Reading : "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners"

Amina TADJER, Mohamed Ali BENREKIA, Aymen DJELID, Abdelkader BENAMARA

Paris Dauphine University

March 31, 2022

Few-Shot Learning is an example of meta-learning, where a learner is trained on several related tasks, during the meta-training phase, so that it can generalize well to unseen (but related) tasks with just few examples, during the meta-testing phase. These learners are called pretrained models. PLMs are language models that have been trained with a large dataset while remaining agnostic to the specific tasks they will be employed on. With a large training data set and compute power, they have proven to be very effective and achieve state-of-the-art performance in a wide range of tasks in natural language processing. However, these big models need a huge amount of computational power, and use billions of parameters, which make them difficult to use by researchers, and result with a large carbon footprint. Thus many researchers and data scientists often conduct studies and experiments to solve this challenging problem. "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners" by Schick, Timo and Hinrich Schütze published on ACL 2021, is one such study that aims to propose a novel solution to the problem with keeping the same performance with state of the art model. In this document we are going to analyze this paper [5]. First, we are going to summarize the paper, and present a brief overview and what we understood from it. Second, we are going to discuss the study by presenting its interesting points and impact, then we'll talk about this work's limitations. Finally, we will sum up with a conclusion to give our honest opinion on this paper.

1 Summary and overview

Recent pretrained language models are ultra-large-scale models with billions of parameters that have demonstrated outstanding performance in various NLP tasks. Compared with previous state-of-the-art finetuning methods, they can achieve competitive results without any or with just a limited quantity of training data. Although studies have shown that scaling up the model improves task-agnostic and few-shot performance, these methods such as GPT-3 are not very usable in real scenarios because of the gigantic model requiring powerful computation machines. This research is based on a new concept which is Pattern-Exploiting Training. PET is a semi-supervised training procedure that reformulates input examples as cloze-style phrases to help language models understand a given task. These phrases are then used to assign soft labels to a large set of unlabeled examples. Finally,

standard supervised training is performed on the resulting training set [4]. The problem with this method is that it only works when the answers to be predicted by the LM correspond to a single token in its vocabulary. The authors claim that the added value of their work is to adapt PET for tasks that require predicting multiple tokens, showing that with fewer parameters, and few hours of training on a single GPU, their model outperforms GPT-3.

Before presenting the technique used, the paper provides an explanation of PET and how it works. PET considers the task of mapping inputs to outputs, requiring a pattern P that maps inputs to cloze questions containing a single mask, and a verbalizer v that maps each output to a single token representing its task-specific meaning in the pattern. The idea is to derive the probability of y being the correct output for x from the probability of $v(y)$ being the "correct" token at the masked position in $P(x)$. It also presents iPET, an iterative variant of PET in which several generations of models are trained on datasets of increasing size that are labeled by previous generations. This preliminary definition was useful to understand the limitation of this method which is that the verbalizer v must map each output to a single token, which is impossible for many tasks. So, the authors introduce PET with multiple masks, by generalizing verbalizers to functions mapping outputs to a token sequence, the authors explain the modifications required in the inference and training phases.

The experiment section uses the benchmark SuperGLUE which contains 8 tasks to compare PET and GPT-3. The authors indicate that they are not using the same training dataset as GPT-3, and they create new training sets by randomly selecting 32 examples for each task using a fixed random seed and create sets of up to 20000 unlabeled examples for each task; this is done by removing all labels from the original training sets. Then, they described the eight NLP tasks (ex: yes or no question/answers). For the setup, they choose ALBERT-xxlarge-v2, the best-performing MLM on SuperGLUE when training is performed on the regular, full size training sets. They use the same model, supplemented by a sequence classification head, as their final classifier. They run PET on the FewGLUE training sets for all SuperGLUE tasks. For some tasks, they use the proposed modification of PET to support verbalizers mapping labels to multiple tokens; for all other tasks, they use regular PET. They also train iPET on some tasks (5 with adapting it when possible).

The results section shows that combining ALBERT with PET performs similarly to the largest GPT-3 model, which is larger by a factor of 785. On average, PET performs 18 points better compared to GPT-3 Med, a model of similar size. iPET brings further improvements for 3 out of the 5 tasks that are used with iPET. Despite PET's strong performance, it still clearly performs worse than a state-of-the-art model trained on the regular, full size SuperGLUE training set. The authors investigate the importance of several factors for few-shot performance: the choice of patterns and verbalizers, the usage of both unlabeled and labeled data, and properties of the underlying language model. They also look into the proposed modification for PET to work with multiple masks and compare it to various baselines. Finally, they measure how choosing different sets of training examples affects performance.

2 Impact and interesting points

First of all, we can say that this paper is trying to solve a real problem in NLP, it's trying to show that we can develop small models with fewer parameters. This not only lowers financial cost, but above

all reduces environmental impact immensely and leads to a much smaller carbon footprint. This as an important contribution to achieving the goal of an environmentally more friendly NLP, and focus on reducing the amount of compute required for few-shot learning is closely related to Green AI.

Second, PET converts the text and label in an example into a fluent sentence, and then uses the probability of generating the label text as the class logit, outperforming GPT3 for few shot learning. Unlabeled data is much easier to obtain than labeled examples for many NLP tasks.

Finally, the implementation, the experiments and the results are well explained and each step is detailed. The code, the datasets and the models are publicly available. The paper was published in 2020 and has already 180 citations, because of its high credibility, impact and its good results.

3 Experiments and limitations

In [2], researchers found that for some classification tasks, the labels can be rephrased with simple rules into sentences. A pretrained language model then judges the label sentence that most likely follows the unlabeled input. An unlabeled review, for instance, might be continued with "It was great/bad" for obtaining binary sentiment labels.

Also, studies in [1] show that this technique does not generalize to all kinds of attributes, such as those that are non-textual (e.g., user IDs that are not text-translatable), multi-labeled (e.g., multiple authors of a paper), and with large vocabularies (e.g., thousands of products available). This method has been successfully applied to pretrained language models, the attributes used to control the text are limited to those that are text-translatable (e.g., topics such as "Technology" or tasks that are described in text) and those with limited vocabulary (e.g., "positive" or "negative" sentiment).

3.1 Method limitations

The pretraining tasks used in prompt learning are token-level, requiring the labels to be mapped to a fixed length token span. On the one hand, when the number of labels grows rapidly, this necessitates a lot of human labor. On the other hand, tasks with variable length options make Left-to-right LM (L2R LM) or masked LM (MLM) difficult to cope with. The length of each candidate entity's description, for example, varies significantly in the entity linking task [6]. So designing high-performing prompts is challenging and requires a very large validation set, leading to the cumbersome manual prompt engineering process.

Another limitation of this method is that manually designed label words is obviously limited with the prior knowledge, which may induced omissions and bias for knowledge expansion [7].

3.2 Experiment limitations

Experiments show that GPT-3 is highly sensitive to the order of the in-context examples: that is, permuting the order of the in-context example led to significant differences in performances. Specifically, Authors in [3] found that providing the in-context examples in descending order of similarity, where the least similar example is presented last, led to a decrease in performance compared to the ascending order. This is likely due to GPT-3's recency bias: more attention is paid to the tokens

closest to the end of the prompt. Although, same authors stated that the problem with this technique is that it requires the manual formulation of multiple prompts, which is costly and sometimes not possible.

4 Conclusion

In conclusion, this paper "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners" effectively accomplished the goal of showing that small LM perform well (even better in some tasks) than large models with billions of parameters. Our analysis of this study revealed a real problem to solve (why), a strong new approach and methodology (novelty), a deep explanation of it (how) and properly reported results. This study largely paved the way for new research in the topic, and a lot of work relies on it. Every methodology has its limitations, and this creates new problems to solve by future work. This analysis allows us to learn new concepts such as few-shot learning, pattern-exploiting training. It also allows us to read more about other related work such as GPT-3, and understand more about transfer learning. To sum up, we can say that this study is very beneficial to learn more about NLP and its challenges, and to learn about alternative methods for Text classification seen in class.

References

[1] R. K. Amplayo, K. M. Yoo, and S.-W. Lee. Efficient attribute injection for pretrained language models. *arXiv preprint arXiv:2109.07953*, 2021.

[2] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*, 2020.

[3] N. Pezzotti. Gpt-3 for few-shot dialogue state tracking. 2021.

[4] T. Schick and H. Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, 2021.

[5] T. Schick and H. Schütze. It's not just size that matters: Small language models are also few-shot learners. *ArXiv*, abs/2009.07118, 2021.

[6] Y. Sun, Y. Zheng, C. Hao, and H. Qiu. Nsp-bert: A prompt-based zero-shot learner through an original pre-training task-next sentence prediction. *ArXiv*, abs/2109.03564, 2021.

[7] Y. Zhu, X. Zhou, J. Qiang, Y. Li, Y. Yuan, and X. Wu. Prompt-learning for short text classification. *ArXiv*, abs/2202.11345, 2022.