3-31-2021

Digitizing Historical Forest Service Data

Floriana Ciaglia
Boise State University

Chinwendum Njoku
Boise State University

Issac Bard
Boise State University

Digitizing Historical Forest Service Data

Abstract

When ecologists are working in the field, they often record their data on datasheets by hand. This hard-won information then tends to remain trapped in physical copies of datasheets which then get stored into filing cabinets, preventing further analysis. We are collaborating with the Sawtooth National Forest Service, which has collected decades of data on historical vegetation and soil conditions in the Sun Valley, Idaho area to digitize their historical data. The goal of this project is to create an Optical Character Recognition (OCR) model able to process the collected handwritten datasheets and generate a digitized version of them. By making nearly a century of environmental data ready for statistical analysis, this project will allow Forest Service and BSU scientists to answer important questions about how some of Idaho's most spectacular landscapes have been affected by climate change, sheep grazing, and natural resource management decisions across areas and timeframes that were previously impractical to tackle.

Digitizing Historical Forest Service Data

Floriana Ciaglia, Chinwendum Njoku, Isaac Bard
Advisor: Dr. Catherine Olschanowsky, Dr. Kelly Hopping

adapt
DATA FLOW OPTIMIZATIONS LAB

BOISE STATE UNIVERSITY
COLLEGE OF ENGINEERING
Department of Computer Science

1. Problem Statement

- Ecologists record vegetation data by hand onto physical paper-sheets.
- Historical Forest Data is inaccessible for further analysis and research.

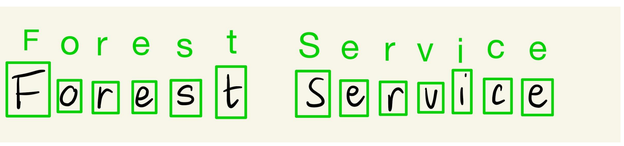
2. Motivation

Vegetation and soil condition data from the Sun Valley, Idaho area has been collected **by hand** and is laying into dusty filing cabinets.

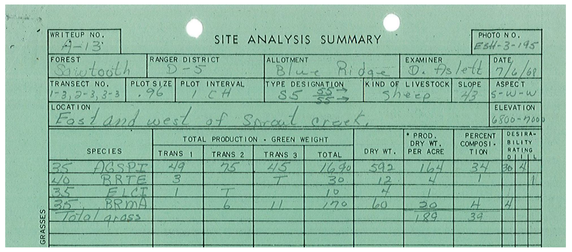
The goal of this project is to **digitize** the data forms to make them available for future scientific research.

3. Optical Character Recognition (OCR)

- Processes image.
- Recognizes ASCII characters in the provided image.
- Extracts the character and saves it into a machine-encoded text.



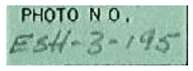
4. Process



Original data format

Ste 1. Identifying sub-fields in the form

- Extract sub-fields from the form using the OpenCV library.
- Each (x, y) coordinate is stored into a JSON file.



Field

Step 2. Bounding box around single characters

- Crop the image around each single character to feed to the model.

Field → Individual Cells

Step 3. Character Classification

- Feed the pre-processed images to a neural network (NN) to classify them.

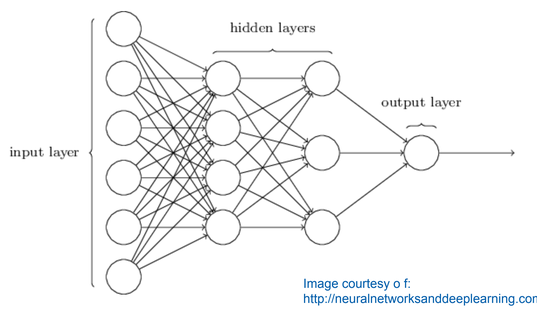


Image courtesy o f: http://neuralnetworksanddeeplearning.com/

5. Models

ResNet

The model used categorical cross entropy loss function, 50 epochs, the SGD optimizer.

The ResNet was trained on the MNIST and the Kaggle datasets.

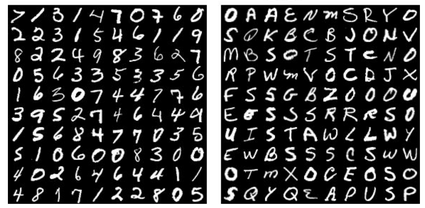


Image courtesy of: https://www.pyimagesearch.com/

Permutations of Convolutional Neural Networks

Five permutations of different models: different amount dense layers, convolutional layers, neurons per layers and dropout, 10 epochs.

The CNN was trained on the EMNIST dataset.

image courtesy of: https://www.researchgate.net/

6. Future Development and Challenges

Pre-process harder words

The letters "S" and "a" are connected by the same hand stroke..

- Letters in hand-written text are often connected by the same hand stroke.
- Learn how to preprocess images where words have connected letters.

Load data into database

- Previous work has been done to create a database where to store the collected digitized data from the Forest Service forms.
- Automate process to load the digitized data into database.

7. Acknowledgements

Boise State's Research Computing Department. 2017. R2: Dell HPC Intel E5v4 (High Performance Computing Cluster). Boise, ID: Boise State University. DOI: 10.18122/B2S41H