Language Models are Awesome-Shot Learners

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
Augmented by GPT-Neo

June 22, 2021

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions - something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

The ability to adapt to domain shifts has long been a central challenge in natural language processing, with language modeling now being a widely used

method for achieving that adaptation. However, this kind of adaptation is more difficult in practice than training from scratch due to the need to handle a large number of languages and an ever increasing number of possible domain shifts. Furthermore, it is often argued that a more practical approach to language modeling is to combine with domain adaptation techniques [1, 2]. For instance, in [1], the authors propose a method called GPT-2, which combines language models with the Gated Associative Memory (GAM) [3, 4]. The authors evaluate the quality of the resulting translations and argue that the method is a useful complement to the standard approach of using GPT-2 only on translation sentences.

In this work, we consider a different problem: to train a language model, or more precisely, a probabilistic graphical model. A language model can be used to generate text or to create grammatical rules which are used to generate text from specific grammatical structures, such as articles, pronouns, verbs, etc. Since only a handful of such structures exist in natural language, language models are difficult to train and hard to compete against the best humans in terms of translation quality. However, if we also require that the model's produced output can be used for text generation, we can achieve state-of-the-art performance with a human-specific, few-shot setting. In the proposed algorithm, we leverage many of the benefits of a few-shot setting, such as the ease of obtaining many training examples, and the fact that the task is not restricted to a few examples per domain, that is, we are not limited by the number of possible NLP tasks.

We first explain our approach for training a sparse language model and describe the hyperparameters used for all experiments. Then, we extend our technique to a few-shot language model, training many models on small batches of examples. We evaluate on three different tasks for which we have released datasets, from the translation of articles, to the generation of grammatical rules for the creation of sentences, and to the generation of grammatical structures for question-answering and cloze tasks. For all tasks, we first train a few-shot language model and then use it in the few-shot setting for translation, question-answering, and cloze tasks. We evaluate the proposed setup on the development and test sets of the IWSLT dataset [17], which contains 6,766 articles [18] and 1,726,821 sentences [15], respectively. We also evaluate on the evaluation set of the BERT dataset [19], which contains 6,828 examples [20], and the largest available dataset for cloze tasks, the Cloze-2 dataset [21], which contains 7,928 examples.

Hyperparameter values

We perform experiments across 200 training iterations, each consisting of five epochs with a batch size...