How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Adam Roberts[*]
Google
adarob@google.com

Colin Raffel[*]
Google
craffel@gmail.com

Noam Shazeer
Google
noam@google.com

Abstract

It has recently been observed that neural language models trained on unstructured text can implicitly store and retrieve knowledge using natural language queries. In this short paper, we measure the practical utility of this approach by fine-tuning pre-trained models to answer questions *without access to any external context or knowledge*. We show that this approach scales surprisingly well with model size and outperforms models that explicitly look up knowledge on the open-domain variants of Natural Questions and WebQuestions.

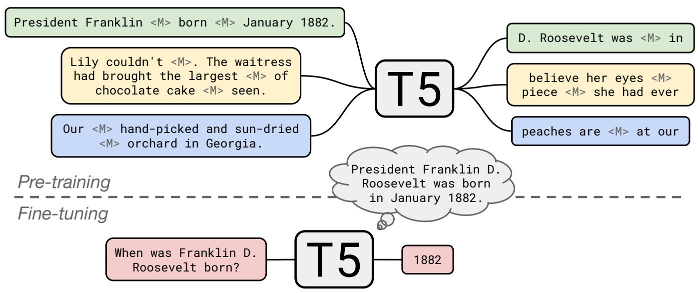

Figure 1: T5 is pre-trained to fill in dropped-out spans of text (denoted by <M>) from documents in a large, unstructured text corpus. We fine-tune T5 to answer questions without inputting any additional information or context. This forces T5 to answer questions based on "knowledge" that it internalized during pre-training.

1 Introduction

Big, deep neural language models[1] that have been pre-trained on unlabeled text have proven to be extremely performant when fine-tuned on downstream Natural Language Processing (NLP) tasks (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Lan et al., 2019; Raffel et al., 2019). Interestingly, it has also recently been observed that these models can internalize a sort of implicit "knowledge base" after pre-training (Petroni et al., 2019; Jiang et al., 2019; Talmor et al., 2019). This behavior is potentially useful because 1) the knowledge base is built by pre-training on unstructured and unlabeled text data, which is freely available in huge quantities on the Internet (Raffel et al., 2019; Wenzek et al., 2019), and 2) it is possible to retrieve information from the knowledge base using natural language queries since these pre-trained language models excel when fine-tuned for natural language understanding tasks.

Past work investigating "language models as knowledge bases" has typically focused on trying to understand the scope of the information stored in the model using synthetic tasks that are similar to the pre-training objective (Petroni et al., 2019; Jiang et al., 2019) or measure reasoning capabilities (Talmor et al., 2019). In this work, we take 评估任务 a different approach by evaluating the capability of language models on the practical task of open-domain question answering – specifically, we fine-tune the model to answer questions *without access to any external knowledge or context*. This requires that the model parse a natural language query and then "look up information" stored in its parameters in order to answer the question. Most past work on question answering either explicitly feeds pertinent information to the model alongside the question (for example, an article that contains the answer (Rajpurkar et al., 2016; Zhang et al., 2018; Khashabi et al., 2018; Clark et al., 2019)) or allows the model to look up informa-

[*] Equal contribution. Noam suggested trying T5 on open-domain QA and coded and ran initial experiments on TriviaQA showing improved performance with model size. Adam wrote the code and ran almost all experiments. Colin set the research scope, wrote the paper, and ran a few experiments.

[1] While "language model" has most often been used to describe a model trained for autoregressive next-step prediction, in this work we use it as a generic term for models trained on text data.

tion in an external knowledge source (Berant et al., 2013; Chen et al., 2017). By giving the model access only to the input question, we can measure how much knowledge it has stored in its parameters while measuring performance on a useful real-world problem.

模型参数量的影响

A separate question we address in this work is whether models with more parameters end up storing more information. It has been shown that transfer learning performance on many downstream tasks tends to improve as the model size and amount of unsupervised pre-training increases (Radford et al., 2019; Liu et al., 2019; Raffel et al., 2019). By measuring knowledge retrieval capabilities on models of various sizes, including models that have an order of magnitude more parameters than considered in past work, we can explore how well this approach scales. Specifically, we leverage the pre-trained "T5" models released by Raffel et al. (2019), the largest of which has 11 billion parameters.

After providing background on question answering and transfer learning for natural language processing in the following section, we present results on three benchmark question-answering datasets in section 3. We conclude with some ideas for future work on using language models as knowledge bases.

2 Background

2.1 Question Answering

Question answering refers to the task of feeding a question to a model as input and training it either to select or output the correct answer. Currently, the most popular variant of this task feeds the model some "context" containing the answer (for example, a paragraph from an encyclopedia article) alongside the question (Rajpurkar et al., 2016; Zhang et al., 2018; Khashabi et al., 2018; Clark et al., 2019). Models can be trained either to indicate the span of the context which contains the answer or output the text of the answer itself. Since this format can be seen as reading some text and answering a question about it, it has also been referred to as "reading comprehension" to distinguish it from other formats of the question-answering task.

A more difficult variant is "open-domain question answering" (Berant et al., 2013; Chen et al., 2017), where the model can be asked arbitrary context-independent questions (e.g. well-known facts or historical details). It is typically assumed that the model can access an external collection of knowledge when answering questions (e.g. a structured knowledge base or unstructured text corpus), but the model is not given any information about where in the collection the answer appears. The reading comprehension task can be considered a simplified version of open-domain question answering where the model is given the oracle context to answer a given question. As an analogy, the open-domain question answering system acts as if it is taking an **open-book** exam where it can find and use information in an external source of knowledge.

In this work, we consider open-domain question answering with the additional constraint that the model is *not* allowed to access any external knowledge whatsoever when answering questions. Instead, the model itself must be pre-trained to store knowledge in its parameters before being fine-tuned to answer questions. In one view, this can be seen as an alternative way to approach open-domain question answering where instead of learning to access the external knowledge the model needs to have "memorized" it in order to answer questions; in another view, this constraint creates a third and potentially more ambitious variant of the question-answering task. A model that answers questions in this way is metaphorically similar to a student taking a **closed-book** exam, where the student must study and memorize all pertinent information before the test.

2.2 Transfer Learning with Language Models

In the past few years, it has become increasingly common to pre-train a language model using an unsupervised objective on a large, unstructured text corpus before fine-tuning it on a downstream task of interest (Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018). The popularity of this form of "transfer learning" is attributable to its empirical success on many NLP tasks (Peters et al., 2018; Devlin et al., 2018; Yang et al., 2019; Lan et al., 2019; Raffel et al., 2019). Loosely speaking, the pre-training step may provide the model with some generally useful awareness of meaning, syntax, and "world knowledge". In question answering in particular, most state-of-the-art systems use some form of transfer learning.

Currently, the most popular model architectures used in transfer learning for NLP are Transformer-based (Vaswani et al., 2017) "encoder-only" models like BERT (Devlin et al., 2018). These models can produce a single prediction for each input token and have been applied to reading comprehension-style question answering by predicting which tokens of the context contain the answer. Encoder-only models are not applicable to closed-book question answering (where models are not provided with any additional context) because it is not possible to choose an answer span from the input. An alternative framework, recently proposed by Raffel et al. (2019), is to treat every NLP task as a text-to-text problem using an encoder-decoder Transformer. When this framework is applied to question-answering, the model is trained to generate the literal text of the answer in a free-form fashion. Despite the potential difficulty of generating rather than identifying the answer, Raffel et al. (2019) demonstrated state-of-the-art results on the SQuAD (Rajpurkar et al., 2016), MultiRC (Khashabi et al., 2018), BoolQ (Clark et al., 2019), and ReCoRD (Zhang et al., 2018) reading comprehension datasets.

The text-to-text framework is directly applicable to closed-book question answering since the model can be trained to generate an answer with or without any additional information in its input. Crucially, fine-tuning a text-to-text model to answer questions without any context requires that the model retrieve information from its parameters that it learned during pre-training. Radford et al. (2019) considered a similar task to evaluate the zero-shot question-answering capabilities of a language model. The concurrent "RELIC" model of Ling et al. (2020) learns representations for an explicitly predefined set of entities and is evaluated on the same closed-book variant of TriviaQA that we consider. Relatedly, Petroni et al. (2019) show that it is possible to manually convert some questions to a fill-in-the-blank format amenable to an encoder-only model (e.g. "Who developed the theory of relativity?" gets mapped to "The theory of relativity was developed by ____").

3 Experiments

3.1 Datasets

We consider the following open-domain question-answering datasets: 开放域QA数据集

Natural Questions (Kwiatkowski et al., 2019) is a dataset of questions from web queries, each of which is accompanied by a Wikipedia article that contains the answer.

WebQuestions (Berant et al., 2013) also comprises questions from web queries that have been matched to corresponding entries in FreeBase (Bollacker et al., 2008).

TriviaQA (Joshi et al., 2017) is a dataset of questions from quiz-league websites. Each question is accompanied by pages from a web search and a Wikipedia search that may contain the answer.

In this work, we only make use of the questions from each dataset – we completely ignore the matching documents supplied for each question.

In terms of evaluation, for WebQuestions and TriviaQA we follow the standard evaluation procedures where each predicted answer is compared to the ground-truth after both are lowercased and stripped of articles, punctuation, and duplicate whitespace (Rajpurkar et al., 2016). Natural Questions is distributed with a much richer set of annotations: Each question can be annotated either as unanswerable (given the oracle context) or with a short or yes/no answer; questions in the validation set can be annotated more than once; and some questions have multiple answers (e.g. "Who are the members of the Beatles?" has four answers). We consider two variants of Natural Questions, both of which omit the "unanswerable" label and the long answers (which are nearly impossible to predict without the oracle context). The first is variant the standard "open-domain" version as used e.g. by (Lee et al., 2019; Min et al., 2019b,a; Asai et al., 2019), where 1) the model is only ever trained to output a single answer; 2) any questions with answers longer than five tokens are ignored; 3) answers are normalized before being compared as in WebQuestions and SQuAD; and 4) a predicted answer is considered correct if it matches any of the answers provided by any of the annotators. The second closely matches the official evaluation procedure used by the Natural Questions leaderboard, where our model is trained to predict all ground-truth answers and is only considered correct if it predicts *all* answers for any one of the annotators. As in the official evaluation, we consider questions with fewer than two non-null annotations unanswerable (given the context), but because we cannot predict unanswerability, we only report the recall score. Further, because our model does not have access to the oracle context,

针对每个数据集，具体的评估方法

we also normalize predicted and ground-truth answers when comparing them.

3.2 Training

T5的预训练

We leverage the pre-trained models provided by Raffel et al. (2019) referred to as the "Text-to-Text Transfer Transformer" (T5). These models were pre-trained on a multitask mixture including an unsupervised fill-in-the-blank task as well as supervised translation, summarization, classification, and reading comprehension tasks. Note that none of the reading comprehension datasets (SQuAD (Rajpurkar et al., 2016), MultiRC (Khashabi et al., 2018), BoolQ (Clark et al., 2019), and ReCoRD (Zhang et al., 2018)) overlap with the question-answering datasets we consider in this paper. In order to measure how closed-book question-answering performance scales with model size, we perform experiments with the Base (220 million parameters), Large (770 million), 3B (3 billion), and 11B (11 billion) variants of T5.

四个规模的预训练模型

微调

For fine-tuning, we follow the procedure used in (Raffel et al., 2019): We use the AdaFactor optimizer (Shazeer and Stern, 2018) with a constant learning rate of 0.001 and a 10% dropout rate. We decode the model's predictions by choosing the most likely token at each timestep. When monitoring performance on the validation split of each dataset we found that performance tended to increase quickly and then plateau. We therefore simply trained each model on 10,000 batches with 196,608 tokens each on the train and validation sets and evaluated the final checkpoint on the test set. Note that Natural Questions does not have a public test set, so standard practice on the open-domain variant is to report performance on the validation set and use a held-out portion of the training set for hyperparameter tuning and model selection.

To map question-answering tasks to the text-to-text format, we simply feed the question with a task-specific prefix into the model as input and train it to predict the literal answer text as output. This was straightforward in all cases except the variant of Natural Questions which includes questions with multiple answers. In that case, we trained the model to output each answer delimited by the text "answer:" (for example, "answer: John Lennon answer: Ringo Starr answer: George Harrison answer: Paul McCartney"). We then split out each answer from the model's predictions as a postprocessing step before evaluating it against the list of answers provided by each annotation.

Table 1: Test set scores for T5 variants and previous results on the open-domain variants of Natural Questions (NQ), WebQuestions (WQ), and TriviaQA (TQA).

	NQ	WQ	TQA
Chen et al. (2017)	–	20.7	–
Lee et al. (2019)	33.3	36.4	47.1
Min et al. (2019a)	28.1	–	50.9
Min et al. (2019b)	31.8	31.6	**55.4**
Asai et al. (2019)	32.6	–	–
Ling et al. (2020)	–	–	35.7
T5-Base	27.0	29.1	29.1
T5-Large	29.8	32.2	35.9
T5-3B	32.1	34.9	43.4
T5-11B	**34.5**	**37.4**	50.1

都显式使用外部知识源

注：在小模型上性能仍不够理想.

3.3 Results

Our results on the open-domain Natural Questions, WebQuestions, and TriviaQA tasks are shown in table 1. Notably, performance on each dataset improves as the model size increase, with T5-11B performing best in every case. Compared to prior results on these datasets, T5-11B achieves state-of-the-art on both Natural Questions and WebQuestions and beats all other methods except Min et al. (2019a) and Min et al. (2019b) on TriviaQA. We note that concurrent work by Guu et al. (2020) outperforms our results on both Natural Questions and WebQuestions (with scores of 40.4 and 40.7 respectively) by training a model to both retrieve relevant documents and answer questions in an end-to-end manner. Importantly, all previous methods except Ling et al. (2020) operate in the "open-book" setting by explicitly retrieving and using information from an external knowledge source. In contrast, we use a fundamentally different approach – our model must internalize knowledge during pre-training because it does not have access to any external information when answering questions.

Having established that our approach is competitive on open-domain question answering, we now evaluate it on the standard (and more difficult) multi-answer variant of Natural Questions. Virtually all models which consider this task are reading comprehension systems that select the correct answer from an oracle context. After fine-tuning, T5-11B achieves a recall of 34.6 on the validation set, which lags behind the state-of-the-art score of 51.9 from (Pan et al., 2019)[2] but outperforms the best baseline published alongside the dataset (which achieved a recall of 33.2 (Kwiatkowski et al., 2019)). This suggests that T5 can perform reasonably well even on questions that have many answers.

4 Conclusion

In this short paper, we have shown that a large language model pre-trained on unstructured text can attain competitive results on open-domain question answering benchmarks without any access to external knowledge. This suggests a fundamentally different approach to designing question-answering systems, whose shortcomings motivate many threads for future work: First, we obtained state-of-the-art results only with the largest model which had 11 billion parameters. This model size can be prohibitively expensive in resource-constrained settings. Second, "open-book" models typically provide some indication of what information they accessed when answering a question that provides a useful form of interpretability. In contrast, our model distributes knowledge in its parameters in an inexplicable way, which precludes this form of interpretability. Finally, the maximum-likelihood objective used to train our model provides no guarantees as to whether a model will learn a fact or not. This makes it difficult to ensure that the model obtains specific knowledge over the course of pre-training. We are excited about combining our approach with ideas like the concurrently proposed Retrieval-Augmented Language Model (Guu et al., 2020), which addresses some of these shortcomings by explicitly training the model to access external knowledge in an end-to-end fashion. We also are interested in measuring and improving performance on more difficult question-answering tasks like DROP (Dua et al., 2019) which require reasoning capabilities.

①.小规模模型上仍然不行.

②.可解释性不行.

Acknowledgments

We thank Kelvin Guu, Kenton Lee, Ming-Wei Chang, Zora Tung, and Ice Pasupat for providing the open-domain question-answering evaluation setup; Roy Frostig and Katherine Lee for comments and suggestions on this manuscript; and Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu for working with us on the research that led to this paper.

[2] Validation set recall scores from (Pan et al., 2019) were reported in private correspondence with the authors.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over Wikipedia graph for question answering. *arXiv preprint arXiv:1911.10470*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Kelvin Guu, Kenton Lee, Zora Tung, Pasupat Panupong, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know? *arXiv preprint arXiv:1911.12543*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning cross-context entity representations from text. *arXiv preprint arXiv:2001.03765*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. A discrete hard EM approach for weakly supervised question answering. *arXiv preprint arXiv:1909.04849*.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Knowledge guided text retrieval and reading for open domain question answering. *arXiv preprint arXiv:1911.03868*.

Lin Pan, Rishav Chakravarti, Anthony Ferritto, Michael Glass, Alfio Gliozzo, Salim Roukos, Radu Florian, and Avirup Sil. 2019. Frustratingly easy natural question answering. *arXiv preprint arXiv:1909.05286*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

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

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzman, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.