

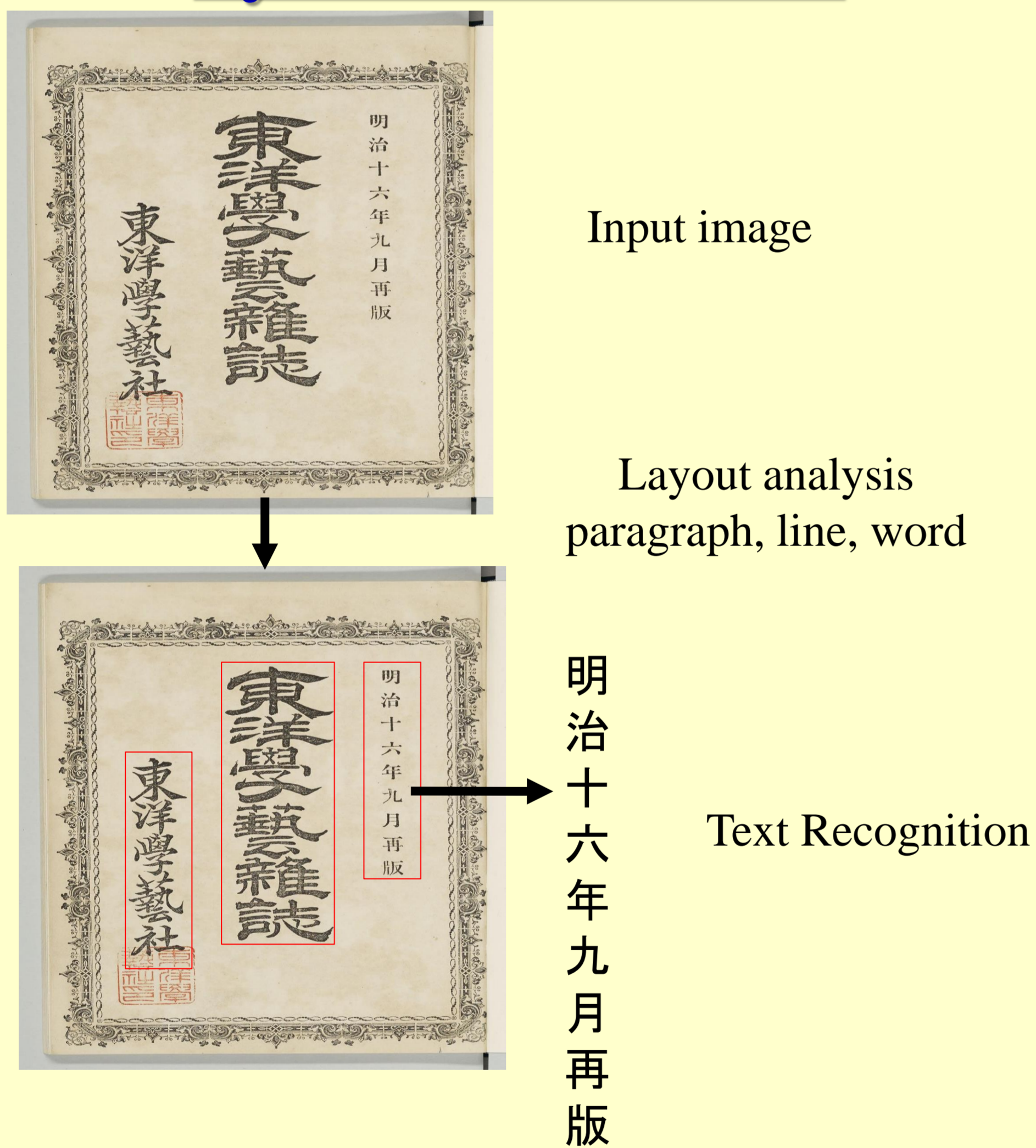
# Recent Achievements of Deep Learning on Recognition of Modern Japanese Magazines

LE DUC ANH データ科学研究系 特任助教

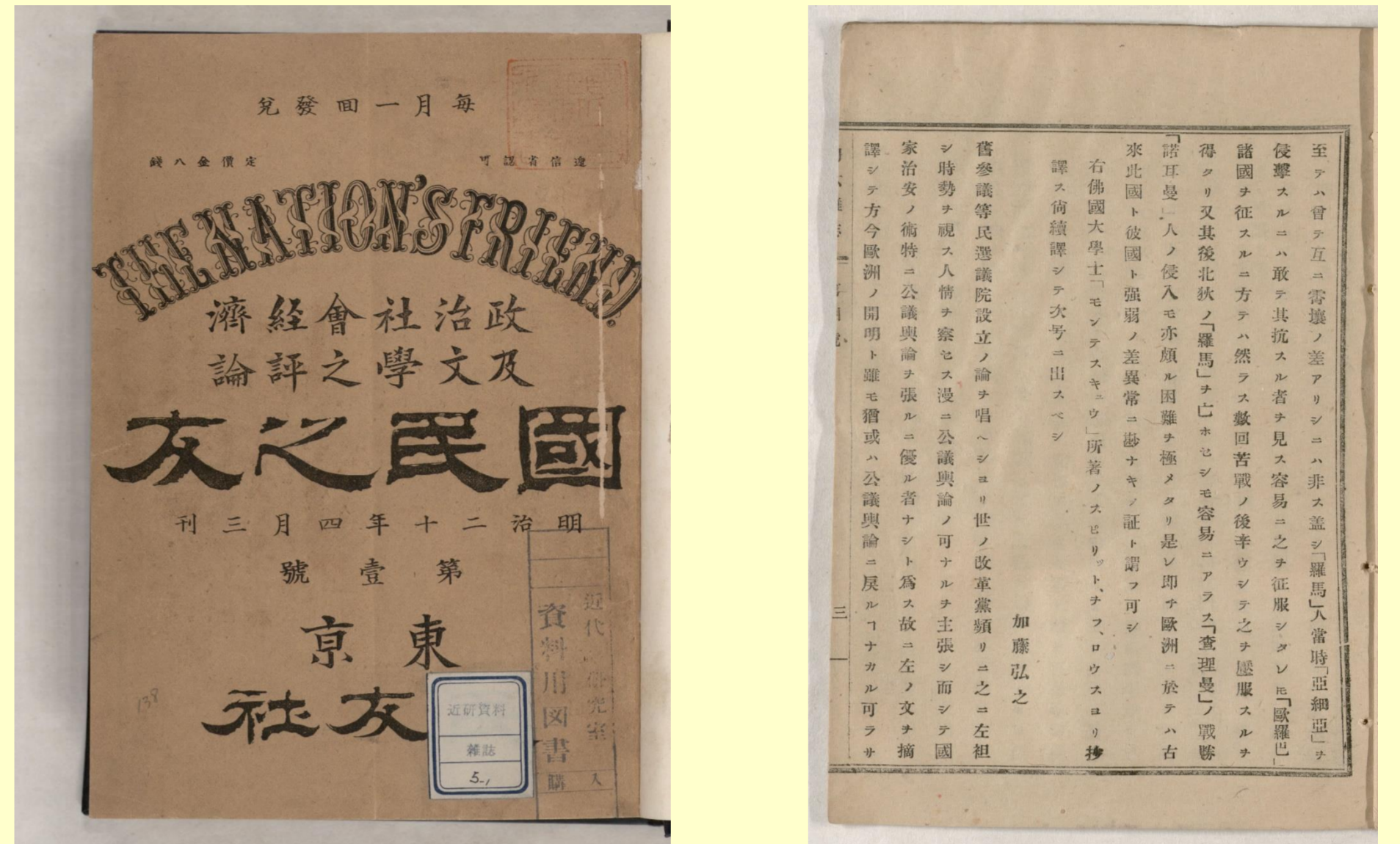
## Introduction

- Since historical documents are an invaluable resource for historians in exploring social aspects, lifestyles, even weather in the previous era, many countries have been preserved their historical documents
- Document analysis and recognition can speed up the transcription process.
- We Employ state-of-the-art technologies in Deep Learning to digitize modern Japanese document
- We aim to build an OCR system for modern Japanese historical documents. The recognition system should be easy to train and reuse by other researchers from document analysis and digital humanities fields

## System Overview



## Dataset & results



Japanese Modern Magazines (~1900)

- Collected dataset: The dataset contains 922 pages from historical magazines in Japan from 1870 to 1945.
- The number of categories is 5,398 which contains many character categories that do not use in current Japanese character system.

### Text Detection

| Precision | Recall | F1   |
|-----------|--------|------|
| 89.6      | 92.0   | 90.8 |

### Text Recognition

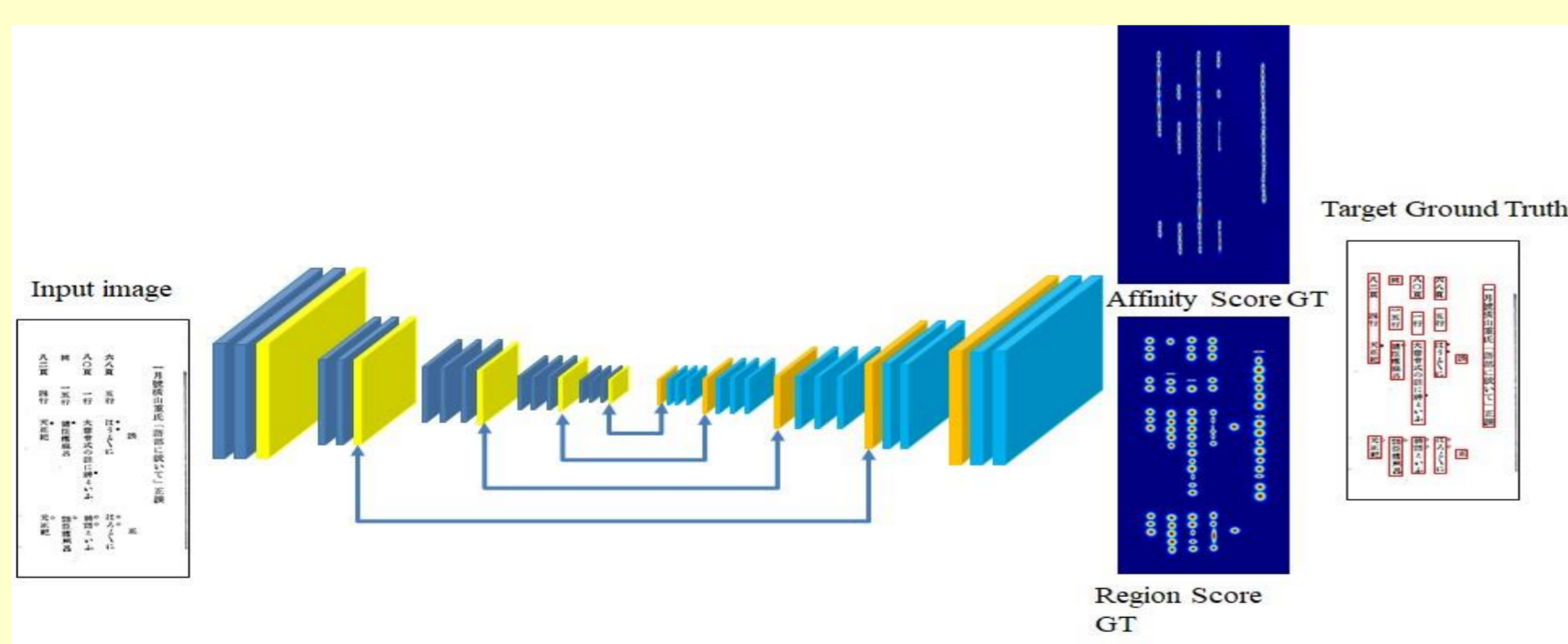
| Character Error Rate |
|----------------------|
| 22.52                |

### Example of recognition result

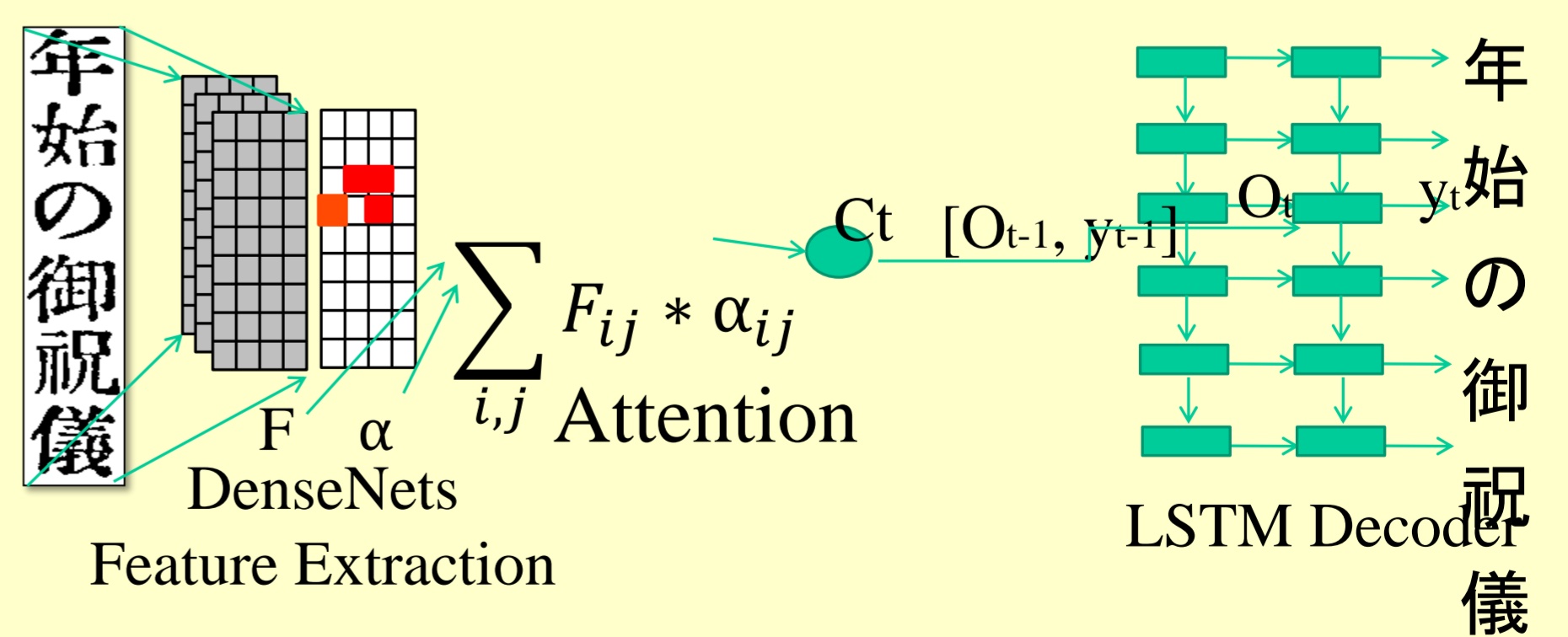


<https://github.com/ducanh841988/Kindai-OCR>

## System Architecture



Text Detection (CRAFT)



Text Recognition (ADE)