MSc Individual Project Progress Report

Michael Song (02405765) on 07 July

Progress before last meeting

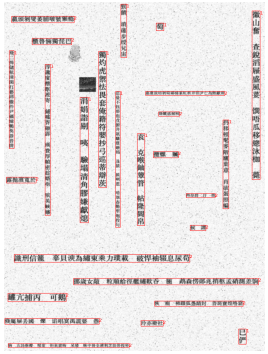
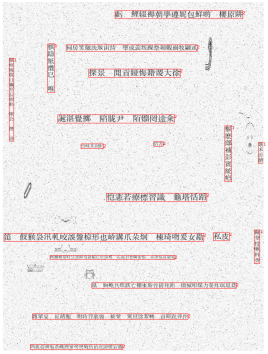
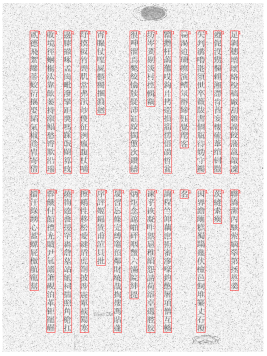
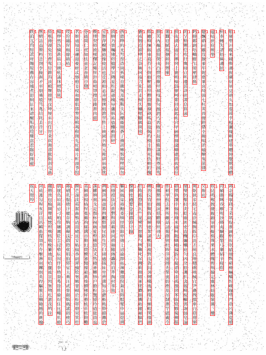
- Background and progress report
- Data collection (journal images with labels)
- Synthetic datasets for training
- Trained following models:
  - Text Line Detector (LD), based on Faster-RCNN
  - Text Line Recognizer (LR), based on CRNN
  - Character Detector (CD), based on Faster-RCNN
  - Character Recognizer (CR), based on Faster-RCNN
  - Character Classifier (CC), based on ResNet
- Compared the performance with following methods:
  - CR
  - CD + CC
  - LD + CR
  - LD + CD + CC
  - LD + LR
- **LD + LR** is better and chosen for further improvement
  - LD performs well on the most common layout.
  - LR only performs well on train/validation set, **but not** on target images.
    - The synthetic training set may not fully matches the target data.
  - Both LD and LR are only trained on vertical text images.

Progress since last meeting

Labelling

- Manual text extraction for 20 target images.

Training

- Re-train **LD** on a new synthetic dataset, which:
  - Add **random layout** images (left two) to simulate multiple text layout (include horizontal text lines), in addition to the most common layout (right two)
  - Integrate icon images from `Icons-50` into the training data with varying sizes, to simulate non-text area in the target data.
  - Apply a wider range of **data augmentation**, including text size, character blank, noise, intensity shift, Gaussian blur etc.
  -    
- Re-train **LR** on a new synthetic dataset, which:
  - Combines vertical and horizontal text.
  - Apply **frequency-based** character selection - character that appears more frequently (data source) is more likely to be chosen in the synthetic images.
  - Enlarge the character **typeface** from 1 to 4 (6k to 24k character images in the database).

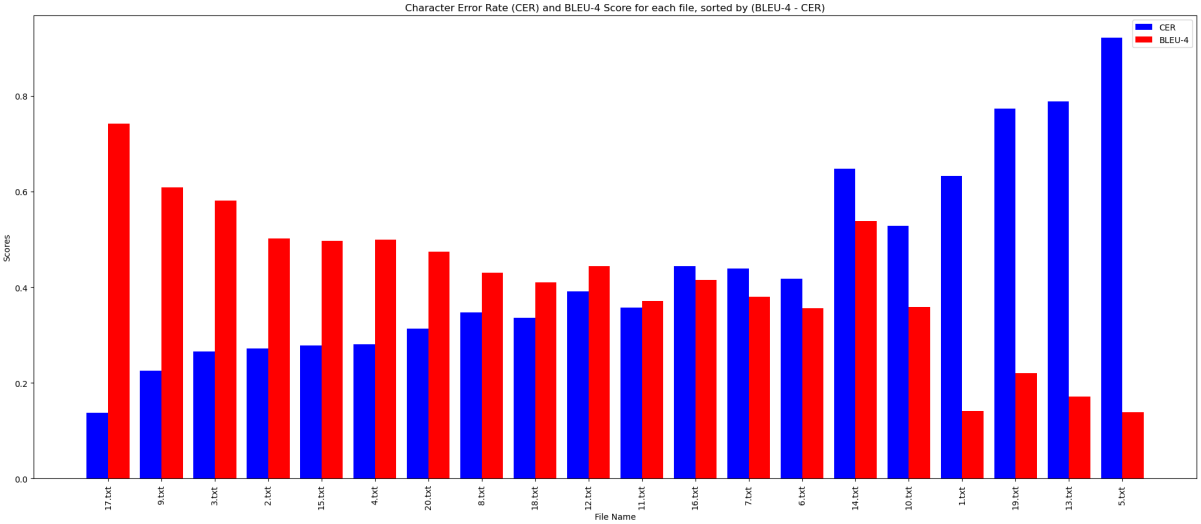
– Apply a wider range of **data augmentation**.

Result

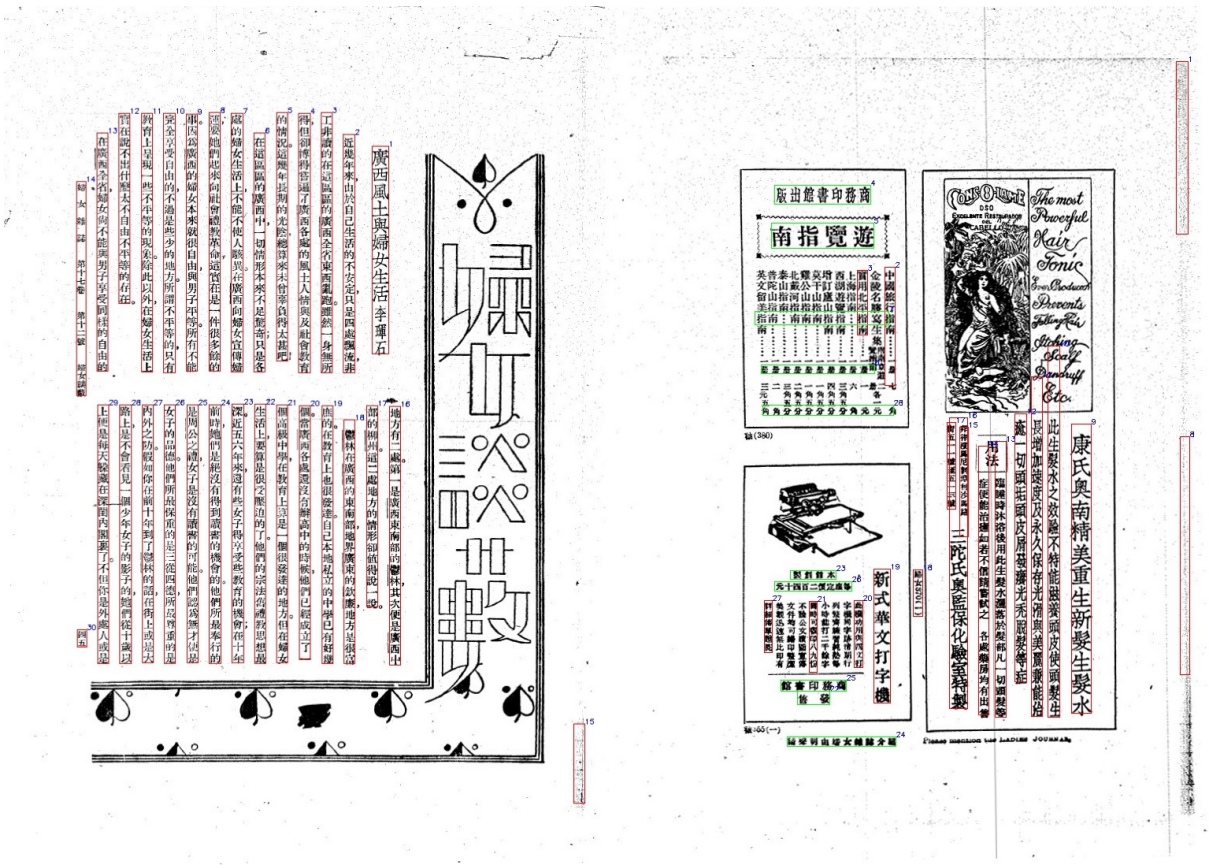
- LD can handle a wider range of text layouts.
- Both LD and LR can handle vertical/horizontal text lines.
- LD is less likely to take non-text area (like images and noise) as text lines.
- Reduced over-fitting for LR.

Metrics

- Inference speed: **0.8 sec/image** in average for the target data, on Nvidia GTX1650
- Character Error Rate (CER): **0.4397** in average on 20 test images.
- BLEU-4 Score: **0.4139** in average on 20 test images.

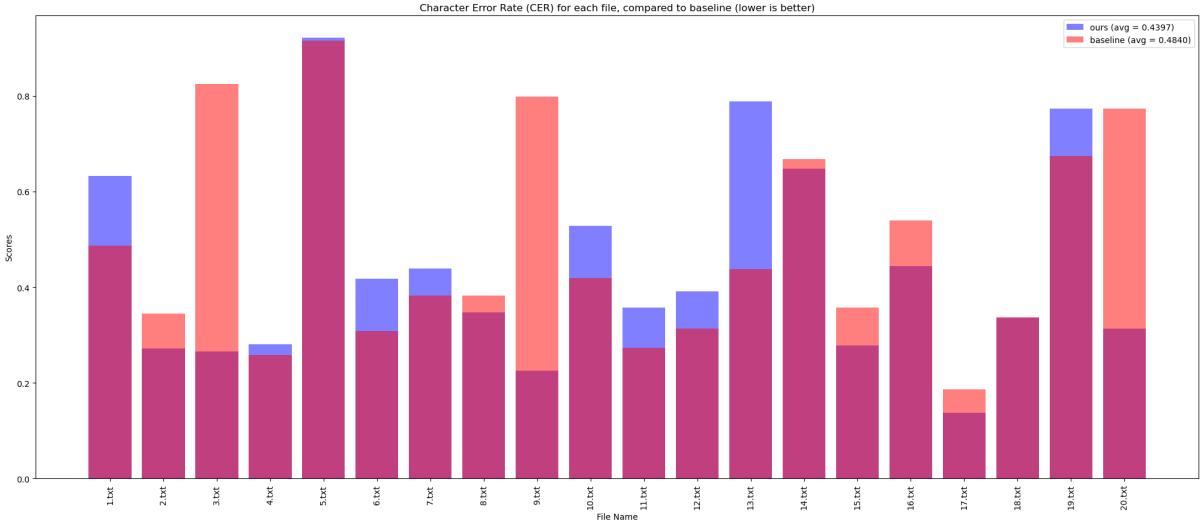


- Regarding (BLEU - CER), the best result is on `17` (left), and the worst is on `5` (right).

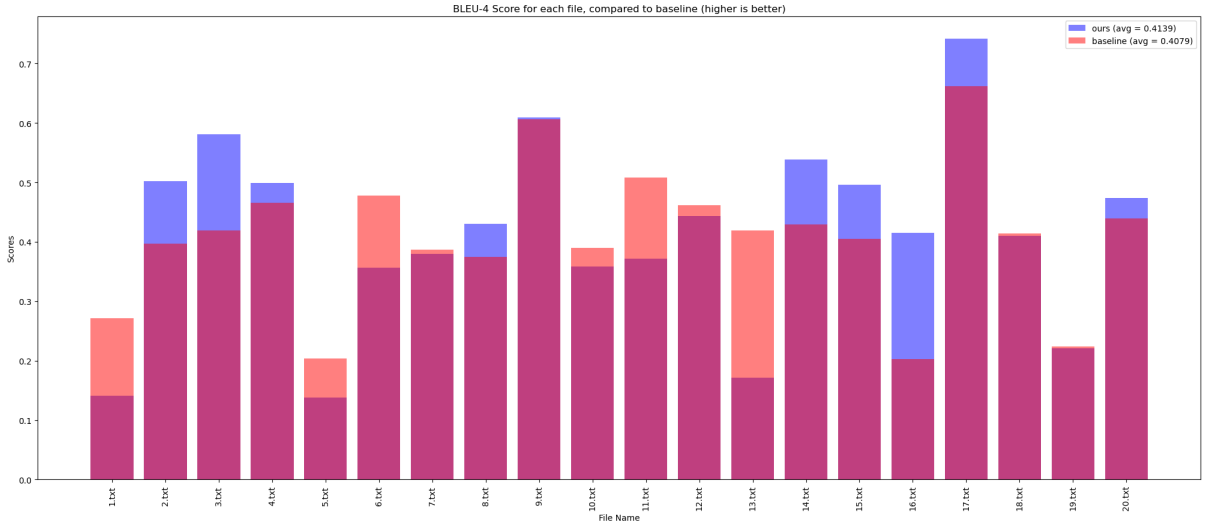


Comparison

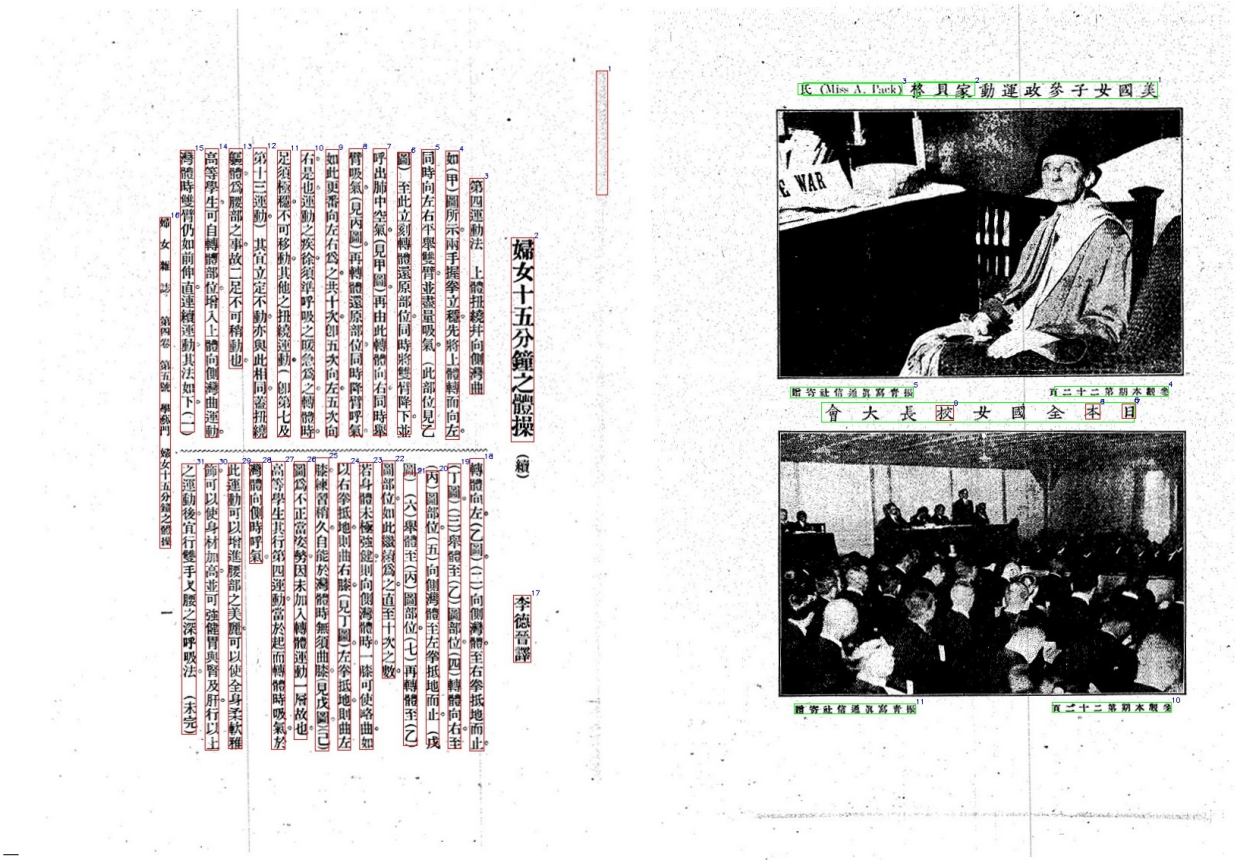
- Baseline: Apple Preview built-in OCR engine
- Character Error Rate (CER)
  - Our model is better in average.
  - On some samples, our model is much better.



- BLEU-4 Score
  - Our model is slightly better in average.



- For example, our model performs much better than the baseline on `3` (left), while the baseline is much better on `13` (right).



Others

- Obtained textual data for all 35,851 journal images in the database using current models, with both traditional and simplified Chinese.

Future work

- Improve the current model.
- Experiment with other LR methods, such as encoder-decoder-based models.