西安电子科技大学计算机科学与技术学院

本科生毕业论文（设计）开题报告

（2016 级）

学生姓名 张浩楠

专　　业 计算机科学与技术

学　　号 16030130106

指导教师 张南

2019 年　11　月 15　日

（本表一式三份，学生、指导教师、学院各一份）

一、论文名称及项目来源

论文名称：基于深度学习的自动机规范挖掘器的设计与实现

指导教师：张南

题目类别：（ √ ）软件　　（ ）硬件　　（ ）软硬结合

二、研究目的和意义

研究目的：为了挖掘更精确的 FSA 模型，提出一种新的对执行轨迹进行深度学习的规范挖掘算法，并解决当下规范挖掘算法所面临的一些问题。

研究意义：由于为了满足用户需求的快速发展，软件应用程序和库经常在没有规范文档的情况下发布，即使有形式化的规范，随着系统软件在短时间内的迅速更新和迭代，它们也可能会过时，于此同时，编写形式化规范文档或说明需要开发人员具备必要的技能和动力，这是一个昂贵且耗时的过程；缺乏规范会对系统的可维护性和可靠性产生负面影响。由于缺少文档化的规范，开发人员可能很难理解一段代码，并且由于错误的假设，软件可能会有更多 bug，与此同时，开发者们难以利用需要形式化规范作为输入的先进的 bug 查找和测试工具。

基于上述所面临的一系列问题，许多自动化方法被提出来帮助开发人员降低手工起草正式规范的成本。对于这项工作，我们着重于关注众多规范挖掘算法，这些算法从执行轨迹中推断基于有限状态自动机（FSA）的规范，并提出一种新的对执行轨迹进行深度学习的挖掘算法——DSM，它能够输出更加精确的 FSA，去解决规范挖掘中面临的一些问题。例如，如果输入执行轨迹中的方法频繁地以一个特定的顺序出现，或者输入轨迹的数量过小，则会导致输出的 FSA 缺失一般性或是过拟合等问题。

如今，Android 是最受欢迎的平台，坐拥数百万个应用和支持的设备，而事实上，Android 很容易成为攻击者的目标。因此，我们提出了一种技术来使用 DSM 输出的 FSA 来构建更为全面的沙箱，该沙箱考虑敏感 API 调用的上下文——即在敏感的应用编程接口调用之前调用的接口方法，来更好地保护安卓用户免受攻击者的攻击。

三、国内外研究现状和发展趋势

（1）k-Tails。该算法最早在 1972 年由 A.W.Biermann 和 J.A.Feldman 提出，用于基于执行轨迹的规范自动挖掘。现有的 k-tail 算法通过合并来自执行轨迹的状态，将库的应用编程接口协议捕获作为只需转换的 MTS。算法将每对状态与接下来 k 次调用的相同序列合并(因此成为“k-tail”)。K-tails 算法依赖 k 值的选取，这种选择通常需要在生成模型的精度(较小的 k 意味着由于范围有限而带来更多的虚假合并)和完全性/召回(较大的 k 意味着较少的合并和较少的泛化)之间进行权衡。现有的基于 k-tails 的技术旨在用较小的 k 来提高算法模型的精度，尽管有一些改进，但仍然有所限制及不足。

（2）IvoKrka,YuriyBrun 和 NenadMedvidovic 针对 Contractors 算法进行增强，提出了 CONTRACTOR++算法，其包括两个部分：①Contractor 核心算法，这是一种最近专门基于程序不变量创建 MTS 模型的算法。②对 Contractor 的优化，使其能够处理推断程序不变量而不是手动指定。

（3）IvoKrka,YuriyBrun 和 NenadMedvidovic 提出了 SEKT 和 TEMI 两种新的算法，这两种算法将执行轨迹与自动推断的程序状态不变量相结合。SEKT 通过添加新的全局合并要求来修改 k-tail 算法：合并状态必须对应相同的抽象程序状态。TEMI 算法由两个阶段组成，第一个阶段，在概念上类似于 CONTRACTOR++，构建一个只有转换的 MTS，捕获对象接口的所有调用序列，称这个模型为基于不变量的模型转换系统；第二个阶段促进在可能的轨迹中观察到的转换。

随着软件系统在短时间内的进一步发展， 自动化规范挖掘出的规范的质量还不完善，通过结合深度学习来实现更为精确的 FSA 模型，进一步提高 F-measures 度量值及算法模型在实际应用中的有效性。

四、主要研究内容、要解决的问题及本文的初步方案

主要研究内容:

设计出一种结合深度学习的自动机规范挖掘器，实现具有较高精准度的 FSA 模型。

要解决的问题:

1. 测试用例的生成和跟踪收集
2. RNNLM 底层架构的选择和网络训练
3. 自动机构造

初步方案:

1. 选择 Randoop 作为测试用例生成工具，并且为了提高方法序列的覆盖率，除了缺省值，还向 Randoop 提供特定特定类的的文本来加快 Randoop 创建新对象的速度，而无需花费时间为构造函数搜索合适的输入值。
2. 选择长短期记忆网络（LSTM）作为 RNNLM 的底层架构，与标准的递归神经网络（RNN）架构相比，LSTM 在学习长期依赖性方面更好。此外，LSTM 对于长序列是可伸缩的。
3. 将 FSA 构建划分为四个子步骤：跟踪采样、特折提取、聚类和模型选择。

   跟踪采样：该子步骤使用通过一定启发式的算法，得出能够代表所有训练轨迹的方法序列子集，而非全部的训练集，以此来降低计算成本。

   特折提取：主要利用前缀树，在所选轨迹中构造一个 PTA，使用训练好的 RNNLM 对每个树节点进行特征提取。

   聚类：主要利用不同参数设置的聚类算法（K-means 和层次聚类），在不同设置的 PTA 节点上运行来创建许多不同的 FSA。

   模型选择：遵循一种特定的算法来预测构造的 FSAs 的 F 度量，并输出具有最高 F 度量的 FSA。

五、工作的主要阶段、进度和完成时间

起止时间	工作内容
2019.12.1—2020.1.7	搜集、查阅资料，阅读相关文献，熟悉掌握规范挖掘、有限状态自动机、RNN、语言模型、PTA 等基本概念和知识；学习 Randoop 的安装使用。
2020.1.8—2020.2.11	研究整体算法的基本框架和步骤，学习特折提取和模型选择中运用到的算法并实现。
2020.2.12—2020.3.17	设计并实现一种基于深度学习的自动机规范挖掘器。
2020.3.18—2020.5.5	撰写论文。
2020.5.6—2020.5.19	修改论文。
2020.5.20—2020.5.31	准备答辩。

六、已进行的前期准备工作

1.搜集并查阅了相关规范挖掘和自动机的资料，对于要实现的规范挖掘器的架构和思想有了初步的认识，尝试安装 Randoop 并学习使用，尝试实现前缀树代码。

2.查找相关资料进一步学习 RNN 以及 LSTM 和 RNNLM 的原理和内容，学习 tensorflow 中的函数调用，以方便后期实现。

3.阅读关于基于深度学习的规范挖掘文献，同时学习以前以及提出的关于规范挖掘的算法和文献，但是在自动机构造的子步骤：对 PTA 的节点进行特折提取、在具有不同设置的 PTA 节点上运行许多聚类算法来创建许多不同的 FSA，不是特别的理解，有待继续学习和尝试。

指导教师签字：