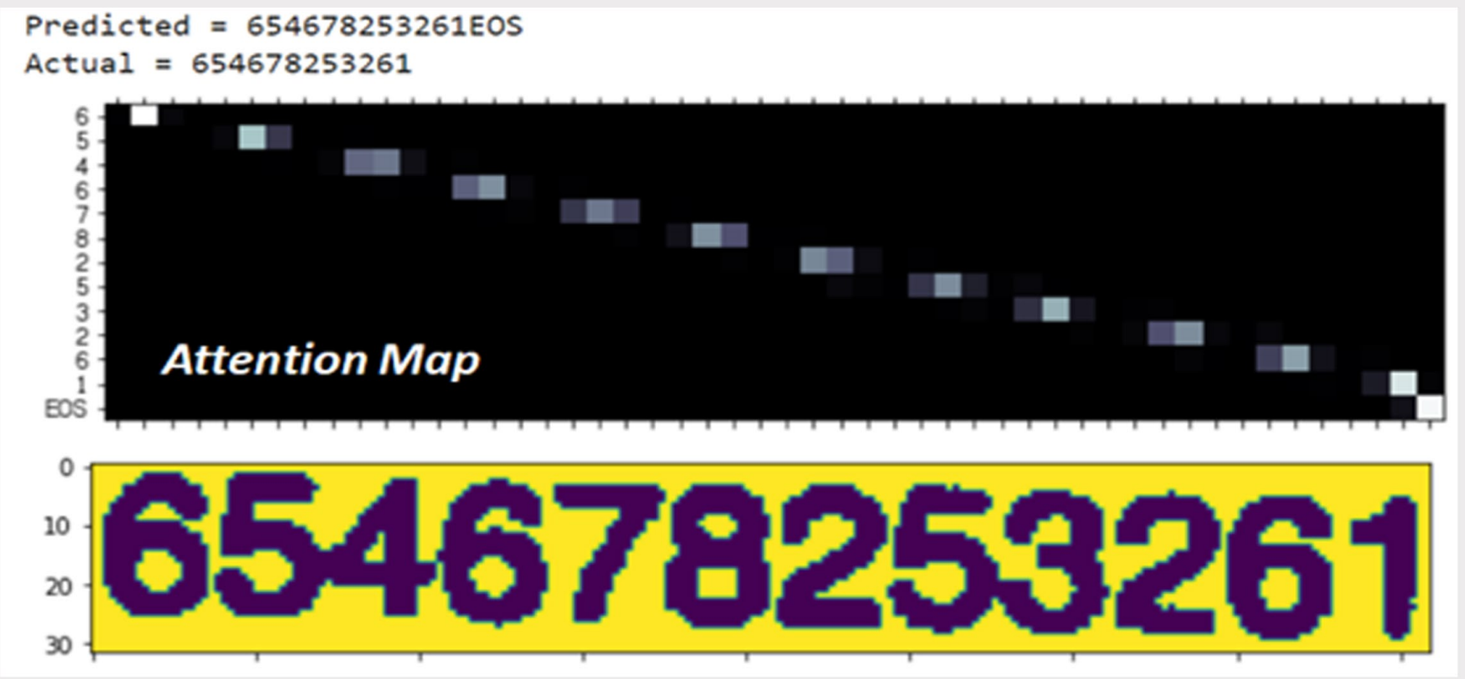
Synthetic Dataset and Data Augmentation



- Three millions synthetic text-line images (H = 32, W > 64)
- Khmer OS font, size 11
- Mixture of numbers, words, phrase and sentences
- Complex Data Augmentation with erosion, dilation, removal, noisy background, etc.

Model Training & Benchmarking





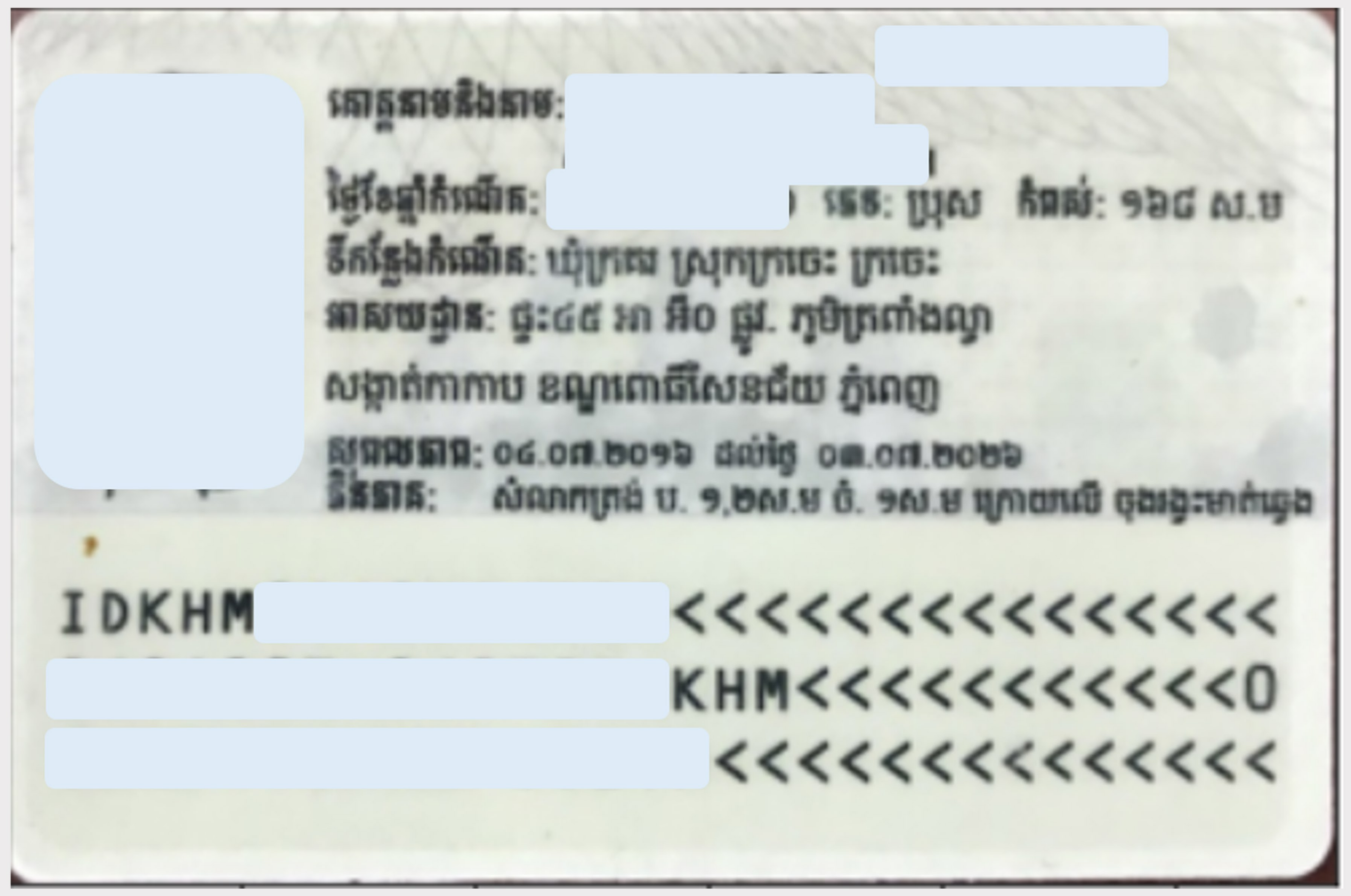
The trained model was benchmarked against Tesseract OCR for Khmer language on the test set of 5000 images.

Result of character error rate :

- Our trained model : 1.5%
- Tesseract OCR : 4.5%

Application to Khmer ID Card

We applied the trained model to extract fields from Khmer ID card which contains texts of different fonts in the same line (excluding sensitive data).



[redacted]
[redacted]
គោត្តនាមនិងនាម: [redacted]
ថ្ងៃខែឆ្នាំកំណើត: [redacted] ភេទ: ប្រុស កំពស់: ១៦៨ ស.ម
ទីកន្លែងកំណើត: ឃុំក្រគរ ស្រុកក្រចេះ ក្រចេះ
អាសយដ្ឋាន: ផ្ទះ៤៥ អា អឺ០ ផ្លូវ. ភូមិត្រពាំងល្ហា
សង្កាត់កាកាប ខណ្ឌពោធិ៍សែនជ័យ ភ្នំពេញ
សុពលភាព: ០៤.០៧.២០១៦ ដល់ថ្ងៃ ០៣.០៧.២០២៦
ភិនភាគ: សំលាកត្រង់ ប. ១,២ស.ម ចំ. ១ស.ម ក្រោយលើ ចុងរង្វះមាត់ឆ្វេង

IDKHM[redacted]<<<<<<<<<<<<<<<<<
[redacted]KHM<<<<<<<<<<<<0
[redacted]<<<<<<<<<<<<<<<<<

Gender: ប្រស
Height: ១៦៨ ស.ម
Place of Birth: ឃុំក្រគរ ស្រុកក្រចេះ ក្រចេះ
Address 1: ផ្ទះ៤៥ អា អឺ0 'ផ្លូវ. ភូមិត្រពាំងល្ហា
Address 2: សង្កាត់កាកាប ខណ្ឌពោធិ៍សែនជ័យ ភ្នំពេញ
Validity: ០៤.០៧.២០១៦ ដល់ថ្ងៃ ០៣.០៧.២០២៦
Symbol: សំលាកត្រង់ ប. ១,២ស.ម ចំ. ១ស.ម ក្រោយលើ
ចុងរង្វះមាត់ឆ្វេង

Conclusion

- We presented an end-to-end Khmer OCR system which utilizes an encoder-decoder (Seq2Seq) network.
- The proposed model was trained a collection of text-line images with degrading effects.
- The experiment results suggested that the proposed solution outperformed the current state-of-art Tesseract engine for Khmer OCR by achieving a CER of 1.5% vs 4.5%.