Accelerating Malware Classification: A Vision Transformer Solution

Shrey Bavishi
Department of Computer Science
Indian Institute of Technology-Bombay
Mumbai, India 400076
200050132@iitb.ac.in

Shrey Modi
Department of Computer Science
Indian Institute of Technology-Bombay
Mumbai, India 400076
200020135@iitb.ac.in

Abstract

The escalating frequency and scale of recent malware attacks underscore the urgent need for swift and precise malware classification in the ever-evolving cybersecurity landscape. Key challenges include accurately categorizing closely related malware families. To tackle this evolving threat landscape, this paper proposes a novel architecture LeViT-MC which produces state-of-the-art results in malware detection and classification. LeViT-MC leverages a vision transformer-based architecture, an image-based visualization approach, and advanced transfer learning techniques. Experimental results on multi-class malware classification using the MaleVis dataset indicate LeViT-MC's significant advantage over existing models. This study underscores the critical importance of combining image-based and transfer learning techniques, with vision transformers at the forefront of the ongoing battle against evolving cyber threats. We propose a novel architecture LeViT-MC which not only achieves state of the art results on image classification but is also more time efficient.

1 Introduction and Background

The modern cybersecurity landscape is fraught with an ever-increasing threat posed by malware, presenting a relentless challenge to security professionals and researchers. Despite tireless efforts within the cybersecurity industry to combat these threats, cyber attackers remain undeterred, continually evolving their tactics and techniques, and devising sophisticated evasive strategies such as polymorphism, metamorphism, and code obfuscations which make the problem very challenging.

This paper introduces LeViT-MC, an innovative approach to malware characterization and analysis utilizing the MaleVis dataset. Our methodology builds upon the insights of Nataraj et al. [2011], which suggest that malware executables can be effectively represented as image-like matrices, revealing significant visual similarities among malware from the same family.

In this study, we will first review the various methods previously employed for malware detection and classification, along with their inherent limitations. Subsequently, in Section 2 , we will propose our novel architecture, LeViT-MC, which leverages the binary classification capabilities of DenseNet and the rapid inference speed of vision transformers. To the best of our knowledge, this represents the first instance of combining DenseNet and vision transformer architectures for the purposes of malware detection and classification. Section 3 details the experimental procedures and results obtained, demonstrating state-of-the-art accuracy and inference times attributable to our novel architecture. We will discuss the limitations of our work in Section 4 , before concluding in Section 5. All code utilized to generate the results discussed in this paper can be found at this link.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1.1 Static and Dynamic Analysis

Early malware detection methods primarily relied on static and signature-based approaches. Zadok et al. [2001] explored the extraction of static features, including byte sequences and Santos et al. [2013] explored the opcode patterns, in conjunction with machine learning techniques for classification. Dynamic analysis methods aimed to understand malware behavior by monitoring network activities and system calls. Imran et al. [2015] introduced a malware classification approach based on Hidden Markov Models (HMMs), analyzing API call sequences. However, dynamic analysis proved inefficient, particularly when malware adapted its behavior during execution.

1.2 Vision-Based Approaches

The transition to vision-based approaches marked a significant advancement in malware detection. Abijah Roseline et al. [2020], Singh et al. [2019], Roseline et al. [2019] began visualizing features like opcode sequences and system calls as images, offering new perspectives on classification. Nataraj et al. [2011] developed an efficient approach to visualising binary files as greyscale images and classfied the images using K-nearest neighbours. Conti et al. [2008] demonstrated the effectiveness of visual methods in classifying binary files and analyzing new file structures.

In their follow-up study, Falana et al. [2022] introduced a malware detection system achieving an impressive 99.8% accuracy on malware detection. These methods talk only about malware detection, that is, a binary classification into beningn and malign classes, but fail to work in detecting the type of malware, which is also very crucial for system security.

1.3 Transformer-Based Model

Recent years have witnessed the rise of transformer-based models in malware detection. Oak et al. [2019], Rahali and Akhloufi [2021] introduced BERT(Bidirectional Encoder Representations from Transformers)-based models, McLaughlin [2022] presented Malceiver, a hierarchical Perceiver model and Hu et al. [2020] introduced hierarchical transformer architectures for malware classification. The application of transformer-based models was expanded beyond assembly code analysis by Sherlock(Seneviratne et al. [2022]) a transformer-based model for Android malware classification. Most of these methods rely on analysing and integrating opcodes, which significantly increases the inference time for detection and classification.

2 Methodology

2.1 Image Representation of Malware

Our approach involves transforming PE binary files into RGB images by reading groups of 3 bytes and arranging them in a 2-dimensional vector space. Each byte, ranging from 0 to 255, is visualized as a pixel value in an image. The three bytes constitute one pixel in each of the three channels (RGB). This method preserves location information between malicious attack patterns and captures the order of these patterns, which is essential for precise classification, especially when dealing with structurally similar malware instances.

2.2 LeViT-MC Architecture

We propose a novel architecture, LeViT-MC, that combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) to create a robust malware classification system. The architecture consists of two main components: Binary Classification Stage: A fine-tuned DenseNet CNN followed by a classification head categorizes input images into benign and malign. Malware Family Classification Stage: For images classified as malign, a LeViT (Lightweight Vision Transformer) further classifies them into specific malware families.

Binary Classification Stage

We utilize a fine-tuned DenseNet CNN for the initial binary classification of images into benign and malign categories. DenseNet has demonstrated excellent performance in binary classification

tasks for malware images as also showed by Falana et al. [2022] . The dense connectivity pattern in DenseNet allows for better feature reuse and improved information flow, making it particularly effective for capturing the intricate patterns present in malware images.

Malware Family Classification Stage

For the more granular task of classifying malign images into specific malware families, we employ the LeViT architecture introduced by Graham et al. [2021]. LeViT is designed for faster inference in image classification tasks, making it particularly suitable for real-time malware detection scenarios. More details about the LeViT architecture are given in Appendix A

2.3 Transfer Learning and Fine-tuning

To leverage the power of pre-trained models and adapt them to our specific task, we employ transfer learning techniques: We initialize the DenseNet component with weights pre-trained on ImageNet. The LeViT component is initialized with weights from a model pre-trained on a large-scale image classification task. We fine-tune both components on the MaleVis(Bozkir et al. [2019]) dataset, allowing the model to adapt to the specific characteristics of malware images. This approach enables our model to benefit from the general feature extraction capabilities learned from large datasets while specializing in the nuances of malware classification.

A complete workflow of our architecture is given in Figure 1

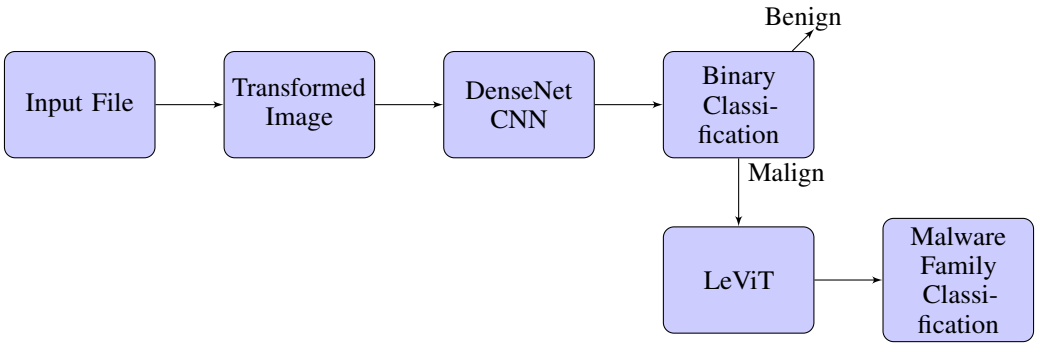

Figure 1: LeViT-MC Workflow

3 Experiments

3.1 Dataset and Implementation Details

We utilized the MaleVis(Bozkir et al. [2019]) dataset, comprising 14,226 RGB images in 224x224 pixel resolution. The dataset is divided into 26 categories: one representing benign samples and 25 representing various malware types. We have included some samples and more information about the dataset in Appendix B. We employed the default partition of 70% training and 30% validation data. Experiments were conducted using an NVIDIA A100-SXM4 GPU with 80GB RAM. Training was performed with a batch size of 32, using the Adam optimizer with an initial learning rate of 1e-5 and a ReduceLROnPlateau scheduler with a decay of 0.1 and patience of 10.

3.2 Results

LeViT-MC demonstrated superior performance in both accuracy and inference speed. We achieved 96.6% accuracy, a groundbreaking multiclass classification accuracy on the MaleVis dataset, outperforming all the previous state-of-the-art models. A comparision with previous multiclass state-of-the-art classification accuraries is given in Table 1

While gaining the highest accuracy, our novel architecture also gains the highest inference speed by utilising the fast inference of LeViT transformers. An inference speed comparision with the average

Table 1: Accuracy comparison with various state-of-the-art models

Study	Accuracy
Nataraj et al. [2011]	91.69%
Agarap [2019]	79.36%
Patil et al. [2021]	93%
Paik and Jin [2022]	93.49%
LeViT-MC	**96.6%**

inference speed for different architectures on the same dataset as quoted in the study by Bianco et al. [2018] is shown in Figure 2. Our model is about has around 10x better inference times than normal vision transformers and around 3x better inference times than the best CNNs like ResNet.

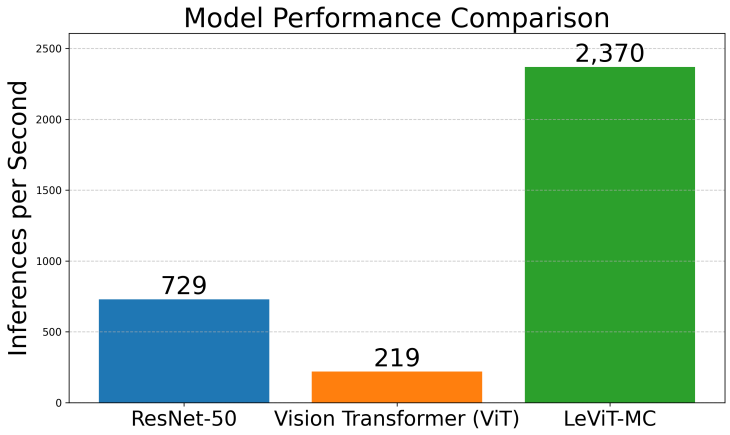

Figure 2: Perfomance Comparision

4 Limitations

While LeViT-MC has demonstrated exceptional performance, a couple of considerations remain:

- **Dataset Constraints:** Our current evaluation was performed on the MaleVis dataset, and while this dataset covers a broad spectrum of malware types, its generalizability to other datasets remains untested.
- **Real-World Deployment and Scalability:** The computational efficiency observed in controlled environments (e.g., with high-performance GPUs) may not fully translate to resource-constrained devices like IoT systems or embedded malware detection platforms.

5 Conclusion

LeViT-MC demonstrates exceptional performance in malware classification, achieving 96.6% accuracy and processing 2370 images per second. These results underscore the potential of combining CNNs and lightweight Vision Transformers in addressing the challenges of rapid and accurate malware detection and classification.

Looking ahead, our future work will emphasize the practical deployment of our methodology in real-world scenarios. Effective implementation in cybersecurity infrastructures is essential to proactively defend against evolving malware threats. We aim to explore integration with existing security systems, ensuring that our approach not only meets performance benchmarks but also adapts to the dynamic nature of cyber threats. Additionally, we will focus on enhancing explainability in our model to foster trust and transparency in its decision-making processes, facilitating its acceptance in operational environments.

References

S. Abijah Roseline, G. Hari, S. Geetha, and R. Krishnamurthy. Vision-based malware detection and classification using lightweight deep learning paradigm. In Neeta Nain, Santosh Kumar Vipparthi, and Balasubramanian Raman, editors, *Computer Vision and Image Processing*, pages 62–73, Singapore, 2020. Springer Singapore. ISBN 978-981-15-4018-9.

Abien Fred Agarap. Towards building an intelligent anti-malware system: A deep learning approach using support vector machine (svm) for malware classification, 2019.

Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018. doi: 10.1109/access.2018.2877890. URL https://doi.org/10.1109%2Faccess.2018.2877890.

Ahmet Bozkir, Ahmet Cankaya, and Murat Aydos. Utilization and comparision of convolutional neural networks in malware recognition. 03 2019. doi: 10.1109/SIU.2019.8806511.

Gregory J. Conti, Erik Dean, Matthew Sinda, and Benjamin Sangster. Visual reverse engineering of binary and data files. In *Visualization for Computer Security*, 2008. URL https://api.semanticscholar.org/CorpusID:12625809.

Olorunjube James Falana, Adesina Simon Sodiya, Saidat Adebukola Onashoga, and Biodun Surajudeen Badmus. Mal-detect: An intelligent visualization approach for malware detection. *Journal of King Saud University - Computer and Information Sciences*, 34(5):1968–1983, 2022. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2022.02.026. URL https://www.sciencedirect.com/science/article/pii/S1319157822000702.

Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference, 2021.

Xiaohui Hu, Rui Sun, Kejia Xu, Yongzheng Zhang, and Peng Chang. Exploit internal structural information for iot malware detection based on hierarchical transformer model. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 927–934, 2020. doi: 10.1109/TrustCom50675.2020.00124.

M. Imran, M. Afzal, and M. Qadir. Similarity-based malware classification using hidden markov model. In *2015 Fourth International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)*, pages 129–134, Los Alamitos, CA, USA, oct 2015. IEEE Computer Society. doi: 10.1109/CyberSec.2015.33. URL https://doi.ieeecomputersociety.org/10.1109/CyberSec.2015.33.

Niall McLaughlin. Malceiver: Perceiver with hierarchical and multi-modal features for android malware detection, 2022.

L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath. Malware images: Visualization and automatic classification. In *Proceedings of the 8th International Symposium on Visualization for Cyber Security*, VizSec '11, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306799. doi: 10.1145/2016904.2016908. URL https://doi.org/10.1145/2016904.2016908.

Rajvardhan Oak, Min Du, David Yan, Harshvardhan Takawale, and Idan Amit. Malware detection on highly imbalanced data through sequence modeling. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, AISec'19, page 37–48, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368339. doi: 10.1145/3338501.3357374. URL https://doi.org/10.1145/3338501.3357374.

Joon-Young Paik and Rize Jin. Malware Family Prediction with an Awareness of Label Uncertainty. *The Computer Journal*, page bxac181, 12 2022. ISSN 0010-4620. doi: 10.1093/comjnl/bxac181. URL https://doi.org/10.1093/comjnl/bxac181.

Shruti Patil, Vijayakumar Varadarajan, Devika Walimbe, Siddharth Gulechha, Sushant Shenoy, Aditya Raina, and Ketan Kotecha. Improving the robustness of ai-based malware detection using adversarial machine learning. *Algorithms*, 14(10), 2021. ISSN 1999-4893. doi: 10.3390/a14100297. URL https://www.mdpi.com/1999-4893/14/10/297.

Abir Rahali and Moulay A Akhloufi. Malbert: Using transformers for cybersecurity and malicious software detection. *arXiv preprint arXiv:2103.03806*, 2021.

S. Abijah Roseline, A. D. Sasisri, S. Geetha, and C. Balasubramanian. Towards efficient malware detection and classification using multilayered random forest ensemble technique. In *2019 International Carnahan Conference on Security Technology (ICCST)*, pages 1–6, 2019. doi: 10.1109/CCST.2019.8888406.

Igor Santos, Felix Brezo, Xabier Ugarte-Pedrero, and Pablo G. Bringas. Opcode sequences as representation of executables for data-mining-based unknown malware detection. *Inf. Sci.*, 231: 64–82, may 2013. ISSN 0020-0255. doi: 10.1016/j.ins.2011.08.020. URL `https://doi.org/10.1016/j.ins.2011.08.020`.

Sachith Seneviratne, Ridwan Shariffdeen, Sanka Rasnayaka, and Nuran Kasthuriarachchi. Self-supervised vision transformers for malware detection. *IEEE Access*, 10:103121–103135, 2022. doi: 10.1109/access.2022.3206445. URL `https://doi.org/10.1109%2Faccess.2022.3206445`.

Ajay Singh, Anand Handa, Nitesh Kumar, and Sandeep Kumar Shukla. Malware classification using image representation. In *International Conference on Cyber Security Cryptography and Machine Learning*, 2019. URL `https://api.semanticscholar.org/CorpusID:31642038`.

E. Zadok, M. G. Schultz, E. Eskin, and S. J. Stolfo. Data mining methods for detection of new malicious executables. In *2012 IEEE Symposium on Security and Privacy*, page 0038, Los Alamitos, CA, USA, may 2001. IEEE Computer Society. doi: 10.1109/SECPRI.2001.924286. URL `https://doi.ieeecomputersociety.org/10.1109/SECPRI.2001.924286`.

A Appendix A: The LeViT Architecture

The LeViT-256 architecture, part of the LeViT family, represents a hybrid approach that combines the strengths of convolutional networks and transformer models for image classification tasks. This architecture addresses the computational inefficiencies of standard transformers by integrating convolutional stages to enhance spatial inductive biases, resulting in improved performance and reduced complexity.

LeViT-256 employs a multi-stage structure consisting of:

1. A convolutional stem for efficient initial image processing and dimension reduction
2. Transformer blocks for capturing global dependencies
3. Attention pooling for effective information aggregation

This design enables LeViT-256 to achieve a balance between accuracy and computational efficiency, making it suitable for low-latency and edge-based deployment scenarios. The architecture processes images through hierarchical down-sampling within its transformer layers, reducing the overall number of operations while maintaining strong performance on image classification benchmarks.

By leveraging multi-scale feature extraction and maintaining smaller feature map sizes, LeViT-256 significantly reduces computational overhead. The model transforms images into patches, processes them through stages of attention layers that capture global contextual features while reducing image resolution, and finally classifies the output using a linear head. This approach allows LeViT-256 to achieve state-of-the-art accuracy while maintaining high computational efficiency, positioning it as a promising solution for real-time image classification applications.

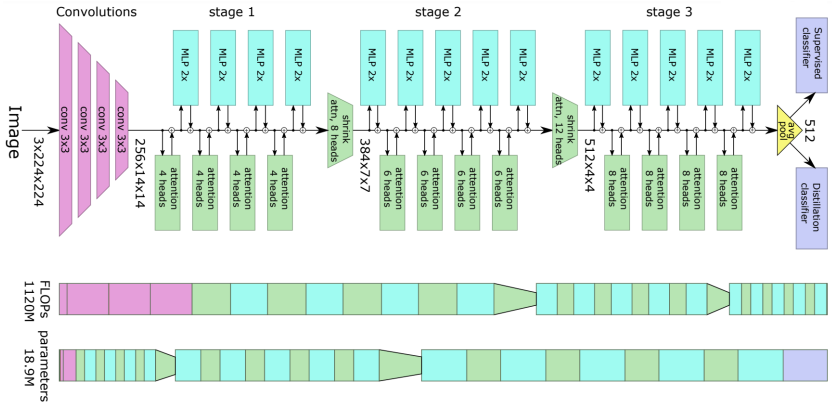

Figure 3: LeViT-256 Architecture

B Appendix B: MaleVis Dataset

In our experimental setup, we employed the MaleVis dataset, short for "Malware Evaluation with Vision," which was publicly released in 2019. This dataset comprises a total of 14,226 images, provided in two square resolutions, specifically 224 by 224 pixels and 300 by 300 pixel, that have been converted to the RGB format. The dataset is divided into 26 distinct categories, with one representing benign samples and the remaining 25 representing various types of malware. Our study utilized images with a 224 by 224 pixels resolution. As illustrated in Figure 5, the dataset exhibits balanced class distribution among various malware types each comprising approximately 500 images. In contrast, the "Normal" class contains more samples, totalling 1832 images. To conduct our experiments, we use the default partition given by the dataset, containing 70% images in the train dataset and 30% in the validation dataset.

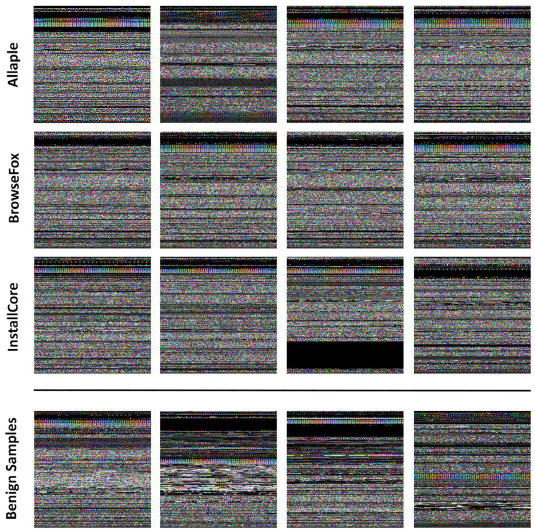

Figure 4: Sample images from the MaleVis Dataset

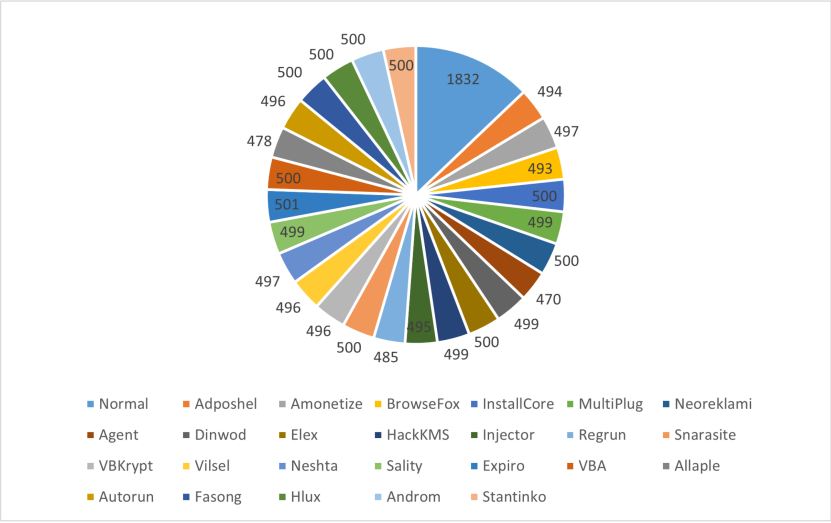

Figure 5: Distribution of the various classes of malwares