# George Mason University

# DAEN 690

# Summer 2020

# **Accure Autoencoder for denoising OCR**

**Report – 6/16/2020**

**Team Autoencoder**

**Gauthami Karavi**

**Jun Wang**

**Matt Machado**

**Stephen Schade**

**Yun Li**

Table of Contents

[1 Problem Definition 3](#_Toc42378948)

[2 Assumptions 3](#_Toc42378949)

[3 Scope 3](#_Toc42378950)

[4 Risks 4](#_Toc42378951)

[5 Possible Data Sources 4](#_Toc42378952)

[6 References 4](#_Toc42378953)

# 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 There are endless uses for this type of technology, from the digital processing of historical documents to the extraction of text in personal photos.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.

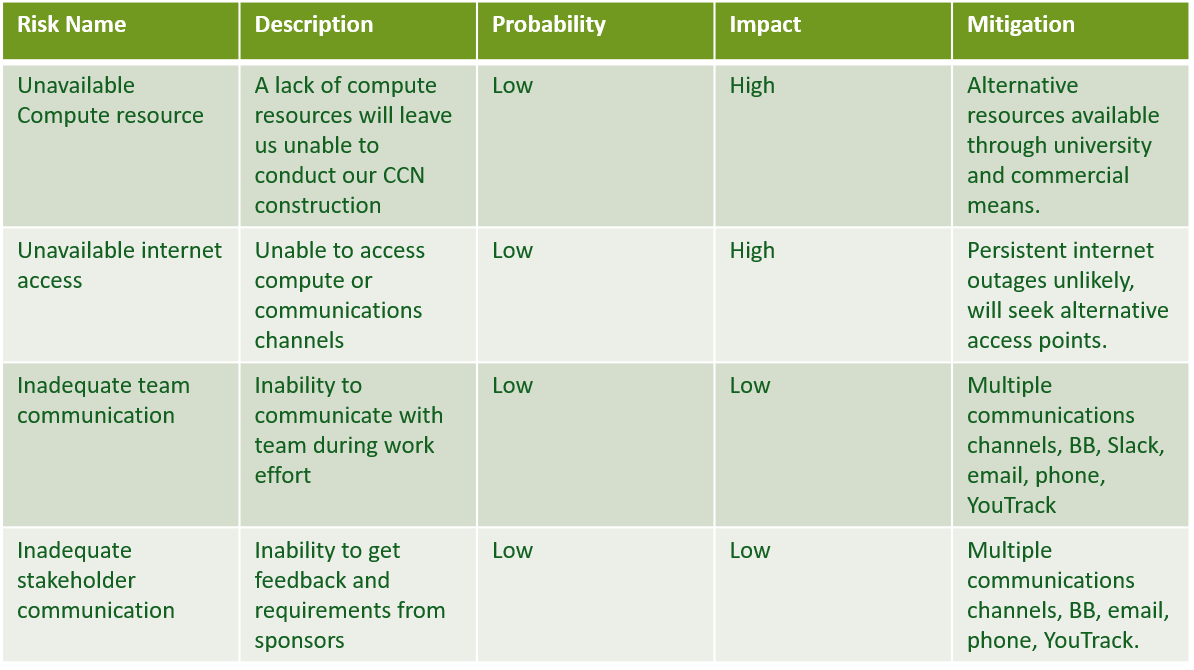
The project will use K-means clustering, Python’s Tesseract library (i.e., PyTesseract), and denoise imager. We plan to leverage Google Colab computing resources. [1]; [2]; [3]; [4]; [5]; [6]; [7]

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**



# Possible Data Sources

The team is in the process of evaluating datasets for our project. Final determination on datasets will be made by completion of Sprint 1.

The University of California, Irvine Machine Learning Repository hosts multiple OCR-related datasets, including handwritten digits and others. <https://archive.ics.uci.edu/ml/datasets.php>

Kaggle hosts datasets associated with a competition on OCR denoising. <https://www.kaggle.com/c/denoising-dirty-documents/data>

The National Institute of Standards and Technology (NIST) host the “Structured Forms Reference Set of Binary Images” dataset. 5,590 images of completed structured form faces. <https://www.nist.gov/srd/nist-special-database-2>

# 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. |