greatlearning
Learning for Life

Additional Case Study

Image Denoising using Autoencoders

Domain

Computer Vision, Image quality

Business Context

Optical Character Recognition (OCR) is the process of getting typed or handwritten documents into a digitised format. If you've read a classic novel on a digital reading device or had your doctor pull up old healthcare records via the hospital computer system, you've probably benefited from OCR.

OCR makes previously static content editable, searchable, and much easier to share. But, a lot of documents eager for digitisation are being held back. Coffee stains, faded sun spots, dog-eared pages, and a lot of wrinkles are keeping some printed documents offline and in the past.

Problem Statement

Getting useful information out of the images which contain a lot of noise is a challenge. In this case study, we will learn how we can use Autoencoders to solve this problem.

Commented [1]: Add Problem statement.

Dataset

Commented [2]: check

Commented [3]: add 2-3 sample images

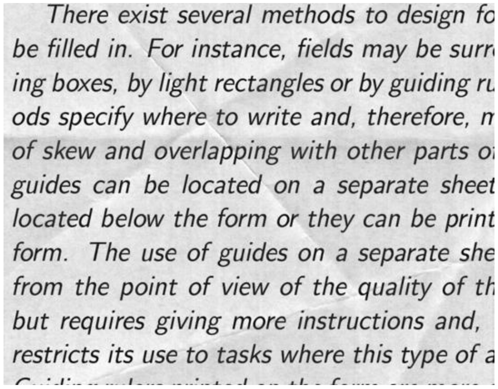
Commented [4]: done

A dataset of images of scanned text. These images contain various styles of text, to which synthetic noise has been added to simulate real-world, messy artefacts. The training set includes the test without the noise (train_cleaned)

- Noisy images: 144
- Cleaned images: 144
- Test images(noisy): 72

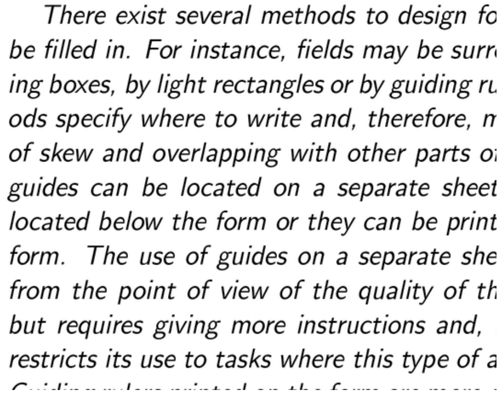
Sample Images

Noisy - Train



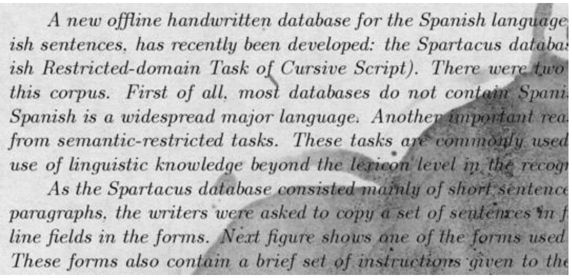
There exist several methods to design fo
be filled in. For instance, fields may be surr
ing boxes, by light rectangles or by guiding ru
ods specify where to write and, therefore, m
of skew and overlapping with other parts o
guides can be located on a separate sheet
located below the form or they can be print
form. The use of guides on a separate she
from the point of view of the quality of th
but requires giving more instructions and,
restricts its use to tasks where this type of a
Guiding rulers printed on the form are more

Cleaned - Train



There exist several methods to design fo
be filled in. For instance, fields may be surr
ing boxes, by light rectangles or by guiding ru
ods specify where to write and, therefore, m
of skew and overlapping with other parts o
guides can be located on a separate sheet
located below the form or they can be print
form. The use of guides on a separate she
from the point of view of the quality of th
but requires giving more instructions and,
restricts its use to tasks where this type of a
Guiding rulers printed on the form are more

Noisy - Test



A new offline handwritten database for the Spanish language
ish sentences, has recently been developed: the Spartacus databa
ish Restricted-domain Task of Cursive Script). There were two
this corpus. First of all, most databases do not contain Spani
Spanish is a widespread major language. Another important rea
from semantic-restricted tasks. These tasks are commonly used
use of linguistic knowledge beyond the lexicon level in the recog
As the Spartacus database consisted mainly of short sentence
paragraphs, the writers were asked to copy a set of sentences in f
line fields in the forms. Next figure shows one of the forms used
These forms also contain a brief set of instructions given to the

Steps

1. Extract data from zip files

2. Look at the dataset in more detail
    a. Set data directory path variables
    b. Check the number of images in train, train_cleaned & test folders
    c. View feature and label
3. Define a function to load the images and save them into NumPy array
    a. Get the NumPy arrays for features and labels using the above function
4. Split data into training and validation
5. Define the autoencoder model
6. Compile the model
7. Summarize the model
8. Fit the model
9. Predict on test data

Explore more

- Explore if you can implement autoencoders on a different dataset
- Build more complex and deep architectures to get better results

Happy Learning!