Please find the links to download the MaleX dataset-related Google Drive files below and DO NOT re-share these links:

- images directory (~28 + 37+ 8 + 5 GB tar files)
  - https://drive.google.com/drive/folders/1KgKcPUXW-E-pdDfH-pDXxKyHdCsvPa2u?usp=sharing
- features_zipped.tar.gz (~2.3 GB)
  - https://drive.google.com/file/d/1xegez-YZ-pfLrDVVrzDbFApG2qNsG7vT/view?usp=sharing
- labels_zipped.tar.gz (~200 MB)
  - https://drive.google.com/file/d/1UdoMDIrkD0iZXBkgAfYOYwaoTju6LV7m/view?usp=sharing

Be sure that you reference this paper when you use the dataset for your research: Malware Detection Using Frequency Domain-Based Image Visualization and Deep Learning (2021)

BibTeX entry:

```
@inproceedings{mohammedmalware,
    title = {Malware Detection Using Frequency Domain-Based Image Visualization and
Deep Learning},
    author = {Mohammed, Tajuddin Manhar and Nataraj, Lakshmanan and Chikkagoudar,
Satish and Chandrasekaran, Shivkumar and Manjunath, BS},
    year = 2021,
    booktitle = {Proceedings of the 54th Hawaii International Conference on System
Sciences},
    pages = 7132
}
```

Please feel free to contact us by filling out the form for any questions and refer to our Github page for more details. We also welcome suggestions and feedback on the quality of the dataset.

Please feel free to also check out our other related published works:

- SPAM: Signal Processing to Analyze Malware (2016)
- OMD: Orthogonal Malware Detection Using Audio, Image, and Static Features (2021)
- HAPSSA: Holistic Approach to PDF Malware Detection Using Signal and Statistical Analysis (2021)
- MalGrid: Visualization Of Binary Features In Large Malware Corpora (2022)

Also, check out our web-accessible service, MalSee which recasts suspect software binaries as images and exploits computer vision techniques to automatically detect malware.

MaleX Dataset README

Overview

The MaleX dataset is a comprehensive collection of malware and benign executable samples designed for cybersecurity research and machine learning applications. It is presented in two versions, each offering a unique set of features and data points to accommodate a wide range of analytical needs.

Version 1 (v1) Highlights:

- Total Samples: 864,669 malware and 179,725 benign executables.
- MIME Types: Primarily "application/x-dosexec".
- Data Points: Includes MD5, SHA256 hashes, and labels from Avast and BitDefender.
- Format: Data is organized in Hierarchical Data Format (.hdf5 files).
- Malware Families: Features a dictionary of 100 malware families, identified by SHA256 hashes, based on Avast labels with at least 750 samples per family.
- Additional Resources: Accompanied by byteplot images and bigram_dct images for further analysis.

Version 2 (v2) Highlights:

- Total Samples: 1,789,632 malware and 306,244 benign executables.
- MIME Types: Includes both "application/x-dosexec" and "application/octet-stream".
- Data Points: Expanded to include MD5, SHA256 hashes, Avast and BitDefender labels, file size, and MIME type.
- Format: Data continues to be available in Hierarchical Data Format (.hdf5 files).

Malware/Benign (MAL/BEN) Labeling Criterion

To facilitate the classification of samples into malware and benign categories, the following criterion is recommended. However, users are encouraged to apply alternative strategies as deemed appropriate:

```
if bflabel == 0 and conf <= 0.3:
    label = 0   # Benign
elif bflabel == 1 or conf > 0.3:
    label = 1   # Malware
else:
    label = 2   # Unknown
```

- bflabel: BitDefender's label (1 indicates malicious).
- conf: The VirusTotal ratio (i.e., the number of positive hits divided by the total number of AV vendors), which is not included in this dataset.

Sample Commands

To compress byteplot images into a .tar.gz archive, use the following command:

```
tar cvzf mal_byteplot_imgs_zipped.tar.gz byteplot_imgs_RxR/
```
