Introduction to Static Analysis

Lecture 11
CS295

Definitions

• Program analysis
  Discovering facts about programs.

• Dynamic analysis
  Program analysis using program executions.

• Static analysis
  Program analysis without running the program.

History

• Static analysis is nearly as old as programming.

• First used in compilers
  - For program optimization
    - Starting with FORTRAN (1954)

• Use for software quality nearly as old
  - Type systems

So What?

• It’s a big field
  - With different approaches
  - And applications
  - And lots of terms

• This lecture aims to sketch basic
  - concepts
  - techniques
  - terminology

Laundry List Outline

• Analysis paradigms
  - Type systems
  - Dataflow analysis
  - Model checking

• Terminology
  - Abstract values
  - Flow insensitive
  - Flow sensitive
  - Path sensitive
  - Local vs. global analysis
  - Verification vs. bug-finding
  - Soundness
  - False positives and false negatives
  - . . .

Type Systems

A Notation for Describing Static Analyses
Type Systems

- Types are the most widely used static analysis
- Part of nearly all mainstream languages
  - Widely recognized as important for quality
- Type notation is useful for discussing concepts
  - We use type notation to discuss type checking, dataflow analysis, and model checking

What is a Type?

- Consensus
  - A set of values
- Examples
  - Int is the set of all integers
  - Float is the set of all floats
  - Bool is the set (true, false)

More Examples

- List[Int] is the set of all lists of integers
  - List is a type constructor
  - A function from types to types
- Foo, in Java, is the set of all objects of class Foo
- Int → Int is the set of functions mapping an integer to an integer
  - E.g., increment, decrement, and many others

What is a Type?

- Consensus
  - A set of values
- A type is an example of an abstract value
  - Represents a set of concrete values
  - Every static analysis has abstract values
- In type systems,
  - Every concrete value is an element of some abstract value
  - I.e., every concrete value has a type

Abstraction

- All static analyses use abstraction
  - Represent sets of values as abstract values
- Why?
  - Because we can’t reason directly about the infinite set of possible concrete values
  - For performance (even just termination), must make such approximations
- In type systems, the approximations are called types

The Next Step

- Now we need to compute with types
  - Actually analyze programs
- Type systems have a well-developed notation for expressing such computations
Rules of Inference

- Inference rules have the form
  *If Hypothesis is true, then Conclusion is true*
- Type checking computes via reasoning
  *If $E_1$ and $E_2$ have certain types, then $E_3$ has a certain type*
- Rules of inference are a compact notation for "If-Then" statements

From English to an Inference Rule

- Start with a simplified system and gradually add features
- Building blocks
  - Symbol $\land$ is "and"
  - Symbol $\Rightarrow$ is "if-then"
  - $x:T$ is "$x$ has type $T$"

From English to an Inference Rule (2)

If $e_1$ has type Int and $e_2$ has type Int, then $e_1 + e_2$ has type Int

$(e_1$ has type Int $\land e_2$ has type Int $) \Rightarrow e_1 + e_2$ has type Int

From English to an Inference Rule (3)

The statement

$(e_1 : \text{Int} \land e_2 : \text{Int}) \Rightarrow e_1 + e_2 : \text{Int}$

is a special case of

Hypothesis$_1 \land \ldots \land$ Hypothesis$_n$ $\Rightarrow$ Conclusion

This is an inference rule.

Notation for Inference Rules

- By tradition inference rules are written
  $\vdash$ Hypothesis$_1 \ldots \vdash$ Hypothesis$_n$ $\vdash$ Conclusion
- Type rules have hypotheses and conclusions
  $\vdash e : T$
- $\vdash$ means "it is provable that . . ."

Two Rules

- $i$ is an integer
  $\vdash i : \text{Int}$ $[\text{Int}]$

- $e_1 : \text{Int}$
  $\vdash e_2 : \text{Int}$
  $\vdash e_1 + e_2 : \text{Int}$ $[\text{Add}]$
Two Rules (Cont.)

- These rules give templates describing how to type integers and + expressions
- By filling in the templates, we can produce complete typings for expressions
- Note that
  - Hypotheses prove facts about subexpressions
  - Conclusions prove facts about the entire expression

Example: 1 + 2

1 is an integer
\[ \vdash 1 : \text{Int} \]
2 is an integer
\[ \vdash 2 : \text{Int} \]
\[ \vdash 1 + 2 : \text{Int} \]

A Problem

- What is the type of a variable reference?

\[ x \text{ is a variable} \]
\[ \vdash x : ? \quad \text{[Var]} \]

- The rule does not carry enough information to give \( x \) a type.

A Solution

- Put more information in the rules!
  - An environment gives types for free variables
    - An environment is a function from variables to types
    - For other static analyses the environment may map variables to other abstract values
    - A variable is free in an expression if it is not defined within the expression

Type Environments

Let \( A \) be a function from Variables to Types

The sentence \( A \vdash e : T \) is read:

Under the assumption that variables have the types given by \( A \), it is provable that the expression \( e \) has the type \( T \)

Modified Rules

The type environment is added to all rules:

\[ A \vdash e_1 : \text{Int} \]
\[ A \vdash e_2 : \text{Int} \quad \text{[Add]} \]
\[ A \vdash e_1 + e_2 : \text{Int} \]
New Rules

And we can write new rules:

\[
\frac{A(x) = T}{A \vdash x : T} \quad [\text{Var}]
\]

Summary

Describe static analyses using logics of the form:

\[
A' \vdash e' : T'

A'' \vdash e'' : T''

A \vdash e : T
\]

The program (or program fragment) to be analyzed.

The abstract value computed for e.

Analysis of expression is recursively defined using analysis of subexpressions.

Assumptions needed for aspects of e that are determined by e’s environment.

The Rule of Signs

- We want to estimate a computation’s sign

- Example: \(-3 \times 4 = -12\)

- Abstraction: \(- \times + = -\)

Abstract Values

- + = \{set of positive integers\}
- 0 = \{0\}
- - = \{set of negative integers\}

- Environment: Variables \rightarrow \{+, 0, -\}

Example Rules

\[
\begin{array}{ll}
A \vdash e_0 : + & A \vdash e_0 : + \\
A \vdash e_1 : - & A \vdash e_1 : + \\
A \vdash e_0 \times e_1 : - & A \vdash e_0 \times e_1 : + \\
\end{array}
\]

\[
\begin{array}{ll}
A \vdash e_0 : 0 & A \vdash e_0 : x \\
A \vdash e_1 : x & A \vdash e_1 : 0 \\
A \vdash e_0 \times e_1 : 0 & A \vdash e_0 \times e_1 : 0
\end{array}
\]
A Problem

\[
\begin{align*}
A & \vdash e_0 : + \\
A & \vdash e_1 : - \\
A & \vdash e_0 + e_1 : ?
\end{align*}
\]

Solution:
Add abstract values to ensure analysis is closed under all operations:

\[
\begin{align*}
+ & \equiv \{ \text{positive integers}\} \\
0 & \equiv \{ 0 \} \\
- & \equiv \{ \text{negative integers}\} \\
\tau & \equiv \{ \text{all integers}\} \\
\bot & \equiv \{
\}
\end{align*}
\]
We don't have an abstract value that covers this case!

Solution:
Add abstract values to ensure analysis is closed under all operations:

\[
\begin{align*}
A & \vdash e_0 : + \\
A & \vdash e_1 : + \\
A & \vdash e_0 + e_1 : + \\
A & \vdash e_0 : 0 \\
A & \vdash e_1 : + \\
A & \vdash e_0 / e_1 : 0 \\
A & \vdash e_0 / e_1 : \bot
\end{align*}
\]

Back To The Main Story . . .

More Example Rules

\[
\begin{align*}
A & \vdash e_0 : + \\
A & \vdash e_1 : - \\
A & \vdash e_0 + e_1 : \tau \\
A & \vdash e_0 : 0 \\
A & \vdash e_1 : + \\
A & \vdash e_0 : 0 \\
A & \vdash e_0 / e_1 : \bot
\end{align*}
\]

A More Complex Rule

\[
\begin{align*}
A & \vdash e_0 : \text{Bool} \\
A & \vdash e_1 : T_1 \\
A & \vdash e_2 : T_2 \\
T_1 & = T_2 \\
A & \vdash \text{if } e_0 \text{ then } e_1 \text{ else } e_2 : T_1 \\
\end{align*}
\]

Comments on Soundness

• Sound analyses are extremely useful
  - If a program has no errors according to a sound system, then no errors of that kind can arise at runtime
  - We have verified the absence of a class of errors

• Verification is a very strong guarantee
  - The property verified always holds

Soundness

\[
\begin{align*}
A & \vdash e_0 : \text{Bool} \\
& \quad e_0 \text{ is guaranteed to be a Bool} \\
A & \vdash e_1 : T_1 \\
& \quad e_1 \text{ and } e_2 \text{ are guaranteed to be of the same type} \\
A & \vdash e_2 : T_2 \\
& \quad T_1 = T_2 \\
A & \vdash \text{if } e_0 \text{ then } e_1 \text{ else } e_2 : T_1 \\
& \quad \text{[If-Then-Else]}
\end{align*}
\]

• An analysis is sound if
  - Whenever \( A \vdash e : T \)
  - And if \( A(x) = T' \) then \( x \) has a value \( v \in T' \)
  - Then \( e \) evaluates to a value \( v \in T \)
Comments on Soundness

- But soundness has a price in many applications
  - Spurious errors/warnings due to abstraction
  - These are false positives
- Alternative is to use unsound analyses
  - Allows some control of false positives
  - Introduces possibility of false negatives
  - Undetected errors

Type systems are sound
- But most analyses used for detecting bugs are not sound
- These are called bug finding analyses

Constraints

\[ A \vdash e_0 : \text{Bool} \]
\[ A \vdash e_1 : T_1 \]
\[ A \vdash e_2 : T_2 \]
\[ T_1 = T_2 \]
Side constraints must be solved
\[ A \vdash \text{if } e_0 \text{ then } e_1 \text{ else } e_2 : T_1 \]

Many analyses have side conditions
- Often constraints to be solved
- All constraints must be satisfied
- A separate algorithmic problem

Algorithm:
1. Input: If-expression and A
2. Analyze \( e_0 \), check it is of type Bool
3. Analyze \( e_1 \) and \( e_2 \) giving types \( T_1 \) and \( T_2 \)
4. Solve \( T_1 = T_2 \)
5. Return \( T_1 \)

Global Analysis

\[ A \vdash e_0 : \text{Bool} \]
\[ A \vdash e_1 : T_1 \]
\[ A \vdash e_2 : T_2 \]
\[ T_1 = T_2 \]
\[ \text{if } e_0 \text{ then } e_1 \text{ else } e_2 \]

The first step requires the overall environment \( A \)
- Only then can we analyze the subexpressions

This is global analysis
- Requires the entire program
- Or we must somehow construct a model of the environment

Local Analysis

Algorithm:
1. Analyze \( e_0 \), inferring environment \( A_1 \)
   check type is Bool
2. Analyze \( e_1 \) and \( e_2 \) giving types \( T_1 \) and \( T_2 \)
   and environments \( A_2 \) and \( A_3 \)
3. Solve \( T_1 = T_2 \) and \( A_1 = A_3 \)
4. Return \( T_1 \) and \( A_1 \)

Another Example

- Consider a recursive function
  \[ f(x) = \ldots f(e) \ldots \]
  - If \( x : T_1 \) and \( e : T_2 \) then \( T_2 \geq T_1 \)
    - Can be relaxed to \( T_2 \geq T_1 \)

  Recursive functions yield recursive constraints
  - Some with loops
    - How hard constraints are to solve depends on constraint language, details of application

  Constraints

\[ A \vdash e_0 : \text{Bool} \]
\[ A \vdash e_1 : T_1 \]
\[ A \vdash e_2 : T_2 \]
\[ T_1 = T_2 \]
\[ \text{if } e_0 \text{ then } e_1 \text{ else } e_2 : T_1 \]

Many analyses have side conditions
- Often constraints to be solved
- All constraints must be satisfied
- A separate algorithmic problem
Local Analysis

In this approach, we first analyze the subexpressions, inferring the needed environment from the subexpression itself. Because the separately computed environments might not agree, they need to be constrained to be equal to get a valid analysis for the entire expression.

Local vs. Global Analysis

• Global analysis
  - Is usually simpler than local analysis
  - But may require a lot of extra engineering to construct models of the environment for partial programs

• Local analysis
  - Allows analysis of, e.g., a library without the client
  - Technically more difficult
    - Requires allowing unknown parameters in environments, which can be solved for later

Flow Insensitivity

Subexpressions are independent of each other

Flow Insensitivity (Again)

Subexpressions are independent of each other

Flow Sensitivity

Rules produce new environments and analysis of a subexpression cannot take place until its environment is available.

Comments on Flow Insensitivity

• Flow insensitive analyses are often very efficient and scalable
• No need for modeling a separate state for each subexpression

Flow Sensitivity

• Analysis of subexpressions is ordered by environments: flow sensitive
• Dataflow analysis is flow sensitive analysis
• The order of statements matters
Comments on Flow Sensitivity

- Example
  - Rule of signs extended with assignment statements
  \[ A \rightarrow e : +, A \]
  \[ A \rightarrow x := e, A[x \rightarrow +] \]
  - \( A[x \rightarrow +] \) means \( A \) modified so that \( A(x) = + \)

- Flow sensitive analysis can be expensive
  - Each statement has its own model of the state
  - Polynomial increase in cost over flow-insensitive

Comments on Path Sensitivity

- Model checking is flow- and path-sensitive analysis
  - In practice, path sensitive analyses are also flow sensitive
    - Although in theory they are independent ideas

- Path sensitivity can be very expensive
  - Exponential number of paths to track
    - But appears to be necessary in many applications

- Often implemented using backtracking instead of explicit merges of environment
  - Explore one path
  - Backtrack to a decision point, explore another path

Path Sensitivity

- Part of the environment is a predicate saying under what conditions this expression is executed.
- At points where control paths merge, still keep different paths separate in the final environment

Summary

- A very rough taxonomy
  - Type systems = flow-insensitive
  - Dataflow analysis = flow-sensitive
  - Model checking = flow- and path-sensitive

- But note
  - These lines have been blurred
  - E.g., lots of flow-sensitive type systems recently