# George Mason University

# DAEN 690

# Summer 2020

# **Accure Autoencoder for denoising OCR**

**Report – 6/23/2020**

**Team Autoencoder**

**Gauthami Karavi**

**Jun Wang**

**Matt Machado**

**Stephen Schade**

**Yun Li**

Table of Contents

[1 Problem Definition 3](#_Toc43656538)

[2 Assumptions 3](#_Toc43656539)

[3 Scope 3](#_Toc43656540)

[4 Risks 4](#_Toc43656541)

[5 Data Acquisition 4](#_Toc43656542)

[5.1 Overview 4](#_Toc43656543)

[5.2 Field Descriptions 5](#_Toc43656544)

[5.3 Data Context 5](#_Toc43656545)

[5.4 Data Conditioning 5](#_Toc43656546)

[5.5 Data Quality Assessment 5](#_Toc43656547)

[5.6 Other Data Sources 6](#_Toc43656548)

[6 References 7](#_Toc43656549)

# Problem Definition

Optical Character Recognition (OCR) technology enables the extraction of text data from images. There are endless uses for this type of technology, from the digital processing of historical documents to the extraction of text in personal photos. It is designed to capture numbers, letters, and punctuation and transfer these characters into data that can be searched, processed, analyzed, and stored.

OCR technology is valuable in any instance where documents or images have text. This is especially useful as society moves away from physical documents and into the digital data space. Using OCR technology, entire document archives once relegated to file cabinets can instead be harvested for digital text. Doing so creates opportunities for non-digital text to be leveraged by other computer-enabled technologies.

This technology is not perfect, and it is difficult to translate character images into a digital text without some inaccuracies or errors. Image “noise,” which can occur if image resolution or coloring issues interfere with the characters, is one reason that OCR technology may have trouble reading character text.

This project aims to build a Convolutional Neural Network (CNN) auto-encoder that reduces image noise during the OCR process. We will evaluate the fidelity of our CNN by conducting OCR text extraction on a dataset before and after the implementation of our method.

# Assumptions

This project assumes the availability of compute resources through Google Colab and the instruction and guidance of commercial and academic stakeholders. Additionally, our ability to assess the usefulness of the constructed auto-encoder relies on the availability of training data considered to be ‘noisy’ with regard to standard OCR methods.

# Scope

Our effort will develop and implement a CNN auto-encoder for application on OCR applications. The project will be based on the development and optimization of an autoencoder that will input a set of character images and classify the images.

The project will use K-means clustering, Python’s Tesseract library (i.e., PyTesseract), and denoise imager. We plan to leverage Google Colab computing resources. [1]; [2]; [3]; [4]; [5]; [6]; [7]

In addition to delivery of a prototype, the project team will evaluate our denoising model with respect to accuracy and performance.

# Risks

This project acknowledges the existence of facets that may negatively affect successful project completion. These include the unavailability of compute resources, internet access, or adequate means for team and stakeholder communication. We consider overall risk to be low.

**Risk Matrix**

A screenshot of a cell phone

Description automatically generated

# Data Acquisition

## Overview

This project will leverage the Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset [8]. This dataset is the product of academic research by Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Their 2015 paper, “Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval,” researched the ability of convolutional neural networks to conduct state-of-the-art document image classification and retrieval [9]. RVL-CDIP is a subset of the IIT-CDIP Test Collection, which was used for a 2006 conference paper titled “Building a test collection for complex document information processing [10].” The IIT-CDIP dataset is itself a subset of the Legacy Tobacco Document Library (TLDL), an archive of 14 million documents related to tobacco industry advertising, manufacturing, marketing, scientific research, and political activities. The full dataset is hosted by the University of California, San Francisco university library [11].

## Field Descriptions

The data consist of 400,000 grayscale images and 16 classes with 25,000 images per class. All images are unique. Data is classified using the following labels.

|  |  |
| --- | --- |
| 0 – letter | 8 – file folder |
| 1 – form | 9 – news article |
| 2 – email | 10 - budget |
| 3 - handwritten | 11 - invoice |
| 4 - advertisement | 12 - presentation |
| 5 – scientific report | 13 – questionnaire |
| 6 – scientific publication | 14 - resume |
| 7 - specification | 15 - memo |

## Data Context

TBD

## Data Conditioning

TBD

## Data Quality Assessment

The RVL-CDIP dataset is an academic creation that is designed to lend itself to computer vision and machine learning efforts. For each image, the largest dimension is not to exceed 1000 pixels. All are image formats and are classified by type of document as well as a predefined train, test, and validation set.

The raw dataset has several outstanding issues that need to be mitigated before final analysis can occur. The first issue is that each image in the dataset is separated into individual file folders making it difficult to extract the image into a tidy corpus for analysis. The second issue is that the data is currently in .TIF format (Tagged Image File) and may be insufficient for analysis in further steps. The third issue presented is the sheer size of the dataset corresponding with the preferred data storage on Google Drive. Google Drive uploads the information slowly and either the data needs to be scaled down for analysis only or another tool needs to be research for proper uploading. Finally, the dataset contains a few “bad” images for our analysis. These images contain little or no text and almost all noise and need to be eliminated from our purposes.

* Completeness: The RVL-CDIP dataset includes 400,000 grayscale images in 16 classes, with 25,000 images per class. Also, it is split into 320,000 training images, 40,000 validation images, and 40,000 test images.
* Uniqueness: The dataset consists 16 different classes, and each image in each class is unique.
* Consistency – Images are sized so not exceed 1000 pixels for largest dimension.
* Integrity – All images are .tif(Tagged Image File), which may be challengeable for analysis in further steps.
* Conformity – Pending confirmation with sponsor
* Accuracy – N/A(Pending)
* Overall Quality:

## Other Data Sources

We have identified several additional datasets that may also be used for our project. These datasets vary in quality, volume of data, and usefulness toward our defined project scope.

TBD

# References

|  |  |
| --- | --- |
| [1] | C. G. K. S. I. Wiraatmaja, "The Application of Deep Convolutional Denoising Autoencoder for Optical Character Recognition Preprocessing," in *International Conference on Soft Computing, Intelligent System and Information Technology*, 2017. |
| [2] | L. H. X. Zheng, "Character Segmentation for License Plate Recognition by K-Means ALgorithm," *G. Maino and G.L. Foresti (Eds.): ICIAP,* vol. II, no. LNCS 6979, pp. 444-453, 2011. |
| [3] | J. L. Wei, "Autoencoders: Neural Networks for Unsupervised Learning," Intuitive Deep Learning : Medium, 18 Febuary 2019. [Online]. Available: https://medium.com/intuitive-deep-learning/autoencoders-neural-networks-for-unsupervised-learning-83af5f092f0b. [Accessed 6 June 2020]. |
| [4] | A. Rosebrock, "Using Tesseract OCR with Python," Pyimagesearch.com, 10 July 2017. [Online]. Available: https://www.pyimagesearch.com/2017/07/10/using-tesseract-ocr-python/. [Accessed 6 June 2020]. |
| [5] | J. P. S. G. S. G. H. S. D. Shriansh Srivastava, "Optical Character Recognition on Bank Cheques Using 2D Convolution Neural Network," *Applications of Artificial Intelligence Techniques in Engineering. Advances in Intelligent Systems and Computing,* vol. 697, pp. 589-596, 2018. |
| [6] | A. I. A. Muna Ahmed Awel, "REVIEW ON OPTICAL CHARACTER RECOGNITION," *International Research Journal of Engineering and Technology,* vol. 6, no. 6, p. June, 2019. |
| [7] | M. M. N. M. Roland Graef, "A Novel Hybrid Optical Character Recognition Approach for Digitizing Text in Forms," *Lecture Notes in Computer Science,* vol. 11491, 2019. |
| [8] | A. U. a. K. G. D. Adam W. Harley, "The RVL-CDIP Dataset," Carneigie Mellon University Department of Computer Science, 2020. [Online]. Available: https://www.cs.cmu.edu/~aharley/rvl-cdip/. [Accessed 21 June 2020]. |
| [9] | A. U. a. K. G. D. Adam W. Harley, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval," *ICDAR,* 2015. |
| [10] | G. A. S. A. O. F. D. G. a. J. H. D. Lewis, "Building a test collection for complex document information processing," in *Proc. 29th Annual Int. ACM SIGIR Conference*, 2006. |
| [11] | University of California, San Francisco, "The Legacy Tobacco Document Library (LTDL)," 2007. [Online]. Available: http://legacy.library.ucsf.edu/.. [Accessed 21 June 2020]. |