Robustness of CNN-based malware byteplot classification under standard image transformations

Author: Tudor Ioan Tănăsescu | Responsible Professor: Tom J. Viering | Supervisor: Akash Amalan

Delft University of Technology | Technische Universiteit Delft

1. Introduction

Malware threatens all information systems [1]:

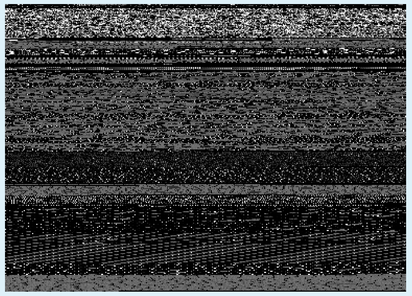
- National healthcare infrastructure
- Financial institutions
- Industrial control systems

The scale of the problem is increasing [2], and traditional detection systems can't keep up [3] [4].

Possible solution: Convolutional Neural Network (CNN) detection:

- Convert binaries to byteplot images [4]
- Train a CNN model that will then classify unseen binaries

Project goal: study the robustness of CNN malware classifiers in the face of standard image transformations

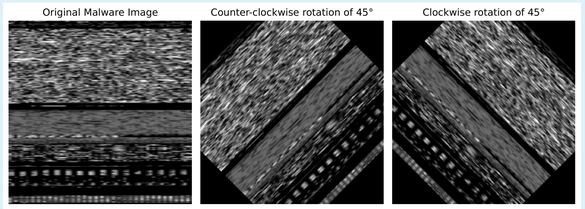


2. Research Question

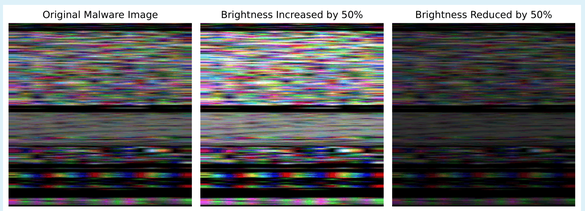
How does the accuracy of a CNN-based malware byteplot classifier degrade under standard image transformations, and what can the pattern of degradation reveal about the visual features the network relies on for classification?

3. Transformations

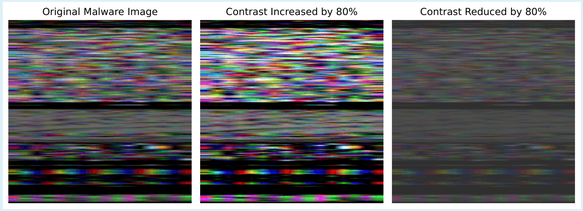
- Rotations



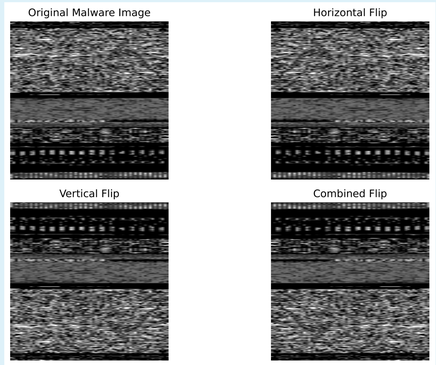
- Brightness shifts



- Contrast Shifts



- Vertical, Horizontal and Combined Flips



4. Dataset & Experimental Setup

The dataset contains 20,020 unique byteplots:

- 10,010 grayscale representation
- 10,010 RGB representation

The experiments were divided in two stages.

Stage 1:

- Train the model on clean images
- Train the model on distorted images (75% clean images, 25% distorted images)

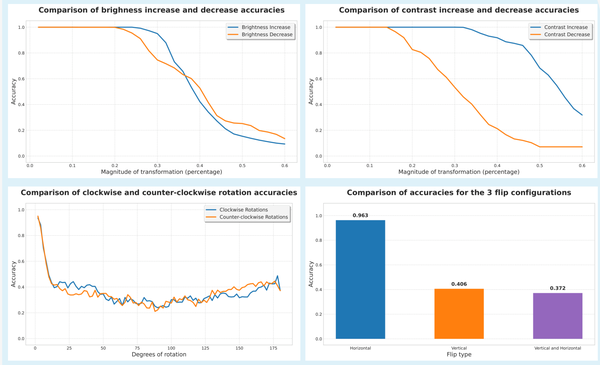
Stage 2: Train a model on each transformation to establish if gains are self-attributing

All evaluations are executed on distorted images:

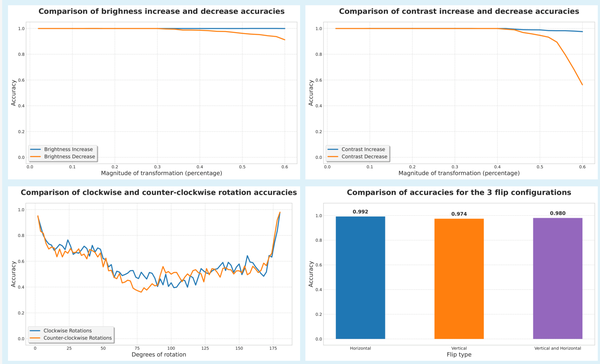
- Rotations: 0° to 180° in 2° increments
- Brightness shifts: up to 60% in 2% increments for Stage 1, up to 100% in 2% increments for Stage 2
- Contrast shifts: up to 60% in 2% increments for Stage 1, up to 100% in 2% increments for Stage 2
- Flips: vertical, horizontal and combined flips are evaluated separately

5. Results

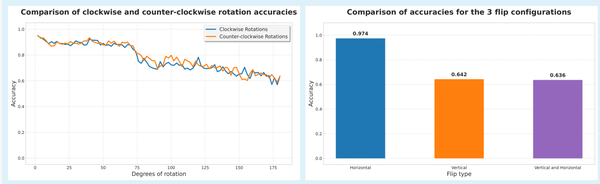
5.1 Model trained on clean images



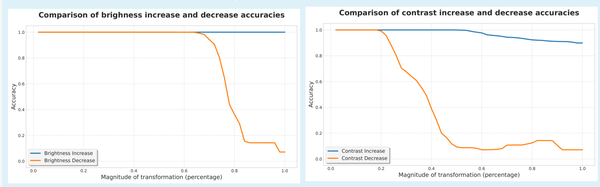
5.2 Model trained on distorted images



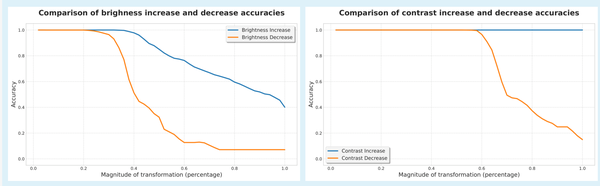
5.3 Model trained on rotations



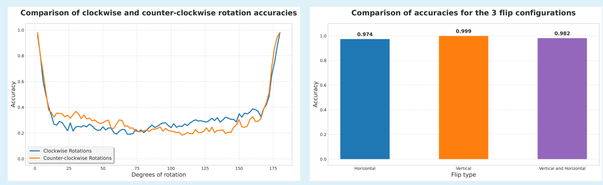
5.4 Model trained on brightness shifts



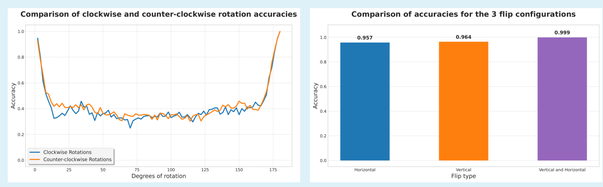
5.5 Model trained on contrast shifts



5.6 Model trained on vertical flips



5.7 Model trained on combined flips



6. Conclusion

- Training on distorted images improves robustness broadly **without sacrificing** clean image performance
- RGB representations **amplify existing trends** in grayscale byteplots
- The classifier relies on **globally positioned features** that collapse the moment the image is reoriented
- Brightness and contrast results indicate the network exploits **relative intensity**, not absolute byte values
- Rotation training partially improves vertical flip robustness, confirming sensitivity to **vertical positional disruption**
- Fragility to disruptions is a **training distribution problem**, not a byteplot expressiveness limit

References

[1] N. Younas, S. Riaz, S. Ali, R. Khan, F. Ali, and D. Kwak, "Detecting malicious code variants using convolutional neural network (CNN) with transfer learning"

[2] F. Cremer, B. Sheehan, M. Fortmann, A. N. Kia, M. Mullins, F. Murphy, and S. Materne, "Cyber risk and cybersecurity: a systematic review of data availability"

[3] N. Marastoni, R. Giacobazzi, and M. Dalla Preda "Data augmentation and transfer learning to classify malware images in a deep learning context"

[4] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, "Malware images: visualization and automatic classification"