Deep Reinforcement Learning for Black-Box Testing of Android Apps

ANDREA ROMDHANA, DIBRIS - Università degli Studi di Genova, FBK-ICT, Security & Trust Unit
ALESSIO MERLO, DIBRIS - Università degli Studi di Genova
MARIANO CECCATO, Università di Verona
PAOLO TONELLA, Università della Svizzera italiana

The state space of Android apps is huge and its thorough exploration during testing remains a major challenge. In fact, the best exploration strategy is highly dependent on the features of the app under test. Reinforcement Learning (RL) is a machine learning technique that learns the optimal strategy to solve a task by trial and error, guided by positive or negative reward, rather than by explicit supervision. Deep RL is a recent extension of RL that takes advantage of the learning capabilities of neural networks. Such capabilities make Deep RL suitable for complex exploration spaces such as the one of Android apps. However, state of the art, publicly available tools only support basic, tabular RL. We have developed ARES, a Deep RL approach for black-box testing of Android apps. Experimental results show that it achieves higher coverage and fault revelation than the baselines, which include state of the art tools, such as TimeMachine and Q-Testing. We also investigated qualitatively the reasons behind such performance and we have identified the key features of Android apps that make Deep RL particularly effective on them to be the presence of chained and blocking activities.

CCS Concepts:

- Software verification and validation → Empirical software validation.

Additional Key Words and Phrases: Deep reinforcement learning, Android testing

ACM Reference Format:

Andrea Romdhana, Alessio Merlo, Mariano Ceccato, and Paolo Tonella. 2021. Deep Reinforcement Learning for Black-Box Testing of Android Apps. ACM Trans. Softw. Eng. Methodol. 01, 1, Article 000 (January 2021), 26 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The complexity of mobile applications (hereafter, apps) keeps growing, as apps provide always more advanced services to the users. Nonetheless, it is of utmost importance that they work properly once they are published on app markets as most of their success/failure depend on the user’s evaluation. Therefore, an effective testing phase is fundamental to minimize the likelihood of app failures during execution. However, automated testing of mobile apps is still an open problem and the complexity of current apps makes their exploration trickier than in the past, as they can contain states that are difficult to reach and events that are hard to trigger.

There exist several approaches to automated testing of mobile apps that aim at maximizing code coverage and bug detection during testing. Random testing strategies [20, 30] stimulate the App

Authors’ addresses: Andrea Romdhana, andrea.romdhana@dibris.unige.it, DIBRIS - Università degli Studi di Genova, FBK-ICT, Security & Trust Unit; Alessio Merlo, alessio.merlo@unige.it, DIBRIS - Università degli Studi di Genova; Mariano Ceccato, mariano.ceccato@univr.it, Università di Verona; Paolo Tonella, paolo.tonella@usi.ch, Università della Svizzera italiana.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.
1049-331X/2021/1-ART000 $15.00
https://doi.org/10.1145/1122445.1122456

ACM Trans. Softw. Eng. Methodol., Vol. 01, No. 1, Article 000. Publication date: January 2021.
Under Test (AUT) by producing pseudo-random events. However, random exploration with no
guidance may get stuck when dealing with complex transitions. Model-Based strategies [4, 21, 39]
extract test cases from navigation models built by means of static or dynamic analysis. If the
model accurately reflects the AUT, a deep exploration can be achieved. Nonetheless, automatically
constructed models tend to be incomplete and inaccurate. Structural strategies [6, 16, 31] generate
coverage-oriented inputs using symbolic execution or evolutionary algorithms. These strategies
are more powerful, since they are guided by a specific coverage target. However, they do not take
advantage of past exploration successes to learn the most effective exploration strategy dynamically.

Reinforcement Learning (RL) is a machine learning approach that does not require a labeled
training set as input, since the learning process is guided by the positive or negative reward
experienced during the tentative execution of the task. Hence, it represents a way to dynamically
build an optimal exploration strategy by taking advantage of the past successful or unsuccessful
moves.

RL has been extensively applied to the problem of GUI and Android testing [34, 36]. However,
only the most basic form of RL (i.e., tabular RL) has been applied to testing problems so far. In tabular
RL the value of the state-action associations is stored in a fixed table. The advent of Deep Neural
Networks (DNN) replaced tabular approaches with deep learning ones, in which the action-value
function is learned from the past positive and negative experiences made by one or more neural
networks. When the state space to explore is extremely large (e.g. when an app has a big amount
of widgets), Deep RL has proved to be substantially superior to tabular RL [8] [37] [27]. In this
respect, we argue that the state space of Android apps is definitely a good candidate for a successful
adoption of Deep RL instead of tabular RL for testing purposes.

This paper presents the first Deep RL approach, ARES, for automated black-box testing of Android
apps. ARES uses a DNN to learn the best exploration strategy from previous attempts. Thanks to
such DNN, it achieves high scalability, general applicability and the capability to handle complex
app behaviors.

ARES implements multiple Deep RL algorithms that come with a set of configurable, often
critical, hyperparameters. To speed up both the selection of the most appropriate algorithm for the
AUT and the fine tuning of its hyperparameters, we have developed another tool, FATE, which
integrates with ARES.

FATE is a simulation environment that supports fast assessment of Android testing algorithms
by running synthetic Android apps (i.e., abstract navigational models of real Android apps). The
execution of a testing session on a FATE synthetic app is on average 10 to 100 times faster than the
execution of the same session on the corresponding real Android app.

We applied ARES to two benchmarks made by 41 and 68 Android apps, respectively. The first
benchmark compares the performance of the ARES algorithms, while the latter one evaluates ARES
w.r.t. the state-of-the-art testing tools for Android.

Experimental results confirmed the hypothesis that Deep RL outperforms tabular RL in the
exploration of the state space of Android apps, as ARES exposed the highest number of faults and
obtained the highest code coverage. Furthermore, we carried out a qualitative analysis showing
that the features of Android apps that make Deep RL particularly effective include, among others,
the presence of concatenated activities and blocking activities, protected by authentication.

To sum up, this paper provides the following advancements to the state of the art:

- ARES, the first publicly available testing approach based on Deep Reinforcement Learning,
  released as open source;
- FATE, a simulation environment for fast experimentation of Android testing algorithms, also
  available as open source;
Deep Reinforcement Learning for Black-Box Testing of Android Apps

• A thorough empirical evaluation of the proposed approach, whose replication package is publicly available to the research community.

2 BACKGROUND

After a general overview on RL, this section presents in more detail Tabular RL and Deep RL.

2.1 Overview on Reinforcement Learning

The objective of Reinforcement Learning [41] is to train an agent that interacts with some environment to achieve a given goal. The agent is assumed to be capable of sensing the current state of the environment, and to receive a feedback signal, named reward, each time the agent takes an action.

At each time step \( t \), the agent receives an observation \( x_t \), takes an action \( a_t \) that causes the transition of the environment from state \( s_t \) to state \( s_{t+1} \). The agent also receives a scalar reward \( R(x_t, a_t, x_{t+1}) \), that quantifies the goodness of the last transition.

For simplicity, let us assume \( x_t = s_t \) (in the general case, \( x_t \) might be a subset of \( s_t \)). The behavior of an agent is represented by a policy \( \pi \), i.e., a rule for making the decision on what action to take, based on the perceived state \( s_t \). A policy can be:

- Deterministic: \( a_t = \pi(s_t) \), i.e. a direct mapping between states and actions;
- Stochastic: \( \pi(a_t|s_t) \), a probability distribution over actions, given their state.

DDPG [29] and TD3 [15] are examples of RL algorithms that learn a deterministic policy, while SAC [22] is a RL algorithm that learns a stochastic policy.

The standard mathematical formalism used to describe the agent environment is a Markov Decision Process (MDP). An MDP is a 5-tuple, \( \langle S, A, R, P, \rho_0 \rangle \), where:

- \( S \) is the set of all valid states,
- \( A \) is the set of all valid actions,
- \( R : S \times A \rightarrow \mathbb{R} \) is the reward function, with \( r_t = R(s_t, a_t, s_{t+1}) \),
- \( P : S \times A \rightarrow P(s) \) is the transition probability function, with \( P(s_{t+1}|s_t, a_t) \) being the probability of transitioning into state \( s_{t+1} \) starting from state \( s_t \) and taking action \( a_t \),
- \( \rho_0(s) \) is the starting state distribution.

Markov Decision Processes obey the Markov property: a transition only depends on the most recent state and action (and not on states/actions that precede the most recent ones).

The goal in RL is to learn a policy \( \pi \) which maximizes the so-called expected return, which can be computed as:

\[
R(\tau) = \sum_{t=0}^{\infty} y^t r_t
\]

A discount factor \( y \in (0, 1) \) is needed for convergence. It determines how much the agent cares about rewards in the distant future relative to those in the immediate future. \( \tau \) is a sequence of states and actions in the environment \( \tau = (a_0, s_0, a_1, s_1...) \), named trajectory or episode. Testing an Android app can be seen as a task divided into finite-length episodes.

To estimate the expected return in practice, it is convenient to express it using a value function and an action-value function. The value function \( V^\pi(s) \) is defined as the expected return starting in a state \( s \) and then acting according to a given policy \( \pi \):

\[
V^\pi(s) = E[R_t|s_0 = s]
\]

The action-value function \( Q^\pi(s, a) \) can be used to describe the expected return after taking an action \( a \) in state \( s \) and thereafter following the policy \( \pi \):
Correspondingly, we can define the \textit{optimal value function}, \( V^*(s) \), as the \( V^*(s) \) that gives the highest expected return when starting in state \( s \) and acting according to the optimal policy in the environment. The \textit{optimal action-value function}, \( Q^*(s,a) \), gives the highest achievable expected return under the constraints that the process starts at state \( s \), takes an action \( a \) and then acts according to the optimal policy in the environment.

A policy that chooses greedy actions only with respect to \( Q^* \) is optimal, i.e., knowledge of \( Q^* \) alone is sufficient for finding the optimal policy. As a result, if we have \( Q^* \), we can directly obtain the optimal action, \( a^*(s) \), via \( a^*(s) = \text{argmax}_a Q^*(s,a) \). The optimal value function \( V^*(s) \) and action-value function \( Q^*(s,a) \) can be computed by means of a recursive relationship known as the \textit{optimal Bellman equations}:

\[
V^*(s_t) = \max_a E[r(s_t, a_t) + \gamma V^*(s_{t+1})]
\]
\[
Q^*(s_t, a_t) = E[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})]
\]

### 2.2 Tabular RL

Tabular techniques refer to RL algorithms where the state and action spaces are approximated using value functions defined by means of tables. In particular, Q-Learning [43] is one of the most adopted algorithms of this family. Q-Learning aims at learning a so-called Q-Table, i.e., a table that contains the value of each state-action pair. A Q-Table represents the current estimate of the action-value function \( Q(s,a) \). Every time an action \( a_t \) is taken and a state \( s_t \) is reached, the associated state-action value in the Q-Table is updated as follows:

\[
Q(s_t, a_t) := Q(s_t, a_t) + \alpha (r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))
\]

where \( \alpha \) is the learning rate (between 0 and 1); \( \gamma \) is the discount factor, applied to the future reward. Typically, \( \gamma \) ranges from 0.8 to 0.99 [35] [29] [15], while \( \max_a Q(s_{t+1}, a) \) gives the maximum value for future rewards across all actions. It is used to update the reward for the current state-action pair.

RL algorithms based on the tabular approach do not scale to high-dimensional problems, because in such cases it is difficult to manually define a good initial Q-Table. In case a good initial Q-Table is not available, convergence to the optimal table by means of the update rule described above is too slow [29].

### 2.3 Deep Reinforcement Learning

In large or unbounded discrete spaces, where representing all states and actions in a Q-Table is impractical, tabular methods become highly unstable and incapable to learn a successful policy[35]. The rise of deep learning, relying on the powerful function approximation properties of deep neural networks, has provided new tools to overcome these limitations. One of the first deep reinforcement learning algorithms is DQN (Deep Q-Networks) [35].

DQN uses convolutional neural networks to approximate the computation of the action-value function \( Q^\pi \). Training of such neural networks is achieved by means of memory replay: the last \( N \) experience tuples are sampled uniformly and replayed when updating the weights of the networks.

While DQN can indeed solve problems with high-dimensional observation spaces, it can only handle discrete and low-dimensional action spaces. The recent advancements over DQN described in the following paragraphs (namely, DDPG, TD3 and SAC) overcome such limitation and allow dealing with high-dimensional action spaces.
Deep Deterministic Policy Gradient (DDPG). DDPG [29] is an Actor-Critic algorithm, i.e., it includes two roles: the Critic, which estimates the value function, and the Actor, which updates the policy \( \pi \) in the direction suggested by the Critic. It is based on a deterministic policy gradient [38] that can operate over continuous action spaces.

More specifically, the Critic has the objective of learning an approximation of the function \( Q^*(s) \). Suppose that the approximator is a neural network \( Q_\phi(s, a) \), with parameters \( \phi \), and that we have a set \( D \) of transitions \( (s_t, a_t, r_t, s_{t+1}, d) \), where \( d=1 \) if \( s_{t+1} \) is a final state; 0 otherwise. Such set \( D \) of previous experiences is called replay buffer. To avoid overfitting, the replay buffer should be large enough. The loss function of the neural approximator is the so-called Mean-Squared Bellman Error (MSBE), which measures how close \( Q_\phi(s, a) \) is to satisfying the Bellman equation:

\[
L(\phi, D) = E[(Q_\phi(s_t, a_t) - (r_t + \gamma(1 - d) \max_{a_{t+1}} Q_\phi(s_{t+1}, a_{t+1})))^2]
\]

The term subtracted from \( Q_\phi(s_t, a_t) \) is named the target, because minimization of MSBE makes the Q-function as close as possible to this value. Since the target depends recursively on the same parameter \( \phi \) to train, MSBE minimization can become unstable. The solution is to use a second neural network, called target network, whose parameters are updated to be close to \( \phi \), but with some time delay that gives stability to the process.

The Actor’s goal is to learn a deterministic policy \( \pi_\theta(s) \) which maximizes \( Q_\phi(s, a) \). Because the action space is continuous and we assume the Q-function is differentiable with respect to the action, we can use gradient ascent with respect to the policy parameter.

Twin Delayed DDPG (TD3). Although DDPG can often achieve good performance, it tends to be susceptible to critical tuning of its hyperparameters. This happens when the hyperparameters of the Q-function make it overestimate the Q-values, which in turn leads to policy failure, as the Actor is actually exploiting such errors as guidance.

TD3 [15] is an algorithm which addresses this issue by introducing three major changes, mostly on the Critic side: (1) Clipped Double Q-Learning: TD3 learns two Q-functions instead of one, and uses the smaller of the two Q-values as the target in the MSBE function. (2) Delayed Policy Update: TD3 updates the policy and the target networks less frequently than the Q-function. (3) Target Policy Smoothing: TD3 adds noise to the target action to make it harder for the policy to exploit Q-function errors, by smoothing out Q across changes of the action.

Soft Actor Critic (SAC). The central feature of SAC [22] is entropy regularization. Using the entropy-regularized method, an agent gets a bonus reward at each time step which is proportional to the entropy of the policy at that time step. In fact, differently from TD3, the policy of SAC is non deterministic and inclusion of entropy in the reward aims at promoting policies with a wider spread of alternatives to choose stochastically from. The RL problem becomes the problem of finding the optimal policy \( \pi^* \) according to the following equation:

\[
\pi^* = \arg\max_\pi E\left[ \sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot|s_t))) \right]
\]

where \( H \) is the entropy, and \( \alpha > 0 \) is the trade-off coefficient.

3 APPROACH

In this section, we describe ARES (Application of REinforcement learning to android Software testing), our approach to black-box Android GUI testing based on Deep RL. The overview of the approach is shown in Figure 1. The RL environment is represented by the Android application under test (AUT), which is subject to several interaction steps. At each time step, assuming the GUI state is \( s_t \), ARES first takes an action \( a_t \). Then, it observes the new GUI state \( s_{t+1} \) of the AUT.
Fig. 1. The ARES testing workflow

and a reward $r_t$ is computed. Intuitively, if the new state $s_{t+1}$ is similar to the prior state $s_t$, the reward is negative. Otherwise, the reward is a large positive value. In this way, ARES promotes the exploration of new states in the AUT, under the assumption that this is useful to test the application more thoroughly.

The reward is used to update the neural network, which learns how to guide the Deep RL algorithm to explore the AUT in depth. The actual update strategy depends on the deep RL algorithm being used (either DDPG, TD3 or SAC).

### 3.1 Problem Formulation

To apply RL, we have to map the problem of Android black-box testing to the standard mathematical formalization of RL: an MDP, defined by the 5-tuple, $\langle S, A, R, P, \rho_0 \rangle$. Moreover, we have to map the testing problem onto an RL task divided into several finite-length episodes.

**State Representation.** Our approach is black-box, because it does not access the source code of the AUT. It only relies on the GUI of the AUT. The state $s_t \in S$ is defined as a combined state $(a_0, \ldots, a_n, w_0, \ldots, w_m)$. The first part of the state $a_0, \ldots, a_n$ is a one hot encoding of the current activity, i.e., $a_i$ is equal to 1 only if the currently displayed activity is the $i$-th activity, it is equal to 0 for all the other activities. In the second part of the state vector, $w_j$ is equal to 1 if the $j$-th widget is available in the current activity; it is equal to 0 otherwise.

**Action Representation.** User interaction events in the AUT are mapped to the action set $A$ of the MDP. ARES infers executable events in the current state by analyzing the dumped widgets and their attributes (i.e., `clickable`, `long-clickable`, and `scrollable`). In addition to the widget-level actions we also model two system-level actions, namely `toggle internet connection` and `rotate screen`. These system-level actions are the only system actions that can be easily tested. In fact, since Android version 7.0, testing other system-level actions (like those implemented in Stoat [39]) would depend on the Android version used [40], [19], and would require a rooted device [17]. In fact, working with a rooted device compromise the testing of protected apps [42], such as apps in the finance category.

Each action $a$ is 3-dimensional: the first dimension represents the widget ARES is going to interact with or the identifier of a system action. The second dimension specifies a string to be used as text input, if needed. Actually, an index pointing to an entry in a dictionary of predefined strings is used for this dimension. The third dimension depends on the context: when the selected widget is both `clickable` and `long-clickable`, the third action determines which of the two actions to
take. When ARES interacts with a scrollable object, the third dimension determines the scrolling direction.

**Transition Probability Function.** The transition function $P$ determines which state the application can transit to after ARES has taken an action. In our case, this is decided solely by the execution of the AUT: ARES observes the process passively, collecting the new state after the transition has taken place.

**Reward Function.** The RL algorithm used by ARES receives a reward $r_t \in R$ every time it executes an action $a_t$. We define the following reward function:

$$r_t = \begin{cases} 
\Gamma_1 & \text{if } act(s_{t+1}) \notin act(E_t) \text{ or crash} \\
-\Gamma_2 & \text{if } pack(act(s_{t+1})) \neq pack(AUT) \\
-\Gamma_3 & \text{otherwise}
\end{cases}$$

with $\Gamma_1 \gg \Gamma_2 \gg \Gamma_3$ (in our implementation $\Gamma_1 = 1000$, $\Gamma_2 = 100$, $\Gamma_3 = 1$).

The exploration of ARES is divided into episodes. At time $t$, the reward $r_t$ is high ($\Gamma_1$) if ARES was able to transition to an activity never observed during the current episode $E_t$ (i.e., the next activity $act(s_{t+1})$ does not belong to the set of activities encountered so far in $E_t$): if a new episode is started at $t+1$, its set of activities is reset: $act(E_{t+1}) = \emptyset$.

Resetting the set of encountered activities at the beginning of each new episode is a technique that encourages ARES to visit and explore the highest number of activities in each episode, so as to continuously reinforce its explorative behaviors. In contrast, giving a positive reward for each new activity only once, or, on the contrary, each time we see an activity change, would result respectively in too sparse, singular reinforcement feedback or in rewarding of cycling behaviors [12].

The reward is high ($\Gamma_1$) also when a faulty behavior (crash) occurs. It is very low ($-\Gamma_2$) when the displayed activity does not belong to the AUT (i.e., the package of the current activity, $pack(act(s_{t+1}))$, is not the package of the AUT), as we aim to explore the current AUT only. In all other cases, the reward is moderately negative ($-\Gamma_3$), as the exploration still remains inside the AUT, even if nothing new has been discovered.

### 4 IMPLEMENTATION

ARES features a custom environment based on the OpenAI Gym [9] interface, which is a de-facto standard in the RL field. OpenAI Gym is a toolkit for designing and comparing RL algorithms and includes a number of built-in environments. It also includes guidelines for the definition of custom environments. Our custom environment interacts with the Android emulator [18] using the Appium Test Automator Framework [3].

#### 4.1 Tool Overview

As soon as it is launched, ARES leverages Appium to dump the widget hierarchy of the GUI in the starting activity of the AUT. The widget hierarchy is analyzed searching for clickable, long-clickable and scrollable widgets. These widgets are afterwards stored in a dictionary containing several associated attributes (e.g., resource-id, clickable, long-clickable, scrollable, etc.) and compose the action vector, i.e., the vector of executable actions in the current state. At each time step, ARES takes an action according to the behavior of the exploration algorithm. Once the action has been fully processed, ARES extracts the new widget hierarchy from the current GUI and calculates its MD5 hash value. If it has the same MD5 value of the previous state, ARES leaves the action vector unchanged. If the MD5 value does not match, ARES updates the action vector. ARES performs the
observation of the AUT state and returns the combined vector of activities and widgets. ARES organizes the testing of each app as a task divided into finite-length episodes. The goal of ARES is to maximize the total reward received during each episode. Every episode lasts at least 250 time steps. Its duration is shorter only if the app crashes. Once an episode comes to an end, the app is restarted and ARES uses the acquired knowledge to explore the app more thoroughly in the next episode.

4.2 Application Environment

The application environment is responsible for handling the actions to interact with the AUT. Since the environment follows the guidelines of the Gym interface, it is structured as a class with two key functions. The first function \texttt{init(desired\_capabilities)} is the initialization of the class. The additional parameter \texttt{desired\_capabilities} consists of a dictionary containing the emulator setup along with the application to be tested. The second function is the \texttt{step(a)} function, that takes an action \texttt{a} as command and returns a list of objects, including observation (current AUT state) and reward.

4.3 Algorithm Implementation

ARES is a modular framework that adopts a plugin architecture for the integration of the RL algorithm to use. Hence, extension with a new exploration algorithm can be easily achieved. In the current implementation, ARES provides five different exploration strategies: (1) random, (2) Q-Learning, (3) DDPG, (4) SAC, (5) TD3. The random algorithm interacts with the AUT by randomly selecting an action from those in the action vector. Compared to Monkey [20], our random approach performs better, since it selects only actions from the action vector. In fact, Monkey generates random, low level events on the whole GUI, which could target no actual widget and then be discarded.

Our Q-Learning strategy implements the algorithm by Watkins and Dayan [43]. The deep RL algorithms available in ARES are DDPG, SAC and TD3. Their implementation comes from the Python library \textit{Stable Baselines} [23], and allow ARES to save the status of the neural network as a policy file, at the end of the exploration. In this way, the policy can be loaded and reused on a new version of the AUT at a later stage, rather than restarting ARES from scratch each time a new AUT version is released. ARES is publicly available as open source software at https://github.com/H2SO4T/ARES.

4.4 Compatibility

ARES has been successfully tested on Windows 10, MacOS 11.1 (and older), Ubuntu 20 (and older) and Scientific Linux 7.5. ARES is fully compliant with parallel execution and enables parallel experiments to be performed on emulators or real devices, handling each instance in a completely separate manner. ARES is also compatible with several Android versions (i.e., it has been successfully tested on Android 6.0, 7.0, 7.1, 8.0, 8.1, 9.0, and 10.0). Moreover, since ARES is based on the standard OpenAIGym, new algorithms and exploration strategies can be easily added to the tool.

5 FAST ANDROID TEST ENVIRONMENT (FATE)

Deep RL algorithms require fine tuning, which is expensive on real apps. Therefore, we developed FATE, a simulation environment for fast Android testing. FATE models only the navigation constraints of Android apps, so it can be used to efficiently compare alternative testing algorithms and to quickly tune their corresponding hyperparameters. After this algorithm selection and tuning phase through FATE is completed, the selected algorithms and their configurations are ported to ARES to test real apps.

ACM Trans. Softw. Eng. Methodol., Vol. 01, No. 1, Article 000. Publication date: January 2021.
5.1 Model-based Prototyping

In FATE, developers model an Android app by means of a deterministic Finite State Machine (FSM) \( F = (\Sigma, S, s_0, \delta, F) \), where \( \Sigma \) is a set of events, \( S \) a set of states with \( s_0 \) the initial state and \( F \) the set of final states, and \( \delta \) the state transition function \( \delta : S \times \Sigma \rightarrow 2^S \). The states \( S \) of the FSM correspond to the activities of the app, while the events \( \Sigma \) trigger the transitions between activities, which in turn are modeled as a transition table \( \delta \). Events represent the clickable widgets (e.g., buttons) available in each activity. Transitions have access to a set of global variables and possess, among others, the following attributes: ID, type, active (boolean attribute), guard (boolean expression that prevents the transition from being taken if it evaluates to false), set (new values to be assigned to global variables), destination (target activity, i.e., value of \( \delta \)).

Figure 2 shows the FATE model of the prototypical app Social Network. To build such model, developers can use Ptolemy [26] to graphically draw a FSM that mimics the behavior of the application. While creating a FSM with Ptolemy is not mandatory in FATE, it simplifies the job of designing a logically correct model, thanks to the checks it performs. Then, FATE translates automatically the Ptolemy model (saved in XML format) into a JSON file that replicates the structure and behavior of the Ptolemy model. The JSON model translation has two main fields: global vars and nodes. The first contains a list of global variables organized by name and value. The latter contains a list of all the activities. Each activity is characterized by a node_id and a list of corresponding node transitions, each including all the respective transition attributes.

The model of Social Network in Figure 2 contains a transition from login to main_act which is subjected to the guard user_pass == real_pass, i.e., the entered password must be correct in order for the transition to be taken. In the JSON model, such transition is coded as:

```json
"transition": {
  "transition_id": 0,
  "type": "button",
  "active": true,
  "guard": "user_pass == real_pass",
  "set": null,
  "destination": "main_act"
}
```
Another example of guarded transition is the one between messages and chat\_x. The guard $\text{count Messages} \geq 1$ checks whether there exists at least one message from which a chat thread can be started. In the Json model this is coded as:

```json
"transition": {
  "transition_id": 0,
  "type": "button",
  "active": true,
  "guard": "count_messages \geq 1",
  "set": null,
  "destination": "main_act"
}
```

In FATE, a Python environment compliant with the OpenAI Gym standard takes as input the Json app model and tests it automatically using the selected algorithm. The available algorithms are: (1) Random, (2) Q-Learning, (3) DDPG, (4) SAC, (5) TD3. FATE was built with modularity in mind and new exploration algorithms can be easily added to the tool.

Compared to testing an Android app through Espresso or Appium, FATE makes test case execution extremely faster, because there is no need to interact with the app via its GUI. Moreover, the application navigation logic is simulated by the transition function $\delta$, which makes it usually much faster to execute. As a consequence, developers can run a large number of experiments, evaluate multiple algorithms, check various algorithm or application configurations, and find the optimal set of hyperparameters, all of which would be prohibitively expensive to execute on a standard Android testing platform. FATE is publicly available as open source software at https://github.com/H2SO4T/ARES.

### 5.2 Representative Family of Models

For fast evaluation of the Deep RL algorithms implemented in ARES, we modeled four Android apps using FATE. Each model is configurable with a variable degree of complexity. To obtain a set of app models that are representative of the most common apps used in everyday life, we inspected AppBrain (a website that aggregates Google Play Store statistics and rankings) [2] and selected four different and representative categories from the top-ten: Music & Audio, Lifestyle, Business, and Shopping. From each category we then selected and modeled in FATE one prototypical app: Player, Social Network, Bank and Market Place.

The simplest scenario is Player. It features a wide number of activities arranged in a tree-like structure. It reflects the generalization of a variety of applications, including apps or app components to manage the settings and to stream/add/remove media contents. Social Network (see Figure 2) starts by prompting a login activity with fields for username and password. Following the login activity we have several activities that replicate the behavior of a standard social network, including a basic chat activity. The Bank model is characterized by the presence of inner password-protected operations. Market Place models a typical app for e-commerce: the user can search for goods, login, purchase products, and monitor the orders. The four representative app models used in this work are publicly available inside the FATE tool.

### 6 EVALUATION

We seek to address the following research questions, split between the following two studies:

**Study 1 (FATE):**
• **RQ1** Which deep RL algorithm and which algorithm configuration performs better on the synthetic apps?

• **RQ2** How does activity coverage vary as the model of the AUT becomes increasingly difficult to explore?

• **RQ3** What are the features of the synthetic apps that allow deep RL to perform better than Q-Learning?

• **RQ4** Are the results of synthetic apps comparable to those of the their translated counterparts?

In Study 1, we take advantage of the fast execution granted by FATE to compare alternative RL algorithms (RQ1) and determine their optimal configuration (see Appendix 1). Since in this study we use synthetic apps generated from models (i.e., Player, Social Network, Bank and Market Place), coverage is measured at the granularity of Android activities. In fact, there is no source code implementing the business logic. Text inputs are chosen from a pool of 20 strings, which include credentials necessary to pass through the login activities. To account for non determinism, we executed each algorithm 60 times for each hyperparameter configuration of the algorithms, and applied the Wilcoxon non-parametric statistical test to draw conclusions on the difference between algorithms and configurations, adopting the conventional $p$-value threshold at alpha. Since multiple pairwise comparisons are performed with overlapping data, the chance to reject true null hypotheses may increase (type I error). To control this problem, we adopt the Holm-Bonferroni correction [24], that consists in using more strict significance levels when the number of pairwise tests increases.

Each run has a length of 4000 time steps, which is close to an hour of testing in a real Android test setting. With FATE, 4000 times steps are executed approximately in 200 seconds.

To answer RQ2 we consider the best performing configuration of the Deep RL algorithms, as selected from RQ1, and gradually increase the exploration complexity of the apps. Specifically, 20_strings, 40_strings, 80_strings indicate an increasing size of the string pool (from 20, to 40 and 80), while augmented_5 and augmented_10 indicate an increasing size of the self navigation links (with 5 or 10 “dummy” buttons) within the login activities. For the assessment, we adopt the widely used metric AUC (Area Under the Curve), measuring the area below the activity coverage plot over time. To account for the non determinism of the algorithms, we repeated each experiment 30 times and applied the Wilcoxon non-parametric statistical test.

In RQ3 we investigate qualitatively the cases where Deep RL is superior to tabular Q-Learning. With RQ4, we want to understand if results obtained on synthetic app models run by FATE correlate with those obtained when executing the same apps, once they are translated into real Android apps in Java code that can be executed by ARES. In particular, we translated 3 synthetic apps to Java/Android: Social Network, Bank and Market Place. We then repeated the scenario 40_strings on them using ARES. We compare the rankings of the algorithms being evaluated that are produced respectively by FATE and by ARES.

**Study 2 (ARES):**

• **RQ5** How do code coverage and time-dependent code coverage compare between Random, Q-Learning, DDPG, and SAC?

• **RQ6** What are the fault exposure capabilities of the alternative approaches?

• **RQ7** What features of the real apps make deep RL perform better than Q-Learning?

• **RQ8** How does ARES compare with state-of-the-art tools in terms of coverage and bug detection?

In Study 2, we use real apps and compare the alternative Deep RL algorithms between each other and with Random and tabular Q-Learning. At last, we compare ARES to state of the art testing tools.
To address RQ5-RQ6 and RQ7, we randomly selected 41 apps among the 500 most starred F-Droid apps available on GitHub. We consider coverage at the source code level and compare both the final coverage as well as the coverage increase over time (RQ5). To obtain coverage data we instrumented each app by means of JaCoCo [33]. As in Study 1, we measured AUC and compared AUC values using the Wilcoxon statistical test with significance level set to 0.05 (with correction). We exclude TD3 from the comparison, since it performed consistently worse than the other RL algorithms on synthetic apps.

In addition to code coverage, we also report the number of failures (unique app crashes) triggered by each approach (RQ6). To measure the number of unique crashes observed, we parsed the output of Logcat and (1) removed all crashes that do not contain the package name of the app; (2) extracted the stack trace; (3) computed the hash code of the sanitized stack trace, to uniquely identify it. With RQ7 we analyze qualitatively the different performance of Deep RL vs Q-Learning on real apps.

To address RQ8, we evaluate and compare ARES and state of the art tools in terms of code coverage and number of crashes, using two different sets of apps under test, RQ8-a and RQ8-b, that accommodate the different requirements and constraints of the tools being compared. As state of the art tools, we selected Monkey [20], Sapienz [32], TimeMachine [14] and Q-Testing [36]. In RQ8-a we compare ARES, Monkey, Sapienz and TimeMachine on a set of 68 apps coming from AndroTest [10]. These apps are instrumented using Emma [1], the same coverage tool that is used in Sapienz and TimeMachine. In RQ8-b, we compare ARES to Q-Testing on a set of ten apps instrumented using JaCoCo, the coverage tool supported by Q-Testing.

All experimental data was generated and processed automatically. Each experiment was conducted with a 1 hour timeout and was repeated 10 times, for a total of 4560 hours (∼190 days). The emulators involved in the study are equipped with 2 GB of RAM and Android 10.0 (API Level 29) or Android 4.4 only for the tool comparison.

### 6.1 Experimental Results: Study 1

| App   | Config | Rand | Q-Learn | TD3  | SAC  | DDPG | Effect Size |
|-------|--------|------|---------|------|------|------|-------------|
| Player | 20_str | 88719| 89022   | 89903| 89943| 90337| L(Rand), M(Q) |
|       | 40_str | 23097| 23097   | 23097| 23097| 23097| S(DDPG)    |
|       | 80_str | 3998 | 3998    | 3998 | 3998 | 3998 | M(Q, Rand) |
|       | aug_5  | 3998 | 3998    | 3998 | 3998 | 3998 | M(Q, Rand) |
|       | aug_10 | 3998 | 3998    | 3998 | 3998 | 3998 |             |
| Social| 20_str | 22894| 21622   | 29139| 28016| 36977| M(Q, Rand) |
|       | 40_str | 9746 | 7750    | 9205 | 16535| 6458 | S(Q, DDPG) |
|       | 80_str | 4335 | 5640    | 5730 | 7254 | 4774 |            |
|       | aug_5  | 3998 | 3998    | 3998 | 3998 | 3998 | -           |
|       | aug_10 | 3998 | 3998    | 3998 | 3998 | 3998 | -           |
| Bank  | 20_str | 12815| 8634    | 8702 | 14914| 11472|            |
|       | 40_str | 6289 | 6289    | 6289 | 6289 | 6289 | -           |
|       | 80_str | 10236| 10236   | 10236| 10236| 10236| -           |
|       | aug_5  | 10236| 10236   | 10236| 10236| 10236| -           |
|       | aug_10 | 10236| 10236   | 10236| 10236| 10236| -           |
| Market| 20_str | 18917| 16377   | 16500| 21208| 16027|            |
|       | 40_str | 15944| 15944   | 15944| 15944| 15944| -           |
|       | 80_str | 18917| 18917   | 18917| 18917| 18917| -           |
|       | aug_5  | 10236| 10236   | 10236| 10236| 10236| -           |
|       | aug_10 | 10236| 10236   | 10236| 10236| 10236| -           |

Table 1. AUC for synthetic apps: effect size between winner (shaded cell) and other algorithms is reported only when p-value is statistically significant (S = Small; M = Medium; L = Large)

Figure 3 shows the coverage growth for the synthetic app Social. Each curve shows the mean over 60 runs. The shaded area around the mean represents the Standard Error of the Mean (SEM = \( \sigma / \sqrt{n} \), where \( \sigma \) is the standard deviation and \( n = 60 \) the number of points). The highest activity coverage is obtained consistently by deep RL algorithms, which correspondingly have also higher AUC values. Table 1 reports the AUC obtained on the synthetic apps in all tested configurations.
Table 1 also shows the Vargha-Delaney effect size in the case of a $p$-value $< \alpha$ (with correction), between the winner algorithm (highest AUC) and the remainders.

Results show that deep RL algorithms achieve higher coverage in most experiments. DDPG performs better in the simplest configuration, 20$_{\text{strings}}$, while SAC performs better in almost all other configurations, including the most complex ones. Q-Learning prevails in only two scenarios belonging to Market Place, but the difference from the other algorithms is not statistically significant ($p$-value $> \alpha$).

RQ1: Results lead us to select DDPG and SAC as best performing deep RL algorithms, to be involved in Study 2 experiments.

DDPG is selected due to its high performance in fairly simple scenarios; SAC because of its ability to adapt and maintain good performance in the majority of scenarios.

RQ2: While in simple situations (e.g., the Player app), all algorithms achieve a high level of coverage, when things get more complex (e.g., when string pool increases), Deep RL algorithms retain higher coverage than all other algorithms.

We have manually inspected the step by step exploration performed by Q-Learning and by the Deep RL algorithms. We found that login activities complicate substantially the exploration performed by Q-Learning. In fact, it is more difficult to reproduce the right username-password combination for a tabular Q-Learning algorithm, which has limited adaptation capabilities, whereas Deep RL algorithms memorize the right combination in the DNN used to guide the exploration. In addition, large action spaces make it challenging for Q-Learning to learn an effective exploration strategy, while the DNNs used by deep RL algorithms can easily cope with large spaces of alternatives.
to choose from. This is confirmed by the performance degradation of Q-Learning as the string pool becomes larger or as new interactive elements ("dummy" buttons) are added, which confuse Q-Learning during its exploration.

**RQ3:** The performance of Q-Learning declines in the presence of blocking activities that require specific input combinations that must be learned from past interactions or when the input/action space becomes excessively large, while Deep RL can learn how to cope with such obstacles thanks to the DNN employed to learn the exploration strategy.

Table 2 shows the ranking of the algorithms produced by ARES vs FATE on the 3 apps that were translated from the synthetic FATE models to Java/Android. Below the ranking, Table 2 shows the AUC values obtained by the respective algorithms. The behaviors of the considered algorithms on synthetic (FATE) vs translated (ARES) apps are very similar. The AUC values are quite close and the Spearman’s correlation between AUC values across algorithms is 0.99 for Social, 0.89 for Bank and 0.99 for Market; it is 0.95 overall. All correlations are statistically significant at level 0.05. ARES required 450 hours to complete the experiments. FATE required around 10 hours, reducing the computation time by a factor 45, while producing similar results as ARES.

| App   | Tool | ARES | Ranking / AUC |
|-------|------|------|---------------|
| Social | Q-Learn | 4: 7788 | 5: 6802 |
|       | DDPG  | 3: 9547 | 2: 9594 |
|       | Rand  | 1: 15101 |
|       | SAC   | 1: 15101 |
|       | TD3   | 1: 15101 |
| Bank  | Q-Learn | 4: 8614 | 5: 7976 |
|       | DDPG  | 3: 9344 | 2: 10932 |
|       | Rand  | 1: 16138 | 2: 19932 |
|       | SAC   | 1: 16138 | 2: 19932 |
|       | TD3   | 1: 16138 | 2: 19932 |
| Market | Q-Learn | 1: 16866 | 2: 16788 |
|       | DDPG  | 4: 15936 | 5: 15930 |
|       | Rand  | 3: 15944 |
|       | SAC   | 3: 15944 |
|       | TD3   | 3: 15944 |

Table 2. Ranking of algorithms produced by ARES vs FATE; AUC values below ranked algorithms

**RQ4:** The results obtained on synthetic apps are comparable to those obtained on their translated counterparts.

### 6.2 Experimental Results: Study 2

Table 3 shows coverage and crashes produced by each algorithm deployed in ARES. The highest average coverage and average number of crashes over 10 runs are shaded in gray for each app. We grouped the apps into three different size categories (Low-Medium, Medium, and High), depending on their ELOC (Executable Lines Of Code). Inspecting the three categories, results show that the Deep RL algorithms arise more often as winners when the ELOC increase. Usually, larger size apps are more sophisticated and offer a richer set of user interactions, making their exploration more challenging for automated tools. We already know from Study 1 that when the action space or the observation space of the apps increase, Deep RL can infer the complex actions needed to explore such apps more easily than other algorithms. Study 2 confirms the same trend.

Overall, SAC achieves the best performance, with 42.61% instruction coverage and 0.3 faults detected on average. DDPG comes next, with 40.09% instruction coverage and 0.12 faults detected.
on average. To further investigate these results, we computed the AUCs reached by each algorithm and we applied the Wilcoxon test to each pair of algorithms. Table 4 shows the AUCs achieved by the 4 algorithms and the Vargha-Delaney effect size between the winner and the other algorithms when the \( p \)-value is less than \( \alpha \). SAC results as the winner on 56% of the considered real apps, followed by Random (34%). Moreover, Table 4 confirms the trend observed in Table 3: as ELOC increase, a higher proportion of Deep RL algorithms produces the highest AUC. Figure 4 shows an example of code coverage over time for the app **Loyalty-card-locker**, averaged on 10 runs. SAC increases its coverage almost until the end of the exploration, while the other algorithms reach a plateau after around 35 minutes of exploration.

Table 3. Average coverage and number of crashes observed on 41 real open-source apps in 10 runs of ARES

| Applications          | ELOC   | Rand Q | SAC | DDPG |
|-----------------------|--------|--------|-----|------|
| Silent-ping-smms      | 463    | 41     | 41  | 41   |
| Drawablelenotepad     | 452    | 20     | 21  | 26   | 25   | 0.7  | 0.6  | 0.7  | 0.1  |
| SmsMatrix             | 466    | 23     | 20  | 24   | 22   | 1.2  | 0    | 1.2  | 0.9  |
| Busybox               | 540    | 75     | 73  | 74   | 76   | 0    | 0    | 0    | 0    |
| WiFiKeyShare          | 627    | 37     | 36  | 37   | 37   | 0    | 0    | 0    | 0    |
| Talalarno             | 1094   | 69     | 71  | 71   | 71   | 0.5  | 0.5  | 0    | 0    |
| AquaDroid             | 1157   | 55     | 55  | 55   | 55   | 1.0  | 0.4  | 0.8  | 0.3  |
| Lexica                | 1215   | 72     | 72  | 74   | 75   | 0.3  | 0.1  | 1.5  | 1.2  |
| Loyalty-card-locker   | 1228   | 41     | 37  | 50   | 41   | 0.5  | 0.4  | 0.8  | 0.1  |
| Dns66                 | 1264   | 58     | 58  | 58   | 58   | 0    | 0    | 0.2  | 0.2  |
| Gpatest               | 1311   | 47     | 46  | 47   | 46   | 0    | 0    | 0    | 0    |
| Memento               | 1336   | 77     | 76  | 74   | 77   | 0    | 0    | 0    | 0    |
| Editor                | 1547   | 50     | 46  | 51   | 50   | 0    | 0    | 0    | 0    |
| AndOTP                | 1560   | 20     | 25  | 27   | 20   | 0.5  | 0.7  | 0.2  | 0.2  |
| BookyMcBookface       | 1595   | 26     | 25  | 25   | 24   | 0    | 0    | 0    | 0    |
| Tuner                 | 2207   | 80     | 74  | 79   | 75   | 0    | 0    | 0    | 0    |
| WiFiAnalyzer          | 2511   | 78     | 75  | 80   | 79   | 0    | 0    | 0    | 0    |
| AdAway                | 3064   | 38     | 37  | 45   | 40   | 0    | 0    | 0.1  | 0.1  |
| Qpiplogger            | 3201   | 36     | 31  | 32   | 28   | 0    | 0    | 0    | 0.1  |
| Connectbot            | 3904   | 26     | 25  | 28   | 20   | 0    | 0    | 0.5  | 0.2  |
| Neurolab              | 3954   | 29     | 28  | 29   | 28   | 0    | 0    | 0.4  | 0.3  |
| Anuto                 | 4325   | 46     | 46  | 47   | 47   | 0    | 0    | 0    | 0    |
| PassAndroid           | 4569   | 1      | 1   | 1    | 1    | 0    | 0    | 0    | 0    |
| Markor                | 4607   | 51     | 43  | 53   | 41   | 0.3  | 0.4  | 0    | 0    |
| Vanilla               | 4747   | 29     | 34  | 41   | 33   | 0    | 0    | 0    | 0    |

| Average               | 45     | 43.84  | 46.76| 44.32| 0.15 | 0.12 | 0.28 | 0.15 |
|-----------------------|--------|--------|-----|------|

| Applications          | Rand Q | SAC | DDPG |
|-----------------------|--------|-----|------|
| Afwall                | 5130   | 12  | 12   | 16   | 13   | 0    | 0    | 0    |
| OpenTracks            | 5260   | 45  | 42   | 44   | 45   | 0    | 0    | 0    |
| Opentasks             | 5772   | 43  | 50   | 53   | 44   | 0    | 0    | 0.2  |
| UserLand              | 5901   | 60  | 60   | 60   | 60   | 0.1  | 0.2  | 0.4  |
| Simple-Solitaire      | 5907   | 10  | 30   | 33   | 31   | 0    | 0.4  | 0.4  |
| Authorizer            | 5925   | 5   | 5    | 5    | 5    | 0    | 0    | 0    |
| YalpStore             | 6734   | 35  | 34   | 38   | 33   | 0    | 0    | 0    |
| CameraRoll            | 6836   | 32  | 31   | 31   | 32   | 0.8  | 0.1  | 1.6  |
| AntennaPod            | 7975   | 46  | 40   | 48   | 38   | 0.5  | 0.1  | 0.8  |
| Phonograph            | 8758   | 16  | 16   | 16   | 16   | 0    | 0    | 0    |

| Average               | 30.4   | 30.5  | 35.4 | 31.5 | 0.14 | 0.08 | 0.34 | 0.09 |

| Applications          | MicroMathematics | LightningBrowser | Firefox-focus | RedReader | Wikipedia | Slide | Average |
|-----------------------|------------------|------------------|--------------|-----------|----------|-------|---------|
|                       | 10505            | 11961            | 12482        | 12958     | 23543    | 30483 | 34.33   | 34.50 | 39.5   | 36.5  | 0.21  | 0.1   | 0.38  | 0.1  |

| Average               | 34.33           | 34.50            | 39.5          | 36.5      | 0.21     | 0.1   | 0.38   | 0.1  |

| Applications          | Total Average    | Unique crashes   |
|-----------------------|------------------|------------------|
|                       | 39.62            | 39.58            | 42.61         | 40.09     | 0.17   | 0.1   | 0.3   | 0.12  |

|                       | 75               | 43               | 102            | 52        |

ACM Trans. Softw. Eng. Methodol., Vol. 01, No. 1, Article 000. Publication date: January 2021.
RQ5: SAC reached the highest coverage in 24/41 apps, followed by DDPG (11 apps), Random (10 apps), and Q-Learning (1 app). SAC has also the highest AUC in 24/41 apps, followed by Random (13 apps), Q-Learning (11 apps) and DDPG (5 apps).

![Fig. 4. Instruction coverage over time for the app Loyalty-card-locker](image)

Table 3 shows that SAC exposed the highest number of unique crashes (102), followed by Random (73), DDPG (52) and Q-Learning (43). The most common types of error exposed by SAC during testing are: RuntimeException (34 occurrences), NullPointerException (14), IllegalArgumentException (13). As shown in Figure 5, there is around thirty percent overlap between the crashes found by SAC and the other algorithms. The overlap with Random is the highest. SAC discovers about 40% of unique crashes found by Random; however, SAC found many new crashes that Random did not find.

RQ6: The SAC algorithm implemented in ARES generates the highest number of crashes, 102, in line with the results on coverage (RQ5), where SAC was also the best performing algorithm.

We have manually inspected the coverage progress of the different algorithms on some of the real apps considered in Study 2. We observed that deep RL algorithms achieve higher coverage than the other algorithms when it is necessary to replicate complex behaviors in order to: (1) overcome blocking activities, e.g., to create an item in order to be able to later access its properties, or to bypass authentication; (2) to pass through concatenated activities without being distracted by already seen activities or ineffective buttons (high dimensional action/observation space); (3) reach an activity located deeply in the app. Such behaviors are possible thanks to the learning capabilities...
| App            | Rand | Q | SAC | DDPG | Effect Size |
|---------------|------|---|-----|------|-------------|
| Silent-ping-sms | 0.36 | 0.36 | 0.38 | 0.37 | S(DDPG), M(Rand), L(Q) |
| Drawerable-notepad | 0.13 | 0.14 | 0.20 | 0.18 | L(Rand, Q) |
| SmsMatrix     | 0.13 | 0.13 | 0.11 | 0.11 | - |
| Busybox       | 0.54 | 0.56 | 0.68 | 0.68 | L(Q, Rand) |
| WiFiKeyShare  | 0.29 | 0.29 | 0.33 | 0.30 | L(DDPG, Q, Rand) |
| Talalarmo     | 0.62 | 0.60 | 0.64 | 0.64 | L(Q) |
| AquaDroid     | 0.529 | 0.526 | 0.531 | 0.522 | L(Rand, Q, DDPG) |
| Lexica        | 0.63 | 0.61 | 0.66 | 0.65 | L(Q, Rand) |
| Loyalty-card-locker | 0.23 | 0.23 | 0.34 | 0.21 | L(DDPG, Q, Rand) |
| Dns66         | 0.51 | 0.51 | 0.47 | 0.45 | L(DDPG, SAC) |
| Gpptest       | 0.40 | 0.40 | 0.39 | 0.36 | L(DDPG) |
| Memento       | 0.64 | 0.65 | 0.64 | 0.65 | - |
| Editor        | 0.42 | 0.37 | 0.37 | 0.37 | L(DDPG, Q, SAC) |
| AndOTP        | 0.16 | 0.18 | 0.23 | 0.15 | M(Rand), L(DDPG) |
| BookyMcBookface | 0.22 | 0.20 | 0.20 | 0.20 | M(DDPG), L(SAC) |
| Tuner         | 0.68 | 0.66 | 0.60 | 0.60 | L(DDPG, SAC) |
| WiFAalyzer    | 0.56 | 0.56 | 0.67 | 0.58 | L(DDPG, Q, Rand) |
| AdAway        | 0.25 | 0.25 | 0.27 | 0.25 | - |
| Gpslogger     | 0.24 | 0.25 | 0.23 | 0.20 | L(DDPG, SAC) |
| Connectbot    | 0.19 | 0.19 | 0.22 | 0.09 | L(DDPG, Q, Rand) |
| Neurolab      | 0.23 | 0.23 | 0.22 | 0.22 | - |
| Anuto         | 0.35 | 0.40 | 0.43 | 0.33 | L(DDPG, Q, Rand) |
| PassAndroid   | 0.018 | 0.017 | 0.017 | 0.017 | - |
| Markor        | 0.40 | 0.40 | 0.43 | 0.25 | L(DDPG, Q, Rand) |
| Vanilla       | 0.17 | 0.23 | 0.26 | 0.23 | L(Rand) |
| Ahsel         | 0.09 | 0.09 | 0.13 | 0.10 | L(DDPG, Q, Rand) |
| OpenTracks    | 0.37 | 0.35 | 0.35 | 0.35 | - |
| Opentasks     | 0.18 | 0.46 | 0.46 | 0.29 | L(DDPG, Rand) |
| UserAnd       | 0.49 | 0.49 | 0.49 | 0.47 | - |
| Simple-Solitaire | 0.06 | 0.21 | 0.19 | 0.18 | L(Rand) |
| Authorizer    | 0.05 | 0.046 | 0.047 | 0.049 | S(SAC), M(Q) |
| YalpStore     | 0.28 | 0.28 | 0.31 | 0.26 | L(DDPG, Q, Rand) |
| Camera-Roll   | 0.26 | 0.37 | 0.25 | 0.25 | L(Rand, DDPG, SAC) |
| AntennaPod    | 0.33 | 0.33 | 0.35 | 0.22 | L(DDPG) |
| Phonograph    | 0.085 | 0.077 | 0.075 | 0.076 | S(Q, DDPG, SAC) |
| MicroMathematics | 0.17 | 0.17 | 0.30 | 0.18 | L(DDPG, Q, Rand) |
| Lightning-Browser | 0.28 | 0.28 | 0.36 | 0.29 | L(DDPG, Q, Rand) |
| Firefox-focus | 0.27 | 0.28 | 0.43 | 0.35 | L(Rand, Q) |
| RedReader     | 0.31 | 0.31 | 0.31 | 0.33 | - |
| Wikipedia     | 0.30 | 0.30 | 0.33 | 0.28 | - |
| Slide         | 0.13 | 0.13 | 0.12 | 0.14 | - |

Table 4. AUCs achieved on real apps; effect size between winner and others when \( p \)-value < \( \alpha \)

of the DNNs used by deep RL algorithms, while they are hardly achieved by the other existing approaches, including tabular Q-Learning.

**RQ7:** In presence of blocking activities or complex concatenated activities (activities with a high number of widgets or located in depth) that require the capability to reuse knowledge acquired in previous explorations, the learning capabilities of deep RL algorithms make them the most effective and efficient exploration strategies.

### 6.2.1 Comparison between ARES and state-of-the-art tools. RQ8-a

Table 5 shows the coverage reached and the faults exposed by each testing tool on 68 Android apps from AndroTest. Coverage data are summarized by means of boxplots in Figure 6. The highest average coverage and average number of crashes over 10 runs are highlighted with a gray background. ARES achieved 54.2% coverage and detected 0.48 crashes on average. TimeMachine achieved 50.4% code coverage and 0.42 faults on average. Sapienz reached a mean code coverage of 48.8%, and discovered 0.22 faults.
Fig. 5. Comparison of total number of unique crashes on the 41 apps involved in RQ5-6-: dark gray areas indicate the proportion of crashes found by both techniques on average. Monkey achieved 43.9% code coverage and discovered 0.11 faults. ARES achieved the highest code coverage on average, followed by TimeMachine, Sapienz and Monkey. TimeMachine detected most unique crashes (179), followed by ARES (171), Sapienz (103) and Monkey (51). Actually, ARES discovered less crashes than TimeMachine mostly because TimeMachine uses a system-level event generator, taken from Stoat [39], which ARES does not support. However, system events present two major drawbacks: a) they vastly change depending on the Android version [40] [19] (despite TimeMachine is compatible with Android 7.1, it uses only system-level actions compatible with Android 4.4 [13]); and b) in order to execute them, root privileges are required. ARES does not require root privileges in order to work properly on any app (i.e., we recall that protected apps do not execute on rooted devices [42]). Analyzing the execution traces of each tool, we searched and identified the faults immediately after the generation of system events not related to the AUT. More than a third (63) of the crashes generated by TimeMachine comes from system-level actions. Figure 7 shows a pairwise comparison of detected crashes among evaluated techniques. Only 20% of unique crashes found by ARES are also found by TimeMachine. This shows that ARES can be used together with other state-of-the-art Android testing techniques (in particular, TimeMachine) to jointly cover more code and discover more crashes.

The good results of ARES in code coverage and exposed faults are due to the reinforcement mechanisms of the RL algorithms and to the reward function that drives the testing tool through states of the app leading to hard-to-reach states. Search-based techniques such as Sapienz typically observe the program behavior over an event sequence that may be very long. Hence, the associated coverage feedback, used to drive search-based exploration, does not have enough fine-grained details to support good local choices of the actions to be taken. In fact, the fitness function used by these algorithms evaluates an entire action sequence and does not consider the individual steps in the sequence. TimeMachine, improved this weakness, by relying on the coverage feedback obtained at an individual state to determine which portion of the app is not still explored. The drawback of this kind of approach is a higher computational cost, that requires a higher testing time. ARES offers a better trade-off between the granularity of the feedback and the computational cost required to obtain it.
RQ8-a: ARES achieved the highest code coverage in 41/68 apps, followed by TimeMachine (12/68), Sapienz (6/68) and Monkey (2/68). ARES triggered crashes more often than other tools in 23/68 apps, followed by TimeMachine (15/68), Sapienz (5/68) and Monkey (0/68). However, TimeMachine generated the highest number of unique crashes (179), but 63 of them come from system-level events. ARES generates 171 faults, Sapienz 103, and Monkey 51.

Fig. 6. Code coverage achieved by ARES, TimeMachine, Sapienz, and Monkey

Fig. 7. Comparison of total number of unique crashes involved in RQ8-a: dark gray areas indicate the proportion of crashes found by both testing tools
| App               | Monkey | Sapienz | TM | ARES |
|-------------------|--------|---------|----|------|
| a2dp              | 38     | 44      | 42 | 42   |
| aagtl             | 16     | 18      | 17 | 19   |
| aardvst           | 11     | 15      | 17 | 17   |
| acal              | 18     | 27      | 27 | 28   |
| addi              | 19     | 20      | 17 | 20   |
| astdroid          | 30     | 38      | 36 | 35   |
| a4gpr             | 45     | -       | 59 | 56   |
| aka               | 65     | 84      | 77 | 84   |
| alarmclock        | 64     | 41      | 60 | 71   |
| aLogCat           | 67     | 71      | 75 | 87   |
| Amazed            | 36     | 66      | 63 | 89   |
| AnyCut            | 63     | 65      | 63 | 73   |
| anytime           | 31     | 50      | 43 | 55   |
| autoanswer        | 12     | 16      | 21 | 15   |
| batterydog        | 63     | 67      | 62 | 69   |
| battery           | 73     | 78      | 77 | 92   |
| bites             | 34     | 41      | 45 | 43   |
| blockish          | 55     | 52      | 68 | 45   |
| bomber            | 76     | 73      | 77 | 84   |
| Book-Catalogue    | 27     | 29      | 27 | 25   |
| CountdownTimer    | 74     | 62      | 77 | 84   |
| dalvik-explorer   | 66     | 72      | 70 | 72   |
| dialer2           | 39     | 42      | 42 | 44   |
| DivideAndConquer  | 48     | 83      | 82 | 80   |
| dflexplorer       | 41     | 49      | 55 | 64   |
| frozenbubble      | 80     | 76      | 75 | 70   |
| gestures          | 37     | 52      | 51 | 55   |
| handroid          | 7      | 15      | 18 | 18   |
| hotdeath          | 75     | 75      | 72 | 74   |
| importcontacts    | 40     | 39      | 40 | 42   |
| junemido          | 53     | 41      | 54 | 63   |
| k9mail            | 6      | 7       | 8  | 8    |
| LNM               | 47     | -       | -  | 75   |
| lockpatterngenerator | 75   | 79      | 74 | 78   |
| LolcatBuilder     | 26     | 25      | 29 | 26   |
| manpages          | 40     | 73      | 70 | 74   |
| mileage           | 38     | 45      | 48 | 45   |
| Mirrored          | 57     | 59      | 62 | 59   |
| mrv               | 41     | 60      | 43 | 56   |
| multismsender     | 34     | 59      | 61 | 73   |
| MunchLife         | 67     | 72      | 71 | 88   |
| MyExpenses        | 41     | 60      | 50 | 63   |
| myLock            | 25     | 31      | 50 | 30   |
| Nectroid          | 34     | 66      | 58 | 57   |
| netcounter        | 43     | 70      | 58 | 69   |
| PasswordMaker     | 53     | 58      | 55 | 59   |
| passwordmanager   | 7      | 8       | 17 | 18   |
| Photostream       | 30     | 34      | 35 | 29   |
| QuickSettings     | 50     | 45      | 46 | 52   |
| RandomMusicPlayer | 53     | 58      | 58 | 63   |
| Ringdroid         | 22     | 29      | 48 | 30   |
| sanity            | 26     | 19      | 31 | 22   |
| soundboard        | 42     | 32      | 59 | 61   |
| SpriteMethodTest  | 58     | 80      | 73 | 88   |
| SpriteText        | 60     | 68      | 57 | 60   |
| swiftp            | 12     | 14      | 13 | 17   |
| SyncMyPix         | 21     | 21      | 23 | 25   |
| tippy             | 75     | 83      | 74 | 85   |
| tomidroid         | 47     | 46      | 51 | 69   |
| Translate         | 49     | 48      | 48 | 59   |
| Triangle          | -      | -       | -  | -    |
| weight-chart      | 63     | 67      | 66 | 71   |
| whohasmystuff     | 61     | 68      | 66 | 81   |
| wikipedia         | 31     | 32      | 33 | 35   |
| Wordpress         | 4      | 5       | 7  | 8    |
| worldclock        | 83     | 88      | 86 | 90   |
| youtube           | 51     | 57      | 56 | 69   |
| znoborns          | 30     | 16      | 35 | 37   |

| Average | 43.9 | 48.8 | 50.4 | 54.2 |
| Sum     | 51   | 103  | 179  | 171  |

Table 5. Results on 68 open-source apps coming from AndroTest
Table 6. Comparison between Q-Testing and ARES

| App          | Coverage ARES | Faults ARES | Coverage Q-Testing | Faults Q-Testing |
|--------------|--------------|-------------|-------------------|-----------------|
| Alogcat      | 84           | 0           | 76                | 0               |
| Antennapod   | 48           | 0.8         | 42                | 0.4             |
| AnyCut       | 72           | 0           | 67                | 0               |
| batterydog   | 65           | 0.5         | 49                | 0.3             |
| Jamendo      | 64           | 0.4         | 46                | 0.7             |
| Multissmsender | 71          | 0.4       | 45                | 0               |
| Myexpanses   | 63           | 0.4         | 36                | 0.2             |
| talalarmo    | 74           | 0.8         | 74                | 0.5             |
| Tomdroid     | 61           | 0.3         | 50                | 0.2             |
| vanilla      | 41           | 0.5         | 40                | 0.4             |
| **Average**  | **64.3**     | **0.41**    | **52.3**          | **0.27**        |
| **Sum**      |              |             |                   |                 |

RQ8-b: Table 6 shows the coverage reached and the faults exposed by ARES and Q-Testing on 10 Android apps instrumented with JaCoCo. ARES achieved 64.3% coverage and exposed 0.41 faults on average; it detected 17 unique crashes. Q-Testing achieved 52.5% code coverage and 0.27 crashes on average, and it detected 16 unique crashes. ARES achieved the highest code coverage on almost all apps, and on average ARES covered 12% more code than Q-Testing. Q-testing generated 6 faults in common with ARES, while other 4 faults are generated using the system-level events of Stoat. In this comparison, the main advantage of ARES seems to be a better reward function, that encourages the tool to visit the greatest number of activities within the same episode. Instead, Q-Testing determines the reward of the Q-Learning algorithm by computing the similarity between Android app states, which does not guarantee an efficient way to overcome blocking activities.

Fig. 8. Code coverage achieved by ARES and Q-Testing.

RQ8-b: ARES reached the highest code coverage on 9/10 apps, the highest average of crashes in 7/10 apps, and the highest number of unique crashes. Q-Testing, the state-of-the art RL Android testing tool, reached the same level of exploration of ARES only in 1/10 app, and the highest average number of crashes on only 1/10 apps.
7 RELATED WORK
7.1 Random Testing
Testing tools in this category generate random events on the AUT GUI. Monkey [20] is one of the most popular black-box Android testing tools. It triggers events by interacting randomly with screen coordinates. This simple random approach worked relatively well on some benchmark applications [11]. Nonetheless, Monkey tests involve a significant number of ineffective or repeated events, as there is no guidance to make the exploration efficient. ARES implements a smarter version of Monkey, which we used as a baseline (Random). Such improved version selects only actions that are actually possible in a given GUI state, thus making the exploration strategy a bit more efficient.

7.2 Model-based Testing
Model-based tools [4] [5] first build navigation models of the Android app by means of static or dynamic analysis, used to explore efficiently the application states, and then they extract test cases from such models, to eventually expose bugs. Stoat [39] uses a stochastic FSM to model the app behavior. The app model is built using dynamic analysis, enhanced with a weighted UI exploration strategy and with the help of static analysis. Compared to Stoat, ARES can be viewed as computing a navigation model implicitly: the MDP model used by the deep RL algorithms. One key advantage of using an implicit model is that we do not have to deal with the combinatorial explosion of its size, which is controlled by Stoat by means of model compaction heuristics.

7.3 Structural Testing
Structural strategies [6, 14, 16, 31, 32] generate coverage oriented inputs using symbolic execution or evolutionary algorithms. Sapienz [32] maximizes code coverage and bug revelation using a Pareto-optimal multi-objective search based approach, which applies genetic operators such as mutation and crossover to produce new test cases. It can generate specific input for text fields by reverse engineering the APK. This process occasionally results in invalid sequences, which are discarded by the fitness functions that reward test cases with high coverage. TimeMachine [14] improves Sapienz by identifying interesting states in the past and restarting the search process from them when the search stagnates. While coverage oriented approaches require the possibility to instrument the app to measure coverage and to possibly execute it symbolically, our approach is black-box and does not suffer from the scalability issues affecting, e.g., symbolic execution.

7.4 Machine Learning Based Testing
Some ML based testing approaches [7] [25] [28] use an explicit, supervised training process to learn from previous test executions. They can reuse previous knowledge, acquired on different apps or on past versions of the app under test. Approaches as QBE [25] make the transfer of knowledge to new apps possible by abstracting the app state in a form that is supposed to hold across different domains and implementations. However, the effectiveness of such transfer learning process depends on the similarity between new and old apps.

One of the first works proposing RL for GUI testing is AutoBlackTest [34]. This approach is based on the simplest form of RL, tabular Q-Learning, whose effectiveness is strongly dependent on the initial values in the Q-Table. On the contrary, ARES learns the action-value function from scratch during the exploration of the AUT. One of the most recent approaches to Android testing based on deep learning is Q-Testing [36]. However, it also uses tabular Q-Learning as backbone, while learning is limited to the computation of the similarity between Android app states, which
determines the reward of the Q-Learning algorithm. ARES instead learns both the state similarity and the action-value function during its interactions with the AUT.

8 CONCLUSION AND FUTURE WORK

We have proposed an approach based on Deep RL for the automated exploration of Android apps. The best exploration strategy is learned automatically as the test progresses. The approach is implemented in the open source tool ARES, which is complemented by FATE, a model based Android testing tool that we developed to support fast execution and configuration of the alternative deep RL algorithms of ARES. The resulting configuration of ARES, in particular when running the SAC deep RL algorithm, outperformed all the considered baselines in terms of coverage achieved over time and exposed bugs.

In our future work, we plan to investigate specific fault categories that are particularly relevant for Android apps, such as security vulnerabilities. In fact, we think that the adaptation and reward mechanisms used by deep RL algorithms to learn the optimal exploration strategy could be particularly effective when the fault to be exposed is a security fault. We also plan to port ARES and FATE to iOS.

REFERENCES

[1] 2006. Emma. Retrieved December 30, 2020 from http://emma.sourceforge.net
[2] 2020. Appbrain. Retrieved November 20, 2020 from https://www.appbrain.com
[3] 2020. Appium. Retrieved September 25, 2020 from http://appium.io
[4] Domenico Amalfitano, Anna Rita Fasolino, Porfirio Tramontana, Salvatore De Carmine, and Atif M Memon. 2012. Using GUI ripping for automated testing of Android applications. In 2012 Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering. IEEE, 258–261.
[5] Domenico Amalfitano, Anna Rita Fasolino, Porfirio Tramontana, Bryan Dzung Ta, and Atif M Memon. 2014. MobiGUITAR: Automated model-based testing of mobile apps. IEEE software 32, 5 (2014), 53–59.
[6] Saswat Anand, Mayur Naik, Mary Jean Harrold, and Hongseok Yang. 2012. Automated concolic testing of smartphone apps. In Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering. 1–11.
[7] Nataniel P Borges, Maria Gómez, and Andreas Zeller. 2018. Guiding app testing with mined interaction models. In 2018 IEEE/ACM 5th International Conference on Mobile Software Engineering and Systems (MOBILESof). IEEE, 133–143.
[8] Justin A Boyan and Andrew W Moore. 1995. Generalization in reinforcement learning: Safely approximating the value function. In Advances in neural information processing systems. 369–376.
[9] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. arXiv preprint arXiv:1606.01540 (2016).
[10] Shauvik Roy Choudhary, Alessandra Gorla, and Alessandro Orso. 2015. Automated test input generation for android: Are we there yet?(e). In 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 429–440.
[11] Shauvik Roy Choudhary, Alessandra Gorla, and Alessandro Orso. 2015. Automated test input generation for android: Are we there yet?(e). In 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 429–440.
[12] Jack Clark and Dario Amodei. 2016. Faulty Reward Functions in the Wild. Retrieved January 4, 2021 from https://openai.com/blog/faulty-reward-functions/
[13] Zhen Dong, Marcel Bohme, Lucia Cojocaru, and Abhik Roychoudhury. 2020. Github Repository: TimeMachine. Retrieved November 20, 2020 from https://github.com/DroidTest/TimeMachine/blob/master/fuzzingandroid/sys_event_generator/sys_event.py
[14] Zhen Dong, Marcel Bohme, Lucia Cojocaru, and Abhik Roychoudhury. 2020. Time-travel Testing of Android Apps. In Proceedings of the 42nd International Conference on Software Engineering (ICSE). 481–492.
[15] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. arXiv preprint arXiv:1802.09477 (2018).
[16] Xiang Gao, Shin Hwei Tan, Zhen Dong, and Abhik Roychoudhury. 2018. Android testing via symbolic synthetic execution. In 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 419–429.
[17] Google. 2019. Broadcasts. Retrieved October 20, 2020 from https://developer.android.com/guide/components/broadcasts
9 APPENDIX

9.1 Study 1-Deep RL

Certain parameters of the algorithms are omitted for simplicity, and can be found in the documentation of Stable Baselines [23].

**Control Policy:**
Deep RL algorithms rely on a policy based on a MLP composed of 2 layers and 64 neurons each.

**Learning rates:**
- DDPG: 0.0001
- SAC: 0.0003
- TD3: 0.0003

| DDPG Configuration | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|--------------------|---|---|---|---|---|---|---|---|
| random_exploration | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.7 | 0.8 | 0.8 |
| nb_train_steps     | 5  | 25 | 5  | 25 | 5  | 25 | 5  | 25 |

| TD3 Configuration  | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|--------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| random_exploration | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.7 | 0.8 | 0.8 |
| train_frequency    | 25  | 100 | 25  | 100 | 25  | 100 | 25  | 100 |

| SAC Configuration  | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|--------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| target_update_interval | 1 | 1 | 2 | 2 | 5 | 5 | 10 | 10 |
| train_frequency    | 1   | 5   | 1   | 5 | 1 | 5 | 1 | 5 |

9.2 Study 1-Tabular RL

| Q-Learning Configuration | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|--------------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| epsilon                  | 0.5 | 0.6 | 0.7 | 0.8 | 0.5 | 0.6 | 0.7 | 0.8 |
| gamma                    | 0.99| 0.99| 0.99| 0.99| 0.9 | 0.9 | 0.9 | 0.9 |
| alpha                    | 0.628| 0.628| 0.628| 0.628| 0.628| 0.628| 0.628| 0.628 |

9.3 Study 2

Some parameters used in Study 2 have been selected from the results of Study 1 (see Section 6).

**DDPG:**
- Control Policy: MLP, 2 layers, 64 neurons
- Learning rate: 0.0001
- nb_train_steps: 10
- random_exploration: 0.7

**SAC:**
- Control Policy: MLP, 2 layers, 64 neurons
- Learning rate: 0.0003
- train_freq: 5
- target_update_interval: 10

**TD3:**
- Control Policy: MLP, 2 layers, 64 neurons
- Learning rate: 0.0003
- train_freq: 10
- random_exploration: 0.8

**Q-Learning:**
- epsilon: 0.8
- gamma: 0.9
- alpha: 0.628