Semester Project

Paintings Denoising and Anomaly Detection Using Auto-Encoders

Part2

Datasets and Dataset Curation

Jiaxin Xu

INTRODUCTION

The task at hand is to build neural networks to remove high-frequency noise from images and to do anomaly detection. The solution will utilize the concept of auto-encoders (AE), which are self-supervised neural networks that can learn to encode and decode input information. The goal is to train neural networks to effectively remove noise from the inputs while preserving the important details and semantic information and to robustly detect abnormal images.

Datasets Description

(1) The Fashion MNIST dataset

The Fashion MNIST is a great dataset to start with. It consists of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image associated with labels from 10 classes. The Fashion MNIST dataset can be directly loaded from the Pytorch module `torchvision.datasets`. However, in order to better understand the "tensor" and "dataloader" concepts in Pytorch, the Fashion MNIST dataset is loaded through two separate csv files for training and testing, which is downloaded from Kaggle (dataset link (*1*) can be found in the reference) and then inheritance from Pytorch Dataset class is implemented. Both the training and testing csv file have 785 columns, where the first column is the label (class), from 0-9, and the remaining columns are the pixel values (from 0-255) represented in a row vector. Fig.1 shows an example data in the training dataset, which is a "Pullover", corresponding to label "2".

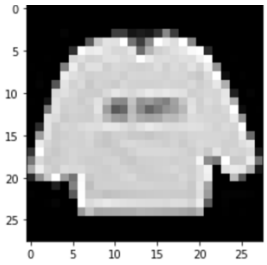

Fig. 1. Example data in the Fashion MNIST.

(2) The Edvard Munch Paintings dataset

The Edvard Munch Paintings dataset (*2*) contains a collection of 1769 color paintings created by Norwegian painter Edvard Munch. It also has a summary table of each painting's name, year, location, technique, and size. Each example has a different image size. With much less and more complicated data (colorful) than the Fashion MINST, this is a good example of a more realistic situation to test the practical application of AE. The images are different sizes, and they need to all be the same size for training. So, some image transformations are carried out as shown in Fig. 2 and Fig. 3, where all the paintings are resized and cropped to a size of 256x256.

Fig. 2. Example of original painting data in Edvard Munch Paintings dataset

Fig. 3. Example of transformed painting data in Edvard Munch Paintings dataset

(3) The Van Gogh Paintings dataset

The Van Gogh Paintings dataset (*3*) contains a collection of 2025 color paintings created by a Dutch post-Impressionist painter Vincent Willem van Gogh. Although it has different classes assigned, it will not be leveraged in this project. This whole dataset set will be used for the anomaly detection task, as an abnormal image mixed with the Edvard Munch Paintings dataset. Same to the Edvard Munch Paintings dataset, some image transformations are carried out as shown in Fig. 4 and Fig. 5.

Fig. 4. Example of original painting data in Van Gogh Paintings dataset

Fig. 5. Example of transformed painting data in Van Gogh Paintings dataset

(4) Dataset split

For the denoising task, the Fashion MNIST data already has a training and testing split, and a random 20% of the training is planned to be further split as validation. The painting data will be randomly split into training:validation:testing = 3:1:1. As for the anomaly detection task on the painting data, 80% of the Edvard Munch Paintings dataset will be used to train the AE and select a threshold of reconstruction loss. And the remaining 20% of the Edvard Munch Paintings dataset will be mixed with a random sampled subset of Van Gogh Paintings dataset (which has the same number of data as the Edvard Munch Paintings test dataset) to form the test data.

References

1. Fashion MNIST, (available at https://www.kaggle.com/datasets/zalando-research/fashionmnist).

2. Edvard Munch Paintings, (available at https://www.kaggle.com/datasets/isaienkov/edvard-munch-paintings).

3. Van Gogh Paintings, (available at https://www.kaggle.com/datasets/ipythonx/van-gogh-paintings).