Deep Learning and Training Dataset Size

CS 298 Proposal, January 29, 2019

Parth Jain (parthjayantilal.jain@sjsu.edu)

Advisor: Dr. Mark Stamp (stamp@cs.sjsu.edu)

Committee Member: Dr. Thomas Austin (taustin22@gmail.com)

Committee Member: Dr. Katerina Potika (katerina.potika@gmail.com)

Committee Member: Fabio Di Troia (fabioditroia@msn.com)

The graph in [1] purports to show that the accuracy of deep learning models continue to increase, as a function of the training dataset size, after other machine learning techniques have reached a plateau. That is, deep learning is thought to be superior when the model is trained on large amounts of data, but for smaller amounts of data, other machine learning models may perform better and require significantly less computational effort. In this research, we intend to investigate the tradeoff between deep learning and other machine learning techniques, as the size of the training dataset varies. We will consider this problem in the context of malware detection, using a recently acquired large malware dataset [2].

The research done in this field have either implemented a single model or compared machine learning with deep learning models, but none of them has compared deep learning and machine learning models over different sizes of the dataset. Like, in [3] and [4] machine learning models are used for for malware classification, in [5] deep neural network is used for malware classification. In [6] deep neural network is compared with random forest, a machine learning algorithm for malware classification.

In this research, we will compare k-nearest neighbors (k-NN), support vector machines (SVM), hidden Markov models (HMM), to multiple variations of deep learning techniques. In each case, we will determine the effectiveness of the models as the size of malware training dataset varies over an extremely wide range.

CS 297 Results

(1) Researched about using KNN, SVM & deep neural network for malware detection and family of malware detection.

(2) Developed a batch script in python to take a dataset as input extract the binaries from it then convert it into bigrams construct a feature vector from it for each file.

(3) Implemented KNN algorithm using K-fold cross validation for malware family detection.

CS 298 Deliverables

(1) Documentation of the CS 298 proposal.

(2) Create a script to take any dataset containing binary files as input and create feature vector from it.

(3) Implement a script to use KNN and SVM classifier for malware family detection over different size of dataset.

(4) Implement deep learning technique to classify malware family over different size of dataset.

(5) Study the nature of the accuracy graph for three algorithms over size of dataset.

(6) Documentation of the CS 298 report.

Schedule

Date Range	Description
01/24/2019 - 01/30/2019	Documentation of the CS 298 proposal
01/24/2019 - 01/30/2019	Create a script to take any dataset containing binary files as input & create feature vector from it
02/30/2019 - 02/05/2019	Implement a script to use KNN classifier for malware family detection over different size of dataset
02/06/2019 - 02/14/2019	Implement a script to use SVM classifier for malware family detection over different size of dataset
02/15/2019 - 02/28/2019	Implement deep learning technique to classify malware family over different size of dataset
03/01/2019 - 03/15/2019	Compare the accuracy of the three algorithms over different size of dataset
03/16/2019 - 03/30/2019	Study the nature of the accuracy graph for three algorithms over size of dataset
03/31/2019 - 04/15/2019	Prepare CS 298 report
04/16/2019 - 04/30/2019	Defense practice and defense

Challenge and Innovation

(1) Collecting a suitable encrypted malware set for conducting experiments.

(2) Selecting the feature and extracting it.

(3) Running experiments with a large number of malware files.

References

[1] Kaggle, "Deep learning vs machine learning efficiency over size of data." https://ibb.co/m2bxcc, Mar 2018. (Accessed on 10/01/2018).

[2] S. Kim, "PE header analysis for malware detection, Master's Project 624, Department of Computer Science, San Jose State University." https://scholarworks.sjsu.edu/etd_projects/624/, 2018. (Accessed on 10/01/2018).

[3] B. Sanjaa and E. Chuluun, "Malware detection using linear svm," in *Ifost*, vol. 2, pp. 136–138, June 2013.

[4] M. Imran, M. T. Afzal, and M. A. Qadir, "Similarity-based malware classification using hidden markov model," in *2015 Fourth International Conference on Cyber Security, Cyber Warfare, and Digital Forensic (CyberSec)*, pp. 129–134, Oct 2015.

[5] B. Cakir and E. Dogdu, "Malware classification using deep learning methods," in *Proceedings of the ACMSE 2018 Conference*, ACMSE '18, (New York, NY, USA), pp. 10:1–10:5, ACM, 2018.

[6] M. Sewak, S. K. Sahay, and H. Rathore, "Comparison of deep learning and the classical machine learning algorithm for the malware detection," *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Jun 2018.