A Study on the Explainability of Multimodal Malware Detection Using XAI

2410046 YU ZHOU

In recent years, multimodal deep learning has demonstrated remarkable success in the field of malware detection. However, the "black box" nature of decision-making processes, a byproduct of increasing model complexity, poses a significant and undeniable risk in security operations where reliability is paramount. In particular, the phenomenon where false negatives in specific classes are obscured by high overall average accuracy metrics represents a critical challenge.

This study focuses on "MALSSL," the core image processing module of the multimodal detection framework "MIDALF," conducting a comprehensive analysis of its behavioral characteristics. Results indicate that while the model performs exceptionally well in benchmark tests, detailed confusion matrix analysis reveals a critical flaw: a complete inability (0% accuracy) to identify specific malware families exhibiting similar textures.

To investigate the root cause of this structural vulnerability, the first phase of this study involved selecting the most appropriate Explainable AI (XAI) algorithm for diagnosing the model's internal behavior. Through an analysis of five primary methods—Grad-CAM, Grad-CAM++, XGrad-CAM, LayerCAM, and SHAP—the results demonstrated that the ResNet-18 backbone, employed as the baseline, possesses insufficient feature extraction capabilities. Consequently, comparative verification across ResNet variants (34, 50, 101, and 152) led to the selection of ResNet-50 for its optimal balance between computational efficiency and feature extraction, successfully

elevating the overall accuracy, including the target families, to 98.7

Furthermore, validation based on three metrics—Fidelity, Complexity, and Robustness—identified LayerCAM as the method most capable of accurately retaining subtle feature information within deep CNN layers, establishing it as the core diagnostic tool.

Additionally, to ensure objective validity beyond visual qualitative assessment, this study developed a novel integrated diagnostic framework combining "LayerCAM + Shannon Entropy" and conducted empirical experiments. Quantitative analysis revealed that ResNet-18 recorded a high entropy value of 6.82 bits in the attention regions, confirming its accurate capture of encrypted payload areas. However, the persistence of classification failure (0% accuracy) implies that the shallower structure of ResNet-18 failed to translate the complex patterns hidden within these high-entropy regions into the feature representations necessary for effective identification. In contrast, ResNet-50 achieved a higher entropy value of 7.04 bits. Mathematically, this increase demonstrates that the deeper architecture of ResNet-50 establishes a stronger and more definitive focus on subtle malicious features within the payload.

Keywords: Malware Detection, Multimodal Learning, ResNet-50, Explainable AI (XAI), Entropy