OCR Project Report

Claudia Wisniewski, Fausto Frisenna, Nikolina Karamarko, Kazem Zhour

September 25, 2024

Contents

Optical recognition is a field that consists of extracting the text out of an image in order for it to be manipulated by a computer. Even though paper material is a natural way of communication for humans, it is not understandable by a computer. For this reason, a lot of insurance companies focus on this type of tool in order to OCR (Optical Character Recognition) process handwritten documents (birth certificates, sale certificates) The goal of this project is to use a Deep Learning algorithm (Computer Vision) to recognize the characters within PDF/PNG files.

1 Report 1: introduce, describe, visualize and manipulate your dataset

1.1 Introduction to the project

Context

- Context of the project's integration into your **business**.
- From a **technical** point of view.
- From an **economic** point of view.
- From a **scientific** point of view.

Objectives

- What are the main objectives to be achieved? Describe in a few lines.
- For each member of the group, specify the level of expertise around the problem addressed.
- Have you contacted business experts to refine the problem and the underlying models? If yes, detail the contribution of these interactions.
- (Are you aware of a similar project within your company or in your entourage?

What is its progress? How has it helped you realize your project? How does your project contribute to improving it?)

1.2 Understanding and Manipulation of Data

Framework

- Which set(s) of data(s) did you use to achieve the objectives of your project?
- Are these data freely available? If not, who owns the data?
- Describe the volume of your dataset.

Relevance

- Which variables seem most relevant to you with regard to your objectives?
- What is the target variable?
- What features of your dataset can you highlight?
- Are you limited by some of your data?

Pre-processing and Feature Engineering

- Did you have to clean and process the data? If yes, describe your treatment process.
- Did you have to carry out normalization/standardization type transformations of your data? If yes, why?
- Are you considering dimension reduction techniques in the modeling part? If yes, why?

Visualizations and Statistics

- Have you identified relationships between different variables? Between explanatory variables? And between your explanatory variables and the target(s)?
- Describe the distribution of these data, distribution, and outliers (pre/post-processing if necessary).
- Present the statistical analyses used to confirm the information present on the graphs.
- Conclude the elements noted above, allowing them to project themselves into the modeling part.