Malware Analysis in Data Science

Title

Malware image classification using deep learning models.

Team

Aaryan Mehta: 20BAI1108

Ladi Jeevan Sai: 20BAI1293

Sai Nikhil: 20BAI1217

Sarvesh Chandak: 20BAI1221

Objective

- To convert a malware into an image and analyse it.
- To develop a deep learning model that can accurately classify malware and non-malware images.
- To compare the performance of the CNN-based approach with other malware detection methods and evaluate its advantages and limitations.
- To improve the detection rate of unknown malware by using visual analysis instead of traditional methods.
- To reduce the number of false positives and false negatives in malware detection.
- To improve the overall efficiency.

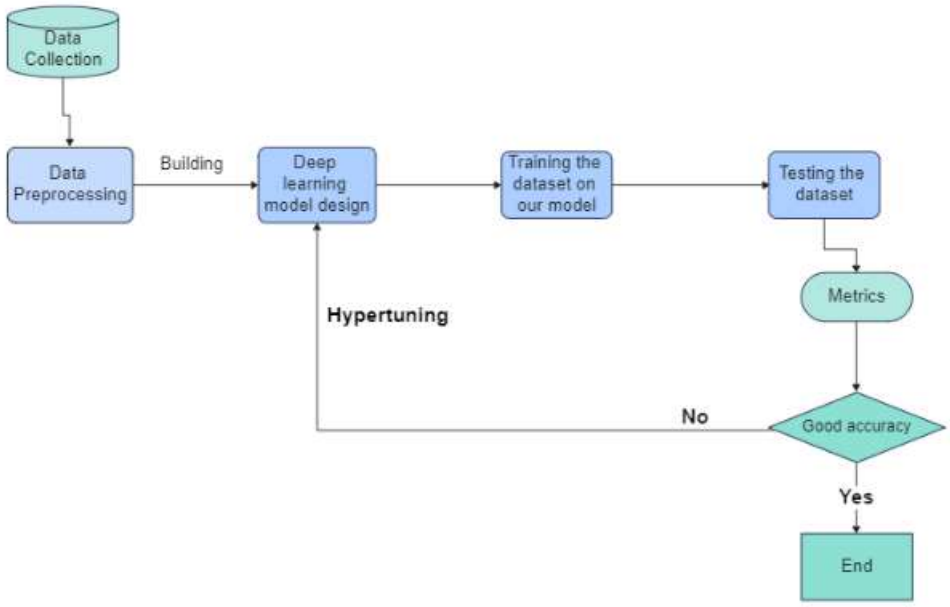
Dataset

The Malimg Dataset contains 9339 malware images, belonging to 25 families/classes. Thus, our goal is to perform a multi-class classification of malware.

Link for the dataset:

https://www.kaggle.com/competitions/malware-classification/data

Architecture Diagram



Modules Planned

- **Data Collection and Preprocessing:**

This module involves collection of datasets of labelled malware and non-malware images, and pre-processing the dataset which we will use for training and testing. This may include tasks such as resizing, augmenting and normalizing the images.

- **Deep learning Model Design:**

In this module, we will use several deep learning algorithms like cnn, vgg, resnet and Inception architectures will be designed and implemented. This may include selecting the appropriate layers, number of filters and kernel size, optimizer, and loss function for the model.

- **Training and Validation:**

Our Deep Learning model will be trained on the pre-processed dataset and fine-tuned to improve its performance. This will be done by splitting the data into training and testing, and training the model on the training set.

- **Testing and Evaluation:**

The trained models will be evaluated using the validation dataset and the test dataset. This will be done to check the accuracy, precision, recall, and other evaluation metric of the model on unseen data. Ultimately, we will select the best model out of all.