Assignment 2 Report

Henrique Lacreta Alves

Abstract

This report contains a brief analysis on Assignment 2 results, the methods used, and points in which better results could be evaluated with given improvements.

1. Introduction

The second Machine Learning assignment had the objective of exploring label-based Machine Learning algorithms by tackling the problem of identifying which camera model was used to take certain pictures. This is a real-world forensics problem, in which certain photograph characteristics must be used to understand from which camera model it came from.

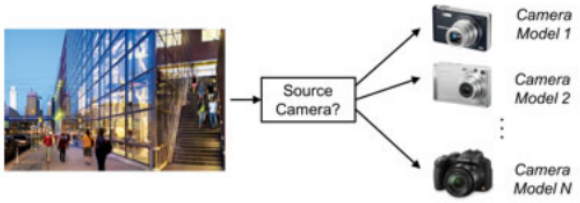

Figure 1. The problem of identifying a photograph camera source.

.

2. Activities

The best prior solutions to this problem are known to use Deep Neural Networks, as this is a good candidate problem for the number of features it presents on each dataset. Other supervised-learning approaches include Logistic Regression, which was used for this assignment.

As said before, this is a problem which can be cursed with a huge number of features if not approached properly - so it's common in the literature to extract only the noise-features from the images, as it is the real information that can be used to analyse from which camera that picture came from (digital cameras from different models have different fabrication process, which make the noise form each work in different set of conditions).

3. Proposed Solutions

The solution proposed on this assignment is to first extract the right features from the image using noise-analysis, than using Logistic Regression on the data extracted to train and predict correctly the pictures-camera model pairs.

4. Feature Extraction

As said before, the features that have the most physical characteristics information from the cameras are the noise from the picture, so first it should be collected from the dataset.

To make the process faster, and after analysing the test data, the first transformation used on each image was to crop it on the center on a 256x256 subimage. The loss of information in this case won't be as high because most pictures are taken with the focused subject in the center of the picture (or in close proximity), and picture noise is a global characteristic on the picture (influenced by luminosity and reflectivity).

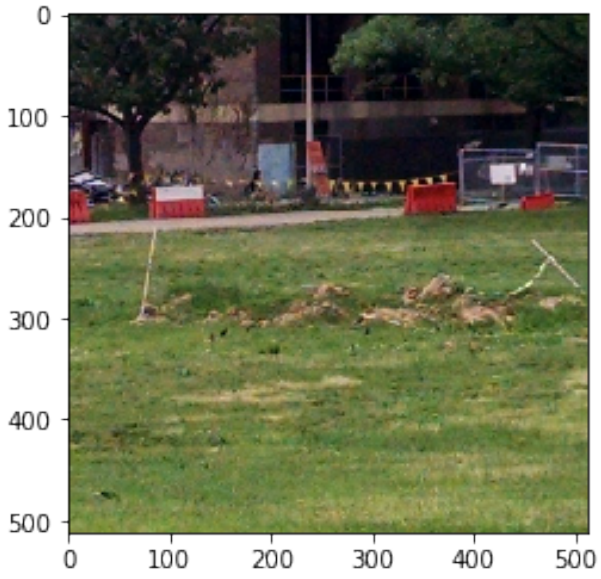

Figure 2. A cropped image example.

.

Second, a median filter is used on each picture channel to create a new subimage without noise. This new subimage is used to subtract from the original one, creating a picture depicting the noise of the original image.

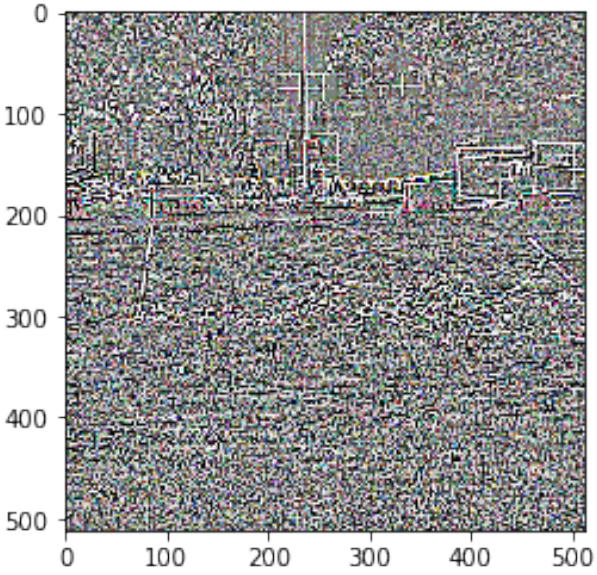

Figure 3. The noise image result.

.

It must be noted that figure 3 was prepared for visualization on the report; most of its values were lower than 0 from the subtraction, and only few spots of noise would be visible. This wouldn't influence the Logistic Regression algorithm, as it would normalize the dataset before anyway.

It was used the Median filter in particular because of its properties of filtering salt-and-pepper kind of noise without losing much of contour information, which would make the result 'noised' image more revealing on this aspect.

After getting the noise image, a discrete wavelet transformation was used on it, resulting on a wavelet filtered image and 3 components for each direction gradient. Finally, statistic information is gathered from each channel from the noise and all images resulted from it: the minimum value, the maximum value, mean, variance, skewness and kurtosis. In total, 90 features are extracted from each individual training image ((noise + 4 x wavelet images) x 3 channels x 6 statistic features).

5. Logistic Regression

Using the features extracted, a Logistic Regression model was trained with regularization strength of 1e-5 and maximum number of iterations 100.

6. Results

Since the test data didn't have the label for each image, the accuracy was tested using 20% of the training data after shuffling the order of the images; the coefficients got from the Logistic Regression made 52.35% of accuracy.

7. Conclusion

After setting the use of Logistical Regression as the model for the problem at hand, the most important focus of the analysis was the extraction of features from the training set. Using wavelets, it was possible to replicate information from the images in good amounts, resulting in a fairly successful accuracy for the training data.

Better results could be evaluated by increasing the number of features extracted; other filters besides median could be used, and increasing the number of wavelet filtering on the subimages. As the number of features increased, the PCA algorithm would become necessary as to reduce the dimensionality of the problem.