Efficient Summary Reuse for Software Regression Verification

Fei He, Qianshan Yu, and Liming Cai

Abstract—Software systems evolve throughout their life cycles. Many revisions are produced over time. Verifying each revision of the software is impractical. Regression verification suggests reusing intermediate results from the previous verification runs. This paper studies regression verification via summary reuse. Not only procedure summaries, but also loop summaries are proposed to be reused. This paper proposes a fully automatic regression verification technique in the context of CEGAR. A lazy counterexample analysis technique is developed to improve the efficiency of summary reuse. We performed extensive experiments on two large sets of industrial programs (3,675 revisions of 488 Linux kernel device drivers). Results show that our summary reuse technique saves 84 to 93 percent analysis time of the regression verification.

Index Terms—Regression verification, program verification, abstraction refinement, summary reuse

1 INTRODUCTION

Along with the widespread use of software in our daily life, there is a growing concern for software reliability. At the same time, market pressure demands quick product introductions. The software companies are required to introduce new features to their software products in shorter release cycles. Since errors may be introduced with new features, the new products must be reverified to ensure their correctness.

Software verification [1] has made great success in recent years. However, it is still very time-consuming. Verifying every revision of the software is impractical. Inspired by the success of regression testing [2], [3], researchers in formal verification community proposed the technique of regression verification [4], [5], [6], [7], [8]. Taking into consideration that many intermediate results are produced during the verification, and the computation of these results is costly, regression verification aims to make use of these intermediate results in the verification of new program revisions.

Different intermediate results have been proposed for reuse, including abstract precisions, state-space graphs, constraint solver solutions, and interpolation-based procedure summaries. Beyer et al. [7] proposed to record the final abstract precision in the previous verification run, and reuse it in the current verification. Henzinger et al. [9] proposed to reuse the state-space graph for incremental checking of temporal safety properties. Visser et al. [10] noticed the important role of constraint solving in software verification, and proposed to reuse the constraints solving results.

Procedure summaries, representing input/output behaviors of procedures, have been proposed in [11] to be reused in incremental upgrade checking. Note that procedure summaries are reasonably small to store, technically easy to process, and do not require much extra computation effort to be reused. Therefore, reusing procedure summaries is a good choice for regression verification.

Inspired by [11], this paper studies the summary-based regression verification for predicate analysis. In [11], the procedure summaries are mainly constructed by interpolations. In this paper, we consider the summaries constructed using abstract states of predicate analysis. Note that these abstract states are by-products of program analysis [12], [13], [14]. Thus, it does not require additional computational effort to generate these summaries. Moreover, different from existing techniques, our approach considers the reuse of not only procedure summaries, but also loop summaries. We build a unified framework for reusing both of them.

Moreover, we consider regression verification in the context of counter-example guided abstraction refinement (CEGAR) [15]. Summary reuse techniques need to be adapted to the CEGAR framework (see Section 5). A lazy counterexample analysis technique is further proposed to address the effectiveness issue of summary reuse (see Section 5.3). Considering that CEGAR is a widely-adopted technique in software verification [16], [17], [18], [19], [20], our approach can be applied to most state-of-the-art software verifiers. To the best of our knowledge, our approach represents a novel attempt to the regression verification with CEGAR.

We implemented our approach on top of CPAchecker [17]. We have performed extensive experiments on two large sets of industrial programs. The first set of programs contains 1,119 real-world program revisions of 62 Linux device drivers, and the second contains 2,556 artificial program revisions (by mutation) of 426 Linux device drivers. In total, there are 6,749 verification tasks, among which 6,064 are regression verification tasks. Experimental results show a very promising performance of our approach. With the set
of real-world programs, in comparison to the standalone verification without reuse, our approach solves 216 more regression verification tasks and saves 93.1 percent of analysis time. With the set of artificial programs, our approach solves 10 more regression verification tasks and saves 84.2 percent of analysis time.

The main technical contributions of this paper are summarized as follows:

- We propose a unified framework for reusing both procedure summaries and loop summaries.
- We propose a fully automatic regression verification model, respectively.
- We implement our approach in the software verification tool CPAchecker. Experimental results show the promising performance of our approach.

The remainder of this paper is organized as follows. Section 2 introduces the necessary backgrounds. Section 3 motivates our approach using a simple example. Section 4 reviews the CEGAR-based program verification and the definition of program summaries. Section 5 presents our CEGAR-based regression verification framework. Section 6 reports evaluation results on our approach. Section 7 discusses related work and Section 8 concludes this paper.

2 Backgrounds

2.1 Abstraction and Refinement

Abstraction plays a central role in software verification. Abstraction omits details of the system behaviors, resulting in a simpler model. We call the model before and after abstraction the concrete and the abstract model, respectively. An abstraction is conservative [21] iff it does not omit any behavior of the concrete model. Conservative abstraction guarantees that the properties (more precisely, the $\text{ACTL}^*$ properties [21]) established on the abstract system also hold on the concrete system. The reverse, however, is not guaranteed: if the abstract model falsifies the property, the concrete model does not necessarily falsify this property.

The abstract precision [7] (for short, precision) defines the level of abstraction of an abstract model. The precision must be at a proper level. A too-coarse precision may fail to verify the property; a too-fine precision, however, may lead to state space explosion. Finding a proper precision appears to require ingenuity.

Counterexample-guided abstraction refinement [15] provides a framework for automatically finding proper precisions. Starting from an initial abstract precision, it iteratively checks if the corresponding abstract model satisfies the desired property. If the property is satisfied, it must also hold on the concrete model, the algorithm terminates and reports “correct”. Otherwise, the checker returns a path on the abstract model that falsifies the desired property. The algorithm then checks if the returned path is valid on the concrete model or not. If it is, the algorithm finds a real bug, it thus terminates and reports “incorrect”. Otherwise, the precision is too coarse, and needs to be refined with the counterexample. Then the above process repeats, until either “correct” or “incorrect” is reported.

The abstract precision does not necessarily keep the same throughout the program [22]. To simplify the discussion, we assume in this paper that the abstract precisions are defined at the level of procedures, i.e., each procedure is associated with a unique abstract precision.

2.2 Software Verification

Model checking and program analysis are two major approaches for software verification. Comparing these two techniques, model checking is more precise with fewer false positives produced, while program analysis is comparatively more efficient and can be applied to more programs. An increasing tendency to software verification is to integrate these two techniques together [23], to get a good balance between accuracy and efficiency.

Predicate abstraction [22], [24] is a widely-adapted abstraction technique [1], [19] for software verification. It creates an abstract model with respect to a set of predicates defined on the program variables. This predicate set defines an abstract precision for predicate abstraction. The state space of the abstract model is only related to the number of predicates in the abstract precision. Finding proper predicates is the key problem for predicate abstraction. One popular technique is based on interpolation computation on the counterexamples [25], [26].

Interprocedural analysis deals with programs with multiple procedures. One simple way of interprocedural analysis is to inline a copy of the callee procedure at each of its call sites. The inlining technique is, however, very expensive and may lead to context explosion for recursive procedures. Another interprocedural analysis technique is to use summaries [12], [13]. A procedure summary (or shortly, a summary) describes the input/output behaviors of a procedure. This technique plugs summaries at each call site of the procedure. Re-analysis of the procedure body at each of its call sites can be avoided using this technique, the efficiency is therefore improved.

There are at least two kinds of procedure summaries in literature: the state-based summaries [12], [13], where each summary is a pair of input and output states of the procedure; and the interpolation-based summaries [11], where the summary is an overapproximation of the procedure’s behaviors.

In this paper, we assume a deterministic, single-threaded program and a safety property. To specify the property, a special error location is introduced in the program. We say the program is correct if and only if the error location is not reachable.

3 A Motivating Example

Fig. 1 shows a simple program that consists of two procedures: main and inc. A while loop is implemented in the main procedure, and in the loop body the inc procedure is invoked. The inc procedure takes two input parameters: $a$ and $sign$, and outputs either $a+1$ if $sign = 0$, or $a-1$ if $sign = 0$. We want to verify that the error location (at line 6) is not reachable in any execution of this program.

Consider an invocation to the inc procedure (at line 3) with parameters $a = 0$ and $sign = 1$, the returned value is $rv = 1$. The pair of this entry state (i.e., $a = 0 \land sign = 1$)
main()
{
    int i = 0, x = 0;
    while (i < 10) {
        if (x <= 5)
            x = inc(x, 1);
        else
            x = inc(x, 0);
        i = inc(i, 1);
    }
    if (x < 5) goto error;
}
inc(int a, int sign)
{
    if (sign) return a + 1;
    else return a - 1;
}

Fig. 1. An example program.

and the exit state (i.e., rv = 1) summarizes this execution of
the inc procedure. Later, when the inc procedure is
invoked again, if its entry state is again a = 0 ∧ sign = 1,
then without entering the inc procedure, we can immediately
determine its exit state as rv = 1.

The execution of a loop can also be summarized by a pair
of an entry state and an exit state. Consider the while loop
in the main procedure, its entry state (i.e., the state exactly
before the program enters the loop at line 1) is i = 0 ∧ x = 0,
and its exit state (i.e., the state when the program exits the
loop at line 6) is x = 6 ∧ i = 10. Similar to the procedure
summary, the pair of these two states also summarizes an
execution of this loop, and is called a loop summary.

Assume that the original program evolves to a new revi-
sion. Apparently, this new revision needs also to be checked
to guarantee its correctness. Assume that in the new revi-
sion, the inc procedure does not change, then the summa-
tories of this procedure, which were generated in the
previous round of verification, can be reused in the new
round of verification. Similarly, if the while loop does not
change in the new revision, the previous-generated summa-
tories for this loop can also be reused. How to efficiently reuse
the previously-generated summaries in regression verification is
the main research problem we want to solve in this paper.

Moreover, in the above discussions, the program is ana-
yzed by tracking its concrete states. The concrete state
space of a program is, however, considerably huge and
often infinite. A practical verification technique (including
the regression verification) needs to be performed on the
abstract state space. How to efficiently combine regression verifica-
tion and abstraction techniques, especially the counterexample-
guided abstraction refinement, is another research target of
this paper.

4 CEGAR-BASED VERIFICATION

In this section, we first review the CEGAR-based program verifica-
tion, upon which our regression verification scheme is based. We then propose a unified definition for procedure
summaries and loop summaries.

4.1 Preliminaries

We begin by introducing the necessary preliminaries for
program verification.

Control-Flow Automata (CFA) [22], [23] were adopted in
many software verification techniques (for example, BLAST
and CPAchecker) for representing programs. Given a pro-
gram P, let L be the set of program locations and St be the
set of statements of P, respectively. The CFA of P is a pair
(L, G), where L is the set of program locations, and G ⊆ L ×
St ∧ L is the set of control flow edges. The CFA is different
from the control flow graph with program statements label-
ing the edges rather than the vertices. For example, CFA of the
main procedure in Fig. 1 is shown in Fig. 2, where l_{err}
represents the error location, and l_0, l_{exit} represent the entry
and exit locations of the main procedure, respectively.

A state of a program is a configuration of the program
location and the set of facts that we know about the program
at that location. Formally, a concrete state of the program P
is a pair (l, u), where l ∈ L is a program location and u is a full
assignment to all variables of P. The assignment u is also
called the concrete data state of P. Let λ be a set of predicates,
representing the current abstract precision. An abstract state
is a pair (l, s), where l is a program location, and s is a valu-
ation to all predicates in λ. The valuation s is also called the
abstract data state of P. In the remainder of the paper, we
denote P^k the abstract model of P with respect to λ.

Consider the CFA of the main procedure in Fig. 2, with
the abstract precision λ_{main} = \{i < 10, x ≤ 5, x < 5\}. An
abstract data state is a valuation to the three predicates in
λ_{main}. During the procedure of the analysis, the value of a
predicate may be true (abbreviated by 0), false (abbreviated
by 1) or non-deterministic (abbreviated by *). We use a vec-
tor to denote an abstract data state. For example, the abstract
data state at l_0 is [*,*,*] (for short, written *[*,*,*]),
indicating that all predicates’ values are non-deterministic
at this location. And when the program transits from l_0 to l_1,
the abstract data state at l_1 is 111, since executing the state-
ments i=0 and x=0 can make the three predicates all true.

A path \( \pi \) of the program is an alternating sequence of
states and program statements, i.e.,

\[ \pi = (l_0, s_0) \xrightarrow{st_0} (l_1, s_1) \xrightarrow{st_1} \cdots \xrightarrow{st_{n-1}} (l_n, s_n). \]

A path \( \pi \) is a concrete path of \( P \) (or an abstract path of \( P^k \)) iff all states on \( \pi \) are concrete states of \( P \) (or abstract states of
A path \( \pi \) is called a CFA path if \( l_0 \) is the entry location of the program, and for each \( i \) with \( 0 \leq i < n \) there exists a CFA edge \( g = (l_i, s_t, l_{i+1}) \). In other words, a CFA path represents a syntactical walk through the CFA. A counterexample \( P \) is a CFA path of \( P \) (or \( P^\lambda \)) that ends at the error location.

Consider the CFA in Fig. 2 with the abstract precision \( \lambda_{\text{main}} = \{ i < 10, x \leq 5, x < 5 \} \). A possible counterexample is shown in Fig. 3, where the abstract data states are labelled beside the corresponding program locations.

We use the strongest post-condition operator \( SP \) to define the semantics of a CFA path. For a formula \( \varphi \) and a statement \( st \), \( SP_{st}(\varphi) \) represents the set of data states that are reachable from any of the states that satisfy \( \varphi \) after the execution of \( st \). Let \( s_1, s_2, \ldots, s_{n-1} \) be the sequence of program statements passed by the CFA path \( \pi \). The semantics of \( \pi \) is the successive application of the \( SP \) operator to each statement of \( \pi \), i.e., \( SP_{\pi}(\varphi) = SP_{s_n}(\ldots(SP_{s_1}(\varphi)\ldots)) \).

**Definition 1.** A CFA path \( \pi \) starting from the abstract state \((l, s)\) is feasible iff \( SP_{\pi}(s) \) is satisfiable.

Note that a feasible path is always a CFA path. Let \((l_0, s_0)\) be the initial state of \( P^\lambda \). An abstract state \((l, s)\) of \( P^\lambda \) is reachable iff there exists a feasible path \( \pi \) of \( P^\lambda \) that ends at the location \( l \) such that \( s \models SP_{\pi}(s_0) \).

**Definition 2.** The abstract model \( P^\lambda \) is correct iff the error location is not reachable in \( P^\lambda \).

### 4.2 CEGAR

We next describe the scheme of the standalone program verification (i.e., without reuse) via predicate abstraction and CEGAR [22], [23].

Let \( \lambda \) be an abstract precision, and \( P^\lambda \) be the abstract model with respect to \( \lambda \). Since the predicate abstraction is conservative [24], to verify the program \( P \), it is sufficient to find a proper abstract precision \( \lambda \) such that \( P^\lambda \) is correct. This can be achieved by the scheme of the counterexample-guided abstraction refinement (Fig. 4). Initially, the abstract precision is set to empty. The abstract precision is then iteratively refined by adding new predicates, until the program is verified.

Each iteration consists of two phases: a model validation phase and a counterexample analysis phase. During the model validation phase, we check if the abstract model \( P^\lambda \) is correct or not. If \( P^\lambda \) is correct, we immediately conclude that \( P \) is also correct. Otherwise, we get a counterexample \( \pi \) that is a CFA path of \( P^\lambda \) ending at the error location.

During the counterexample analysis phase, the counterexample \( \pi \) is semantically analyzed to determine whether it is feasible or not. If it is feasible, we find a real execution of the program \( P \) that reaches error, and we thus conclude that \( P \) is incorrect. Otherwise, \( \pi \) is a spurious counterexample, and the proof of its infeasibility can be used to refine the abstract precision [22]. The refinement is performed by adding new predicates in the abstract precision, such to eliminate the spurious counterexample from the refined model. After the refinement, the next iteration continues.

### 4.3 Summaries

We now introduce a unified definition for the procedure and loop summaries.

Let \( \varphi \) be a program fragment (either a procedure or a loop). An *entry state* of \( \varphi \) is a state at the entry location of \( \varphi \), and an *exit state* of \( \varphi \) is a state at its exit location. A pair of an entry state and an exit state summarizes the input/output behavior of one execution of \( \varphi \) [12], [13].

**Definition 3.** A summary of a program fragment \( \varphi \) is a triple \( \langle \lambda, \varphi_{in}, \varphi_{out} \rangle \), where \( \lambda \) is an abstract precision, \( \varphi_{in} \) and \( \varphi_{out} \) are Boolean combinations of predicates in \( \lambda \), representing an entry state and a set of exit states of \( \varphi \), respectively.

A summary states that if the entry state of \( \varphi \) satisfies \( \varphi_{in} \), its exit state must satisfy \( \varphi_{out} \). This definition is particularly suitable for predicate analysis. Let \( \lambda \) be the current abstract precision, and \( P^\lambda \) be the abstract model of \( P \) with respect to \( \lambda \). The model validation is essentially to traverse the state space of \( P^\lambda \) [17] to find that if the error location is reachable or not. Let \((l, s)\) be the current abstract state, and \( \varphi \) be the program fragment to be executed, the model validation algorithm needs to traverse all possible paths of \( \varphi \) under the abstract precision \( \lambda \) to compute the set of exit states. Let \( \phi \) be the formula representing the entry state \((l, s)\), and \( \phi' \) be the formula representing the set of exit states, the triple \( \langle \lambda, \phi, \phi' \rangle \) is a summary of \( \varphi \).

With the above definition, a summary corresponds to a subset of paths in \( \varphi \). The main advantage of using the state-based summaries is the efficiency. Consider a state-based
summary \(\langle \lambda, \phi_{in}, \phi_{out}\rangle\), the abstract precision \(\lambda\) is determined at each iteration of CEGAR; the entry state \(\phi_{in}\) and the set \(\phi_{out}\) of exit states are computed in the model validation process. In conclusion, all ingredients of this summary are by-products of CEGAR. There needs no additional computation to generate the state-based summaries.

Note that an abstract precision must be specified in the summary, since the entry state and the exit state must both be defined over the predicates in the abstract precision. Recall that we assume a unique abstract precision throughout a procedure or a loop. Thus only one abstract precision needs to be specified here. Otherwise, if the abstract precision differs in different points of \(\varrho\) (for example, as in the lazy abstraction [22]), we need to specify an abstract precision for \(\phi_{in}\) and \(\phi_{out}\), respectively.

All the generated summaries are maintained in a summary cache \(\mathcal{E}\). During the program analysis, whenever a program fragment is encountered, the verifier seeks in \(\mathcal{E}\) for an applicable summary. Let \(\lambda_i\) be the current abstract precision, and \(s_i\) be the current abstract data state. A summary \(\langle \lambda, \phi_{in}, \phi_{out}\rangle\) of \(\varrho\) is called applicable iff \(\lambda_i \subseteq \lambda\) and \(s_i \Rightarrow \phi_{in}\). If any applicable summary exists, the verifier directly uses \(\phi_{out}\) of this summary as the exit state of \(\varrho\). Otherwise, the verifier needs to conduct a heavy fix-point computation [27] on the fragment \(\varrho\) to compute its exit state.

5 CEGAR-BASED REGRESSION VERIFICATION

Summaries convey important information about the verification. In this section, we propose some efficient summary reuse techniques for regression verification.

5.1 Overview

An overview of our CEGAR-based regression verification is shown in Fig. 5. Besides the program \(P\), a set \(\mathcal{E}'\) of the previously-generated summaries is also provided for the regression verification. Note that these summaries are produced by the previous revisions, and may not be applicable to the current revision. Similar to [11], we propose a summary selection step to guarantee the safe reuse of summaries (see Section 5.2).

As in a standalone verification, each iteration of CEGAR for a regression verification also consists of two phases: a model validation phase and a counterexample analysis phase. The former phase is exactly the same as in the standalone verification. The counterexample analysis, however, requires more careful handling, which will be discussed in Section 5.3.

Note that our summary reuse does not depend on the verification result. A summary here represents an execution of the corresponding program fragment. No matter whether the verification result is “correct” or “incorrect”, as long as the fragment does not change semantically, the summary can be reused. This is very different from the interpolation-based summaries [11], where the summaries are related to the property to be verified, and can only be reused when the verification result is “correct”.

5.2 Summary Selection

Let \(P'\) be the old revision of \(P\). For each fragment \(\varrho\) (either a procedure or a loop) of \(P\), let \(\varrho'\) be its previous version in \(P'\). If \(\varrho'\) does not exist in \(P'\), i.e., \(\varrho\) is a newly added fragment in \(P\), we simply let \(\varrho' = \text{NULL}\).

In the summary selection (Algorithm 1), we check for each \(\varrho\) of \(P\) if \(\varrho\) is semantically equivalent to \(\varrho'\) or not. If it is, the summaries of \(\varrho'\) are selected. Otherwise, these summaries are abandoned. These selected summaries are then reused in the regression verification to initialize the summary cache \(\mathcal{E}\).

Algorithm 1. Summary Selection

\begin{verbatim}
for all \(\varrho \in P\) do
    Let \(\varrho'\) be its previous version in \(P'\);
    if \(\varrho \equiv \varrho'\) then
        Select summaries of \(\varrho'\);
    end
end
Use selected summaries to initialize \(\mathcal{E}\);
\end{verbatim}

Note that the semantic equivalence checking \(\varrho \equiv \varrho'\) is very expensive in computation, we thus choose to perform syntactic checking instead. The syntactic checking is less precise, i.e., may miss some semantically equivalent fragments, but preserves the soundness, i.e., the fragments that pass the syntactic checking must be semantically equivalent.

5.2.1 Syntactic Checking

In the following, we first discuss two notions, i.e., syntactically unchanged fragments and semantically equivalent fragments, then, we propose our syntactic checking technique.

Definition 4. Let \(\varrho\) be a program fragment in \(P\) and \(\varrho'\) be its previous version in \(P'\), we say \(\varrho\) is syntactically unchanged if all statements of \(\varrho\) and \(\varrho'\) are same, and syntactically changed, otherwise.

In the following, we consider only syntactically unchanged fragments for possible summary reuse. Note that a syntactically changed procedure may still be semantically equivalent to its previous version. We ignore this case since the semantic equivalence checking is too expensive.

On the other hand, a syntactically unchanged fragment is not necessarily to be semantically equivalent. Comparing only statements in a fragment is not enough for checking the semantics of this fragment. For example, the call statement in a fragment may lead the execution to a statement outside the fragment (in the called procedure). If the called procedure

Fig. 5. CEGAR-based regression verification.
changes, even all statements in the fragment under consideration remain the same, its semantics has changed.

Consider the program in Fig. 1, we assume that in the new program revision the main procedure remains the same while the inc procedure changes, then the semantics of the main procedure is regarded as changed.

To summarize, the syntactic changes of a fragment may lead the semantics of another fragment to be changed. We thus have the following definition.

Definition 5. Let $g_1, g_2$ be two program fragments of $P$, we say $g_1$ impacts $g_2$, written $g_1 \prec g_2$, if either of the following statements is satisfied:

- $g_1$ is a procedure and is called in $g_2$, or
- $g_1$ is a loop and is nested in $g_2$.

The above impact relation extends the call relation over procedures by taking loops and their nesting structures into consideration.

The impact relation is transitive, i.e., if $g_1 \prec g_2$ and $g_2 \prec g_3$, then $g_1 \prec g_3$. Let $\prec^*$ be the transitive closure of $\prec$. If $g_1 \prec^* g_2$, i.e., there exists $g_1 = g_{1_1} \prec \cdots \prec g_{1_k} = g_2$, we call $g_2$ reachable from $g_1$. Moreover, we define the set of forward reachable fragments from $g_1$ as $FReach(g_1) = \{g_2 \mid g_1 \prec^* g_2\}$. Apparently, $FReach(g_1)$ is the maximal set of fragments that $g_1$ can impact. We also define the set of backward reachable fragments from $g_1$ as $BReach(g_1) = \{g_2 \mid g_2 \prec^* g_1\}$, which is the maximal set of fragments that have impact on $g_1$.

Consider the three fragments in the program in Fig. 1: the main procedure (denoted as $\varrho_{main}$), the inc procedure (denoted as $\varrho_{inc}$) and the loop in the main procedure (denoted as $\varrho_{loop}$). They satisfy: $\varrho_{inc} \prec \varrho_{loop}, \varrho_{inc} \prec \varrho_{main}$ and $\varrho_{loop} \prec \varrho_{main}$. Thus we have $FReach(\varrho_{inc}) = \{\varrho_{loop}, \varrho_{main}\}$, $BReach(\varrho_{inc}) = \emptyset$. In other words, any change in $\varrho_{inc}$ can impact the semantics of $\varrho_{loop}$ and $\varrho_{main}$, and no other fragment can impact the semantics of $\varrho_{inc}$.

The concepts of forward/backward reachable sets can be lifted to a fragment set. Let $\tau$ be a set of fragments, $FReach(\tau) = \bigcup_{\varrho \in \tau} FReach(\varrho)$, and $BReach(\tau) = \bigcup_{\varrho \in \tau} BReach(\varrho)$.

Definition 6. Let $\varrho$ be a program fragment in $P$ and $\varrho'$ its previous version in $P'$, we say $\varrho$ is globally syntactically unchanged if

1) $\varrho$ is syntactically unchanged, and
2) all fragments in $BReach(\varrho)$ are syntactically unchanged.

Lemma 1. A globally syntactically unchanged fragment is semantically equivalent to its previous version.

Proof. This is a direct conclusion by the definition of globally syntactically unchanged fragment.

To find the globally unchanged fragments of $P$, we syntactically compare each fragment of $P$ to its previous version. According to the comparing results, fragments in $P$ are divided into two parts: the syntactically unchanged set $\tau_1$ and the syntactically changed set $\tau_2$. Then we have the following lemma.

Lemma 2. Let $\tau_1$ and $\tau_2$ be the set of syntactically unchanged and syntactically changed fragments of $P$, respectively. Fragments in $\tau_1 \setminus FReach(\tau_2)$ are all globally syntactically unchanged.

Proof. Assume that the lemma does not hold, i.e., there is a fragment $g_1 \in \tau_1 \setminus FReach(\tau_2)$ that is not globally syntactically unchanged. By $g_1 \in \tau_1$ and Definition 6, there must be a fragment $g_2 \in BReach(g_1)$ such that $g_2$ is syntactically changed. By $g_2 \in BReach(g_1)$, we have $g_1 \in FReach(g_2)$. By $g_2$ being syntactically changed, we have $g_2 \in \tau_2$, and thus $g_1 \in FReach(\tau_2)$. This is contradicted with the assumption. Thus the assumption does not hold, and the lemma holds.

Let $\tau^* = \tau_1 \setminus FReach(\tau_2)$. The semantic equivalence checking $\varrho \equiv \varrho'$ (on Row 4 of Algorithm 1) is implemented as checking whether $\varrho \in \tau^*$. If $\varrho \in \tau^*$, the summaries of $\varrho$ are selected, and otherwise they are abandoned. Note that the computations of $\tau_1, \tau_2$ and $\tau^*$ involve only syntactic checking of $P$ and $P'$. According to Lemmas 1 and 2, all selected summaries are semantically equivalent to its previous version, and thus they can be safely reused in the regression verification.

5.3 Counterexample Analysis

Given a counterexample returned by the model validation process, we need to check if this counterexample corresponds to a real bug or not. Summary reuse makes this process intricate.

Consider the counterexample in Fig. 3 that contains two procedure calls. During the program verification, these two procedure calls are replaced by two abstract summaries, which are defined over predicates and may introduce spurious behaviors over the program’s concrete semantics. Therefore, to check the feasibility of this counterexample, the inner paths in the inc procedure that correspond to these two summaries must be restored.

Consider the procedure call $\text{inc}(x,1)$ between ($l_3,111$) and ($l_5,11+$) for example. If the summary for this procedure call is generated in the current verification run, the inner path in $\text{inc}$ that leads from ($l_3,111$) to ($l_5,11+$) is available [17]; otherwise, if the summary is inherited from the previous verification, there is no information for the inner path. Then, we have to rely on the heavy fix-point computation [27] to reproduce this path. In other words, with the counterexample checking, the saved analysis on the inc procedure is getting back. The benefits of summary reuse are thus significantly weakened.

5.3.1 Holes

Definition 7. A summary on a path is called a hole if it is inherited from the previous verification runs.

Let $\pi$ be a counterexample path with holes. Replacing a hole with the corresponding inner path is called an expansion. For example, Fig. 6 shows the expanded version of the counterexample in Fig. 3. We call a path holeless if it contains no hole.

Let $H$ be the set of holes on a path $\pi$. The path $\pi$ is split by these holes into $|H| + 1$ path segments. Each of these segments is a holeless path. The semantics of a path with holes is defined as the conjunction of the semantics of its holeless segments.

Theorem 1. Let $\pi$ be a path with holes $H$, and $\Pi$ the set of segments of $\pi$ split by $H$, 
5.3.2 Lazy Analysis Algorithm

A brute-force algorithm for counterexample analysis is to directly expand all holes of the counterexample, and then check its feasibility. This algorithm is correct but inefficient. Recall that to expand a hole, we need to perform a heavy analysis on the program fragment, which usually costs lots of computation. In the worst case, it needs to traverse all paths of the fragment to reproduce the path from a given input state to a given output state. We therefore propose a technique to avoid unnecessary hole expansions.

Our lazy analysis algorithm is shown in Algorithm 2. The basic idea is to expand holes on demand, so as to avoid unnecessary hole expansions. The main body of the algorithm is a while loop. At the beginning of each iteration of the loop, the algorithm checks whether the current path is holeless \((\text{isHoleless}(\pi))\), and whether the current path is infeasible \((\text{isInfeasible}(\pi))\). If both checks return false, the loop continues by expanding one hole in \(\pi\). Otherwise, the current path \(\pi\) must be either infeasible or holeless. For the former case, the algorithm returns \(\pi\); and for the latter case, the algorithm reports “incorrect”.

```
Algorithm 2. lazyAnalysis(\(\pi\))

Input: A finite abstract path \(\pi\)
Output: The expanded path of \(\pi\) if it is infeasible; or “unsafe” if it corresponds to a real path.

while \((\neg \text{isHoleless}(\pi) \lor \text{isInfeasible}(\pi))\) do

let \(h\) be a hole in \(\pi\);
\(\pi \leftarrow \text{expandHole}(\pi, h)\);

end

if \(\text{isInfeasible}(\pi)\) then return \(\pi\);
else return incorrect;
```

Comparing to the brute-force approach, our lazy algorithm needs more feasibility checking. However, it is still beneficial. First, with the lazy approach, the computational efforts for unnecessary hole expansions (which in many cases are expensive) are saved. Second and more importantly, the returned path by the lazy approach is often much shorter than the fully expanded one. Note that the refinement is a heavy step in CEGAR [26]. With a shorter counterexample, the computation efforts for the refinement (for example, the interpolation-based refinement [26]) can often be significantly reduced.

The lazy analysis algorithm can be easily adapted to the existing CEGAR framework (for example, CPAChecker [17]) in the following way. When the verifier in the existing framework returns a counterexample, our algorithm is applied to check if this counterexample is spurious or not. In case of a spurious counterexample, our algorithm returns a (partially) expanded path and gives it to the existing refiner in the framework. The returned path by our algorithm may contain holes. Treating these holes as value assignments, these paths can be directly processed by most of the existing refinement techniques, for example, the interpolation-based refinement [26].

5.4 Precision Reuse

In the beginning of the regression verification (in Fig. 5), the abstract precision is set to be empty, which is in fact not necessary. Dirk Beyer et al. [7] showed that the abstract precision can also be reused in the regression verification. To take this idea, we simply use the final abstract precision \(\lambda\) in the previous verification to initialize the current abstract precision \(\lambda\), i.e., replace \(\lambda \leftarrow \emptyset\) with \(\lambda \leftarrow \lambda\) in Fig. 5.

With precision reuse, the amount of summaries that need to be recorded at the end of the verification run is also reduced. Assume that the abstract precision of the regression verification is initialized as \(\lambda\). Then during the regression verification, the abstract precision is iteratively refined by CEGAR. In other words, summaries with a smaller abstract precision than \(\lambda\) are useless in the regression verification. Thus, we need only to output the summaries with the final precision at the end of each verification run.

6 EXPERIMENTAL EVALUATION

We implemented our regression verification technique on top of CPAChecker [17]. CPAChecker provides a configurable...
framework for software verification, with both predicate abstraction and counterexample-guided abstraction refinement supported. To support regression verification, we realized the following functionalities in CPAchecker:

- summary dumping, i.e., dump all summaries to an external file at the end of the verification,
- summary selection, i.e., load and select summaries from an external file at the beginning of the verification (Section 5.2), and
- lazy counterexample analysis (Section 5.3).

Moreover, precision reuse (Section 5.4) was also integrated into our implementation. In the following, we call our enhanced implementation CPAchecker++.

Experiments are designed to answer the following research questions:

- RQ1. Is summary reuse efficient enough for regression verification?
- RQ2. What is the impact of precision reuse on the efficiency of our approach?
- RQ3. What is the impact of summary types on the efficiency of our approach?
- RQ4. What is the impact of the lazy counterexample analysis on the efficiency of our approach?

6.1 Experimental Setup

We prepared two industrial benchmarks with 3,675 program revisions of 488 Linux device drivers to evaluate our approach:

1) The first benchmark, obtained from [7], consists of 1,119 real-world program revisions of 62 Linux device drivers.
2) The second benchmark, prepared by ourselves, consists of 2,556 program revisions of 426 Linux device drivers, where all drivers were collected1 from the “SystemsDeviceDriversLinux64Reach-Safety” category of the 6th International Competition on Software Verification (SV-COMP’17) [28]. For each driver, we make the program in [28] as the base revision, and use a state-of-the-art mutation tool MiLu [29] to randomly generate 5 artificial new revisions.

Recall that all program revisions in the first benchmark were obtained from the official Linux kernel repositories [7], and they are real revisions implemented by the experienced programmers. We use this benchmark to evaluate our approach on real program changes. In contrast, program revisions in the second benchmark were obtained by adding mutations to the base revision of each device driver. An advantage of the second benchmark is that the mutants generated by random pattern can involve much more unpredictable modifications than those written by human programmers. We use the second benchmark to evaluate our approach on a broader range of program changes.

All experiments are performed on a machine with Intel Xeon E5-2620 CPU of 2.4 GHz 24 cores and 32 GB RAM. We use Ubuntu 16.04 (64-bit) with Linux 4.4.0 and jdk1.8.0. The CPAchecker is configured using the predicateAnalysis-ABE option. Each verification run is limited to 300 seconds (total CPU time), 6 GB of Java heap size and 6 CPU cores.

6.2 Overall Results on Real Revisions (RQ1)

This experiment evaluates our approach on real program revisions. We compare the performance of CPAchecker+ with CPAchecker on the first benchmark. All regression verification techniques, including summary reuse, precision reuse, and lazy counterexample analysis are enabled for CPAchecker+ in this experiment. For simplicity, in the following, we refer to our approach as “Reuse”, and the standard CPAchecker as “no Reuse”.

In our experiments, a verification task is to verify a program revision against a specification. Note that a device driver may have multiple program revisions and also multiple specifications. A pair of a device driver and a specification involves a sequence of verification tasks, where the base revision is verified from scratch, while the other revisions are verification in an incremental way, i.e., as regression verifications. In total, there are 259 driver/specification pairs and 4,193 verification tasks in the first benchmark. Among all tasks, 3,934 are regression verification tasks.

Experimental results are listed in Table 1. Due to page limitation, we restrict this table to the 40 best and 10 worst cases out of the total of 259 driver/specification pairs (sorted by the “Speedup” column). The first two columns (“Driver” and “Specification”) list the device driver name and the specification name, respectively. The third column “LoC” shows the lines of code for the base revision of each device driver.

The fourth column (“#T”) shows the number of regression verification tasks (i.e., the number of revisions minus 1) for each driver/specification pair. The fifth column (“T1st”) lists the analysis time for verifying the first revision which is not a regression verification task. This “T1st” gives us the information on the complexity of verifying each device driver.

Recall that each row in the table corresponds to a driver/specification pair. The last two rows (“Sum” and “Avg”) show the total number of abstract successor computations “#abs_succ”2 and the total analysis time “Tvy” (in seconds) for each driver/specification pair. To conduct a fair comparison, “Tvy” and “#abs_succ” are limited to regression verification tasks that are solved by both approaches. The “Speedup” column shows the average speedup of “Reuse” approach over the “no Reuse” approach, calculated by: 1 − Tvy2/Tvy1. The last “AvgFSize” and “RSR” columns report the average size (in Kilobytes) of summary files, and the average reusable summary ratio, among all revisions of each driver/specification pair, respectively. For each regression verification task, RSR is the proportion of summaries that are kept after the summary selection.

1. The selection strategy is as similar as in [7]. We limit our selection to drivers of Linux 3.4 kernel and with the mutex lock/unlock specifications, and skipped programs whose total CPU time is less than 0.5s and those that need no refinement.

2. An abstract successor is a successor of the current state on the abstract model. Abstract successors computation needs to invoke a SMT solver and is considered as the most time-costly operation in predicate abstraction [23].
take the total and average amount of all rows in the table, respectively.

From Table 1, we observed that our method outperforms “no Reuse” in vast majority of cases. Considering “Speed-up” column, among 259 driver/specification pairs, only one pair that our method is slower. Comparing “#solved” columns of both approaches, 216 more regression verification tasks were solved with our approach. This witnesses the value of summary reuse for regression verification.

Among the common 3,581 regression verification tasks that both approaches can verify, “no Reuse” takes 151.8 thousand seconds of analysis time while our “Reuse” approach finishes in 10.5 thousand seconds. The overall time speedup of our approach is 93.1 percent.

Comparing the numbers of abstract successor computations (“#abs_succ.”) required by “Reuse” and “no Reuse” for each spec/driver pair, we found that our method cuts down the amount significantly (about 98 percent reduction), which can explain the reason for the speedup of analysis time.

Let us look at the “AvgFSize” column. The average size of summary files among all regression revisions is 17.4 KB (the median is 3.8 KB). The added overhead by our approach in storage is acceptable.
TABLE 2
Overall Experimental Results on Larger Changes

<table>
<thead>
<tr>
<th>Revs.</th>
<th>Avg. Diff. Lines</th>
<th>#T</th>
<th>#solved</th>
<th>T_{rv1}</th>
<th>Reuse</th>
<th>#solved</th>
<th>T_{rv2}</th>
<th>Speed-up</th>
<th>RSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3954</td>
<td>511</td>
<td>3588</td>
<td>151.8K</td>
<td>3804</td>
<td>10.5K</td>
<td>93.1%</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td>898</td>
<td>1242</td>
<td>812</td>
<td>32.7K</td>
<td>870</td>
<td>4.5K</td>
<td>86.1%</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>2Revs</td>
<td>259</td>
<td>1562</td>
<td>240</td>
<td>7.7K</td>
<td>245</td>
<td>1.5K</td>
<td>80.8%</td>
<td>0.4</td>
<td></td>
</tr>
</tbody>
</table>

6.2.1 Scaling With Larger Changes
The changes between adjacent revisions may not be very significant. We further use the following settings to evaluate our approach on larger changes:

- 4th: we set up a regression verification task every 4 revisions of the program, and
- 2Revs: we set up a regression verification task for the last revision of each program.

Finally, we get respectively 898 and 259 regression verification tasks using the above two settings.

Experimental results are listed in Table 2. The original setting that incrementally verifies each adjacent revision of the program is referred to as All in the table. The average numbers of changed lines for regression verification tasks using the above three settings are 511, 1242 and 1562, respectively. The increasing avg. diff. lines lead to decreasing RSR, which is reasonable since more program changes can of course lead to less summaries to be reusable.

Observe that our approach can still get considerable performance improvements (86.1 and 80.8 percent of speedups) with the 4th and 2Revs settings, which show the effectiveness of our approach on larger changes.

6.3 Overall Results on Artificial Revisions (RQ1)
The second experiment evaluates our approach on artificial program revisions in the second benchmark. This experiment contains 2,556 verification tasks, involving 426 driver/specification pairs. Among all tasks, 2,130 are regression verification tasks.

Results of this experiment are listed in Table 3. Again, we limit this table to the 40 best and 10 worst cases out of all 426 driver/specification pairs (sorted by the “Speedup” column). Each column is with the same meaning as in Table 1. Note that there is only one specification for each driver in this benchmark, the “Spec.” column is thus skipped.

From this table, we observed similar results as in Table 1. Among all 426 driver/specification pairs, our approach wins on 389 pairs. In total, our approach solved 10 more verification tasks, and the average speedup is 84.2 percent.

6.4 Comparison With Existing Tools (RQ1)
To further demonstrate the efficiency of our approach, two more experiments were conducted to compare CPAchecker\(^+\) with the existing regression verification tools:

- eVolCheck [30], a regression verification tool that implements the technique of interpolation-based procedure summaries [11], and
- UAutomizer\(^+\) [31], a regression verification extension of the famous software verification tool UAutomizer [19].

6.4.1 Comparison With eVolCheck
This experiment was conducted on the set of real-world programs. Before this experiment, some of the programs need to be modified to adapt to the input format of eVolCheck, e.g., replacing “ldv_error()” by “assert(0)”. Note that eVolCheck [30] is just an experimental implementation, and is not fully optimized. Among all 4,193 verification tasks, eVolCheck failed on 3,646 tasks due to various parsing and runtime errors. The comparative experiment was conducted on the remaining 547 verification tasks.

The comparison results are listed in Table 4, where the T_{rv1} and T_{rv2} columns report the total regression verification time without and with reuse, respectively, and the “Speedup” is calculated by 1 – T_{rv2}/T_{rv1}. Note that eVolCheck employs the bounded model checking technique, and its unwinding factor was set to 15. Among the 547 verification tasks, eVolCheck solved (i.e., the underlying SMT solver returned a result) 143 tasks within the time limit of 300 seconds, whereas our CPAchecker\(^+\) solved all. In comparison of the efficiency of the employed reuse techniques, the speedups of eVolCheck and CPAchecker\(^+\) are 75.2 and 89.7 percent, respectively. These results demonstrate the efficiency of our summary reuse technique.

6.4.2 Comparison With UAutomizer
This experiment was conducted on the real-world benchmark, too. Excluding the programs that UAutomizer fails to parse, there are totally 1,177 verification tasks that belong to 90 driver/specification pairs.\(^4\)

Note that the adopted verification techniques are very different in these two tools: UAutomizer\(^+\) uses the trace abstraction, while our CPAchecker\(^+\) uses the predicate abstraction. Moreover, the regression verification techniques implemented in these two tools are also different: one attempts to reuse the previously generated Floyd-Hoare automata [31], while another attempts to reuse the previously generated state-based summaries. To compare the efficiency of their adopted regression verification techniques, we compare the speedups of these two tools (with reuse over without reuse).

The comparison results are listed in Table 5. Note that UAutomizer\(^+\) implements two reuse strategies, i.e., Eager and Lazy. Results for both strategies are reported. From Table 5, CPAchecker\(^+\) achieves a speedup of 90.8 percent,\(^3\)

\(^3\) A successor version of this tool was recently released at: http://verify.inf.usi.ch/upprover
\(^4\) In their original paper [31], the UAutomizer\(^+\) was evaluated on the same set of programs.
and UAutomizer + gets a speedup of 82.8 percent or 81.4 percent. This result demonstrates the efficiency of our summary reuse technique. Note that we cannot conclude the superiority of our reuse technique over UAutomizer + from this result, since they are used in different verification frameworks.

### 6.5 Impact of Reuse Strategies (RQ2)

In the former two experiments, both summary reuse and precision reuse were enabled for our approach. In this experiment, we switch off “precision reuse” and “summary reuse” to evaluate the impact of summary reuse alone.

#### TABLE 3

<table>
<thead>
<tr>
<th>Run set</th>
<th>Device</th>
<th>LoC</th>
<th>#T</th>
<th>#T复活</th>
<th>#solved</th>
<th>#abs_succ</th>
<th>T1</th>
<th>T2</th>
<th>Speedup</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>iuu_phoenix</td>
<td>10258</td>
<td>5</td>
<td>214.5</td>
<td>5</td>
<td>136306</td>
<td>202.2</td>
<td>5</td>
<td>105</td>
<td>2.1</td>
<td>99.0%</td>
</tr>
<tr>
<td>broadcom1</td>
<td>4931</td>
<td>5</td>
<td>75.6</td>
<td>5</td>
<td>244123</td>
<td>72.0</td>
<td>5</td>
<td>77</td>
<td>1.1</td>
<td>88.8%</td>
</tr>
<tr>
<td>sungemphi</td>
<td>5287</td>
<td>5</td>
<td>184.3</td>
<td>5</td>
<td>249324</td>
<td>115.2</td>
<td>5</td>
<td>270</td>
<td>2.0</td>
<td>89.8%</td>
</tr>
<tr>
<td>paride-bpce</td>
<td>9886</td>
<td>5</td>
<td>8544</td>
<td>5</td>
<td>85737</td>
<td>89.1</td>
<td>5</td>
<td>106</td>
<td>1.7</td>
<td>91.1%</td>
</tr>
<tr>
<td>r12830</td>
<td>4457</td>
<td>5</td>
<td>111.9</td>
<td>5</td>
<td>16172</td>
<td>107.9</td>
<td>5</td>
<td>69</td>
<td>2.3</td>
<td>97.9%</td>
</tr>
<tr>
<td>mwifxpecc</td>
<td>11367</td>
<td>5</td>
<td>95.1</td>
<td>5</td>
<td>128270</td>
<td>99.0</td>
<td>5</td>
<td>164</td>
<td>2.3</td>
<td>97.7%</td>
</tr>
<tr>
<td>paride-epc</td>
<td>7695</td>
<td>5</td>
<td>106.0</td>
<td>5</td>
<td>22426</td>
<td>103.4</td>
<td>5</td>
<td>271</td>
<td>2.7</td>
<td>97.4%</td>
</tr>
<tr>
<td>ttytmased</td>
<td>10237</td>
<td>5</td>
<td>67.1</td>
<td>5</td>
<td>135512</td>
<td>66.4</td>
<td>5</td>
<td>120</td>
<td>1.9</td>
<td>97.2%</td>
</tr>
<tr>
<td>paride-kbfc</td>
<td>8009</td>
<td>5</td>
<td>77.9</td>
<td>5</td>
<td>88482</td>
<td>74.9</td>
<td>5</td>
<td>190</td>
<td>2.2</td>
<td>97.1%</td>
</tr>
<tr>
<td>tpm_nsc</td>
<td>4791</td>
<td>5</td>
<td>35.5</td>
<td>5</td>
<td>10087</td>
<td>36.2</td>
<td>5</td>
<td>69</td>
<td>1.1</td>
<td>97.1%</td>
</tr>
<tr>
<td>mxll11sf1demod</td>
<td>82388</td>
<td>5</td>
<td>32.2</td>
<td>5</td>
<td>249425</td>
<td>31.7</td>
<td>5</td>
<td>41</td>
<td>1.0</td>
<td>97.0%</td>
</tr>
<tr>
<td>paride-on26</td>
<td>17745</td>
<td>5</td>
<td>134.6</td>
<td>5</td>
<td>80837</td>
<td>137.1</td>
<td>5</td>
<td>443</td>
<td>3.9</td>
<td>97.8%</td>
</tr>
<tr>
<td>dib3000mb</td>
<td>7868</td>
<td>5</td>
<td>120.3</td>
<td>5</td>
<td>26216</td>
<td>123.1</td>
<td>5</td>
<td>447</td>
<td>4.6</td>
<td>96.2%</td>
</tr>
<tr>
<td>paride-on20</td>
<td>7747</td>
<td>5</td>
<td>40.9</td>
<td>5</td>
<td>20099</td>
<td>41.3</td>
<td>5</td>
<td>106</td>
<td>1.6</td>
<td>96.1%</td>
</tr>
<tr>
<td>paride-dstr</td>
<td>6604</td>
<td>5</td>
<td>34.3</td>
<td>5</td>
<td>16655</td>
<td>33.8</td>
<td>5</td>
<td>84</td>
<td>1.4</td>
<td>95.8%</td>
</tr>
<tr>
<td>cmd640</td>
<td>6406</td>
<td>5</td>
<td>38.5</td>
<td>5</td>
<td>36946</td>
<td>37.7</td>
<td>5</td>
<td>93</td>
<td>1.6</td>
<td>95.7%</td>
</tr>
<tr>
<td>it913x</td>
<td>5947</td>
<td>5</td>
<td>69.1</td>
<td>5</td>
<td>14337</td>
<td>53.6</td>
<td>5</td>
<td>82</td>
<td>2.4</td>
<td>95.6%</td>
</tr>
<tr>
<td>serqt_usb21</td>
<td>10282</td>
<td>5</td>
<td>56.8</td>
<td>5</td>
<td>33423</td>
<td>56.4</td>
<td>5</td>
<td>166</td>
<td>2.6</td>
<td>95.3%</td>
</tr>
<tr>
<td>tuners-g10101</td>
<td>7043</td>
<td>5</td>
<td>54.1</td>
<td>5</td>
<td>2645</td>
<td>47.9</td>
<td>5</td>
<td>79</td>
<td>2.3</td>
<td>95.3%</td>
</tr>
<tr>
<td>cfg12864b</td>
<td>1493</td>
<td>5</td>
<td>15.0</td>
<td>5</td>
<td>5278</td>
<td>14.8</td>
<td>5</td>
<td>13</td>
<td>0.6</td>
<td>95.1%</td>
</tr>
<tr>
<td>paride-comm</td>
<td>5408</td>
<td>5</td>
<td>24.9</td>
<td>5</td>
<td>13211</td>
<td>23.8</td>
<td>5</td>
<td>77</td>
<td>1.2</td>
<td>94.9%</td>
</tr>
<tr>
<td>cp210lix</td>
<td>7113</td>
<td>5</td>
<td>186.9</td>
<td>5</td>
<td>274622</td>
<td>166.8</td>
<td>4</td>
<td>1526</td>
<td>10.8</td>
<td>94.8%</td>
</tr>
<tr>
<td>paride-ktti</td>
<td>2886</td>
<td>5</td>
<td>15.1</td>
<td>5</td>
<td>6299</td>
<td>15.3</td>
<td>5</td>
<td>52</td>
<td>0.8</td>
<td>94.7%</td>
</tr>
<tr>
<td>paride-frag</td>
<td>8979</td>
<td>5</td>
<td>40.8</td>
<td>5</td>
<td>12680</td>
<td>40.1</td>
<td>5</td>
<td>146</td>
<td>2.2</td>
<td>94.6%</td>
</tr>
<tr>
<td>google-gsmi</td>
<td>6826</td>
<td>5</td>
<td>20.8</td>
<td>5</td>
<td>5294</td>
<td>19.9</td>
<td>5</td>
<td>48</td>
<td>1.1</td>
<td>94.3%</td>
</tr>
<tr>
<td>specpax</td>
<td>6913</td>
<td>5</td>
<td>29.6</td>
<td>5</td>
<td>1796</td>
<td>29.3</td>
<td>5</td>
<td>50</td>
<td>1.3</td>
<td>94.2%</td>
</tr>
<tr>
<td>t2c-diolan</td>
<td>5308</td>
<td>5</td>
<td>30.2</td>
<td>5</td>
<td>13083</td>
<td>30.8</td>
<td>5</td>
<td>200</td>
<td>1.8</td>
<td>94.1%</td>
</tr>
<tr>
<td>sil164</td>
<td>6298</td>
<td>5</td>
<td>16.8</td>
<td>5</td>
<td>6906</td>
<td>16.6</td>
<td>5</td>
<td>66</td>
<td>1.0</td>
<td>94.0%</td>
</tr>
<tr>
<td>metro-usb</td>
<td>6288</td>
<td>5</td>
<td>25.3</td>
<td>5</td>
<td>20810</td>
<td>25.0</td>
<td>5</td>
<td>73</td>
<td>1.5</td>
<td>94.0%</td>
</tr>
<tr>
<td>paride-fit3</td>
<td>4459</td>
<td>5</td>
<td>19.5</td>
<td>5</td>
<td>10691</td>
<td>19.7</td>
<td>5</td>
<td>72</td>
<td>1.2</td>
<td>93.9%</td>
</tr>
<tr>
<td>wms31xldo</td>
<td>3720</td>
<td>5</td>
<td>19.9</td>
<td>5</td>
<td>15462</td>
<td>18.7</td>
<td>5</td>
<td>129</td>
<td>1.2</td>
<td>93.3%</td>
</tr>
<tr>
<td>paride-fit2</td>
<td>2593</td>
<td>5</td>
<td>13.6</td>
<td>5</td>
<td>5301</td>
<td>13.1</td>
<td>5</td>
<td>50</td>
<td>0.9</td>
<td>93.3%</td>
</tr>
<tr>
<td>sdtrchrs</td>
<td>1633</td>
<td>5</td>
<td>12.4</td>
<td>5</td>
<td>7725</td>
<td>13.2</td>
<td>5</td>
<td>38</td>
<td>0.9</td>
<td>93.2%</td>
</tr>
<tr>
<td>tdc24m</td>
<td>5039</td>
<td>5</td>
<td>30.9</td>
<td>5</td>
<td>12666</td>
<td>26.7</td>
<td>5</td>
<td>121</td>
<td>1.9</td>
<td>92.8%</td>
</tr>
<tr>
<td>hidzydacron</td>
<td>3211</td>
<td>5</td>
<td>10.6</td>
<td>5</td>
<td>2855</td>
<td>10.2</td>
<td>5</td>
<td>40</td>
<td>0.7</td>
<td>92.9%</td>
</tr>
<tr>
<td>max9890</td>
<td>1736</td>
<td>5</td>
<td>10.2</td>
<td>5</td>
<td>2984</td>
<td>10.3</td>
<td>5</td>
<td>27</td>
<td>0.7</td>
<td>92.8%</td>
</tr>
<tr>
<td>paride-epia</td>
<td>5518</td>
<td>5</td>
<td>28.1</td>
<td>5</td>
<td>10999</td>
<td>27.5</td>
<td>5</td>
<td>187</td>
<td>2.0</td>
<td>92.8%</td>
</tr>
<tr>
<td>wms31xdec</td>
<td>3420</td>
<td>5</td>
<td>21.7</td>
<td>5</td>
<td>18191</td>
<td>22.7</td>
<td>5</td>
<td>149</td>
<td>1.6</td>
<td>92.8%</td>
</tr>
<tr>
<td>paride-frpw</td>
<td>8046</td>
<td>5</td>
<td>33.2</td>
<td>5</td>
<td>9374</td>
<td>33.4</td>
<td>5</td>
<td>182</td>
<td>2.6</td>
<td>92.2%</td>
</tr>
<tr>
<td>leds-55521</td>
<td>5443</td>
<td>5</td>
<td>6.4</td>
<td>5</td>
<td>526</td>
<td>6.4</td>
<td>5</td>
<td>12</td>
<td>0.6</td>
<td>91.4%</td>
</tr>
</tbody>
</table>

#### TABLE 4

<table>
<thead>
<tr>
<th>#computable</th>
<th>T复活</th>
<th>T1</th>
<th>T2</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>eVolCheck</td>
<td>143</td>
<td>2382.7</td>
<td>591.9</td>
<td>75.2%</td>
</tr>
<tr>
<td>CPAchecker+</td>
<td>244</td>
<td>2074.2</td>
<td>213.2</td>
<td>89.7%</td>
</tr>
</tbody>
</table>

#### TABLE 5

<table>
<thead>
<tr>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPAchecker+</td>
</tr>
<tr>
<td>UAutomizer+-Eager</td>
</tr>
<tr>
<td>UAutomizer+-Lazy</td>
</tr>
</tbody>
</table>
reuse" respectively, and evaluate the efficiency of our approach under different reuse strategies, i.e., “no Reuse”, “Precision Reuse”, “Summary Reuse” and “both Reuse”. This experiment was conducted on the second benchmark.

Table 6 shows the results of this experiment. Every row sums up results of all 2,130 regression verification tasks. For each row, we report the total regression verification time “$T_{rv}$” (in seconds), the total number of solved regression tasks “#solved”, and the average memory usage “Mem” (in gigabytes). Note that we only list the time usage in “no Reuse” row. For the other three rows, we show the speedup against “no Reuse” approach.

From Table 6, we found that every reuse technique outperforms “no Reuse”. They are not surprising since the precision reuse can significantly reduce the CEGAR iterations [7], and thus cuts down the verification time (73.5 percent speedup); and the summary reuse can save the repeated computation of summaries, and thus also reduces the verification time (69.5 percent speedup). Moreover, “Summary Reuse” solves 2 more tasks than “Precision Reuse”, illustrating that the former technique is more robust than the latter one. Precision reuse and summary reuse are two orthogonal techniques. By integrating these two techniques, “both Reuse” get the best performance, not only in analysis time (84.2 percent speedup), but also in the number of verified regression tasks.

The “Mem” column shows that all reuse strategies save the memory usage meetly. Again, “both Reuse” saves the most on memory consumption.

### 6.6 Impact of Summary Types (RQ3)

This experiment investigates the efficiency of our approach on different types of summaries. We evaluate the efficiency of our approach with loop summaries reused, procedure summaries reused and all summaries reused, respectively. Note that “precision reuse” and “lazy counterexample analysis” are switched off in this experiment.

We accumulate the analysis time on different summary types. Results are illustrated in Fig. 7, where the X-axis indicates the number of device drivers, and the Y-axis represents the accumulated analysis time. From Fig. 7, we observed that “Procedure Summaries” outperforms “Loop Summaries”, and “All Summaries Reuse” performs the best. The main reason is that our benchmark contains fewer loop statements than procedures (1,273 versus 11,417).

### 6.7 Impact of Lazy Counterexample Analysis (RQ4)

The final experiment evaluates the efficiency of lazy counterexample analysis. Note that this technique is mainly relative to the refinement process, we measure the refinement time and counterexample length in this experiment.

Fig. 8a shows the results on counterexample length in a scatter diagram. Note that each device driver may involve several regression verification tasks, and each regression verification task may require many CEGAR iterations. The reported counterexample length is the accumulated length of all counterexamples generated in all CEGAR iterations among all regression verification tasks of a device driver. In Fig. 8a, each point represents the accumulated counterexample length of a device driver, the X and Y axes indicate our approach with and without lazy counterexample analysis, respectively. Both X and Y axes are logarithmic. A point below the reference line indicates a case where the lazy counterexample analysis is beneficial.

Fig. 8b show results on refinement time. Again, the reported refinement time is the accumulation of time spent on all refinement iterations among all regression verification tasks of all the currently-tested device drivers. In Fig. 8b, the X-axis catalogs the number of device drivers and the Y-axis shows the accumulated refinement time.
On the whole, the lazy counterexample analysis performs better on refinement, saving about 29.9 percent of time. It is also observed that for 415 of 426 device drivers, the corresponding data points in Fig. 8a are above the reference line, indicating that the accumulated counterexample lengths were reduced with this technique. These results conform to our algorithmic analysis in Section 5.3. With the lazy analysis technique, the counterexample is not necessarily to be fully expanded. And a shorter counterexample can usually reduce the refinement efforts.

7 RELATED WORK

Regression verification was investigated mainly in two directions, the verification of differences, and the reuse of previously computed results. We also discuss the summarization and symbolic execution techniques in this section.

Verification of Differences. In this line of research, one attempts to establish the correctness of the new program by proving its (conditional) equivalence to an old and verified program.

Many techniques have been proposed in this line of research. The technique for proving conditional equivalence of two programs by abstraction and decomposition of procedures is proposed in [4], [32]. Backes et al. [5] proposed to distinguish the program behaviors that are impacted by the changes. Only the impacted program behaviors needed to be considered during the regression verification. Beyer et al. [33] proposed the conditional model checking, which outputs a condition such that the program satisfies the specification under this condition. Böhme et al. [6] proposed a partition-based regression verification technique. Instead of proving the absence of regression errors for the entire input space, this approach continuously verifies the input space in a gradual manner. Felsing et al. [34] reduced the equivalence proving of two related imperative integer programs to Horn constraints over uninterpreted predicates, and then solved the constraints using an Horn solver.

Moreover, Rungta et al. [35] presented a technique for interprocedural change impact analysis. Yang et al. [36] introduced an incremental approach for checking the conformance of code against different properties. Trostanetski et al. [37] analyzed the semantic difference between successive revisions. Mora et al. [38] performed modular symbolic execution to prove the equivalence between different versions of libraries with respect to the same parts of codebase (client program).

Reuse of Intermediate Results. In this line of research, one studies the reuse of previously-generated results to the current verification. A variety of information has been proposed for reuse.

Some researchers [9], [39], [40], [41] proposed to keep the reached state space and reuse them in the further verification runs. The rationale of these techniques is that state spaces of consecutive versions tend to be similar. However, recording and reusing reached state space may be costly, and these techniques may not be applicable to large-scale programs. For example, [40] points out 6 times more on memory usage in the worst case. The lines of code of single revision of [40] are less than 1,000.

Visser et al. [10] noticed the importance of constraint solving for symbolic execution. They proposed to cache and reuse the results of constraint solving. This approach was further improved in [42], [43] from different aspects. This group of techniques is orthogonal to our approach. These techniques can be applied to enhance our approach.

Beyer et al. [7] proposed to use abstract precisions as the intermediate results. An abstract precision defines the level of abstraction, which conveys important information on the current verification. They proposed to record the final abstract precision and to reuse it as the initial abstract precision of the current verification. With this technique, the number of refinements can often be reduced. Note that the precision reuse and our summary reuse are orthogonal to each other. It is possible to combine these two reuse techniques together. We have already combined this technique with ours. The combined technique shows a very promising performance.

Fedyukovich et al. [44] offered a regression verification technique for checking property directed equivalence. The safe inductive invariants across program transformations were migrated and established. Rothenberg et al. [31] proposed to reuse the sequence of Floyd-Hoare automata learned during the trace abstraction. Two reuse strategies, eagerly and lazily, were developed in this paper. This technique has been realized in UAutomizer, a well-known software verification tool.

The work most relevant to ours is [11], [30], where a regression verification technique by means of interpolation-based procedure summaries was proposed. Our idea of summary reuse was inspired by these two papers. The main difference lies in the way in which the summaries are constructed during the verification. In [11], [30], the authors use a logical formula $\varphi_A$ to encode the behaviors of the procedure $\rho$, and another logical formula $\varphi_B$ to encode its calling context. Then they compute the interpolation of $\varphi_A$ and $\varphi_B$ and use that as the summary of $\rho$. In contrast, we use the abstract states in predicate analysis to construct the program summaries. Each summary in our paper consists of an entry state and a set of exit states of $\rho$. Our state-based summaries can be completely integrated into the framework of CEGAR. All ingredients of a state-based summary are by-products of CEGAR. There needs no additional computation for generating this kind of summaries. Experimental comparison in Section 6.4 demonstrates the practicability of our approach.

Pastore et al. [45] proposed a method to validate that an already tested code has not been broken by an upgrade. It maintains a test suite that can be used to revalidate the software as it evolves. Different from our approach, this technique is respect to regression testing. The verification technique is used there, as an aid, to validate dynamic properties (or invariants). In contrast, we aim to provide a new regression verification technique via reusing summaries.

Summarization. In this line of research, one tries to replace program fragments with summaries. A summary can usually be represented as an input-output pair of a program fragment. Procedure summaries have been long studied and there are also many studies on loop summaries recently.
Many researchers [11, 30, 46, 47] proposed to use interpolation-based method to generate procedure summaries. [47] combines function summaries with the expressiveness of satisfiability modulo theories (SMT), which makes summaries smaller and more human-readable. [11, 30, 46] implement a procedure summarization approach for software bounded model checking, and uses interpolation-based procedure summaries as over-approximation of procedure calls.

Kroening et al. [48] proposed the idea to substitute a loop with a conservative abstraction of its behavior, constructing abstract transformers for nested loops starting from the innermost loop. They also applied this method in termination analysis. Seghir et al. [49] used various inference rules for deriving summaries based on control structures. However, this approach can only compute precise loop summaries for restricted classes of programs depending on inference rules. Xie et al. [50] proposed a general framework for summarizing multi-path loops. It classifies loops according to the patterns of values changes in path conditions and the interleaving of paths within the loop. A disjunctive summarization is constructed for all the feasible executions in the loop. Different from our method, [50] cannot summarize loops containing non-induction variables, array variables, and nested loops. Godefroid et al. [51] investigated an alternative approach based on automatic loop-invariant generation. This approach can (partially) summarize a loop body during a single dynamic symbolic execution, which can ease the path explosion in dynamic test generation.

Symbolic Execution. In recent years, a great deal of effort has been focused on regression symbolic execution, which takes advantage of the previous analysis of symbolic execution to speedup the current analysis.

Person et al. [52] used a form of overapproximating symbolic execution to skip portions of the program that are provably identical across the versions. In [53], Person et al. presented a regression symbolic execution technique for Java programs, based on the Symbolic PathFinder. It analyzes the CFA of two program versions, computes the locations affected by the program changes, and then applies the symbolic execution to the affected code only. Further more, Guo et al. [54] investigated the symbolic execution technique for multi-threaded programs.

8 Conclusion

We proposed in this paper a fully automatic regression verification technique in the context of CEGAR. Abstract summaries are reused across different abstract precisions and different program revisions. We proposed a unified framework for reusing both procedure summaries and loop summaries. A lazy counterexample analysis algorithm was further proposed to reduce the unnecessary path expansion efforts. We implemented our approach in the software verification tool CPAchecker. Experimental results show the promising performance of our technique.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (61672310, 61527812), the National Key R& D Program of China (2018YF1308601).

References

HE ET AL.: EFFICIENT SUMMARY REUSE FOR SOFTWARE REGRESSION VERIFICATION 1431


For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.