Contrastive finetuning of generative language models for informal premise selection

Jesse Michael Han[1,2], Tao Xu[1], Stanislas Polu[1],
Arvind Neelakantan[1], and Alec Radford[1]

[1] OpenAI
[2] University of Pittsburgh

Introduction

Premise selection [6] is a classic problem in automated theorem proving (ATP) which asks how to select the most relevant lemmas useful for proving a given theorem. As such, it is firmly situated in the domain of formal mathematics and has long been a target for machine learning methods in ATP [9, 3, 4, 5, 10, 1]. In this work, we consider *informal* premise selection, where the statements of premises and theorems are in natural language and labels are given by references to premises in ground truth informal proofs. The NaturalProofs dataset, recently introduced in [12], frames informal premise selection as an information retrieval task.

We explore the applications of pretrained generative language models finetuned on a CLIP-style [8] contrastive objective for retrieval over informal mathematics corpora. We show that WebMath pretraining [7] leads to significant performance gain compared to pretraining only on the same data as GPT-3 [2]. We achieve a new state-of-the-art on the NaturalProofs dataset [12], improving on the previous state-of-the-art by up to 80% while using causal rather than bidirectional transformers and fewer parameters overall.

Methodology

We use decoder-only transformers similar to GPT-3 [2] with $n_{\text{layers}} = 12$, $d_{\text{model}} = 768$, $n_{\text{head}} = 12$, and $d_{\text{head}} = 64$, totalling to $125M$ trainable parameters. After pre-training on the autoregressive language modelling task, we adapt our models for embedding-based retrieval as follows. Given a query/document $\mathbf{x}$, we compute an embedding $\widehat{\mathbf{x}} \in \mathbb{R}^{d_{\text{model}}}$ for $\mathbf{x}$ by taking $\widehat{\mathbf{x}}$ to be the activations for the end-of-text (EOT) token. We finetune our models using an InfoNCE loss [11] exactly analogous to the objective used by CLIP [8]. That is, given a batch of N positive (query, document) pairs, we train the encoder to maximize the cosine similarity of the N positive examples while minimizing the cosine similarity of the $N^2 - N$ negative examples. At test time, we retrieve documents for a given query by maximizing the cosine similarity of their embeddings. We test our methodology on the NaturalProofs dataset [12], which comprises (theorem, premise) pairs extracted from proofs of theorems on ProofWiki. We use the same theorem-wise train/test split in this work.

Unlike CLIP [8] or the BERT-based model studied in NaturalProofs [12], we use the same encoder to embed both queries (theorems) and documents (premises). Since "X is useful to prove Y" is an asymmetric relation and we use a CLIP-style symmetric cross-entropy loss, the encoder must be allowed to distinguish between theorems and references. We do this by simply formatting the inputs to the transformer as

```
Theorem title: <title> <newline> Theorem statement: <statement>

Reference title: <title> <newline> Reference statement: <statement>.
```

During contrastive finetuning, we sample batches of $N = 2048$ pairs by first sampling N theorems from the NaturalProofs train split, and then further sampling a positive reference from the proof of each theorem in the batch. All our models are trained for approximately 7000 steps with the Adam optimizer, using 32 V100 GPUs.
We study three pretraining regimes for the NaturalProofs informal premise selection task:

- **No pretraining.** The model is randomly initialized and only learns theorem/premise representations through contrastive training.
- **GPT-3 style pretraining.** The model is pretrained for 300B tokens on the same data (a mix of filtered CommonCrawl, WebText, books, and Wikipedia) as GPT-3 [2].
- **WebMath pretraining.** Starting from the final snapshot of the previous model, we train for another 72B tokens on the WebMath dataset [7], comprising a mix of math arXiv, Python, Math StackExchange, Math Overflow, and PlanetMath.

We refer to our methodology for informal premise selection as contrastive theorem-premise training (CTPT) and denote the three models above by `ctpt-no-pretrain`, `ctpt-webtext`, and `ctpt-webmath`.

Results and discussion

	recall@10	recall@100	avgp@100	full@100	full@1K
BERT	**20.27**	**59.44**	**14.01**	**27.39**	**70.52**
ctpt-no-pretrain	23.76	54.01	11.91	23.75	56.32
ctpt-webtext	34.39	65.45	17.97	34.76	64.51
ctpt-webmath	**36.92**	**70.39**	**21.53**	**39.49**	**73.52**

Table 1: Our models' performance on the NaturalProofs test set alongside results from [12].

Our main results are displayed in Table 1. The model `ctpt-webmath` outperforms the previous state-of-the-art on all metrics. Our models also utilize 43% fewer parameters since the BERT-based model embeds theorems and references with separate copies of `bert-base-cased` (110M params). It is possible that the `webtext` data contains ProofWiki, but WebMath does not and we consider the significant performance gap between `ctpt-webtext` and `ctpt-webmath` to be of primary interest. We speculate that the models studied in [12] are severely undertrained due to using only 200 randomly sampled negatives for each positive example.

Future directions The results discussed in this extended abstract are preliminary, albeit promising. We plan to ablate the effect of including various components of the pretraining (e.g. Python vs informal math in WebMath, the necessity of `webtext`), as well as the zero-shot performance of our models (i.e. no contrastive finetuning) and potential methods for unsupervised retrieval. We consider the applications of our methodology to premise selection in the formal setting (e.g. inside an ITP or ATP) to also be a promising future direction.

Acknowledgements We thank Raul Puri, Harrison Edwards, Yuhuai Wu, Sean Welleck, and Christian Szegedy for helpful discussions.

References

[1] Kshitij Bansal, Sarah M. Loos, Markus N. Rabe, Christian Szegedy, and Stewart Wilcox. Holist: An environment for machine learning of higher order logic theorem proving. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 454–463. PMLR, 2019.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[3] Geoffrey Irving, Christian Szegedy, Alexander A. Alemi, Niklas Eén, François Chollet, and Josef Urban. Deepmath - deep sequence models for premise selection. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2235–2243, 2016.

[4] Cezary Kaliszyk, François Chollet, and Christian Szegedy. Holstep: A machine learning dataset for higher-order logic theorem proving. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[5] Daniel Kühlwein, Jasmin Christian Blanchette, Cezary Kaliszyk, and Josef Urban. Mash: Machine learning for sledgehammer. In Sandrine Blazy, Christine Paulin-Mohring, and David Pichardie, editors, *Interactive Theorem Proving - 4th International Conference, ITP 2013, Rennes, France, July 22-26, 2013. Proceedings*, volume 7998 of *Lecture Notes in Computer Science*, pages 35–50. Springer, 2013.

[6] Daniel Kühlwein, Twan van Laarhoven, Evgeni Tsivtsivadze, Josef Urban, and Tom Heskes. Overview and evaluation of premise selection techniques for large theory mathematics. In Bernhard Gramlich, Dale Miller, and Uli Sattler, editors, *Automated Reasoning - 6th International Joint Conference, IJCAR 2012, Manchester, UK, June 26-29, 2012. Proceedings*, volume 7364 of *Lecture Notes in Computer Science*, pages 378–392. Springer, 2012.

[7] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. *CoRR*, abs/2009.03393, 2020.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[9] Josef Urban. MPTP 0.2: Design, implementation, and initial experiments. *J. Autom. Reason.*, 37(1-2):21–43, 2006.

[10] Josef Urban, Geoff Sutcliffe, Petr Pudlák, and Jirí Vyskocil. Malarea SG1- machine learner for automated reasoning with semantic guidance. In Alessandro Armando, Peter Baumgartner, and Gilles Dowek, editors, *Automated Reasoning, 4th International Joint Conference, IJCAR 2008, Sydney, Australia, August 12-15, 2008, Proceedings*, volume 5195 of *Lecture Notes in Computer Science*, pages 441–456. Springer, 2008.

[11] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[12] Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun

Cho. Naturalproofs: Mathematical theorem proving in natural language. *CoRR*, abs/2104.01112, 2021.