

Deep Learning - Case Study

Noise Remover from Text Image

Name : Deepak Shah

Enrollment Number : 18012011101

BATCH : DL1

1. Introduction

Image denoising is to remove noise from a noisy image, so as to restore the true image. However, since noise, edge, and texture are high frequency components, it is difficult to distinguish them in the process of denoising and the denoised images could inevitably lose some details. Overall, recovering meaningful information from noisy images in the process of noise removal to obtain high quality images is an important problem nowadays.

2. Tools and Technology

AutoEncoder

Autoencoders are neural network architectures that consist of two sub-networks, namely, encoder and decoder networks, which are tied to each other with a latent space. Autoencoders were first developed by Geoffrey Hinton, one of the most respected scientists in the AI community, and the PDP group in the 1980s. Hinton and the PDP Group aimed to address the “backpropagation without a teacher” problem, a.k.a. unsupervised learning, by using the input as the teacher. In other words, they simply used feature data both as feature data and label data.

3. Dataset

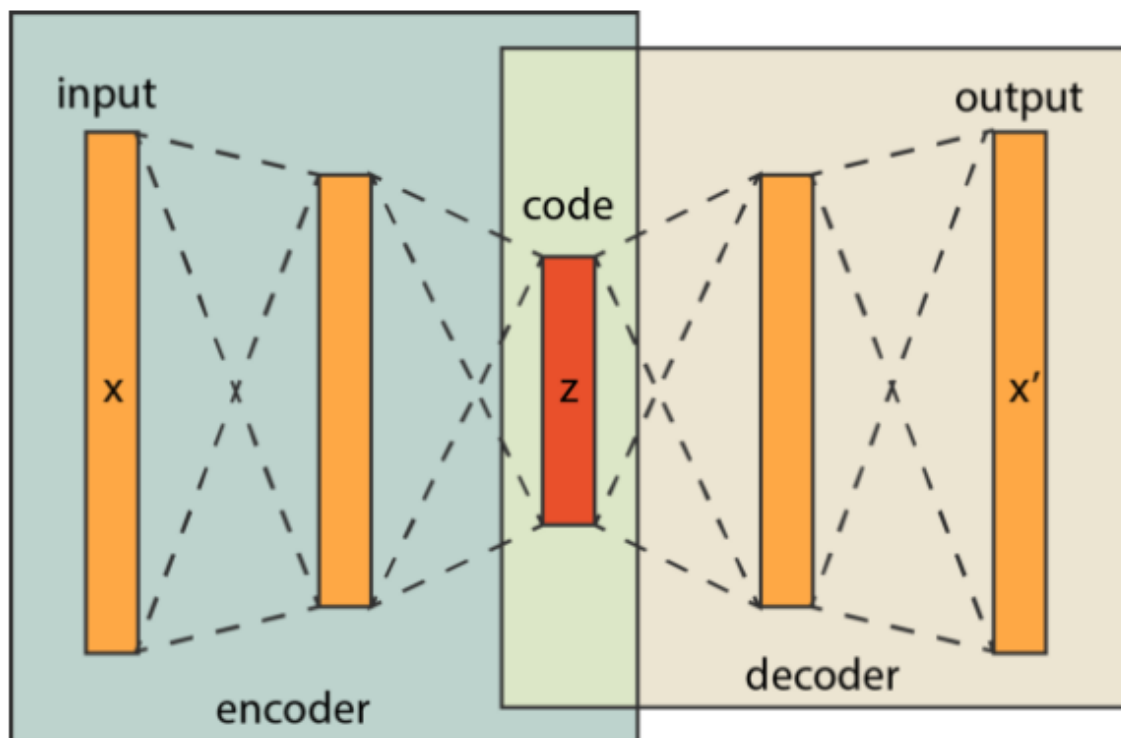
This dataset contains 600 scanned images that are not straight and very noisy. The scanned images are separated into 500 training and 100 test set images - the names of which are provided, as a starting point for building a predictive model.

This dataset consists of 600 scanned page images. The scans are documents containing text written in Latin with the header 'Surety' at different line justifications. Although the text on each document scan differs somewhat, all the documents have two things in common: they have been scanned at non-vertical angle and contain sparse arrangements of speckles across their pages. With some preprocessing, the noise of the scanned documents can be fixed.

Additionally, the dataset contains the scanned angles of the 500 / 600 of the scanned documents meaning that a predictive model can be design to predict the rotation angles of unlabelled scans based on the labelled ones.

4. AutoEncoder Architecture

Autoencoders consists of an encoder network, which takes the feature data and encodes it to fit into the latent space. This encoded data (i.e., code) is used by the decoder to convert back to the feature data. In an encoder, what the model learns is how to encode the data efficiently so that the decoder can convert it back to the original. Therefore, the essential part of autoencoder training is to generate an optimized latent space.



5. Code

https://github.com/Deepak0612/Noise-remover-from-text-image/blob/main/Noise_Remover_from_Text_Image.ipynb

6. Output

A new offline handwritten database for the Spanish language (Spanish Restricted-domain Task of Cursive Script). There were two main goals in creating this corpus. First of all, most databases do not contain Spanish. Spanish is a widespread major language. Another important reason was the need for semantic-restricted tasks. These tasks are commonly used for the evaluation of the use of linguistic knowledge beyond the lexicon level in the recognition of handwritten text. As the Spartacus database consisted mainly of short sentences and paragraphs, the writers were asked to copy a set of sentences in five-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

A new offline handwritten database for language, which contains full Spanish sentences, has been developed: the Spartacus database. The Spanish Restricted-domain Task of Cursive Writing (SRTWC) were two main reasons for creating this corpus. The most databases do not contain Spanish sentences. Spanish is a widespread major language. A reason was to create a corpus from semantic tasks. These tasks are commonly used in practice of linguistic knowledge beyond the lexicon and syntax process.

A new offline handwritten database for language, which contains full Spanish sentences, has been developed: the Spartacus database. The Spanish Restricted-domain Task of Cursive Writing (SRTWC) were two main reasons for creating this corpus. First, most databases do not contain Spanish sentences. Second, Spanish is a widespread major language. A third reason was to create a corpus from semantic and syntactic tasks. These tasks are commonly used in practice to evaluate the use of linguistic knowledge beyond the lexicon in the recognition process.

A new offline handwritten database for language, which contains full Spanish sentences, has been developed: the Spartacus database (Spanish Restricted-domain Task of Cursive Acquisition). There were two main reasons for creating this corpus. First, most databases do not contain Spanish sentences. Second, Spanish is a widespread major language. A third reason was to create a corpus from semantic and syntactic tasks. These tasks are commonly used in practice to evaluate the acquisition of linguistic knowledge beyond the lexicon and grammar acquisition process.