EP3140/PH6130: Data Science Analysis
Benford's Law and Adversarial Images

Anita Dash - MA20BTECH11001 Teerth Raval - ME20BTECH11046

May 1, 2024

Abstract

This report is based on a group project part of EP3140/PH6130 Data science analysis course. Benford's Law is an observation about the frequency distribution of leading digits in many real-life datasets. This phenomenon has applications in detecting fraud, assessing data quality, and verifying the authenticity of natural datasets. Adversarial attacks involve manipulating inputs to machine learning models by adding imperceptible perturbations that drastically change the cost function. This causes the model to behave unexpectedly and misclassify input images. This project aims to check if Benford's Law can be used as a test to differentiate between original images and adversarial images, and run a classification algorithm.

1 Introduction

Adversarial attacks are a way to manipulate machine learning models, particularly Convolutional Neural Networks (CNNs), by introducing carefully crafted input data that is designed to drastically change the cost function. These attacks exploit the vulnerabilities or blind spots in the model's decision-making process. In case of image classification tasks these manipulated inputs are called adversarial images, which do not look any different to the human eye but get misclassified by the model due to the perturbations in the pixel values. In fields like finance, where an adversarial attack can cause malicious activities getting undetected, or healthcare where such an attack can cause misdiagnoses, identification of adversarial inputs becomes a huge concern.

Benford's Law outlines a distribution for the first digits in a real-life dataset, stating that it follows a logarithmic distribution. We wish to see whether Benford's law can be used to identify these adversarial images, by comparing how closely a distribution derived from an image fits Benford's law versus that of an adversarial image. We check if this deviation from Benford's Law can be used as a parameter for a logistic regression model to classify original and adversarial images.

2 Benford's Law

Benford's Law is an observation that in many real life datasets, the first digits of the numbers follow a specific distribution. In particular, the probability of a digit d being the

first digit is:

$$P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}(1 + \frac{1}{d})$$

The leading digits thus form the following distribution:

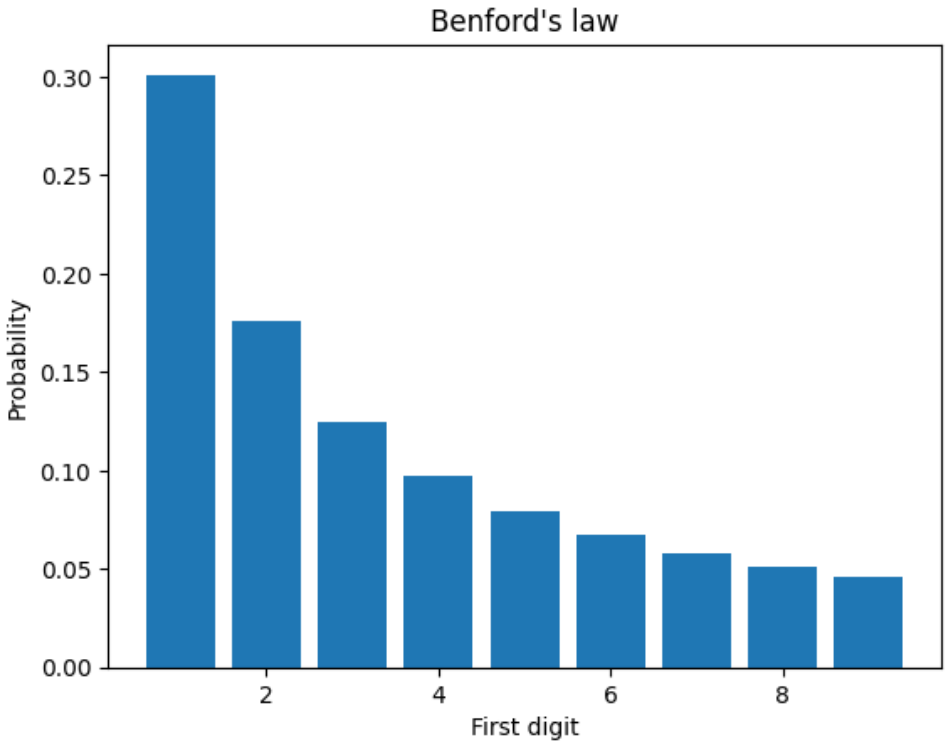


Figure 1: Benford's Law Distribution

Benford's Law applies to a huge variety of datasets: population and related data, bills, stock prices, and even some mathematical constants. Though we do not know exactly what all types of datasets follow Benford's Law, it is observed that usually when the numbers in a dataset span over a range of orders of magnitude, they are likely to follow Benford's Law.

Benford's Law is not followed by datasets predisposed to have specific digits more likely to be the first one, e.g. human body measurements.

3 Image Transformation

According to (1)(2) image luminance possess a histogram that does not admit a closed form and hence it is unlikely an image will follow Benford's law. However different results are obtained when the images are transformed using the following transformations.(4)

3.1 Gradient Magnitude Transformation

(4) Gradient Magnitude Transformation is a transformation that enhances the edges and boundaries of objects within an image.

In an image $x \in \mathbb{R}^{n \times m}$, the gradient $G(x)$ is given by:

$$G(x)_{i,j} = \sqrt{G_{e_1}(x)_{i,j}^2 + G_{e_2}(x)_{i,j}^2}$$

where

$$G_{e_1}(x) = x \star K_{e_1}$$
$$G_{e_2}(x) = x \star K_{e_2}$$
$$K_{e1} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ 1 & 0 & 1 \end{bmatrix}, K_{e2} = \begin{bmatrix} -1 & -2 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

$x \star K_{e1}$ represents the convolution operation, given by

$$(x \star K_{e1})_{i,j} = \sum_{o=1}^{O} \sum_{p=1}^{P} x_{i-o,j-p} K_{o,p}$$

3.2 Direct Cosine Transformation

(3) Discrete Cosine Transformation or DCT converts an image from the pixel domain into the frequency domain, representing it in terms of cosine functions of different frequencies.

$$F(u,v) = \frac{2}{N} C(u)C(v) \sum_{x=1}^{N-1} \sum_{y=1}^{N-1} cos\left[\frac{\pi(2x+1)u}{2N}\right] cos\left[\frac{\pi(2y+1)u}{2N}\right]$$

for $u = 0, ..., N-1$ and $v = 0, ..., N-1$, where $N = 8$.

$$C(k) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0, \\ 1 & \text{otherwise.} \end{cases}$$

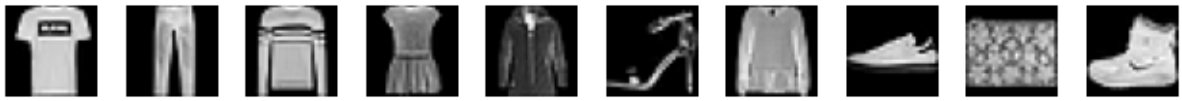
4 Methodology

We want to determine whether we can use Benfords law to distinguish between original images and adversarial images. Our approach is to perform an adversarial attack to generate adversarial images from a set of randomly selected images, we then use benford's law analysis to calculate the deviation of the images from benfords law and verify if the deviation is significant enough to classify original images from adversarial images.

4.1 Dataset

For our project we use the MNIST and Fashion MNIST datasets. Both the datasets consist of images of size 28x28 pixels. The MNIST dataset consists of images of handwritten digits from 0 to 9, while the Fashion MNIST dataset consists of images of clothing items.

(a) MNIST Dataset



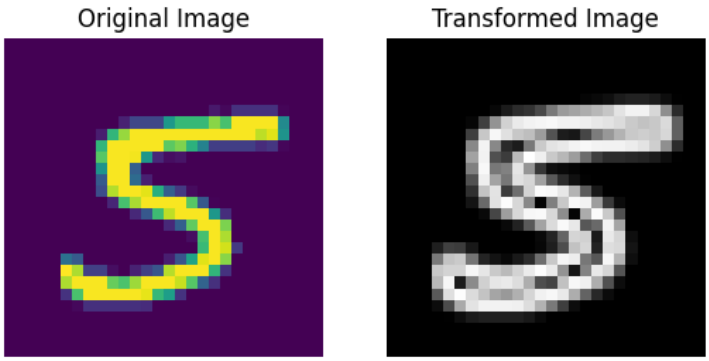
(b) Fashion-MNIST Dataset

4.2 Benford's Law Analysis

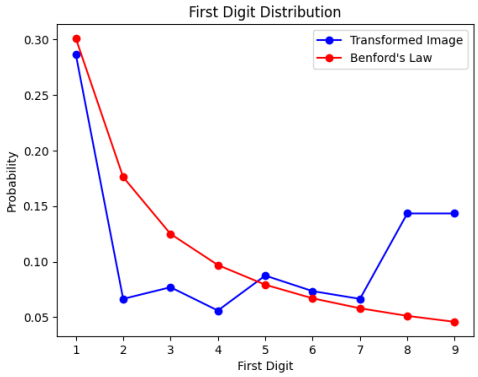
We transform the images in the dataset using both gradient magnitude and dct transformation. We calculate the first digit distribution (FDD) of the transformed images. We then use the Kolmogorov-Smirnov (KS) statistic to calculate the deviation between the FDD of the transformed images and the Benford's Law distribution, using the formula given below:

$$D(p, q) = sup|F(p) - F(q)|$$

Where $F(p)$ and $F(q)$ are the cumulative distribution functions of the FDD and the Benford's Law distribution respectively.

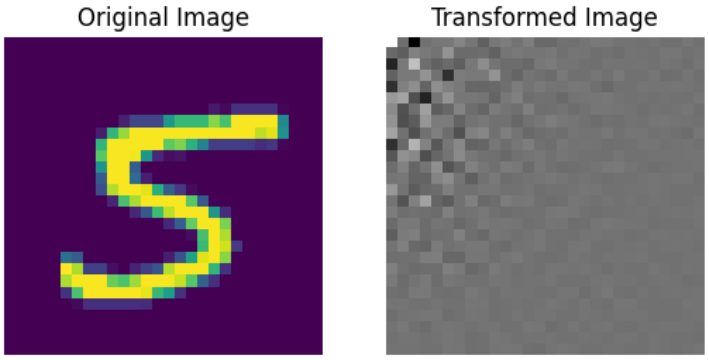


(a) Gradient Transformation

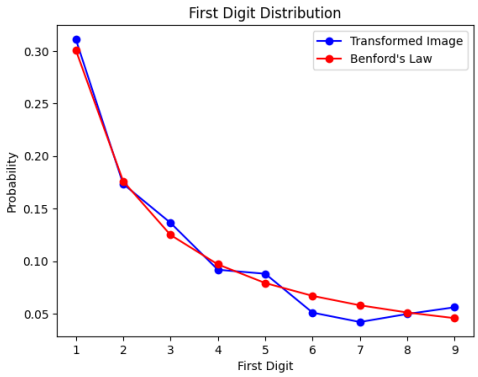


(b) FDD

Figure 3: Benford's Law analysis on a random image transformed via gradient magnitude transformation.



(a) DCT Transformation



(b) FDD

Figure 4: Benford's Law analysis on a random image transformed via DCT transformation.

4.3 Adversarial Attack

(4) Adversarial attacks are a class of attacks on machine learning models that are designed to fool the model into making incorrect predictions. These attacks are designed to exploit the vulnerabilities of the model and are often used to test the robustness of the model. In our project the we attack a simple CNN model trained on the MNIST dataset and the Fashion MNIST datase.

4.3.1 CNN Model

The CNN model consists of two convolutional layers with a layer of max pooling after each convolutional layer. followed by three fully connected layers.

Layer Type	**Dimension**
Conv1	(6, 5, 5)
Conv2	(16, 5, 5)
Max Pooling 1	(2, 2)
Max Pooling 2	(2, 2)
FC1	(120,)
FC2	(84,)
FC3	(10,)

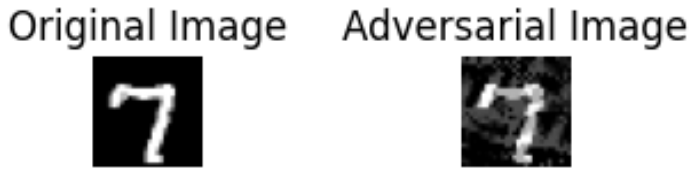
Table 1: Summary of Layer Types and Dimensions in the CNN Architecture.

The training accuracy on the model trained on the MNIST dataset is 0.98 and the test accuracy is 0.98. The training accuracy on the model trained on the Fashion MNIST dataset is 0.91 and the test accuracy is 0.84.

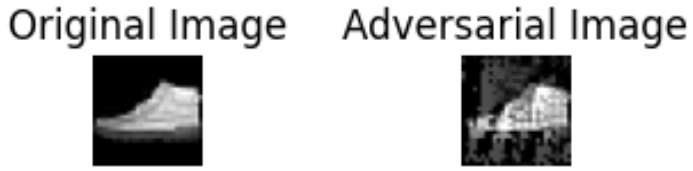
4.3.2 PGD Attack

We use the Projected Gradient Descent (PGD) attack to generate adversarial images. The PGD attack is a white-box attack that is designed to generate adversarial images that are close to the original images. The attack is performed by taking multiple small steps in the direction of the gradient of the loss function with respect to the input image.

We randomly select 1000 images from the test set of the MNIST dataset and the Fashion MNIST dataset respectively and generate adversarial images using the PGD attack. The test accuracy of the model on the adversarial images is 0.0 for both the MNIST and Fashion MNIST datasets.



(a) MNIST



(b) Fashion MNIST

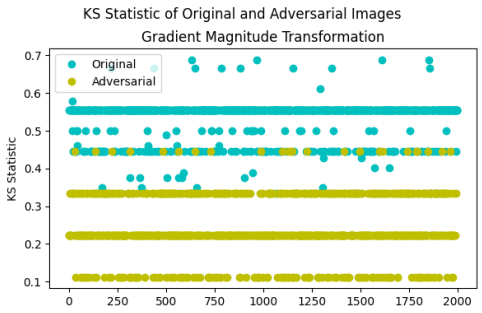
Figure 5: Adversarial attack on images

4.4 Adversarial Image Detection

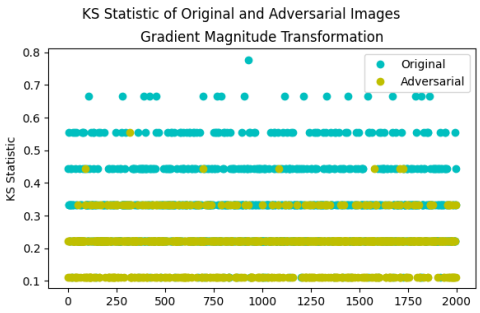
(4) We transform both the original and adversarial images using the gradient magnitude transformation and the direct cosine transformation. We then calculate the KS statistic for the transformed images with respect to the Benford's Law distribution.
We use the KS statistic as the input parameter to a logistic regression model to classify the images as real or adversarial. The logistic regression model is trained on the KS statistic of the original images and the adversarial images. The model is trained on 80% of the images and tested on the remaining 20%.

Dataset	**Accuracy**	**Precision**	**Recall**
MNIST	0.9725	0.9703	0.9751
Fashion MNIST	0.7825	0.7500	0.8507

Table 2: Performance Metrics on Images Transformed via Gradient Magnitude.



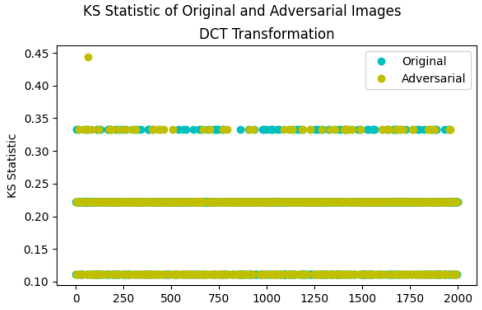
(a) MNIST



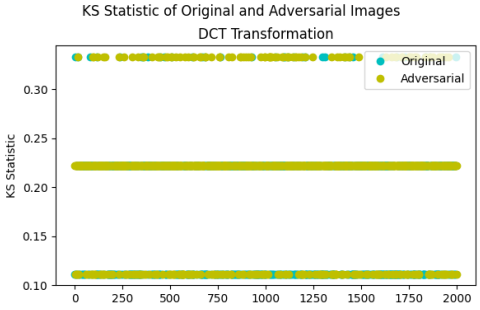
(b) Fashion MNIST

Figure 6: Plot of KS-Statistics of images transformed via gradient magnitude transformation

Based on figure 6 and the results shown in Table 2, When we transform the images using the gradient magnitude transformation, the logistic regression model is able to classify the images as real or adversarial with high accuracy for both the datasets. The model has a high precisions and recall implying that the model is able to correctly classify the images as real or adversarial.



(a) MNIST



(b) Fashion MNIST

Figure 7: Plot of KS-Statistics of images transformed via DCT transformation

Dataset	**Accuracy**	**Precision**	**Recall**
MNIST	0.5175	0.5426	0.2537
Fashion MNIST	0.6075	0.5738	0.8507

Table 3: Performance Metrics on Images Transformed via DCT.

Based on figure 7 and the the results shown in Table 3, When we transform the images using the direct cosine transformation, the logistic regression model is not able to classify the images as real or adversarial properly. the model trained on the MNIST dataset has low precision and low recall implying the model is not reliable. However the fashion MNIST dataset has a high recall but low precision, implying even though there are a lot of false positives, the model is able to correctly classify the adversarial images, but the accuracy is still low.

5 Conclusion

The results of our analysis show that using benford's law as a parameter to distinguish between real and adversarial images works when the images are transformed using the gradient magnitude transformation. Using a logistic regression model with just one parameter as the input to classify the images as real or adversarial is less computationally expensive compared to CNN models that do the same classification. However, the model is not able to classify the images when the images are transformed using the direct cosine transformation. Further analysis is required to determine the reason for this discrepancy.

6 Code

Link to the codes: `https://github.com/anitadash/DSA_Project`

References

[1] JM Jolion. Images and benford's law. *Journal of Mathematical Imaging and Vision*, 14, 2001.

[2] Fernando Perez-Gonzalez, Greg L. Heileman, and Chaouki T. Abdallah. Benford's lawin image processing. 1:I – 405–I – 408, 2007.

[3] Vishnu U. Deepfake detection using benford's law and distribution variance statistic. *IRJET*, 2021.

[4] João G. Zago, Eric A. Antonelo, Fabio L. Baldissera, and Rodrigo T. Saad. Benford's law: What does it say on adversarial images? *Journal of Visual Communication and Image Representation*, 93:103818, May 2023.