Background

As Android apps are playing a more important role in enhancing human life by carrying out various tasks, it is essential to guarantee that they are of high quality with testing conducted. One of the popular approaches to ensure the quality of Android apps is User Interface (UI) testing. However, there are only a few tools for applicability evaluation of industrial apps, while the majority of performance evaluation subjects include only open-source apps.

Introduction

Our empirical study aims to provide informative analysis on the advantages and disadvantages of existing test generation tools, helping app and tool developers improve their design and implementation of tools, as well as the ability of handling more realistic tasks.

Study Methodology

A. Industrial-App Selection

From each category, we sampled on average five top-recommended apps with the most downloads, and eventually obtained 69 industrial apps which are compatible with Android 4.4. The most recent version of Android supported by all the test generation tools under study as well as most of the top-recommended apps.

B. Coverage/Crash Measurement

For code-coverage measurement, we use Ella [12] to instrument all the industrial apps, and collect statistics of method coverage during testing.

C. Study Setup

Each test generation tool was run continuously for 3 hours on each of their applicable industrial apps under study. Our study is conducted on official Android x86 emulators and 4 real phones, all running Android 4.4.

Table II: Overview of Android test generation tools under study

<table>
<thead>
<tr>
<th>Tool</th>
<th>Evaluation Subjects</th>
<th>#Industrial Apps</th>
<th>#Open-source Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monkey</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WCtester</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sapienz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DroidBot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3E-Depth-First</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Practicality of Existing Tools’ Evaluations

We notice that some existing test generation tools are evaluated on only open-source apps while some other tools are also evaluated on industrial apps at the time when the tools are published. Table III presents an overview of the categorization of evaluation subjects for each of the 10 test generation tools at the time when the tool is published. As shown by our study results, open-source apps and large industrial apps can actually have different characteristics with respect to usage patterns, richness of features, and manipulability. Such differences could potentially lead to different evaluation results. Thus, it is both necessary and important for researchers and practitioners to consider large industrial apps when they use real-world apps as evaluation subjects.

Conclusion and future work

In our research, we have done an empirical study of applying existing available Android test generation tools on industrial apps. By studying the effectiveness of these tools on both open-source and industrial apps and analyzing the coverage achieved by each applicable test generation tool, we investigated the characteristics of covered activities achieved by those applicable test generation tools. To be more specific, we have applied all applicable test generation tools on 68 open-source apps and 5 industrial apps provided by Choudhary et al. [11] and have compared the resulting statistics to check the effectiveness of test generation tools as well as the consistency of conclusions under different settings. Moreover, we have analyzed transition depths and have ranked covered activities to better study the effectiveness of test generation tool.

Analysis of Covered Activities

The average distribution of activity transition depths on industrial and open-source apps for each test generation tool under study are presented in Fig.1 and Fig.2. As shown in the figures above, most of the covered activities on industrial apps are depth-1 and depth-2 activities, while the majority of the covered activities on open-source apps are depth-0 and depth-1 activities. What’s more, according to the ATGs, the deepest activity observed on industrial apps is 5 transitions away from launcher activities, while the activity transition depths are all no more than 3 on open-source apps. This inconsistency reveals the fact that industrial apps have more sophisticated activity structures than open-source apps.

Acknowledgement