Master's Program in Data science
Semester 2022

Redaction detection

Author: Ammar Alhashmi

Author: Ammar Alhashmi
Main Supervisor: Dr. Maarten Marx

Contents

1 Methodology

This chapter is about methodology and discusses research strategies, data collection methods and data analysis methods.

1.1 Dataset

The dataset consists of 365 PDF files which contain a total of 32 thousand textual pages. The documents can be found and downloaded from the dutch government website webcovid19 (https://wobcovid19.rijksoverheid.nl/). The pdf documents contain information such as decisions from Dutch ministries concerning corona regulations. The pdf documents mostly contain textual data but due to the privacy laws such as article 10 [1] that protects people personal data, non-linguistic elements are also present in the documents. Non-linguistic elements such as redaction regions that cover personal information.

1.2 Dataset preparation

To prepare the dataset, a few steps needed to be done. The first step was to write a script that would download the pdf documents from the government website. The next step would be to convert the downloaded pdf documents to images. Considering the relatively huge number of pdf documents, a common laptop would take up to two days to download the pdf documents and turn them into images. After that, the next step was to write a script that create a training and testing dataset from the images and extract the image segments. The script should be able to detect shapes such as quadrilaterals or pentagonal shapes because most redaction regions have such shapes.

1.3 Redaction detection model implementation

A 3-layer CNN model will be used on the images dataset, with the current image dataset a CNN model would have a high accuracy [2] . To optimize the output of the model, some regularization and an Adam are needed such as relu or sigmoid functions. For higher accuracy, the network should be trained independently on each type of image sample. If a new image shape is introduced, a new network can be established.

1.4 Experimental setup

Before building and training a CNN model that detects redaction in a dataset that consists of documents. The documents must be converted into images because models such as CNN work better on image features. After having the dataset ready. The dataset should be randomly split into a training, validation, and test set. To reduce overfitting, the hyperparameter epochs can be set to 100 and the training will be stopped if the validation loss does not improve after 10 epochs. While building the CNN model, 3 layers must be added after that regularization functions and Adam must be added to optimize the output. The model was implemented using Python, TensorFlow, and Keras software.

1.5 Evaluation

To measure the performance of the mode that handles segmentation problems, the panoptic quality(PQ) is used. The PQ is a combination of the IoU(Intersection over Union) and AP(Average Precision). The PQ consists of two steps. The first step is matching the predicted and the ground truth segment. The second step is mathematically calculating the PQ using the PQ formula. [3]

Bibliography

[1] Ministry of the Interior and Kingdom Relations. De nederlandse Grodwet. 2022.

[2] Praffil Mishra. Why are Convolutional Neural Networks good for image classification? 2019.

[3] Harshit Kumar. Introduction to Panoptic Segmentation: A Tutorial. 2019.