Retrospective Reports (1)

a) setup of the retrospective meeting

When happened: 14/08/2021 5pm

Who participated: Fan Zhang, David Wu, Tianlei Qi, Siqi Sun

Key discussion items:

- The performance summary of last semester
- The feedback from the client:
  1. Paper "RepresentationLearningFromMulti_domain" should be included in our scope
  2. A detailed plan is needed by utilizing a gantt Gannt chart
  3. Time consumption for model training, model accuracy, resource occupancy rate, and operating efficiency are should be considered as the performance of the models

b) What went well,

Last semester:

- The paper "Vulnerability Discovery with Function Representation Learning from Unlabeled Projects" has been successfully replicated as a baseline for our future work
- The resulting model trained by LSTM and Random forest has the same trend with the resulting showing in the paper
- We have completed the LSTM and Random forest methods as a vulnerability detection method.

This semester:

- We had successfully replicated one data combination trained by method DDAN which was presented in the paper "*A New Approach for Cross Project Software Vulnerability Detection*". And the other method dual-ddan has been replicated as well.

- We developed a high-level understanding of the Semi-supervised Code Domain Architecture Network (SCDAN) explored in the research paper: *"Deep Domain Adaptation for Vulnerable Code Function Identification"*

- We had successfully set up the dual-dan repository and configured the model parameters to fit the requirements for the SCDAN Architecture.

- The issue Can't train an accurate model by dan.py is fixed

c) what was lacking

- Don't have the full preprocessed data set from the supplied repository which means we need to preprocess the lacking data according to the paper by ourselves.

- Do not have access to a capable machine for running the SCDAN algorithm. My local match does not meet the system requirement for training datasets. Even though the parameters are configured I was not able to replicate the method at the moment.

- Still missing the second component of the SCDAN configurations: Enforcing semi-supervised learning by ensuring the source classifier adheres to the clustering assumption

d) what did you learn

Last semester:

1. We have a comprehensive understanding of the progress for training and testing the model.
2. We learned various methods to pre-process the raw code into numerical vectors for model training.
3. We have learned to utilize external tools for project management such as TeamGantt as Gantt Chart tool.
4. The purpose of cross-domain within the context of deep learning-based vulnerability detection: providing 2 data sets, a labeled source, and an unlabelled target, leverages the unlabelled target data set via transfer learning to address the issue regarding the scarcity of labeled source code data.
5. The process of paper screening emphasized keyword searching, publish data filtering, relevance definition and contrastive analysis. We have learned to use multiple online resources to search papers.
6. We have learned to set a benchmark system to evaluate selected papers.

This semester:

1. DAN Deep learning method and applications
2. The 3 major components of the SCDAN Architecture: Generator, Classifier, and Discriminator.
   Generator: This component consists of the bidirectional RNN for training both the source and target data set with both resulting models being fed to the joint space.
   Classifier: within the joint space consist of the trained source data. applies transfer learning from the source domain to the target. Target data will be labeled.
   Discriminator: make the source and target data set indistinguishable from one other
3. How to configure and run the dual-dan algorithm and how to interpret the results.

e) What you are planning to do to improve the obstacles for the next sprint. These reflections can be around progress on implementation of user/technical stories, planning, group work/communication.

1. By using Gantt Chart, define our deadlines and milestones. Therefore, we could have a clear view of where we are and what we should do next.
2. Assign individual tasks and team overall tasks based on the user stories. This could be associated with Gantt Chart tools by setting up assignees.
3. For group work, we will continually use Google Documents and Slides. The online editor could help us on version control and check other team members' progress.
4. We have had regular weekly group meetings since last semester. In this semester we will continue this meeting, through this meeting to exchange ideas and work progress. For regular communication, we will use Slack for communication with clients, and Messenger for team chat.