How to compile pattern matching

Jules Jacobs

April 26, 2021

Abstract

This is a note about compiling ML-style pattern matching to decision trees. I try to keep it as simple as possible while still presenting the key ideas of state of the art pattern match compilation.

1 Introduction

Our goal is to compile ML-style pattern matching:

$$
\begin{aligned}
&\textbf{match } a \textbf{ with} \\
&\mid \mathsf{Add}(\mathsf{Zero}, \mathsf{Zero}) && \implies e_1 \\
&\mid \mathsf{Mul}(\mathsf{Zero}, x) && \implies e_2 \\
&\mid \mathsf{Add}(\mathsf{Succ}(x), y) && \implies e_3 \\
&\mid \mathsf{Mul}(x, \mathsf{Zero}) && \implies e_4 \\
&\mid \mathsf{Mul}(\mathsf{Add}(x, y), z) && \implies e_5 \\
&\mid \mathsf{Add}(x, \mathsf{Zero}) && \implies e_6 \\
&\mid x && \implies e_7
\end{aligned}
$$

In order to keep things straight, I'll use green variables a for bound local variables that have a value and red variables x for variables that will be given a value by pattern matching.

The naive way to compile this is to test the patterns from top to bottom, and try to match a against each pattern. This is inefficient: if the outer constructor of a is an Add but the first pattern fails, then the second pattern will test if a is a Mul, even though it is already known to be an Add at this point. Then the third pattern will test once more whether a is an Add, which is redundant. We'd like to compile pattern matching to code that does **no unnecessary tests**.

This does mean that we *must* start by trying to match the first pattern: if the first pattern succeeds then any work trying to match other patterns is unnecessary. So our strategy will be to emulate the naive strategy that tries to match the first pattern completely, but when we fail in the middle of the matching process, we remember what we learnt to compile a specialised version of the remaining pattern matches that takes what we learnt into account.

2 Compiling the example

Let's represent the pattern matching problem as a list of equations to be tested:

$$
\begin{aligned}
&\mid a = \mathsf{Add}(\mathsf{Zero}, \mathsf{Zero}) && \implies e_1 \\
&\mid a = \mathsf{Mul}(\mathsf{Zero}, x) && \implies e_2 \\
&\mid a = \mathsf{Add}(\mathsf{Succ}(x), y) && \implies e_3 \\
&\mid a = \mathsf{Mul}(x, \mathsf{Zero}) && \implies e_4 \\
&\mid a = \mathsf{Mul}(\mathsf{Add}(x, y), z) && \implies e_5 \\
&\mid a = \mathsf{Add}(x, \mathsf{Zero}) && \implies e_6 \\
&\mid a = x && \implies e_7
\end{aligned}
$$

Our goal will be an algorithm that takes as input such a list of equations, and outputs a tree of simple pattern matches that test against a single constructor:

$$\textbf{match } a \textbf{ with}$$
$$\mid \mathsf{C}(a_1, \ldots, a_n) \implies A$$
$$\mid \_ \qquad\qquad \implies B$$

Let's see how to do this for the example. We start working on the first equation by testing a against the Add constructor with the following pattern match:

$$\textbf{match } a \textbf{ with}$$
$$\mid \mathsf{Add}(a_1, a_2) \implies A$$
$$\mid \_ \qquad\qquad \implies B$$

Then we have to solve the following sub problem for A:

$$\mid a_1 = \mathsf{Zero}, a_2 = \mathsf{Zero} \implies e_1$$
$$\mid a_1 = \mathsf{Succ}(x), a_2 = y \implies e_3$$
$$\mid a_1 = x, a_2 = \mathsf{Zero} \implies e_6$$
$$\mid a = x \qquad\qquad \implies e_7$$

Notice how we now have multiple equations per clause. This is the sub problem for B:

$$\mid a = \mathsf{Mul}(\mathsf{Zero}, x) \implies e_2$$
$$\mid a = \mathsf{Mul}(x, \mathsf{Zero}) \implies e_4$$
$$\mid a = \mathsf{Mul}(\mathsf{Add}(x, y), z) \implies e_5$$
$$\mid a = x \qquad\qquad \implies e_7$$

We can simplify these problems by pushing bare equations such as $a_2 = y$ into the right hand sides. Then the sub problem for A becomes:

$$\mid a_1 = \mathsf{Zero}, a_2 = \mathsf{Zero} \implies e_1$$
$$\mid a_1 = \mathsf{Succ}(x) \implies \mathsf{let}\ y = a_2\ \mathsf{in}\ e_3$$
$$\mid a_2 = \mathsf{Zero} \implies \mathsf{let}\ x = a_1\ \mathsf{in}\ e_6$$
$$\mid \qquad\qquad \implies \mathsf{let}\ x = a\ \mathsf{in}\ e_7$$

(actually, we could have done this for the last equation already in the previous step)

We now continue matching by trying to match $a_1 = \mathsf{Zero}$, by generating the pattern match:

$$\textbf{match } a_1 \textbf{ with}$$
$$\mid \mathsf{Zero} \qquad \implies C$$
$$\mid \_ \qquad\quad \implies D$$

And continuing recursively for C and D.

3 The general algorithm

Let's generalise and see what's going on. Given a list of clauses to generate a pattern matching tree for, we use this algorithm:

1. Push bare equations $a = y$ into the right hand sides, so that all the remaining equations test constructors.
2. Select one of the equations $a = \mathsf{C}(\mathscr{P}_1, \ldots, \mathscr{P}_n)$ in the first clause using some heuristic.
3. Generate this pattern match:

$$\textbf{match } a \textbf{ with}$$
$$\mid \mathsf{C}(a_1, \ldots, a_n) \implies A$$
$$\mid \_ \qquad\qquad \implies B$$

4. Create the two sub problems A and B as follows by iterating over all the clauses. One of three cases can happen:
    (a) The clause contains an equation $a = C(\mathcal{P}_1, \ldots, \mathcal{P}_n), \ldots \text{REST} \ldots$.
        Add the expanded clause $a_1 = \mathcal{P}_1, \ldots, a_1 = \mathcal{P}_1, \ldots \text{REST} \ldots$ to A.
    (b) The clause contains an equation $a = D(\mathcal{P}_1, \ldots, \mathcal{P}_n), \ldots \text{REST} \ldots$ where $D \neq C$.
        Add this clause to B unchanged.
    (c) The clause contains no equation for a. Add this clause to both A and B.
5. Recursively generate code for A and B.

The recursion has two base cases:
- The list of clauses is empty: all patterns failed, so generate an error:
  "Error: Non-exhaustive pattern match."
- The first clause is empty (has zero equations): the first clause succeeded to match, so simply return the right hand side e_i.

4   Heuristic

How do we decide which equation to pick from the first clause to branch on? Any equation works, but we'd like to generate a compact pattern matching tree. Whenever we are in case (c) for one of the other clauses, we have to add that clause to both A and B. That can cause code explosion. To avoid it we want to select the equation that causes the least code explosion. We therefore select an equation that is present in the maximum number of other clauses.

5   Discussion

This algorithm is a bit different than the algorithms in the literature. It is based on the following observations and considerations.

Some of the literature spends effort on avoiding exponential code explosion, and opts to generate redundant tests instead [Aug85]. Exponential code explosion doesn't occur in practice [SR00]. Therefore the right approach seems to me to never generate redundant tests, and to try to avoid code duplication using a heuristic.

In fact, the literature shows that for real world code, different pattern matching algorithms generate almost identical code [SR00, Mar08]. The simple approach of (1) always work on matching the first clause first, to avoid unnecessary tests and (2) greedily try to minimize duplication using the heuristic isn't exactly what is in the literature, but it combines some aspects of the best heuristics in [Mar08].

Generating binary tests instead of n-way tests that simultaneously branch on several possible constructors, may be necessary to avoid duplication:

$$\begin{aligned}
&\textbf{match } (a, b) \textbf{ with} \\
&\mid (A, \_) &&\Longrightarrow e_1 \\
&\mid (\_, A) &&\Longrightarrow e_2 \\
&\mid (\_, B) &&\Longrightarrow e_3 \\
&\mid (\_, C) &&\Longrightarrow e_4 \\
&\mid (\_, D) &&\Longrightarrow e_5 \\
&\mid (B, E) &&\Longrightarrow e_6 \\
&\mid (C, F) &&\Longrightarrow e_7
\end{aligned}$$

If we do a 3-way branch on $a = A, B, C$, then we have to duplicate the middle clauses for both B and C. On the other hand, if we simply test if $a = A$, and then test $b = A, B, C, D, E, F$, then we have no duplication.

Executing an n-way branch may be compiled more efficiently using a jump table than a series of 2-way branches. It is not difficult to detect a series of 2-way branches on the same variable, and convert those back into an n-way branch.

Some of the literature uses decision DAGs instead of decision trees. The easiest way to generate a decision DAG is simply to generate a decision tree and then compress it to a DAG using hash

consing [Mar08]. A simpler though less optimal solution is to only do this for the right hand sides, and not for the internal nodes of the decision tree: we could generate a separate basic block for each unique right hand side, and replace the right hand sides by a jump to the corresponding basic block.

To take advantage of types, you can keep track of the remaining possible constructors of each variable. This way you can avoid generating wildcard branches of the decision tree when there are no remaining possible constructors.

6 Code

Scala code that implements the algorithm can be found at
https://julesjacobs.com/notes/patternmatching/pmatch.sc

References

[Aug85] Lennart Augustsson. Compiling pattern matching. In *Proc. of a conference on Functional programming languages and computer architecture*, pages 368–381, Berlin, Heidelberg, January 1985. Springer-Verlag.

[Mar92] Luc Maranget. Compiling lazy pattern matching. In *Proceedings of the 1992 ACM conference on LISP and functional programming*, LFP '92, pages 21–31, New York, NY, USA, January 1992. Association for Computing Machinery.

[Mar08] Luc Maranget. Compiling pattern matching to good decision trees. In *Proceedings of the 2008 ACM SIGPLAN workshop on ML*, ML '08, pages 35–46, New York, NY, USA, September 2008. Association for Computing Machinery.

[Pet92] Mikael Pettersson. A Term Pattern-Match Compiler Inspired by Finite Automata Theory. In *Proceedings of the 4th International Conference on Compiler Construction*, CC '92, pages 258–270, Berlin, Heidelberg, October 1992. Springer-Verlag.

[Ses96] Peter Sestoft. ML pattern match compilation and partial evaluation. In Olivier Danvy, Robert Glück, and Peter Thiemann, editors, *Partial Evaluation*, pages 446–464, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.

[SR00] Kevin Scott and Norman Ramsey. When Do Match-compilation Heuristics Matter? Technical Report, University of Virginia, USA, 2000.