BT4222 Mining Web Data for Business Insights Academic Year 2022/2023

Semester 1

Project Proposal

Utilising Generative Model to Synthesise Writer Unique Handwriting Dataset

Lam Cheng Jun	A0183311R
Nguyen An Khanh	A0200693L
Tang Kai Yi, Karin	A0221233Y
Tey Kai Cong	A0203512Y
Tong Chen Rong	A0215126R

1. Problem and Motivation

In Singapore, manually grading handwritten examinations are often tedious and adds on to teachers' intense workload [1]. Fortunately, handwriting recognition eases this process. However, handwriting recognition performs poorly with different writing styles and as such are rarely used in examinations. Therefore, we propose a solution that aims to improve the performance of recognizing individual students' handwriting such that it is reliable enough to be used.

2. Business Value and Contribution

Our proposed solution allows handwriting recognition to be used reliably in examination settings, which will drastically improve efficiency of grading. Also, handwriting recognition enables standardised collection of digital handwritten data that can be used for various other business use cases [2].

3. Proposed Solution

We propose a new method of training Optical Character Recognition (OCR) models as seen in Figure 1:

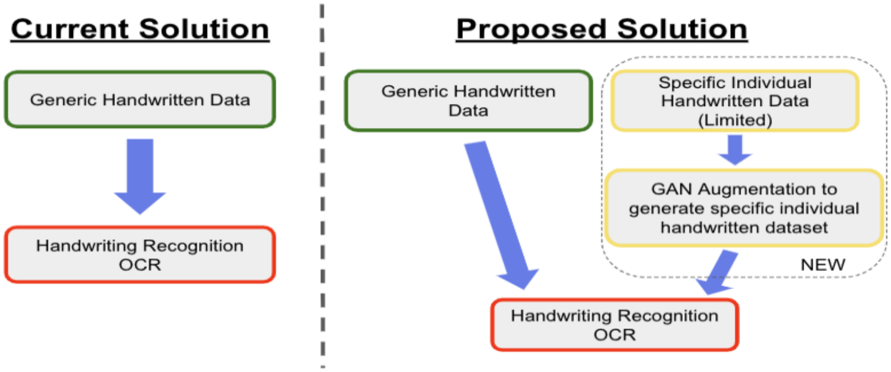
Current Solution
Generic Handwritten Data
Handwriting Recognition OCR
Proposed Solution
Generic Handwritten Data
Specific Individual Handwritten Data (Limited)
GAN Augmentation to generate specific individual handwritten dataset
NEW
Handwriting Recognition OCR

Fig. 1 *illustrating our proposed solution (right) and how it differs from the conventional solution (left)*

Rationale: Currently, generalised datasets containing different writers' handwriting are used for training, and this leads to over generalisation and poor performance regarding individual handwriting. Our solution seeks to improve the training process by training on a dataset with a specific writer's handwriting via transfer learning.

GAN Data Augmentation: We propose using a Generative Adversarial Model (GAN) to generate synthetic handwriting data that is representative of the specific writer's handwriting that we wish to recognise to combat the limited handwritten data from individuals.

4. Machine Learning (ML) & Business Success Metrics

From the ML perspective, the performance of an OCR model is measured by the Character Error Rate (CER), which is defined as the percentage of characters which are predicted incorrectly by the model [3]. From the business perspective, adoption of our pipeline depends on how the digital text output is being used: the highest precision would be demanded when used for grading an examination, whereas some errors may be permitted when archiving handwritten responses. As a proof of concept, our team claims success when an OCR model, further trained on a writer's synthetic handwriting, outperforms the baseline OCR model when tasked with predicting the same writer’s handwriting. If this concept is proven, it may be the necessary improvement OCR models require to achieve the performance threshold where educational institutions would adopt it.

5. Possible Hypotheses

The following hypotheses are listed according to their dependency, with the first hypothesis answering the second and finally answering the third hypothesis.

1. Adaptive GAN can be used to generate synthetic handwriting images that are representative of a writer’s actual handwriting.
2. Writer’s real handwriting data (median of 292 words) is insufficient to further train our OCR model to achieve better prediction outcomes.
3. OCR models can predict handwriting images with greater performance when trained only using an individual’s handwriting instead of a generic handwriting dataset.

6. Dataset and Interesting Features

The proposed dataset accumulates the IAM-OnDB dataset with newly collected samples. It contains data from 294 unique authors, for a total of 85,181 word instances and 406,956 handwritten characters [4]. Most importantly, it can provide more fine-grained word (B) and character-level (C) annotations as compared to the sentence-level (A) annotations in the original IAM-OnDB dataset (Figure 2). Moreover, the dataset offers temporal information on how each stroke is composed, which makes it more appropriate for the case study as it enables the GAN model to recognize styles in handwritings [4].

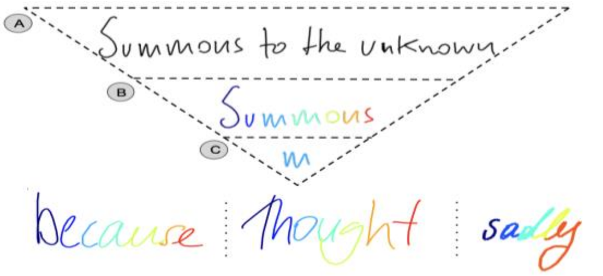

***Fig. 2**: Sequence (A), Word (B), and Character (C) level annotations provided by the dataset and a colour-coded visualisation of a sample sentence being segmented into these annotation levels [4]*

7. Data Collection

The dataset is assembled from two batches of handwritings. The first 200 authors stem from the IAMOnDB dataset [5]. Simple data cleaning was performed by removing illegible handwriting or missing annotations.

To introduce annotations of a lower granularity, samples were segmented down to character level using a commercial tool, in order to obtain their ASCII labels. In addition, a web tool was developed to collect samples from the remaining 94 authors. Each subject was tasked to write extracts of the Lancaster-Oslo-Bergen (LOB) text corpus and the data is again segmented at the character level.

8. Feature Engineering and Machine Learning Pipeline

Every writer has a unique style, and each writer displays much intra-variability in style. For example, the appearance of the letter "a" written alone and within a word may appear different. We believe that when given the task of recognising a single individual's handwriting, an OCR model needs to be trained to recognise the specific style of this writer rather than incorporating the general style of multiple writers. As such, the GAN model would generate synthetic handwriting that would contain features representative of the writer's actual style.

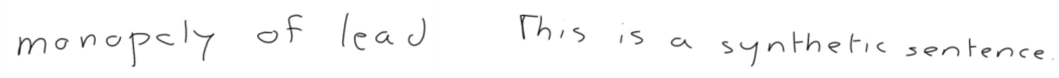

Fig. 3 *illustrating Real handwriting (left); Synthetic handwriting generated using DeepWriting Model (right)*

In conclusion, our machine learning pipeline could be illustrated by the flowchart below (Figure 4). Metadata on raw handwriting could be collected and converted into the desired format via the provided pre-processing scripts. The new input data is to be appended to the proposed dataset and the entire dataset is to be used for training. Once artificial handwritings for the user have been generated, these handwritings are fed into the OCR model and compared against the vanilla OCR model trained on a generic handwriting dataset. Finally, the more performant model would be deployed for our business use case.

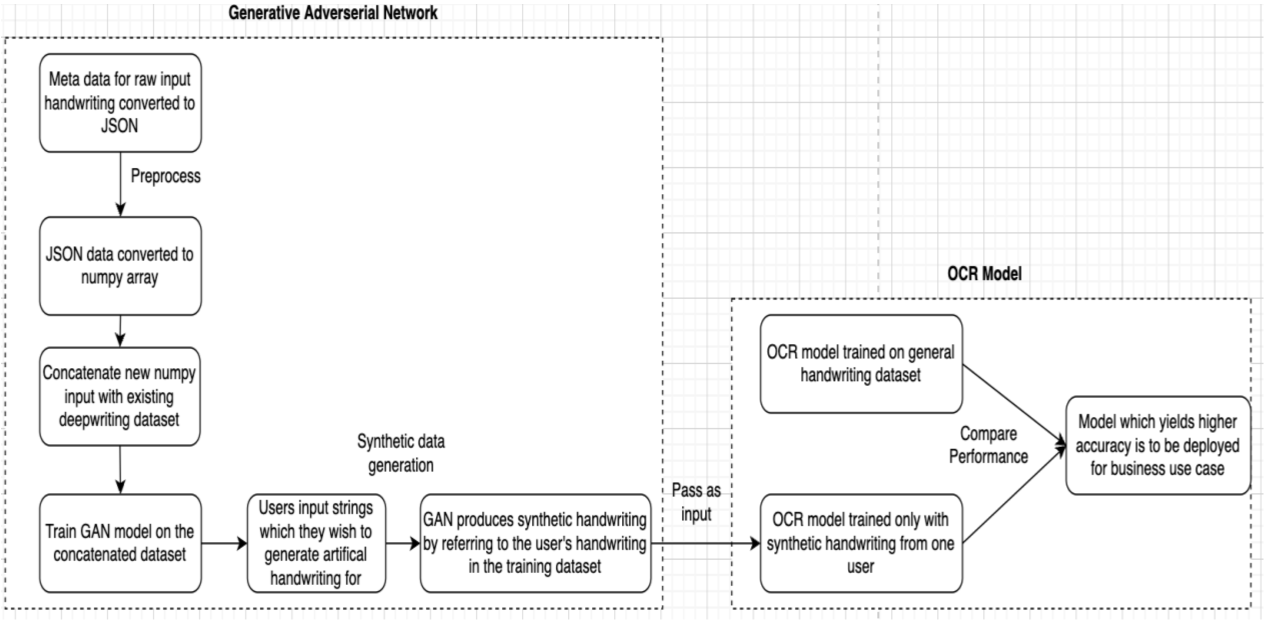
Generative Adverserial Network
Meta data for raw input handwriting converted to JSON
Preprocess
JSON data converted to numpy array
Concatenate new numpy input with existing deepwriting dataset
Train GAN model on the concatenated dataset
Synthetic data generation
Users input strings which they wish to generate artifical handwriting for
GAN produces synthetic handwriting by referring to the user's handwriting in the training dataset
Pass as input
OCR Model
OCR model trained on general handwriting dataset
Compare Performance
OCR model trained only with synthetic handwriting from one user
Model which yields higher accuracy is to be deployed for business use case

Fig. 4: *Pipeline illustrating different types of input data to be fed into the models separately*

References

[1] L. Lee, “Teachers say pay rise doesn’t solve high stress levels, workload issues - TODAY,” TODAY, Aug. 2022. [Online]. Available: https://www.todayonline.com/singapore/moe-pay-increase-teachers-workload-stress-1971726. [Accessed: Oct. 06, 2022]

[2] D. M. Adrian and HK. Emlyn, “Analysis of an automatic grading system within first year Computer Science programming modules,” Analysis of an automatic grading system within first year Computer Science programming modules, Jan. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3437914.3437973. [Accessed: Oct. 06, 2022]

[3] A. Dirk, “Word Error Rate & Character Error Rate – How to evaluate a model,” Word Error Rate & Character Error Rate – How to evaluate a model, Oct. 28, 2019. [Online]. Available: https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/word-error-rate-character-error-rate-how-to-evaluate-a-model/. [Accessed: Oct. 06, 2022]

[4] E. Aksan, F. Pece, and O. Hilliges, “DeepWriting: Making Digital Ink Editable via Deep Generative Modelling,” Jan. 2018, doi: https://doi.org/10.1145/3173574.3173779. [Online]. Available: arXiv:1801.08379

[5] INF, University of Bern, “IAM On-Line Handwriting Database,” *Research Group on Computer Vision and Artificial Intelligence & mdash; Computer Vision and Artificial Intelligence*. [Online]. Available: https://fki.tic.heia-fr.ch/databases/iam-on-line-handwriting-database. [Accessed: Oct. 06, 2022]