Thor: Wielding Hammers to Integrate Language Models and Automated Theorem Provers

Albert Qiaochu Jiang, Wenda Li, Szymon Tworkowski, Konrad Czechowski, Tomasz Odrzygóźdź, Piotr Miłos, Yuhuai Wu, Mateja Jamnik

UNIVERSITY OF CAMBRIDGE Department of Computer Science and Technology

UNIVERSITAS VARSOVIENSIS

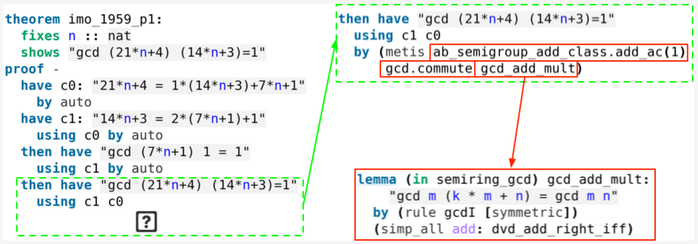
PAN POLISH ACADEMY of SCIENCES

Google Research

NEURAL INFORMATION PROCESSING SYSTEMS

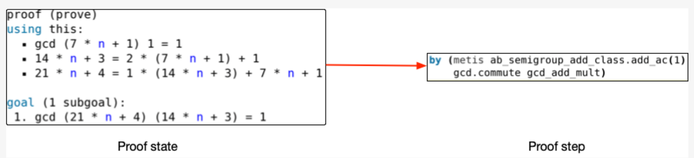
Proof assistant and premise selection

- Proof assistants (e.g, Isabelle) allow mathematical theorems to be proved rigorously.
- We need to select premises (lemmas and definitions) when proving theorems. You don't want to prove every theorem from scratch.



Premise selection is a bottleneck for language models to prove theorems

- Language Models (LMs) predict the next proof step given the proven facts and the remaining objectives. They achieved SoTA in multiple proof assistants [1,2,3].



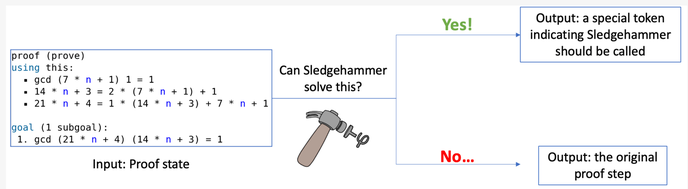
- Premise selection remains a hard retrieval problem for LMs: thousands of lemmas and definitions cannot be crammed into their context windows.
- Empirically, when a LM tries to select premises, 98.1% of the time it produces syntax errors.

Sledgehammer: premise selection based on modern automated theorem provers (ATPs)

- Automated theorem provers can solve SAT, SMT, TPTP, etc. problems that contain millions of variables.
- Proof assistants use Sledgehammer to tackle the premise selection problem by relegating it to ATPs.
- But Sledgehammer is not great with problems involving high-level reasoning, e.g., induction.

Thor: Combining language models and ATPs

- Language models are good at high-level steps (e.g., induction), while ATPs excel at 'low-level' ones (e.g., premise selection).
- Train language model to decide whether to hammer with ATPs.



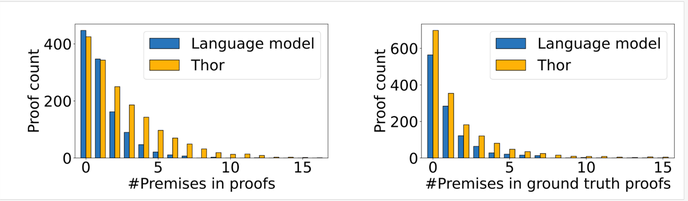
Experiments

- Language model: decoder-only transformer with 700M parameters. Pre-trained on the Github+arXiv subsets of the Pile dataset.
- Sledgehammer: default setup in Isabelle2021.
- Data: Isabelle standard library + the Archive of Formal Proofs (200K theorems, 3M LoC).
- Test set: a suite of 1000 problems from the Archive of Formal Proofs.

Method	Success rate on the test suite (%)
Baselines	
Sledgehammer	25.7
Vanilla language model	39.0
Sledgehammer U Vanilla language model	48.8
Main results	
Thor	57.0

Table 1, Thor solves 8.2% more problems than Sledgehammer and language model naively combined

Is Thor better at premise selection?



Takeaways

- Premise selection is important and difficult. Vanilla pre-trained language models are no good at it.
- Don't throw away the symbolic tools when you have a language model! Integrate them together for better performance.

Let's talk about:

- What's the next step in machine mathematics?
- How could other symbolic tools be integrated with language models?

References

[1] Jiang, Albert Qiaochu, et al. "Language models of Isabelle proofs." *6th Conference on Artificial Intelligence and Theorem Proving*. 2021.

[2] Polu, Stanislas, and Ilya Sutskever. "Generative language modeling for automated theorem proving." *arXiv preprint arXiv:2009.03393* (2020).

[3] Han, Jesse Michael, et al. "Proof Artifact Co-Training for Theorem Proving with Language Models." *International Conference on Learning Representations*. 2021.