Machine Learning Engineer Nanodegree
Capstone Proposal: Banknote Identification of Fake Currency

Anurag Agarwal

April 22, 2017

Domain Background

Every major economy in the present times has to deal with the menace of counterfeit currency. Counterfeit currency is the imitation currency produced without the legal sanction of the state or government. Some of its ill-effects on society include a reduction in the value of real money, and increase in prices (inflation) due to more money getting circulated in the economy. This is the reason that fake currency has been used by governments to wage economic wars against each other. Examples include a high-profile counterfeit scandal which came to light in Hungary in 1926, when several people were arrested in the Netherlands while attempting to procure 10 million francs' worth of fake French 1000-franc bills which had been produced in Hungary [1]. According to the US Department of Treasury, an estimated $70 million in counterfeit bills are in circulation in America [2]. The fake currency being produced today is of very high quality and hard to distinguish from real currency.

Therefore, governments round the globe are adopting anti-counterfeiting measures to fight this menace and remove such illegitimate currency from the system. Use of computers and technology has also facilitated its easy identification. Research is being carried out so that it becomes easy for people to distinguish fake currency from real one with help of their mobile devices [3]. This will be a deterrent for people or governments who are indulging in such malpractices. Therefore, a project in this direction is a necessity.

Problem Statement

Given the huge threat posed by the surge in large amount of high quality counterfeit currency bills, it is desirable to investigate ways to distinguish them using image processing and machine learning approach. Therefore, we intend to develop a supervised machine learning model which can help in the identification of the fake currency. The trained model should be able to correctly identify between fake and real currency with good accuracy. The training of the model will be done using labeled datasets that has samples from both types of currencies. Through this training, the model will learn patterns that differentiate one type of currency from other. After the training and validation of the model is completed, it can be utilized to distinguish whether a new currency bill is real or counterfeit.

Datasets and Inputs

To investigate this problem, we are using Banknote Authentication dataset available at UCI Machine Learning Repository [4]. The data were extracted from images that were taken for the evaluation of an authentication procedure for bank notes. These images consisted of genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tools were used to extract features from images. The features extracted after wavelet transformation have following attribute information:

1. Variance of wavelet transformed image.
2. Skewness of wavelet transformed image.
3. Kurtosis of wavelet transformed image.
4. Entropy of image
5. Class of the currency (genuine/fake)

Features 1-4 are continuous, while feature 5 is an integer. Class 1 represents genuine currency,

and class 0 represents fake currency. The dataset has 1372 instances of features extracted from wavelet transformed images. This dataset will be helpful in training a model which can classify fake currency from real one.

Solution Statement

The solution to the problem will be a trained model which can classify new input images based on features extracted after wavelet transformation. The component of processing the image and doing wavelet transformation is not part of the project. This component can be developed as another project. But our goal in this project is to emphasize on the machine learning aspect of the problem. Hence, we are working with the transformed images. To test the performance of our model, we will holdout a test set which will also be labelled, but not visible to our model. After the training is complete, we will test and measure the performance of our model using metrics like accuracy and coefficient of determination. We will also make a validation set which will be used to compare the performance of different algorithms during training and choose the one which gives the highest score. The parameters of the model can be saved and reproduced again, whenever we need to identify of a new currency bill as genuine or fake. So, we do not require training the model again, and reproduce it whenever required.

Benchmark Model

We will compare the performance of our model with a naïve classifier, which classifies all banknotes as fake. The accuracy of this naïve classifier will be number of fake banknotes to the total number of notes in the data. Our model should perform far better than the naïve approach to be any worthy for intended use. Since, this naïve prediction model does not consider any information to substantiate its claim, it helps establish a benchmark for whether a model is performing well. That been said, using a naïve prediction would be pointless as will predict all notes counterfeit and would not identify any note as genuine.

Evaluation Metrics

We will use two evaluation metrics to quantify the performance of our solution as well as the benchmark model, viz., accuracy and F-beta score. Accuracy is defined as the number of samples correctly classified to the total number of samples. This can be a good metric, but suppose, if it's more detrimental for us if are not able to correctly classify a fake note, than a real one. Therefore, the model's ability to recall all fake currency notes is more important than the model's ability to make precise prediction. In this scenario, we can use, F_β score with $\beta = 2$, as it weighs recall more than precision. The general form is: $F_\beta = (1 + \beta^2) . \frac{precision . recall}{(\beta^2 . precision) + recall}$

Project Design

The workflow for the design of a solution to the given problem begins by acquiring the dataset from the UCI Machine Learning repository and analyze it for any missing values or outliers. On inspection of the dataset, it is visible that the data does not have any missing values and all the attributes are continuous, except the target attribute which is the label of the sample as fake or genuine (0 or 1). It captures four features of the wavelet transformed image, viz., variance, skewness, kurtosis and entropy, all varying from a negative to a positive value. So, the next step would be to normalize these attributes in the range of 0 to 1. Normalization or scaling ensures that all features are treated equally when applying supervised learners. We would also look if any feature is skewed, i.e., it has a sample with vastly different value than other samples. In case any feature is skewed we can apply log transform before scaling the feature.

Once our data is in good form, we will split our data into training and test sets. Thereafter, we will investigate different supervised learning algorithms and determine which is best at modeling the data. The various algorithms that

we may consider for evaluation include, Gaussian Naïve Bayes, Decision Tree, Logistic Regression or Support Vector Machines. and Each of the algorithm has to be compared with naïve predictor to evaluate its performance. We will use our evaluation metrics to do the comparison. We will analyze and discuss each algorithm that we chose for modeling our data in detail in the context of the given problem and identify the best among them. We will support our choice with appropriate visualizations and statistics to make out a case in support of best algorithm we choose.

The next step will be to fine tune the chosen model on few of its important parameters using grid search. This will give us an optimized model tuned over some of its parameters. We will then compare the unoptimized model with the optimized one using our evaluation metrics and look for any improvement in the evaluation scores. This is the entire workflow of our project design.

References

1. https://en.wikipedia.org/wiki/Counterfeit_money
2. https://en.wikipedia.org/wiki/Counterfeit_United_States_currency
3. Lohweg, Volker, et al. "Banknote authentication with mobile devices." *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013.
4. https://archive.ics.uci.edu/ml/datasets/banknote+authentication