Using SVD-based Image Watermarking For ML Datasets

Using SVD based image watermarking techniques to watermark numerical ML datasets.

Authors

Palle Maesen
P.Maesen@student.tudelft.nl

Supervisors

Dr. Zekeriya Erkin
Devris Isler

Concepts

Watermarking: Embedding data into host data to prove ownership, check data tampering, manage copy control, etc.
Singular Value Decomposition: Decomposition of an original matrix A into 3 matrices: U **Σ** and V.
Machine Learning: Giving a machine a certain task. The machine is said to have learned, if its measurable performance has improved from experience.

Motivation

The **media** watermarking technique domain has had the last **30 years** to develop itself. The **non-media** side, however, is a **newer** sub-domain [1].
Images and ML datasets are quite similar in structure. Both are **N-dimensional matrices** containing, often numerical, data.

Research question

How can image watermarking techniques be applied to classification algorithms datasets, without degrading the dataset's quality?

Methodology

The goal of the watermarking technique is to prove ownership of machine learning datasets. The main requirements for this goal are robustness and imperceptibility.

The final technique is the algorithm described by chang et al. in "SVD-based digital image watermarking scheme" [2].

- SVD-based algorithms are often the basis of other matrix decomposition based algorithms.
- Using thresholds to increase the robustness and decrease the imperceptibility adds versatility.

The technique is applied to an ML dataset and is tested on:

- robustness against update, zero-out delete insertion and multi-faced attacks.
- Imperceptibility.

Results

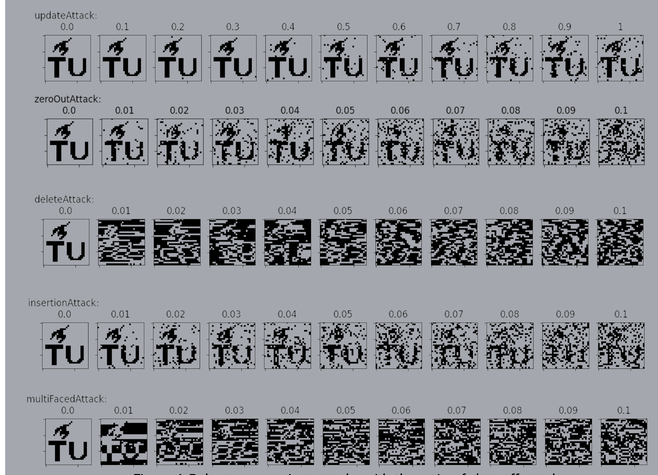

Figure 1: Robustness against attacks with the ratio of data affected

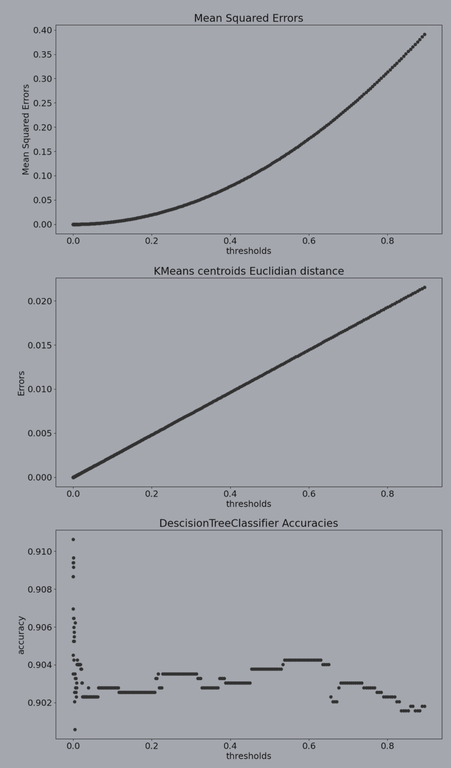

Figure 2: Imperceptibility (Mean Squared, kmeans centroid and decision tree errors) of the watermark for increasing thresholds

Conclusions

- Robustness against any but the deletion and multi-faced attacks are sufficient.
- Robustness against deletion attacks is below satisfactory.

The technique originally makes use of the format limitations of images. Deleting even one entry has a massive impact on the image quality. Deleting one percent of an ML dataset has a limited impact on the dataset's quality. This explains its lackluster robustness against deletion attacks.

It is possible to apply the technique described in [2] to an ML dataset. However, the inability to deal with deletion or reordering attacks completely nullifies the usefulness of the watermarking technique in real applications

Future work

Find an image watermarking technique that is impervious to deletion attacks and apply this technique to an ML dataset.

Citations

[1] A. S. Panah, R. G. van Schyndel, T. K. Sellis, and E. Bertino, "On the properties of non-media digital watermarking: A review of state of the art techniques," IEEE Access, vol. 4, pp. 2670–2704, 2016
[2] C. Chang, P. Tsai, and C. Lin, "Svd-based digital image watermarking scheme," Pattern Recognit. Lett., vol. 26, no. 10, pp. 1577–1586, 2005.

CSE 3000