PROJECT PROPOSAL - RESEEPT: AN EXPENSE TRACKER THAT SORTS RECEIPTS.

Szu-Yu (May) Chen
Student# 1007091629
mayc.chen@mail.utoronto.ca

Xiangyu (Vincent) Chen
Student# 1008421824
xiangyu.chen@mail.utoronto.ca

Changhao (Tony) Weng
Student# 1006804898
changhao.weng@mail.utoronto.ca

Jijia (Georgia) Chen
Student# 1007655985
georgia.chen@mail.utoronto.ca

ABSTRACT

There are neural networks to perform optical character recognition (OCR), and neural networks to determine whether some input images should be categorized as receipts. However, these provide little help to someone who wants to track their spending over time, only has access to physical receipts, and does not wish to store the physical receipts long-term or as bulky digital images. This proposal describes ReseePT, an expense tracker project that is capable of directly scanning and logging receipt information. ReseePT will use a CNN to perform OCR, and an RNN for extraction of key information (date, time, shop name, total spending). Lastly, it will provide the users with a summary of their spending sorted by time (daily, weekly, and monthly), total spending and shop name. —-Total Pages: 7

1 INTRODUCTION

Tracking personal expenses can be a tedious task, our project aims to develop a receipt scanner that facilitates the automated extraction of expenditure details, including the amount spent, date of spending, and merchant details, with the primary objective of assisting users in efficiently tracking their monthly expenses and sorting expenses by their respective dates.

The aim of this project is to find a data set that consists of over 800 samples, which include various receipt layouts, in order to encompass a wide range of real-world scenarios. This diverse data set will not only serve to enhance the model's accuracy but also enable it to effectively handle different receipt structures and formats commonly encountered in practical applications.

The intended method is to construct a Convolutional Neural Network (CNN) that will analyze receipt images with the capability to accurately delineate the textual content, including words, numbers and individual characters, within the receipts. Building a foundation for using optical character recognition (OCR) on receipts.

As the primary objective of this project revolves around the extraction and categorization of words and numbers from receipts, multiple existing works in the market that are proficient in extracting textual and numerical data from images or can efficiently classify information into distinct categories have been identified.

The model's success can be gauged through metrics such as its accuracy and processing speed. For this project, attaining a 95% accuracy in text extraction is the threshold for success, and processing speed should not exceed 10 seconds to meet the success criteria.

The project's overall success hinges on effective collaboration and communication within our team. Success is defined as maintaining group chat responsiveness within 8 hours for all team members. Additionally, the team should have team members absent from weekly meetings less than twice by

the end of the term. Adherence to the project plan and meeting internal deadlines are also critical indicators of success.

The team should engage in a discussion of project requirements before task assignment, ensuring a clear understanding of the workload associated with each task. Once all tasks are identified, team members have the flexibility to express their task preferences. In cases where multiple members express interest in the same task and no mutual accommodation can be reached, the group will collectively decide on task assignments, possibly through an anonymous voting process. Equitable distribution of workload is paramount. If any team member perceives an uneven workload, they are encouraged to communicate this concern. The team will then evaluate the situation and consider task reassignments as necessary to ensure a balanced workload for all.

2 ILLUSTRATION/ FIGURES:

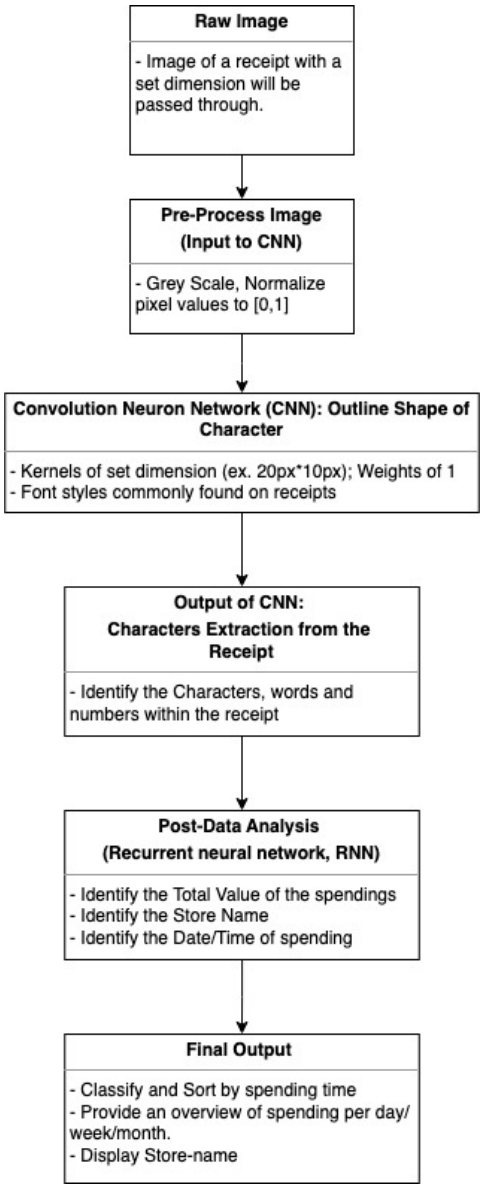

Figure 1: Flow of the Neuron Network

3 BACKGROUND & RELATED WORK

3.1 RELATED WORK 1: DEXT PREPARE

The first relevant reference identified is Dext Prepare, which employs Optical Character Recognition (OCR) and templating technology to extract written details from receipts, invoices, and financial documents. It then presents this information in an electronic format while also extracting the expense date and categorizing the expenditure. Dext (2023)

3.2 RELATED WORK 2: VERIFY

The second related reference discovered is Verify. Verify utilizes Receipt OCR technology, involving the electronic or mechanical conversion of receipt images, printed or handwritten text on receipts, and invoice documents into machine-encoded text through software. This process extracts data from images or scans of documents, such as receipts or invoices, and subsequently presents a digital version of the scanned document, preserving all the original information from the paper document.Verify (2023)

3.3 RELATED WORK 3: DOCUMENT INTELLIGENCE STUDIO

The third discovery is the Document Intelligence Studio. This tool employs Optical Character Recognition (OCR) to interpret and extract both printed and handwritten text from PDF documents and scanned images. It possesses the capability to detect paragraphs, text lines, words, geographic locations, and languages. The Document Intelligence Studio scans receipts, extracting transaction times, merchant details, tax amounts, and total expenses from them.Azure (2023)

3.4 RELATED WORK 4: CORNELL UNIVERSITY - VISUAL DOCUMENT UNDERSTANDING METHODS

This is an article from Cornell University which discusses Visual Document Understanding (VDU) methods, particularly focusing on the limitations of using Optical Character Recognition (OCR) engines to read text in documents. The paper introduces an OCR-free VDU model called Donut, designed with a simple yet effective Transformer architecture and a cross-entropy loss pre-training objective. Donut is shown to outperform existing methods in terms of speed and accuracy in various VDU tasks, addressing issues like high computational costs, language flexibility, and OCR error propagation. Kim et al. (2022)

3.5 RELATED WORK 5: ONLINE TRANSLATION APPLICATION - YOUDAO FANYI

The fifth related work is an online translation application called YouDao Fanyi. Users can upload an image they wish to scan for text extraction. The app processes the image through Optical Character Recognition (OCR) to extract all the text, followed by offering a bilingual comparison for translation.

4 DATA PROCESSING

The data that will be used contains over 900 receipts images found in the dataset online. The dataset contains 973 scanned dated receipts, This is the grouped and organized dataset of the original ICDAR 2019 SROIE competition dataset. In this dataset, the receipts record various types of expenses. For example, some receipts record the expenses from grocery stores while others record from restaurants. So, the different expenses will be recorded with their location. Also, these receipts are recorded on different dates. In this way, the expenses will be sorted by their respective dates. Self-collected data might be added(receipts from shopping, eating out, etc.) if the dataset proves to be insufficient. The bias embodied in our shopping preferences will not affect training of the OCR network, since it does operate on the textual content of the receipts.

In practical use of our end application, input noise may arise from factors such as insufficient lighting, dirty or broken camera lens, distorted text due to bad scan angle or wrinkled paper, faded text

due to low printer ink or subsequent abrasion, and blurry text due to water soaking. The ICDAR data contains most of those types of input noise, except for dirty/broken camera lenses and bad scan angles. Two types of input noise will be introduced into the transformed copies of the data (blur and shear respectively), using image processing software, as well as additional dimming to model lower light conditions.

And after clearing the input noise of data, all image will gray-scaled. After gray-scaling, all images will contain only the intensity values which will be easier to train them in CNN. For faster training with ReLU, the pixels will be normalized to the range [-1, 1]. Also, the algorithm will scale the characters to have the same font size, then the kernel size will be better determined. If necessary, the large rectangular chunks of white space will be detected and removed to optimize convolution run-time.

The output of the OCR network will be used to train our information extraction network. A template receipt text could be used with randomly generated item names and prices so that the extraction network can be trained in parallel. Knupleš (2021)

5 ARCHITECTURE

The neural network that will be utilized is the convolutional neural network (CNN) as it is independent of input size, preserves spatial relationships in the image, and is well suited to capture image patterns. The kernel sizes will be chosen to be smaller than the largest font size in the pre-processed input images. Since the date, time, and total spending are critical information that can cause trouble if logged incorrectly, the neural network must obtain results with high confidence, and perform at least as well as a human looking at the image. As such, the final receipt-scanning application should only log the results if the confidence exceeds a strict threshold, and prompt the user to re-scan (or enter the fields manually) otherwise. To achieve low confidence for noisy input, the loss function should penalize high confidence predictions when the input is noisy. To extract date, time, store name, and total spending, a recurrent neural network (RNN) that operates on the output string of network 1 will be used. For example, it should be able to extract the same date for all of the following representations: Dec. 24, 2023; December 24, 2023; 12/24/2023; 12-24-2023; 24-12-2023

Finally, the end application will allow the user to search and sort receipts by any of date, total spending, and shop name (alphabetical by default). A sorting algorithm such as radix or counting sort may be used.

6 BASELINE MODEL

The baseline model will be a hard-coded algorithm that applies convolution on the input using predefined kernels. These kernels will be approximately 20px high and 14px wide, each containing weights of 1 painted in the shape of the character it will extract (in a font (or several) commonly found on receipts), and 0s elsewhere. To account for various font sizes, the model will apply convolution using each kernel at multiple scales. The scaled kernels will be generated using traditional image processing techniques.

Since this model does not account for the multitude of fonts that receipts can contain, it is expected to perform worse than the OCR network in almost all cases.

Furthermore, since each kernel matches an exact combination of font, font size, and character, this algorithm will be much slower than the primary method. If this algorithm proves too slow to yield results, the input resolution will be lowered or run on cropped input which contains fewer pixels.

For date, time, total spending, and shop name extraction, a rule-based algorithm will be used, matching specific keywords and locations within the text string. For example, the shop name usually appears in a relatively fixed position in the string.

7 ETHICAL CONSIDERATION

An ethical concern arises when the system extracts sensitive information from publicly shared images without consent, potentially infringing on privacy, especially when scanning and extracting

data from social media images. This usage, devoid of proper consent and ethical considerations, can lead to data privacy and security issues. The data set that will be used is open to the public, therefore it does not violate any privacy issues.

The model may face limitations in safeguarding data privacy. For instance, the network's process of extracting the merchant's name from receipts, consequently revealing the places where users shop, raises significant privacy concerns. Storing or processing sensitive financial data, such as receipts that contain credit card information or personal identifiers, can also pose significant privacy risks

In addition, The United Nations Educational, Scientific and Cultural Organization (UNESCO) has created a recommendation guideline for the Ethics of Artificial Intelligence. This will be referenced and referred to throughout the process of the development of the model, and high-risk ethical risks will be highlighted and considered.

8 PROJECT PLAN

Regular meetings will be held every Thursday from 4-5 PM it will be held online. In the Meeting Schedule below, there is an outline of current planned meetings and additional meetings will be arranged if needed. (Table 8.2) The main communication platform is WeChat, questions and concerns will be addressed in the group chat and additional meetings will be arranged. To maximize effective communications, the expectations for team members are that messages sent during weekdays (between 9:00 to 22:00), will be addressed within 2-3 hours, during weekends and holidays, messages need to be addressed within 5 hours. If messages are sent later than 22:00, the message should be addressed before 12:00 the next day. The documents and figures will be stored on the shared Google Drive. Git will be used for version control, and post each person's code responsibilities in WeChat to minimize merge conflicts. The detailed project plan is in Table 8.1 below, and the meeting schedule will be in Table 8.2.

9 RISK REGISTER

There are several potential risks associated with this project which will be discussed in detail in the following section. The effects of these risks and the mitigation strategy of the team are considered.

9.1 RISK 1: TEAM MEMBERS LEAVING THE COURSE

The impact of this risk leads to increased workload and reallocating responsibilities within the group and is usually unexpected and might cause shock and conflict within the team. This will lead to a delay in the project plan and the overall progress of the project. This will be prevented by holding regular check-up meetings throughout the semester to minimize issues that might contribute to a member leaving the team. To mitigate the negative effects of this risk, the leaving member must hold a meeting with all the team members to provide up-to-date information for the most efficient transition, and also provide feedback and address potential issues that caused this. Afterwards, the team must hold a meeting to reallocate the responsibilities and communicate concerns.

9.2 RISK 2: LOW MODEL PERFORMANCE - INSUFFICIENT DATA

The dataset that is selected by the team consists of 900+ data receipts and viewing the quality of the receipts and the corresponding correct information. The data that is insufficient will lead to inaccuracy, biased predictions and high uncertainty by the CNN model, the team will refer to backup datasets or datasets that will be collected by the team as a supplement. The current dataset can also be expanded by introducing other datasets (different, fonts or receipt formats) to increase the variability for the model to learn from.

9.3 RISK 3: SECURITY AND PRIVACY OF SENSITIVE USER INFORMATION

This risk is already addressed in the ethical consideration section above. However, it is still a potential risk, and its effect and mitigation strategy will be addressed. This might lead to ethical issues and a violation of personal privacy. The model will minimize data collection from users and only

8.1 TABLE 1: PROJECT PLAN

Task Status	Title	Task	Internal Deadline	Member Responsible
Completed	Formation of Group (4)	Set up Git repository and Shared Drive	10/10/2023	Vincent
Completed	Project Proposal (Due: 10/13, 11:59pm)	Introduction, Ethical Consideration, Background and Related Work	10/13/2023	Georgia
Not Started		Data Processing	10/13/2023	Tony
Not Started		Risk Register, Illustration and Project Plan	10/13/2023	May
Not Started		Architecture and Baseline Models	10/13/2023	Vincent
Not Started	Project: Developing Neuron Network	Obtain, Organize and Cleaned Data	10/31/2023	Tony
Not Started		Complete Reasonable Baseline Model (Rough Draft)	10/31/2023	Georgia, Vincent and May
Not Started		Produce >1 qualitative and quantitative result	10/31/2023	
Not Started		Adjust Project Plan according to current progress (Meeting)	10/31/2023	All Members
Not Started	Progress Report (Due: 11/03/2023, 11:59PM)	Brief Project Description	11/2/2023 (4 PM)	May
Not Started		Notable Contribution - Data Processing	11/2/2023	Tony
Not Started		Notable Contribution - Baseline Model	11/2/2023	Vincent
Not Started		Notable Contribution - Primary Model	11/2/2023	Georgia
Not Started		Individual Contributions and Responsibilities	11/2/2023	All Members
Not Started	Project Final Report (Due: 12/01/2023, 11:59PM)	Introduction, Background and Related Work, Ethical Consideration	11/26/2023	Georgia
Not Started		Data Processing, Evaluate the Model	11/26/2023	Tony
Not Started		Architecture and Baseline Model, Discussion	11/26/2023	Vincent
Not Started		Illustration and Figures, Results	11/26/2023	May
Not Started		Submission (Template)	11/30/2023	All Members
Not Started	Project Presentation (Due: 12/01/2023, 11:59PM)	Outline and Plan Project Presentation	11/26/2023	All Members
Not Started		Record Project Presentation and Submit	11/28/2023	All Members

8.2 TABLE 2: MEETING SCHEDULE

Date	Topic of Meeting
10/07/2023	Project Proposal Discussion: Decide on a topic and allocate sections of the project proposal.
10/12/2023	Project Proposal Check-in: Check in on the progress, address any questions that have arose. Discussing and confirming the basic outline of the neuron network.
10/13/2023	Project Proposal: Finalize proposal, input into latex template, and submit the proposal.
10/15/2023	Initial Meeting: Discussion on details of approach and allocate responsibilities of developing baseline and primary model
10/19/2023	Check-in Meeting: Address any questions and concerns, make adjustments to project structure if needed
10/26/2023	Check-in Meeting:: Discussion of Project Progress Report, Address any questions and concerns, make adjustments to project structure if needed
11/2/2023	Progress Report: Check the baseline model approach and results. Finalize progress report and submit.
11/9/2023	Check-in Meeting: Address the current status of the model and make adjustments , and check results
11/16/2023	Check-in Meeting: Address the current status of the model and make adjustments , and check results
11/23/2023	Project Final Report and Presentation: Allocation of sections in the presentation and finalize data
11/30/2023	Final Adjustments and Submissions

training data will be applied to improve the hyperparameter, to prevent data breach from the model.. Also, the platform should have a choice for users to opt out of data collection and receive consent from users to collect data.

9.4 RISK 4: TIME CONSTRAINTS

This risk will more likely occur during exam weeks when there are large amounts of due dates and tests that will affect the ability to meet internal deadlines. This might cause a delay in submission and internal deadlines, and it might cause conflict within our team. The risk of time constraints will be mitigated by proposing a complete project plan and internal deadlines with a regular meeting schedule. This will prevent any unexpected workload and the internal deadlines are set to be 1-2 days prior to the deadline, this can give the team a chance to complete work that did not meet the internal deadline. The member who is unable to meet the internal deadline should notify the team prior to the due date.

REFERENCES

Azure, 2023. URL https://formrecognizer.appliedai.azure.com/studio.

Dext. what-technology-does-dext-prepare-use, 2023. URL https://help.dext.com/en/s/article/what-technology-does-dext-prepare-use.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022.

Urban Knupleš. Sroie datasetv2, DEC 2021. URL https://www.kaggle.com/datasets/urbikn/sroie-datasetv2/.

Verify, Sep 2023. URL https://www.veryfi.com/receipt-ocr-api/.