Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs

The divide between informal and formal mathematics

Informal mathematics
- Reasoning with flexibility
- Abundant data
- Flexible reasoning
- Verification in limited circumstances
- Prone to error and false positives

Formal mathematics
- Reasoning with rigour
- Signal in the middle of a proof
- Can potentially verify all mathematical domains
- Limited data
- Fairly rigid reasoning with still not-so-perfect automation

The best of both worlds
Reason informally, and prove formally.
- Humans are extremely good at informal reasoning (though imperfect).
- Language models (Minerva) have also shown impressive informal mathematical reasoning capabilities.

Sketching with few-shot learning
A formal sketch is a sequence of formal conjectures expressing the high-level ideas of the proof. It is well-aligned with the informal proof.

- Codex input:
  - Informal statement 1
  - Informal proof 1
  - Formal statement 1
  - Formal sketch 1
  - Informal statement 2
  - Informal proof 2
  - Formal statement 2
  - Formal sketch 2
  - Informal statement 3
  - Informal proof 3
  - Formal statement 3

- Codex output:
  - Formal sketch 3

Proving open conjectures in the sketches
- To verify the correctness of the formal sketches, we need to close the “gaps” in them.
- We use a symbolic automated theorem proving tool (Sledgehammer + heuristics), but in principle any off-the-shelf prover can be used.

Experimental results
- We experiment on the miniF2F dataset [2], a collection of 488 high-school competition level mathematical problems.
- It is divided into a validation and a test set, but we do not differentiate them in this work.
- We generate 100 informal proofs from each language model and sketch once per proof.

<table>
<thead>
<tr>
<th>Baselines</th>
<th>minIF2F-valid</th>
<th>minIF2F-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sledgehammer</td>
<td>9.9%</td>
<td>10.4%</td>
</tr>
<tr>
<td>Sledgehammer + heuristics</td>
<td>18.6%</td>
<td>20.9%</td>
</tr>
<tr>
<td>Thor (Jiang et al., 2021)</td>
<td>29.3%</td>
<td>29.9%</td>
</tr>
<tr>
<td>Thor + expert iteration (Wu et al., 2022)</td>
<td>37.3%</td>
<td>35.2%</td>
</tr>
</tbody>
</table>

- DSP almost doubles the automated prover’s performance.

Previous SOTA

Best performance on test
Best performance on valid

Let’s talk about
- The further synergy between informal and formal mathematics.
- How to apply AI in maths education?

References