Winter Progress Report – Midterm

1. Project Purpose and Goals

Our project, CDK Data Stream AI, is concerned with the construction of a program that can arbitrarily detect if a document has a signature line and if it has been signed. To accomplish this primary task the program must make use of an open source AI platform, function as a black box, and be modular enough that parts can be added or removed from the program. Accompanying this primary goal is a series of stretch goals that, should our primary concern be easy to implement, become our focus for the remainder of the coding section. These stretch goals include license validation, data security, image categorizing, and vehicle image processing. Should these be accomplished, additional interests have been stated in regards to automatic construction of consumer portfolios based on submitted forms.

A major concern for our client is that we do not get caught up on peripheral issues or stretch goals before the core task is completed. We’ve been explicitly requested not to worry about container software, cloud operation, or user interface in particular. Another issue of which we must remain mindful is data sensitivity; any documents or forms we’re permitted to use for testing purposes cannot be made public. They are sanitized for us to legally view, but the forms are not open to the public. Lastly, CDK Global is aware that this project may not be feasible and in lieu of a completed product, they will accept a report indicating why we were unable to complete the project.

2. Recap of last terms work and progress

All required documentation for fall term has been completed and submitted. All documents requiring client verification have been cleared except for the design document which was pushed into winter term. Regular meetings have been scheduled with our TA every week for 20 minutes as well as 30 minute meetings with our client. Communication has been setup for intergroup contact through the use of group text/chat services, and contact with our client has been established through the university-provided emails. Plans have been outlined for the duration of winter break with a strong focus on code writing.

The program itself has also made some progress throughout the term. We have selected OpenCV as our launching platform for the project and will be using its C++ variant as our leading choice. Our group has successfully completed setup of OpenCV on our personal machines and student FLIP server. This includes adding the ability to view images from PUTTY. Lastly, our most recent development for OpenCV is constructing an image window for any submitted image. Plans have been made to ensure progress continues throughout winter break with the intent to have PDF to JPEG conversion included into our program as well as the ability to construct reduced windows in OpenCV to pinpoint where a signature line is located.

3. Projected goals of Winter term

Winter is the primary window for coding to be completed in and as such our lead goal for the term is to have the primary objective of this project completed by the end of the season. To accomplish this task there are several subsections we constructed towards reaching this goal to help facilitate workflow. The three core breaks came in the form of establishing a file conversion software, construction of an OCR program to isolate and crop images, and a neural network to then judge the provided image. While we allotted the entirety of the term to completing this core function we also made tentative plans for setbacks as well as stretch goals. Some problems we anticipated encountering included: insufficient language for OpenCV, neural network instantiation issues, and inability to accurately crop images down. To protect against these problems, we readied a plan to transition to Python if necessary as well as numerous resources for starting up networks as well as different ways of isolating the image without the use of OCR.

4. OCR Tesseract

//Why we are using it

//What it has been able to do

//What we intend to have it do

5. Convolutional Network

//Same as OCR

6. Problems and Solutions

//OCR difficulties on setup

//OCR reading limitations

//Transition in network design

//Acquisition of data set

7. Progress of Winter term

7.1 Week 1

7.1.1 Activities

For our first week back, the primary focus was establishing what our new schedules looked like as well as determining what meeting times needed to be adjusted. In addition, we received instruction from our instructors that the usual physical meetings for the class would be largely removed to facilitate more time to work on coding. As of this point in time we already had file conversion functioning and had isolated potential problems with the other two components. There was a very short update regarding the difficulties being had for OCR Tesseract setup and that became a focus for the team.

7.1.2 Problems

The primary source of conflict for this week was the change in scheduling. Where before we had largely similar time commitments we now have a very small window of available time to do meetings. This not only meant making inter-group meetings difficult but also meant we needed to change our meeting times with both our client and teaching assistant. The OCR issue stems from the fact our normal workplace being the school servers lacked the required installations to sufficiently run the program we needed to make.

7.1.3 Solutions

For this first week we managed to re-establish communication with each other and determine future meeting times. This included the changing of our meeting times with our client to Thursdays at 2:30pm. Our TA’s schedule also had to change so that portion of the scheduling had to wait until next week. Work also started on finding alternative methods of installation for OCR Tesseract that focused on the creation of a local virtual machine to run it.

7.2 Week 2

7.2.1 Activities

Week two was largely focused on finalizing the schedule changes that had occurred, this also meant we solidified our meeting time with the TA for 1:00pm. To handle the lack of physical meeting times for this term we constructed a new schedule for Monday, Wednesday, and Friday from 3:00-4:00pm to have our group meet to keep on track. Due to scheduling errors we did not have time to go over the neural network this week, but work did continue on setting up OCR.

7.2.2 Problems

The primary problem for this week came in the form of OCR setup again. Efforts to complete setup through a personal VM ended in failure, several version mismatches as well as limitations on the VM’s we chose meant it did not function correctly. Another method of installation that was underway included dual booting our machines into Ubuntu and installing it from there. Problems did surface when attempting this however as a result of BIOS conflicting issues that prevented Ubuntu from launching correctly.

7.2.3 Solutions

The virtual machine concept was ultimately placed on hold, two different variants of VM’s had been tested and both received problems at differing stages. The dual boot ultimately worked out, currently requiring changes in BIOS settings on startup but was able to launch correctly. From there we learned that the most current version of OCR has severe setup issues and a previous version would have to be used for our project.

7.3 Week 3

7.3.1 Activities

This was one of the more productive weeks for our group, neural network was successfully instantiated, and OCR Tesseract was now up and running. It was determined that a traditional network would be insufficient for what we needed and that a convolutional one would be more effective. This new network structure was able to correctly handle several test samples and return a strong accuracy rating. From here the primary concern with the network was adjusting its data types to handle our images and then providing it sufficiently large data sets to train off of. As for OCR it was now correctly running and able to run basic word recognition on sample texts. With that accomplished the goal for OCR was to now correctly detect where a signature is located on a form.

7.4 Week 4

7.4.1 Activities

The network was successfully shared and functioning across all machines at this point in time. We have correctly isolated an issue regarding OCR’s reading capacity when subjected to forms with extensive use of lines to segment, OCR will try and read them. Setup for OCR has been postponed for this week in an effort to prioritize more working components to showcase to our client. We’ve also established that the difference in OCR’s use of Leptonica completely isolates it from what OpenCV is doing. This will assist in modularizing the program but also means that linking them may be more difficult than we originally intended.

7.4.2 Problems

OCR will require modifications to permit the reading of forms with lines present within them as well as the changing fonts and line spacing. In addition, the modularity of the program may be too extreme, concerns have been expressed towards our ability to correctly link the OCR and network together correctly. Lastly, while it was postponed for more pressing concerns, the need for OCR to be running on more than one machine is critical.

7.5 Week 5

7.5.1 Activities

Most of this week was focused on researching how the network operates as well as determining a workaround to the problems previously noted for OCR. The network has been constructed and can handle a given data set but we are not sure yet on how the labels should be constructed and how to convert the images to a readable format for the network. Plans have been mapped out for preparing our poster and progress report, will be receiving input from client to confirm nothing they want secret is revealed. It was also decided that our GitHub repository needed to be restructured to help with organization and legibility.

7.5.2 Problems

With the alpha deadline approaching our primary concern was ensuring the link between OCR and the network could be established and that OCR can read the forms with sufficient accuracy to call an alpha state. In addition, communication with client needed to be maintained to ensure that our poster was acceptable and to the standard that CDK Global expects.

7.5.3 Solutions

While not fully tested, OCR workarounds have been found for reading the document correctly. Effort was put into determining the difficulty of removing the lines, which was deemed to great, and the use of a new function call we had previously missed. The new function call provides a line segmentation that can circumvent the previous problems and manage to isolate the signature window. This does require a few manual inputted values at present so that is now a new concern.