Keyword Identification from Handwritten Doctor's Prescription

Akanksha Bhalla[1], Dhananjay Arora[1], Naman Manocha[1], Shubham Gupta[1]

[1]Team – Mavericks | ITCS-6156, University of North Carolina at Charlotte

Introduction

Doctors have illegible handwriting and it is difficult for patients to understand what is written in their medical prescription. Even pharmacists find it difficult to decipher the handwriting written in medical reports. However, few doctors have started to provide digitized prescriptions to maintain records, but most doctors still provide traditionally handwritten prescriptions on their printed letterhead. Thus, a lot of important data gets lost or does not get reviewed because documents(prescription) never get transferred to digital format.

There are many existing models which convert handwritten images to digital text, but our aim is to go an extra mile and identify the keywords or medical terms from the prescriptions such as the disease, medicines prescribed, medical tests suggested, etc.

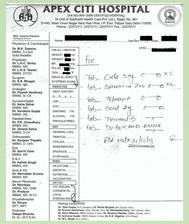

Fig : A sample Doctor's Prescription. Patient's info is hidden as medical data is very sensitive

Motivation

- All of us have come across many instances in our past, where we were not able to understand what's written in our medical report, even when we knew the context of the report as we went to doctor with known illness.
- The use of medical jargons especially describing the medical tests and medicines is a Blackbox for the patients.
- Sometimes even pharmacists find it difficult to read the handwritten notes which possess a great risk of providing wrong medicines to the patients.

Dataset

We used the IAM Handwriting Dataset for the initial training of our model. This dataset is very large, and it consisted of images of handwritten text segregated by words and sentences which we used. Apart from this, we built our own dataset using our past prescriptions. Medical Dataset is very sensitive and kept confidential. So, we planned not to share it with anyone in the future or upload to any repository.

Challenges

- Medical Data is not readily available due to sensitivity. We had to build our own dataset from scratch.
- A medical prescription consists of printed letter-head and handwritten doctor notes.
- We needed to segregate printed and handwritten text and process the handwritten text.
- Since the text is handwritten, it was skewed and slanted

Our Approach

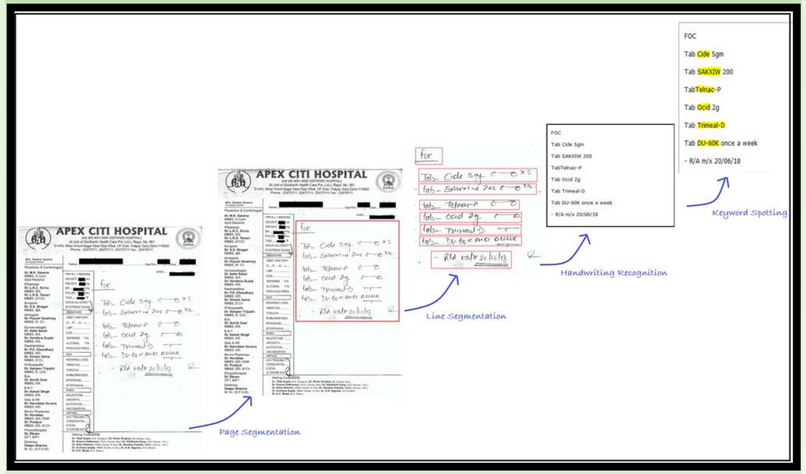

Fig: Model Overview

Preprocessing

- Rescaling
- Skew Correction
- Slant Correction
- Erosion and Dilation
- Image Thresholding

Page Segmentation

We used Hybrid block neural networks of Gluon, consisting of pre-trained ResNet-34 for the features extraction and further used convolution 2D layer with 7X7 kernel size having 3X3 padding and strides of 2X2.

Line Segmentation

For the segmentation of paragraph into the lines, we used Single Shot MultiBox Detector (SSD) with horizontal anchor boxes to identify each line from the paragraph. SSD predicts the bounding boxes and class as it processes the image simultaneously.

Handwritten Text Recognition

We built a model that generates the image features from the input image through CNN, and fed those features to LSTM sequentially. We used an encoding layer for extracting the image features from CNN and down sampled so that the features of the image are reduced. We also created a feature extraction network. Most of the previous work was done with multi-dimensional LSTM to recognize the handwriting. But we implemented that using single dimensional LSTM as it is computationally much cheaper than multidimensional.

Keyword Spotting

- We have built a repository having common medicine names, medical tests and diseases.
- Looked up closest matching words from the repository using the difflib library.
- Voila! the keyword is spotted.

Handwritten Character Recognition Model

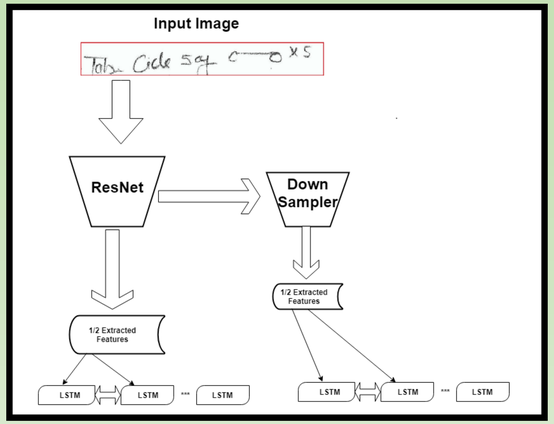

Fig: Handwriting Character Recognition Model Architecture

Future Work

- The keywords can be correlated with the Doctor's background. Right now, we did not focus much on printed text about Doctor's Qualification.
- The keyword spotting can be done in real-time. Only those words are converted to digital text, which is a medical keyword.
- This will save cost of iterating the entire list.
- We built a small prototype medical repository for this project. For future, this can be taken up as a different project which will be very helpful working in Medical Domain.

References

[1] A. Mukherjee, A. Halder, S. Nath, S. Sarkar. "A New Approach to Information Retrieval based on Keyword Spotting from Handwritten Medical Prescriptions", Advances in Industrial Engineering and Management, American Scientific Publishers, Vol. 6, No. 2 (2017),90-96, https://pdfs.semanticscholar.org/0af8/ec347b87569e3f68da04891fafce0c3f1fc8.pdf

[2] D. Nasien, H. Haron, S. Yuhaniz "Support Vector Machine (SVM) for English Handwritten Character Recognition", 2010 Second International Conference on Computer Engineering and Applications, IEEE COMPUTER SOC (2010), https://ieeexplore.ieee.org/document/5445830

[3] B. Balci, D. Saadati, D. Shiferaw "Handwritten Text Recognition using Deep Learning" (2017), http://cs231n.stanford.edu/reports/2017/pdfs/810.pdf

[4] Harald Scheidl [@githubharald], SimpleHTR, (2013), GitHub repository, https://github.com/githubharald/SimpleHTR

[5] Pablo Ruiz Ruiz, "Understanding and visualizing ResNets", Oct 8, 2018, Towards Data Science, https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8

[6] Lamy JB, Venot A, Duclos C. PyMedTermino: an open-source generic API for advanced terminology services. Stud Health Technol Inform 2015;210:924-928