Peerreview: Automatic Vulnerability Detection through Machine Learning

Summary

The paper gives an introduction to the Machine Learning and how it can be used to find vulnerabilities in software. There is the possibility to analyze a software build and to analyze the source code of a software. To analyze them it is needed to format them in special ways, like a building a vector of the usage of words in a source code. For Build based analyzes there is the possibility to create graphs of the control flow in the program. To detect a vulnerability it is important to give the machine source code/builds of software which are then analyzed to train the machine learning model. After the learning phase other source code/builds can be analyzed and the machine makes decisions, if these have vulnerabilities. At the end the author compares build based models and source based models with source based models being the most accurate.

Strength

+ Figures give a good overview on stuff just explained by the author.
+ The introduction to machine learning is giving an overview of the important stuff while not going to deep into it.
+ In the evaluation different models are compared and the best ones (source-based models) are mentioned.
+ The Related Work part gives a good overview of already existing solutions to the problem of vulnerability detection

Weakness

- The paper is lacking a drawn out examples to understand how a vulnerability appears and is detected by humans or machines.
- The Discussion is quite short and does not mention the other side of vulnerability detection potentially being used to hack software.
- The benefits of that methodology are only short mentioned in the introduction and the conclusion.
- The abstract is mentioning specific terms without explaining
- There are abbreviations used while providing an explanation later.
- There is no Background despite the topic probably needing one.
- Figure 1 is not explained.
- The need to set up and configure models for vulnerability detection is better suited in the Discussion rather than the conclusion.
- The title is a bit misleading as it is not quite automated yet.

Comments

- Lookup the structure of a sentence.
- The abstract is introducing a lot of special words. These words are required to understand the topic however you are not explain them there. Try to explain machine learning without such words.
- In the second sentence in the introduction it is unclear who tries to steal what.
- In the introduction you mention 3342 million dollar, which can probably be rounded to 3.3 billion.
- You give a reason why vulnerability Detection is important in the introduction, but you do not mention it later on except for the conclusion
- The source for software errors in aerospace is quite old. Boing 737 max as an example.
- Use the template provided list feature ()
- Give more insight into the figure, make an example with them.
- Explain TextCNN, it is used in the abstract without explanation and later as well.
- You do not explain what Candr is, but you are not mentioning it afterwards.
- In IV.A, you mention that IRs are divided, this seems to be happening while compiling. It is not explained however until later.
- In IV.A, last paragraph, you mention a lot of different variables but do not examine on this topic despite being important.
- In IV.B, you mention that a weakness gets detected but not how exactly, give a small example on the decision making of a machine
- Control Flow Graphs are mentioned as CFG and later explained, switch that up.
- In IV.D you use: different you introduce TextCNN without much explanation of Neural Networks and even TextCNN.
- The Machine learning introduction is a good fit as background.
- Why is 100% accuracy unachievable?