It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners [1]

Maxime BONNIN, Paul CHAUVIN, Colin LENOBLE

May 17, 2023

MVA: Speech and Natural Language Processing

Introduction i

- Pretraining large-scale language models on vast corpora has been a significant breakthrough in NLP [2]; [3]; [4]; [5].
- The project aims to address the limitations of existing approaches like GPT-3.

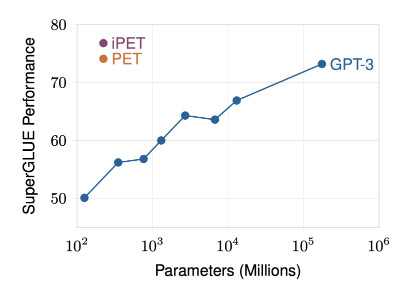

Figure 1: Comparison of GPT3, PET and iPET

Introduction ii

- The proposed method, iPET, predicts multiple tokens for tasks that require such predictions [6].
- iPET outperforms GPT-3 on the SuperGLUE benchmark with only 0.1% of its parameters and a few hours of training on a single GPU [1] thanks to multiple task formulations, robustness to challenging wording, effective utilization of labeled data, and characteristics of the underlying LLM [6]; [1].

Pattern-Exploiting Training

- An MLM is fine-tuned on each pattern-verbalizer pairs, and unlabeled examples are annotated with soft labels based on the ensemble of fine-tuned MLMs. This improves performance.
- The iterative variant of PET, called iPET, is implemented to enhance performance further. Several generations of models are trained on increasing-sized datasets labeled by previous generations, allowing models trained on different patterns to learn from one another over time.
- The proposed method is memory-efficient, as each model's predictions can be computed sequentially. It is expected to demonstrate promising results and be applicable to tasks that require mapping inputs to outputs, especially in scenarios where a large development set is not available.

Experiments - selected tasks

- BoolQ [7] is a QA task where each example consists of a passage p and a yes/no question q.
  p. Question: q? Answer: _.
  p. Based on the previous passage, q? _.
  Based on the following passage, q? _.
- For WiC [8], given a word w and two sentences s_1 and s_2 in which it occurs, the task is to decide if w is used with the same sense in both sentences.
  "s1" / "s2". Similar sense of "w"? _.
  s1 s2 Does w have the same meaning in both sentences? _
  w. Sense (1) (a) "s1" (_) "s2"
- CB [9] is a textual entailment task like MNLI, so we use PVPs similar to [6].
  h? — _ , p , h? — _ , p , h? — _. p , h? — _. p

Experiments - set up

- The Albert-base-v2 model [10] was chosen as the underlying language model for PET due to its good performance on SuperGLUE with standard training sets.
- The final classifier utilized the same Albert-base-v2 model with a sequence classification head added.
- PET was run on FewGLUE training sets for three SuperGLUE tasks without using development sets for hyperparameter optimization, following the setup and hyperparameters from [6].

Experiments - set up

- iPET was trained on the three tasks since their unlabeled sets contained fewer than 1,000 examples.
- The evaluation framework compared few-shot learning approaches, including iPET, PET, and GPT3 [11]. Due to resource limitations, GPT2 [12] was used instead of GPT3, with all layers except the last one frozen and trained on the tasks.
- The performance of the frozen GPT2 model is expected to be lower than that presented in [1] due to the limitations of the method for transformer-based models.

Experiments - Training set up

- All models were trained using the official repository
- Models were trained using NVIDIA RTX 3090 and A5000
- Albert-V2-base: 8 hours per epoch
- GPT2: 12 hours per epoch
- Only 3 epochs for each training

Results - AlbertV2-Base

	AlbertV2-Base (11M)		
	PET	iPET	Sequence-Classifier
BoolQ (acc)	78.2	74.3	75.6
WiC (acc)	51.	-	67.1
CB (acc/F1)	73.2 / 59.5	74.4 / 63.6	66.1 / 55.2

Table 1: Performances (accuracy of F1-score) on BoolQ, WiC and CB tasks using a pretrained Albert-V2-Base (11M) as backbone

	AlbertV2-XXL (223M)		
	PET	iPET	SotA
BoolQ (acc)	79.1	81.2	91.2
WiC (acc)	50.7	49.3	76.9
CB (acc/F1)	87.2 / 60.2	88.8 / 79.9	93.9 / 96.8

Table 2: Paper's performances (accuracy or F1-score) on BoolQ, WiC and CB

Results - GPT2-Base

	GPT2-base (117M)		
	PET	iPET	Sequence-Classifier
CB (acc/F1)	63.5 / 56.3	67. / 58.3	51.2 / 40.9

Table 3: Performances (accuracy or F1-score) on CB tasks using a pretrained GPT2-Base (117M) as backbone

	AlbertV2-XXL (223M)		
	PET	iPET	SotA
BoolQ (acc)	79.1	81.2	91.2
WiC (acc)	50.7	49.3	76.9
CB (acc/F1)	87.2 / 60.2	88.8 / 79.9	93.9 / 96.8

Table 4: Paper's performances (accuracy or F1-score) on BoolQ, WiC and CB tasks wth an Albert-V2-XXL (223)M) as backbone

Discussion - Conclusion

- Pet approach seems to be efficient to reproduce some specific task
- Differences in the complexity of the tasks, which can be seen both in the results and in the difficulties encountered in the project

Timo Schick and Hinrich Schütze.
It's not just size that matters: Small language models are also few-shot learners, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever.
Improving language understanding by generative pre-training.
2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: Pre-training of deep bidirectional transformers for language understanding.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
Roberta: A robustly optimized bert pretraining approach, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu.
Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

Timo Schick and Hinrich Schütze.
Exploiting cloze-questions for few-shot text classification and natural language inference.
In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April 2021. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova.
BoolQ: Exploring the surprising difficulty of natural yes/no questions.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados.
WiC: the word-in-context dataset for evaluating context-sensitive meaning representations.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

References v

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser.
The CommitmentBank: Investigating projection in naturally occurring discourse.
2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut.
ALBERT: A lite BERT for self-supervised learning of language representations.
CoRR, abs/1909.11942, 2019.

- 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.
  2020.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever.
  Language models are unsupervised multitask learners.
  2019.