Light OCR

Abstract:

OCR(Optical Character Recognition) has become one of the most important and quickly developing technique given that it can easily filter and extract digital infor- mation from scanning paper. In order to realize OCR recognition with a light-weight model, we propose a deep learning based OCR method. First we train a Yolo detection model to crop paragraph (a chunk of word) from scanning image. By implementing OpenCV and image morphology method we extract character from cropped paragraph. Finally, we developed a convolu- tional neural network to classify extracted character.

Introduction:

OCR (Optical Character Recognition) engine, referred to as text recognition, extracts and presents data from scanned documents pdfs. The engine detects letters or numbers on the pdf, puts them into words and puts words into sentences. It enables users to view the pdf more clearly and edit the origin pdf files which cannot be modified. Therefore, OCR has become more and more important and a quickly developing technique. This project aims to develop a light OCR model which is less than 20M and it can reach an accuracy on the dataset indicated in the next chapter to about 70%. Deep learning models such as FCNN (Fully Convolutional Neural Network) and YOLO (You Only Look Once) will be utilized to develop the model.

Dataset Description:

Key ideas:

1. Doing the Data Preprocessing on dataset
2. Using FCNN (Fully Convolutional Neural Network) and YOLO (You Only Look Once) to train the dataset
3. Implementing some improvements on the networks (Dropout, Regularization)
4. Obtaining the suitable value for some parameters (Learning, Batch Size, and so on)
5. Reducing the model size and reaching a 70% accuracy
6. Classifying which kind the pdf is, such as resume and answer sheet

Timeline:

- First Week: Write the proposal and start training on the dataset with FCNN and YOLO
- Second Week: Improve the accuracy by adjusting network architecture and hyperparameter values
- Third Week: Complete the midterm report
- Fourth Week: Reduce the model size to be smaller than 20M
- Fifth Week: Complete the final report

Team members:

Chengtao Luo (cl6418@nyu.edu)

Kun Xiao (kx2090@nyu.edu)
Haotian Zheng (hz2687@nyu.edu)

Related Work:

Text Detection as Object Detection:

As we knew, relying on hand-crafted characteristics, Text detection is used to recognize characters literally. Deep learning has a great contribution in object detection and the method that how to tackle the text detection. When practitioners are developing the efficient text detectors, they might use Faster R-CNN models, a well-performing object detectors from conventional computer vision literature. In order to adapt the object detector to text, lengthy default boxes are introduced. Rotation-Sensitive Regression Detector is introduced as well.

End-to-end Model:

For the end-to-end methods, both text detection and transcription are used to improve. Once the text prediction has a low likelihood, it means that the detected box might not catch the complete word. Under this circumstance, an end-to-end strategy might be preferable. For instance, TextDragon, an end-to-end model developed by Feng et al. performs well on distorted text.

Dataset:
We will use the dataset FUNSD from Guillaume Jaume for our model. This dataset includes:

1. Images of scanned documents that contain a wide distribution of texts.
2. Json files (Fig 1) which store arrays of data for the scanned documents. Each entry contains a list of structures that contain data “box”, “text”, “label”, “words”, “linking”, and “id”. “Box” is used to describe the bounding box of the entire structure. The data “text” stores a string of words and/or digits, and their tokenized substrings are described inside the data “words”. The data “word” contains each substring tokenized in “text” and a set of four bounding box vertices to cover the corresponding tokenized string. The data “label” represents the category of the text. The data “linking” simply builds the connection between two different structures in the Json file, and the data “id” is used to name the structure.

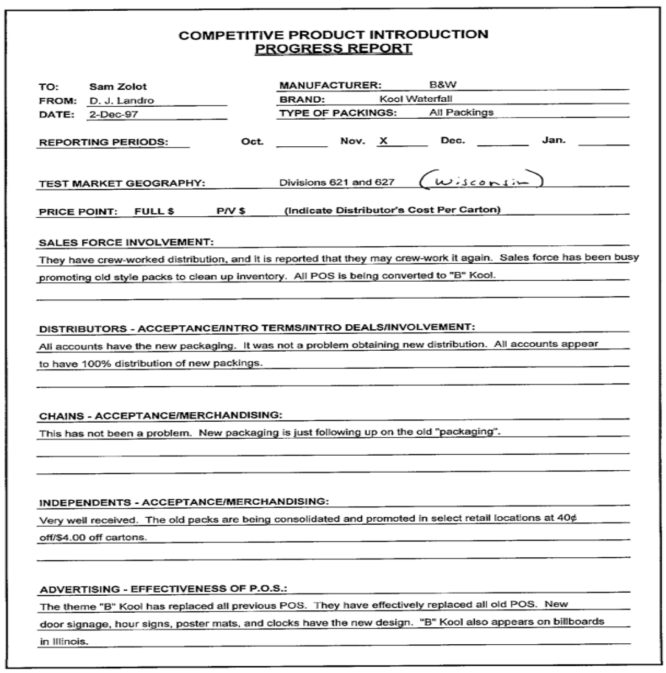
Deliverables:

1. The accuracy is about 70%
2. The model file is less than 20M

Dataset: https://guillaumejaume.github.io/FUNSD/dataset.zip

Fig 1:


```
{
    "form": [
        {
            "box": [
                94,
                200,
                114,
                214
            ],
            "text": "TO:",
            "label": "question",
            "words": [
                {
                    "box": [
                        94,
                        200,
                        114,
                        214
                    ],
                    "text": "TO:"
                }
            ],
            "linking": [
                [
                    0,
                    13
                ]
            ],
            "id": 0
        },
```

COMPETITIVE PRODUCT INTRODUCTION
PROGRESS REPORT

TO: Sam Zolot
FROM: D. J. Landro
DATE: 2-Dec-97

MANUFACTURER: B&W
BRAND: Kool Waterfall
TYPE OF PACKINGS: All Packings

REPORTING PERIODS: Oct. _____ Nov. X Dec. _____ Jan. _____

TEST MARKET GEOGRAPHY: Divisions 621 and 627 (Wisconsin)

PRICE POINT: FULL $ P/V $ (Indicate Distributor's Cost Per Carton)

SALES FORCE INVOLVEMENT:
They have crew-worked distribution, and it is reported that they may crew-work it again. Sales force has been busy promoting old style packs to clean up inventory. All POS is being converted to "B" Kool.

DISTRIBUTORS - ACCEPTANCE/INTRO TERMS/INTRO DEALS/INVOLVEMENT:
All accounts have the new packaging. It was not a problem obtaining new distribution. All accounts appear to have 100% distribution of new packings.

CHAINS - ACCEPTANCE/MERCHANDISING:
This has not been a problem. New packaging is just following up on the old "packaging".

INDEPENDENTS - ACCEPTANCE/MERCHANDISING:
Very well received. The old packs are being consolidated and promoted in select retail locations at 40¢ off/$4.00 off cartons.

ADVERTISING - EFFECTIVENESS OF P.O.S.:
The theme "B" Kool has replaced all previous POS. They have effectively replaced all old POS. New door signage, hour signs, poster mats, and clocks have the new design. "B" Kool also appears on billboards in Illinois.

PAGE 1 OF 2

82250337