Language Models are Unsupervised Multitask Learners

语言模型是无监督多任务学习器

Alec Radford Jeffrey Wu Rewon Child David Luan Dario Amodei Ilya Sutskever

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples.The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linearfashion across tasks. Our largest model, GPT-2,is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

自然语言处理任务，例如问答、机器翻译、阅读理解、摘要提取，通常在特定任务的数据集上进行监督学习。我们证明，当在一个叫做WebText的新数据集上训练时，语言模型开始在没有任何明确的监督的情况下学习这些任务。当把文档和问题作输入，语言模型在CoQA上产生的答案F1值达到了55，达到或超越了4个基线模型中的3个，并且是在没有用到127000+个训练样本的情况下。语言模型的容量对于零样本任务迁移至关重要，而且提升它可以在任务上，以对数线性的方式提升性能。我们最大模型GPT-2是一个15亿参数量的Transformer，在8个测试中，在零样本设置下，有7个达到了最先进的水平，在WebText上表现不佳。模型的样本反应了这些改进，它是包含连贯的文本。这些发现为构建语言处理系统提供了有希望的道路，可以从自然发生的演示中学习。

1、引言

Machine learning systems now excel (in expectation) at tasks they are trained for by using a combination of large datasets, high-capacity models, and supervised learning(Krizhevsky et al., 2012) (Sutskever et al., 2014) (Amodeiet al., 2016). Yet these systems are brittle and sensitive to slight changes in the data distribution (Recht et al., 2018) and task specification (Kirkpatrick et al., 2017). Current systems are better characterized as narrow experts rather than competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

机器学习系统目前通过大量数据集、高容量模型和监督学习(Krizhevsky et al., 2012) (Sutskever et al., 2014) (Amodeiet al., 2016)在他们训练的任务上表现出色。然而，这些系统很脆弱，对数据分布和任务性质很敏感。当前的系统更多地被定性为专才而非通才。我们希望朝着更通用的系统方向发展，这个系统可以运行很多任务，最终无需手动为每个任务创建和打标数据。

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al.,2017), reading comprehension systems (Jia & Liang, 2017),and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the short comings of this approach.

创建机器学习系统的主要方法是收集一些训练数据集（体现目标任务的正确行为），训练一个系统来模拟这些行为，然后在同分布的独立测试集上评估。这在专业领域很有效。但是像字幕模型、阅读理解系统、图像分类这些输入不确定，行为不规则，这些凸显了这个这个方法的缺点。

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

我们怀疑，在单一数据上进行单一任务的训练是当前系统缺乏泛化能力的主要原因。要想用当前的结构使系统变得健壮，有可能需要在一系列的领域和作任务上训练和评估。当前几种基线已被提出用以研究这个，比如GLUE(Wang et al., 2018)和decaNLP(McCann et al., 2018)。

Multitask learning (Caruana, 1997) is a promising frame work for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements (Yogatama et al.,2019) and the two most ambitious efforts to date have trained on a total of 10 and 17 (dataset, objective)pairs respectively (McCann et al., 2018) (Bowman et al.,2018). From a meta-learning perspective, each (dataset,objective) pair is a single training example sampled from the distribution of datasets and objectives. Current ML systems need hundreds to thousands of examples to induce functions which generalize well. This suggests that multitask training many need just as many effective training pairs to realize its promise with current approaches. It will be very difficult to continue to scale the creation of datasets and the design of objectives to the degree that may be required to brute force our way there with current techniques.This motivates exploring additional setups for performing multitask learning.

多任务学习是一种有前途的框架，可以提升泛化能力。然而，NLP领域的多任务训练目前仍属新生领域。最近的工作（Yogatama et al.,2019）有了一定的效果提升，目前还有两个很不错的尝试

(McCann et al., 2018) (Bowman et al.,2018)，分别在10个和17个数据集上进行了训练。从元学习的角度来说，每个数据集都是所有数据集分布抽样而来的。当前的机器学习需要几百到上千的数据集来生成有泛化能力的函数。这表明多任务训练按当前的方法来的话，需要尽可能多得训练数据。要想达到一个需要的水准，如果单独增加数据量和目标函数量会比较困难，需要用当前的方法进行暴力破解。这促进了探索多任务学习的其它方法。

The current best performing systems on language tasks utilize a combination of pre-training and supervised fine tuning. This approach has a long history with a trend towards more flexible forms of transfer. First, word vectors were learned and used as inputs to task-specific architectures (Mikolov et al., 2013) (Collobert et al., 2011), then the contextual representations of recurrent networks were transferred (Dai & Le, 2015) (Peters et al., 2018), and recent work suggests that task-specific architectures are nolonger necessary and transferring many self-attention blocks is sufficient (Radford et al., 2018) (Devlin et al., 2018).

当前语言任务表现最好地是将预训练和监督学习结合起来。这个方法有了很长的历史了，而且朝着更灵活的迁移方向发展。首先，学习到词向量然后送进特定任务的框架 (Mikolov et al., 2013) (Collobert et al., 2011)，让后来，用上下文的rnn网络表示来迁移 (Dai & Le, 2015) (Peters et al., 2018)，最近的工作表明特定任务的网络架构不再是必须的了，迁移自注意力块就已经足够了(Radford et al., 2018) (Devlin et al., 2018)。

These methods still require supervised training in order to perform a task. When only minimal or no supervised data is available, another line of work has demonstrated the promise of language models to perform specific tasks,such as commonsense reasoning (Schwartz et al., 2017) and sentiment analysis (Radford et al., 2017).

这些方法仍然需要监督训练。当仅很少或没有有效监督数据时，另一项工作展示了语言模型在处理特殊任务的前景，比如常识推理（Schwartz et al., 2017）和情感分析（Radford et al., 2017）。

In this paper, we connect these two lines of work and continue the trend of more general methods of transfer. We demonstrate language models can perform down-streamtasks in a zero-shot setting – without any parameter or architecture modification. We demonstrate this approach shows potential by highlighting the ability of language models to perform a wide range of tasks in a zero-shot setting. We achieve promising, competitive, and state of the art results depending on the task.

在本文，我们将这两项工作结合起来，继续研究更泛化能力的迁移方法。我们证明了语言模型可以零样本设置下在下游任务上表现良好，而不用改变任何参数或网络架构。我们通过突出语言模型在零样本设置下在各种任务上的表现，来证明这个方法的潜力。我们在任务上，得到了有前景的、有竞争力的和优秀的结果。

2 方案

At the core of our approach is language modeling. Language modeling is usually framed as unsupervised distribution estimation from a set of examples (x1, x2, ..., xn) each composed of variable length sequences of symbols(s1, s2, ..., sn). Since language has a natural sequential

ordering, it is common to factorize the joint probabilities over symbols as the product of conditional probabilities (Jelinek& Mercer, 1980) (Bengio et al., 2003):

我们方案的核心就是语言模型。语言模型一般表示为从样本（x1, x2,..., xn）中进行无监督的分布估计，其中每个样本是由可变长度的字符（s1, s2, ..., sn）组成。因为语言自然就具有顺序性，因此将联合概率表示为条件概率之积也就很正常不过。

This approach allows for tractable sampling from and estimation of p(x) as well as any conditionals of the form p(sn−k, ..., sn|s1, ..., sn−k−1). In recent years, there have been significant improvements in the expressiveness of models that can compute these conditional probabilities, such as self-attention architectures like the Transformer (Vaswaniet al., 2017).
这个方法可以以任意的条件概率形式p(sn−k, ..., sn|s1, ..., sn−k−1)进行抽样和估计p(x)。近些年，可以计算这些条件概率的模型能力有了重要的提升，例如Transformer（Vaswaniet al., 2017）。
Learning to perform a single task can be expressed in a probabilistic framework as estimating a conditional distribution p(output|input). Since a general system should be able to perform many different tasks, even for the same input, it should condition not only on the input but also on the task to be performed. That is, it should model p(output|input, task). This has been variously formalized in multitask and meta-learning settings. Task conditioning is often implemented at an architectural level, such as the task specific encoders and decoders in (Kaiser et al., 2017)or at an algorithmic level such as the inner and outer loop optimization framework of MAML (Finn et al., 2017). But as exemplified in McCann et al. (2018), language provides a flexible way to specify tasks, inputs, and outputs all as a sequence of symbols. For example, a translation training example can be written as the sequence (translate to french, english text, french text). Like wise, a reading comprehension training example can be written as (answer the question, document,question, answer). McCann et al. (2018) demonstrated it was possible to train a single model, the MQAN to infer and perform many different tasks on examples with this type of format.
学习一个任务可以理解为在概率框架下估计一个条件概率p（output|input）。一个通用系统应该可以适用很多不同的任务，即使对于相同的输入，也应该可以根据任务的不同进行调整。这意味着，模型可以表示为p(output|input, task)。这个模型已经在多任务和元学习中有了多种形式的应用。任务调整通常在架构层面实现，比如特定任务的编码器和解码器（Finn et al., 2017），或者在算法层面，比如内外循环优化框架MAML（Finn et al., 2017）。但是正如McCann等人指出，语言提供了一种灵活的方式来指定任务、输入和输出，所有这些都是符号序列。例如，一个翻译训练样本可以写成序列（translate to french, english text, french text）。类似地，一个阅读理解训练样本可以写成序列（answer the question, document,question, answer）。McCan等人（2018）证明了可以训练一个模型MQAN，然后用这种格式的样本，在不同的任务上进行推理和应用。
Language modeling is also able to, in principle, learn the tasks of McCann et al. (2018) without the need for explicit supervision of which symbols are the outputs to be predicted.

Since the supervised objective is the the same as the unsupervised objective but only evaluated on a subset of the sequence, the global minimum of the unsupervised objective is also the global minimum of the supervised objective. In this slightly toy setting, the concerns with density estimation as a principled training objective discussed in (Sutskeveret al., 2015) are side stepped. The problem instead becomes whether we are able to, in practice, optimize the unsupervised objective to convergence. Preliminary experiments confirmed that sufficiently large language models are able to perform multitask learning in this toy-ish setup but learning is much slower than in explicitly supervised approaches。

语言模型也有可能在理论上学习McCann等人提出的任务，同时不需要显式的学习，比如指定要预测的输出字符。因为无监督学习目标函数和监督学习的目标函数是相同的，仅在序列子集上评估，所以无监督和监督学习的全局最小值是一样的。在这个略不正规的设置下，(Sutskever et al., 2015) 讨论的关于密度估计作为一种有原则的训练目标的问题被绕过了。问题变成了我们能否在实际中优化这个无监督目标到收敛。初步的实验证实，足够大的语言模型是可以在这种玩具化的设置下进行多任务学习的，但就是学习速度与显示的监督学习相比有点慢。

While it is a large step from the well-posed setup described above to the messiness of “language in the wild”, Weston(2016) argues, in the context of dialog, for the need to develop systems capable of learning from natural language directly and demonstrated a proof of concept – learning a QA task without a reward signal by using forward prediction of a teacher’s outputs. While dialog is an attractive approach,we worry it is overly restrictive. The internet contains a vast amount of information that is passively available without the need for interactive communication. Our speculation is that a language model with sufficient capacity will begin to learn to infer and perform the tasks demonstrated in natural language sequences in order to better predict them,regardless of their method of procurement. If a language model is able to do this it will be, in effect, performing unsupervised multitask learning. We test whether this is the case by analyzing the performance of language models in a zero-shot setting on a wide variety of tasks.

Weston(2016)指出，从上述清晰的设置到“野外语言”的混乱，是一个很大的跨越。在对话的背景下，他主张开发能够直接从自然语言中学习的系统，并且展示了一个概念验证——通过使用老师输出的前向预测，学习一个没有奖励信号的QA任务。虽然对话是一种有吸引力的方法，但是我们担心它过于受限。互联网包含大量的信息，可以被动的获取，而不需要主动沟通。我们猜测，具有足够能力的语言模型将开始推断和执行自然语言任务，而不必关注它们的来源。如果一个语言模型可以做到这一点，它将能进行无监督多任务学习。我们通过在一系列任务分析零样本设置下的表现，来测试这是否可行。

2.1 训练数据

Most prior work trained language models on a single domain of text, such as news articles (Jozefowicz et al., 2016),Wikipedia (Merity et al., 2016), or fiction books (Kiroset al., 2015). Our

approach motivates building as large and diverse a dataset as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible.
大多数前人工作都是在单一文本领域训练模型，比如新闻文章 (Jozefowicz et al., 2016)、维基百科 (Merity et al., 2016)，或小说书籍 (Kiroset al., 2015)。我们的方法鼓励收集尽可能大和多的数据集，这样可以收集可能多的领域语言信息。
A promising source of diverse and nearly unlimited text is web scrapes such as Common Crawl. While these archives are many orders of magnitude larger than current language modeling datasets, they have significant data quality issues.Trinh & Le (2018) used Common Crawl in their work on common sense reasoning but noted a large amount of documents “whose content are mostly unintelligible”. We observed similar data issues in our initial experiments with Common Crawl. Trinh & Le (2018)’s best results were achieved using a small subsample of Common Crawl which included only documents most similar to their target dataset,the Winograd Schema Challenge. While this is a pragmatic approach to improve performance on a specific task, we want to avoid making assumptions about the tasks to be performed ahead of time.
一个多样和海量的文本数据是网络爬取，比如Common Crawl。虽然这些数据相较于当前的语言模型数据集大了好几个数量级，但是它们有重大的数据质量问题。Trinh & Le (2018)在他们的常识推理工作中使用了Common Crawl，但是注意到，其实大量的文本其实“内容是无法理解的”。最初，我们使用Common Crawl数据时，也发现了这个数据问题。Trinh & Le (2018)的最好结果是使用Common Crawl的一部分数据，也就是和目标任务Winograd Schema Challenge最相似的数据集。然而，这只是程序化的一种方法来提升任务性能，我们想避免提前对任务做出假设。
The resulting dataset, WebText, contains the text subset of these 45 million links. To extract the text from HTML responses we use a combination of the Dragnet (Peters &Lecocq, 2013) and Newspaper content extractors. All results presented in this paper use a preliminary version of WebText which does not include links created after Dec 2017 and which after de-duplication and some heuristic based cleaning contains slightly over 8 million documents for a total of 40 GB of text. We removed all Wikipedia documents from WebText since it is a common data source for other datasets and could complicate analysis due to overlapping training data with test evaluation tasks.
最终用的数据集WebText，实际上它4500万的链接中的子集。为了提取文本内容，我们结合使用了Dragnet (Peters &Lecocq, 2013) 和Newspaper内容提取器。本文所展示的结果是使用了WebText的早期版本，不包括2017年12月之后的链接，且经过了去重和启发式清洗，最后包括了800万的文档，共40G的文本。我们删掉了维基百科的文档，因为这是其它数据集的常规来源，在做测试评估时会和训练数据冲突，这将会使分析复杂化。
2.2 输入表示
A general language model (LM) should be able to compute the probability of (and also generate) any string. Current large scale LMs include pre-processing steps such as lower-

casing, tokenization, and out-of-vocabulary tokens which restrict the space of model-able strings. While processing Unicode strings as a sequence of UTF-8 bytes elegantly fulfills this requirement as exemplified in work such as Gillicket al. (2015), current byte-level LMs are not competitive with word-level LMs on large scale datasets such as the One Billion Word Benchmark (Al-Rfou et al., 2018). We observed a similar performance gap in our own attempts to train standard byte-level LMs on WebText.
一般的语言模型应该可以计算任何字符串（包括生成的字符串）的概率。当前的大规模语言模型都包含了预处理步骤，比如小写处理、分词处理、词库外字符处理，这些步骤限制了入模字符串的空间。虽然将Unicode字符串作为UTF-8字节序列处理优雅地解决了这个问题，但正如Gillicket al. (2015)在工作所示，在大规模数据集上，如One Billion Word Benchmark (Al-Rfou et al., 2018，当前字节级的语言模型不如词级的语言模型。我们在WebText训练字节级语言模型也证实了这一差异。
Byte Pair Encoding (BPE) (Sennrich et al., 2015) is a practical middle ground between character and word level language modeling which effectively interpolates between word level inputs for frequent symbol sequences and character level inputs for infrequent symbol sequences. Despite its name, reference BPE implementations often operate on Unicode code points and not byte sequences. These implementations would require including the full space of Unicode symbols in order to model all Unicode strings. This would result in a base vocabulary of over 130,000 before any multi-symbol tokens are added. This is prohibitively large compared to the 32,000 to 64,000 token vocabularies often used with BPE. In contrast, a byte-level version of BPE only requires a base vocabulary of size 256. However, directly applying BPE to the byte sequence results in sub-optimal merges due to BPE using a greedy frequency based heuristic for building the token vocabulary. We observed BPE including many versions of common words like dog since they occur in many variations such as dog. dog!dog? . This results in a sub-optimal allocation of limited vocabulary slots and model capacity. To avoid this, we prevent BPE from merging across character categories for any byte sequence. We add an exception for spaces which significantly improves the compression efficiency while adding only minimal fragmentation of words across multiple vocab tokens.
字节对编码（BPE）（Sennrich et al., 2015）是字符和词级语言建模之间的一个实用的折中方案，它有效地在频繁的符号序列的词级输入和不频繁的符号序列的字符级输入之间进行插值。尽管名字这么叫，BPE的实现更多地是在Unicode代码点上操作，而不是字节序列。这些实现需要包含Unicode符号的全部空间，以便对所有Unicode字符串进行建模。这将产生一个超过13万的基础词库（多符号token添加前）。这与通常BPE所使用的32000和64000的词库相比非常大了。作为对比，一个字节级版本的BPE仅需要256大小的词库。然而，直接将BPE用以字节序列将会导致次优合并，因为BPE使用了基于启发式的贪心频率方法来构建tokeny库。我们观察到BPE包含了许常用词的版本，比如dog，有很多变体，比如dog.dog!dog?这个导致有限的词库和模型容量之间的矛盾。为了避免这种情况，我们阻止BPE在任何字节序列中跨字符类别合并。我们为空格添加了一个例外，这显著提高了压缩效率，同时只增加了很少的词语在多个词汇标记之间的分割。

This input representation allows us to combine the empirical benefits of word-level LMs with the generality of byte-level approaches. Since our approach can assign a probability to any Unicode string, this allows us to evaluate our LMs on any dataset regardless of pre-processing, tokenization, or vocab size。

这种输入表示允许我们将词级LM的经验优势与字节级方法的普遍性相结合。由于我们的方法可以为任何Unicode字符串分配一个概率，这使得我们可以在任何数据集上评估我们的LM，而不考虑预处理、分词或词汇量大小。

2.3 模型

We use a Transformer (Vaswani et al., 2017) based architecture for our LMs. The model largely follows the details of the OpenAI GPT model (Radford et al., 2018) with a few modifications. Layer normalization (Ba et al., 2016)was moved to the input of each sub-block, similar to a pre-activation residual network (He et al., 2016) and an additional layer normalization was added after the final self-attention block. A modified initialization which accounts for the accumulation on the residual path with model depth is used. We scale the weights of residual layers at initialization by a factor of 1/√N where N is the number of residual layers. The vocabulary is expanded to 50,257. We also increase the context size from 512 to 1024 tokens and a larger batch size of 512 is used.

我们的语言模型使用了Transformer架构。模型很大程度地沿袭了OpenAI GPT模型的细节，仅有少量改动。层标准化操作移动了每个子块的输入端，这与预激活的残差网络相似，而且新加了一个层标准操作到最后一个自注意力块后面。还有，修改了初始化的方法，这考虑了残差的累积。我们在初始化时，按照1/√N缩放了残差层的权重，其中N是残差层的数量。词库拓展到了50527。我们还增加了输入文本长度，从512扩展到了1024个token，而且用了更大的btch size512。

3 实验

We trained and benchmarked four LMs with approximately log-uniformly spaced sizes. The architectures are summarized in Table 2. The smallest model is equivalent to the original GPT, and the second smallest equivalent to the largest model from BERT (Devlin et al., 2018). Our largest model, which we call GPT-2, has over an order of magnitude more parameters than GPT. The learning rate of each model was manually tuned for the best perplexity on a 5% held-out sample of WebText. All models still underfit WebText and held-out perplexity has as of yet improved given more training time.

我们训练并对四个对数变换后服从均匀分布的语言模型进行了基准测试。结构见表2. 最小的模型和原始GPT等价，次小的和最大的BERT等价 (Devlin et al., 2018)。我们最大的模型称为GPT-2，比GPT超了一个数量级的参数量。每个模型的学习率都是手动调整，以在5%的保留样本上获得最佳的困惑度。所有模型仍然不能很好地拟和WebText，而且保留样本困惑度在给了更多训练时间时，仍然没有改善。

表2 四个模型的网络架构超参

3.1 语言模型

As an initial step towards zero-shot task transfer, we are interested in understanding how WebText LM’s perform at zero-shot domain transfer on the primary task they are trained for – language modeling. Since our model operates on a byte level and does not require lossy pre-processing or tokenization, we can evaluate it on any language model benchmark. Results on language modeling datasets are commonly reported in a quantity which is a scaled or ex-ponentiated version of the average negative log probability per canonical prediction unit - usually a character, a byte, or a word. We evaluate the same quantity by computing the log-probability of a dataset according to a WebText LM and dividing by the number of canonical units. For many of these datasets, WebText LMs would be tested significantly out-of-distribution, having to predict aggressively standardized text, tokenization artifacts such as disconnected punctuation and contractions, shuffled sentences, and even the string which is extremely rare in WebText - occurring only 26 times in 40 billion bytes. We report our main results in Table 3 using invertible de-tokenizers which remove as many of these tokenization / pre-processing artifacts as possible. Since these de-tokenizers are invertible, we can still calculate the log probability of a dataset and they can be thought of as a simple form of domain adaptation. We observe gains of 2.5 to 5 perplexity for GPT-2 with these de-tokenizers.

作为零样本迁移的初始步骤，我们感兴趣的是WebText语言模型在零样本迁移上的表现。因为我们的模型是在字节级操作的，不需要有损失的预处理或分词，因此我们可以在任何语言模型基准测试上进行评估。语言模型的结果一般表示为规范预测单元的负对数概率的平均值的缩放或扩展形式，基本单位包括：字符、字节或单词。我们WebText LM的评估方法是首先对整个数据集计算对数概率，然后除以基本单元的数量，整个计算量与一般的方法是是一致的。对于许多测试数据集，WebText LM将会在显著得超出分布地情况下测试，必须预测一些极度标准化的文本、人为分词（比如断开的标点和缩略词）、打乱的句子、甚至罕见的字符串-40亿字节中只出现了26次。我们在表3报告了主要的测试结果，使用的是可逆分词器，同时尽可能多地去除分词和预处理引起的不规范符号。因为这些分词器是可逆的，我们仍然可以计算数据集的对数概率，他们可以看作领域适应的一种形式。我们观察到使用这些分词器，GPT-2的困惑度由2.5提升到了5。

表3 多个数据集上的零样本迁移结果。这些结果都没有经过训练或微调。PTB和WikiText-2的结果来自(Gong et al., 2018)。CBT的结果来自(Bajgar et al., 2016)。LAMBADA的准确率结果来自(Hoang et al., 2018)。LAMBADA困惑度结果来自(Grave et al., 2016)。其它结果来自 (Dai et al., 2019)。

WebText LMs transfer well across domains and datasets,improving the state of the art on 7 out of the 8 datasets in a zero-shot setting. Large improvements are noticed on small datasets such as Penn Treebank and WikiText-2 which have only 1 to 2 million training tokens. Large improvements are also noticed on datasets created to measure long-term dependencies like LAMBADA (Paperno et al., 2016) and the Children’s Book Test (Hill et al.,

2015). Our model is still significantly worse than prior work on the One BillionWord Benchmark (Chelba et al., 2013). This is likely due to a combination of it being both the largest dataset and having some of the most destructive pre-processing - 1BW’s sentence level shuffling removes all long-range structure.

WebText语言模型在跨领域和数据集上表现很好，在零样本迁移任务上，8个有7个都达到了优秀的结果。在小数据集，比如Penn Treebank和WikiText-2上取得了大的提升，这些数据集只有1到2百万个训练tokens。还有长依赖数据集上也取得了大的进度，比如LAMBADA (Paperno et al., 2016)和the Children’s Book Test (Hill et al., 2015)。我们的模型在One BillionWord Benchmark (Chelba et al., 2013)上效果比之前要差很多。这可能有两个原因，一是它是最大的数据集，二是它进行了破坏性的预处理-句子级的洗牌去掉了长文本结构。

3.2. Children’s Book Test

The Children’s Book Test (CBT) (Hill et al., 2015) was created to examine the performance of LMs on different categories of words: named entities, nouns, verbs, and prepositions. Rather than reporting perplexity as an evaluation metric, CBT reports accuracy on an automatically constructed cloze test where the task is to predict which of 10 possible choices for an omitted word is correct. Following the LM approach introduced in the original paper, we compute the probability of each choice and the rest of the sentence conditioned on this choice according to the LM, and predict the one with the highest probability. As seen in Figure 2 performance steadily improves as model size is increased and closes the majority of the gap to human performance on this test. Data overlap analysis showed one of the CBT test set books, The Jungle Book by Rudyard Kipling, is in WebText, so we report results on the validation set which has no significant overlap. GPT-2 achieves new state of the art results of 93.3% on common nouns and 89.1% on named entities. A de-tokenizer was applied to remove PTB style tokenization artifacts from CBT.

Children’s Book Test (CBT) (Hill et al., 2015) 是用来检验语言模型在不同类词上的表现的：实体识别、名词识别、动词识别、介词识别。CBT使用准确率来评估完形填空，这个完形填空是从10个可能词中选一个。如前文所述，我们计算每个选项的概率，和在此选项条件件，剩余句子的概率，概率最高的即为预测的选项。如图2所示，随着参数增加，模型表现逐步提升，与人类表现的差距逐步变小。数据重叠分析显示，CBT测试的书籍上有一本书叫The Jungle Book，作者 Rudyard Kipling，这本书在WebText上出现过，所以我们在验证集上做了测试，也就没有明显的重叠现象，因为验证集上没有这本书。GPT-2在名词识别和实体识别上达到了最先进的水平，结果分别为93.3%和89.1%。我们使用逆分词器，来去除CBT中PTB风格的分词痕迹。

图2 Children’s Book测试表现，x轴为模型参数量。人类表现出自Bajgar et al. (2016)，不是原始论文中的更低水平。

3.3 LAMBADA

The LAMBADA dataset (Paperno et al., 2016) tests the ability of systems to model long-range dependencies in text. The task is to predict the final word of sentences which require at least 50 tokens of context for a human to successfully predict. GPT-2 improves the state of the art from 99.8 (Grave et al., 2016) to 8.6 perplexity and increases the accuracy of LMs on this test from 19% (Dehghani et al.,2018) to 52.66%. Investigating GPT-2’ s errors showed most predictions are valid continuations of the sentence, but are not valid final words. This suggests that the LM is not using the additional useful constraint that the word must be the final of the sentence. Adding a stop-word filter as an approximation to this further increases accuracy to 63.24%,improving the overall state of the art on this task by 4%. The previous state of the art (Hoang et al., 2018) used a different restricted prediction setting where the outputs of the model were constrained to only words that appeared in the context.For GPT-2, this restriction is harmful rather than helpful since 19% of answers are not in context. We use a version of the dataset without preprocessing.

LAMBADA数据集测试模型的长文本依赖能力。任务是预测句子的最后一个字，这个字人类要想成功预测，都需要前面至少有50个字作参考。GPT-2将困惑度从99.8降到了8.6，准确率从19%提到了52.66%，达到了最先进的水平。调查GPT-2的错误样本发现，许多预测值是句子的有效延续，但不是最后一个字。这表明语言模型并没有使用约束，即预测词必须是句子的最后一个字。添加了一个停止词过滤后作为这个约束的近似，准确率提升到了63.24%，在这个任务上提高了4%的整体水平。早期的最先进水平的方案是使用了一个限制措施，将输出控制在上下文中出现的词。对GPT-2来说，这个限制有害的，因为19%和答案都不在上下文中。还有一点，我们使用的是一个没有预处理过的数据集版本。

3.4. Winograd Schema Challenge

The Winograd Schema challenge (Levesque et al., 2012) was constructed to measure the capability of a system to perform commonsense reasoning by measuring its abilityto resolve ambiguities in text. Recently Trinh & Le (2018)demonstrated significant progress on this challenge using LMs, by predicting the resolution of the ambiguity with higher probability. We follow their problem formulation and visualize the performance of our models with both full and partial scoring techniques in Figure 3. GPT-2 improves state of the art accuracy by 7%, achieving 70.70%. The dataset is quite small with only 273 examples so we recommend reading Trichelair et al. (2018) to help contextualize this result.

Winograd Schema挑战（Levesque等，2012）是为了衡量一个系统进行常识推理的能力而构建的，它通过测量系统解决文本中歧义的能力来实现这一目的。近期，Trinh & Le (2018)使用语言模型在这一挑战上取得了很大的进展，方法是预测歧义的更高概率的答案。我们沿用了他们的问题设定，并可视化了我们的模型结果，有完整评分和部分评分两种形式，在图3中展示。GPT-2将准确率提到了70.70%，提升了7%。数据集非常小，只有273个样本，我们建议阅读一下Trichelair et al. (2018)的原文，以帮助对这一结果的理解。

图3. Winograd Schema挑战的表现，x轴为模型参数量。

3.5. Reading Comprehension

The Conversation Question Answering dataset (CoQA) Reddy et al. (2018) consists of documents from 7 different domains paired with natural language dialogues between a question asker and a question answerer about the document.CoQA tests reading comprehension capabilities and also the ability of models to answer questions that depend on conversation history (such as “Why?”).

对话式问答数据集(CoQA) Reddy et al. (2018)包含了7个不同领域的文档，以及文档相关的自然语言问答对，这是由一个提问者和一个回答者进行的对话。CoQA测试了阅读理解能力，也测试了基于对话历史进行回答问题的能力（比如“为什么”）。

Greedy decoding from GPT-2 when conditioned on a document, the history of the associated conversation, and a final token A: achieves 55 F1 on the development set. This matches or exceeds the performance of 3 out of 4 baseline systems without using the 127,000+ manually collected question answer pairs those baselines were trained on. The supervised SOTA, a BERT based system (Devlin et al.,2018), is nearing the 89 F1 performance of humans. While GPT-2’s performance is exciting for a system without any supervised training, some inspection of its answers and errors suggests GPT-2 often uses simple retrieval based heuristics such as answer with a name from the document in response to a who question.

GPT-2基于文档、历史对话和最后一个token提示，使用贪婪解码方法，在开发集上F1达到了55。这个达到或超越了4个基线系统中的3个，同时没有使用他们训练所使用的12.7万个人工收集的问答对。监督学习的最好的表现是BERT (Devlin et al.,2018)，F1达到了89。虽然，GPT-2的表现是令人激动的，因为它没有进行监督训练，但是它一些答案和错误表明，GPT-2经常使用一些简单的启发式搜索，比如从问题是：谁问的，答案则从相应的文档中找到一个名字。

3.6 Summarization

We test GPT-2’s ability to perform summarization on the CNN and Daily Mail dataset (Nallapati et al., 2016). To induce summarization behavior we add the text TL;DR: after the article and generate 100 tokens with Top-k random sampling (Fan et al., 2018) with k = 2 which reduces repetition and encourages more abstractive summaries than greedy decoding. We use the first 3 generated sentences in these 100 tokens as the summary. While qualitatively the generations resemble summaries, as shown in Table 14, they often focus on recent content from the article or confuse specific details such as how many cars were involved in a crash or whether a logo was on a hat or shirt. On the commonly reported ROUGE 1,2,L metrics the generated summaries only begin to approach the performance of classic neural baselines and just barely outperforms selecting 3 random sentences from the article. GPT-2’s performance drops by 6.4 points on the aggregate metric when the task hint is removed which demonstrates the ability to invoke task specific behavior in a language model with natural language.

我们 CNN and Daily Mail dataset (Nallapati et al., 2016)上测试了GPT-2的文档摘要提取能力。为了进行文本摘要提取，我们在文章后面加了TL;DR（太长不看），然后生成100个tokens，生成方法是随机从前k（k=2）中选一个，这样可以减少重复，同时能生成更抽象的摘要。我们从这100个tokens中选了前面3个句子。虽然，从质量上来说，这些生成的摘要很像摘要，如表14所示，但是它们更多地是关注文档最后的内容，或者是会混淆一些细节，比如车祸中涉几了几辆车，logo是在帽子上还是在衣服上。在常用的ROUGE 1,2,L指标上，生成的摘要只是开始接近于经典的基线模型，仅仅比从文章中随机选择3句话略好一点。去掉提示TL;DR后，GPT-2的表现在平均指标上降了6.4个点，这证明了自然语言中提示符的重要性。

表4 CNN和Daily Mail数据集，使用ROUGE F1指标上的摘要生成表现，Bottom-Up Sum(Gehrmann et al., 2018)是目前的最好水平

3.7 Translation

We test whether GPT-2 has begun to learn how to translate from one language to another. In order to help it infer that this is the desired task, we condition the language model on a context of example pairs of the format english sentence = french sentence and then after a final prompt of english sentence = we sample from the model with greedy decoding and use the first generated sentence as the translation. On the WMT-14 English-Frenchtest set, GPT-2 gets 5 BLEU, which is slightly worse than a word-by-word substitution with a bilingual lexicon inferred in previous work on unsupervised word translation (Conneau et al., 2017b). On the WMT-14 French-Englishtest set, GPT-2 is able to leverage its very strong English language model to perform significantly better, achieving11.5 BLEU. This outperforms several unsupervised machine translation baselines from (Artetxe et al., 2017) and (Lampleet al., 2017) but is still much worse than the 33.5 BLEU of the current best unsupervised machine translation approach(Artetxe et al., 2019). Performance on this task was surprising to us, since we deliberately removed non-English webpages from WebText as a filtering step. In order to confirm this, we ran a byte-level language detector on WebText which detected only 10MB of data in the French language which is approximately 500x smaller than the monolingual French corpus common in prior unsupervised machine translation research.

我们测试了GPT-2的翻译能力。为了帮助它认识这是翻译任务，我们先准备了几个《英文=法文》对，然后加上待翻译的英文名子和=符号，然后使用贪婪解码方法，也就是每次生成概率最大的字，然后使用第一个生成的句子作为翻译后的法文。在WMT-14英法数据集上，GPT-2BLEU分数为5，这比照着字典单字翻译 (Conneau et al., 2017b)的结果还差。在WMT-14法英数据集上，GPT-2可以利用它很强大的英语能力表现的更好，BLEU达到了11.5。这超越了几个无监督翻译基线(Artetxe et al., 2017) and (Lampleet al., 2017)，但是仍然比(Artetxe et al., 2019)的方法差，(Artetxe et al., 2019)是最好的无监督结果，为33.5。这个结果对我们来说很震惊，因为我们特意将非英文的语料删掉了，为了确认这一点，用了一个字节级的语言检测器，发现WebText只有10M的法语数据，这比之前无监督翻译所用到的单语法语语料小了500倍。

3.8. Question Answering

A potential way to test what information is contained with in a language model is to evaluate how often it generates the correct answer to factoid-style questions. Previous showcasing of this behavior in neural systems where all informationis stored in parameters such as A Neural Conversational Model (Vinyals & Le, 2015) reported qualitative results due to the lack of high-quality evaluation datasets. The recently introduced Natural Questions dataset (Kwiatkowski et al., 2019) is a promising resource to test this more quantitatively. Similar to translation, the context of the language model is seeded with example question answer pairs which helps the model infer the short answer style of the dataset.GPT-2 answers 4.1% of questions correctly when evaluated by the exact match metric commonly used on reading comprehension datasets like SQUAD.3 As a comparison point, the smallest model does not exceed the 1.0% accuracy of an incredibly simple baseline which returns the most common answer for each question type (who, what, where,etc...). GPT-2 answers 5.3 times more questions correctly,suggesting that model capacity has been a major factor in the poor performance of neural systems on this kind of task as of yet. The probability GPT-2 assigns to its generated answers is well calibrated and GPT-2 has an accuracy of 63.1% on the 1% of questions it is most confident in. The 30 most confident answers generated by GPT-2 on development set questions are shown in Table 5. The performance of GPT-2 is still much, much, worse than the 30 to 50%range of open domain question answering systems which hybridize information retrieval with extractive document question answering (Alberti et al., 2019).

测试语言模型中包含了什么信息的一种可能方法是，评估事实型问题的正确回答能力。之前有神经系统研究过这一行为，它的信息是储存在参数中，比如神经网络对话模型(Vinyals & Le, 2015)，因为缺乏高质量的评估数据，它只报告了一些定性的结果。最近引入的自然语言问题数据集是一个有希望的的资源，可以更定量地评估。与翻译任务类似，语言模型用问答对来做种子，这可以帮助模型了解到短答案风格。GPT-2在阅读理解数据集SQUAD上，精确匹配的评估标准上，得到了4.1%的淮确率。作为一个比较点，最小的模型没有超越很简单地基线模型的1%，该基线返回每种问题类型（谁，什么，哪里，等等）的答案。GPT-2的正确率高了5.3倍，这表明了模型容量一直是神经网络模型表现不佳的原因。GPT-2分配给它生成的答案的概率是很好校准地，在其最有信息的1%的问题上，有63.1%的准确率。表5展示了30个最小有信息的答案。GPT-2的表现仍然比开放域问答系统的30%-50%的准确率要差好多好多，后者是将信息检索和抽取式文档问答相结合(Alberti et al., 2019)。

4. Generalization vs Memorization

recent work in computer vision has shown that common image datasets contain a non-trivial amount of near-duplicate images. For instance CIFAR-10 has 3.3% overlap between train and test images (Barz & Denzler, 2019). This results in an over-reporting of the generalization performance of machine learning systems. As the size of datasets increases this issue becomes increasingly likely which suggests a similar phenomena could be happening with

WebText. Therefore it is important to analyze how much test data also shows up in the training data.
近期计算机视觉的工作表明，普通图片数据集包含大量的近似重复的图片。比如，CIFAR-10的训练和测试图片有3.3%的重复 (Barz & Denzler, 2019)。这导致了机器学习系统的泛化能力的过高报告。随着数据量的大小提升，这种重复现象在WebText上也极有可能会发生。因此，分析有多少测试数据会出现在训练数据中非常重要。
To study this we created Bloom filters containing 8-grams of WebText training set tokens. To improve recall, strings were normalized to contain only lower-cased alphanumeric words with a single space as a delimiter. The Bloom filters were constructed such that the false positive rate is upper bounded by 1/10-8 . We further verified the low false positive rate by generating 1M strings, of which zero were found by the filter.
为了研究这个问题，我们创建了包含WebText训练集中8-gram的Bloom过滤器。为了提高召回率，字符串被规范化为只包含小写的字母数字单词，以单个空格作为分隔符。Bloom过滤器的构造使得假阳性率上限为1/10-8。我们进一步通过生成1M个字符串来验证低假阳性率，其中没有一个被过滤器发现。（话外音：简单地说，就是作者用一种特殊的数据结构来存储和检索一些文本片段，这种数据结构可以节省空间并且准确地判断一个文本片段是否在训练集中出现过。）
These Bloom filters let us calculate, given a dataset, the percentage of 8-grams from that dataset that are also found in the WebText training set. Table 6 shows this overlap analysis for the test sets of common LM benchmarks. Common LM datasets' test sets have between 1-6% overlap with WebText train, with an average of overlap of 3.2%. Somewhat surprisingly, many datasets have larger overlaps with their own training splits, with an average of 5.9% overlap.
给定一个数据集，这些Bloom过滤器可以计算8-grams在WebText训练集中的比例。表6展示了常见语言模型基线测试集的重叠分析。常见的LM测试集和WebText训练集的重合度在1-6%之间，平均为3.2%。有些惊讶的是，许多数据测试集跟它们的训练集竟有更高的重合率，平均5.9%。

表6 测试集与训练集的8-grams重合率

our approach optimizes for recall, and while manual inspection of the overlaps shows many common phrases, there are many longer matches that are due to duplicated data. This is not unique to WebText. For instance, we discovered that the test set of WikiText-103 has an article which is also in the training dataset. Since there are only 60 articles in the test set there is at least an overlap of 1.6%. Potentially moreworryingly, 1BW has an overlap of nearly 13.2% with itsown training set according to our procedure.
我们的方法优化了召回率，虽然手动检查显示有很多重复是由于短语相同，但还是有很多更长的匹配是因为重复造成的。这并非WebText所独有的问题。例如，我们发现WikiText-103测试集上有一篇文章是在训练集上的。因为至少有60篇文章在测试集，因此至少有1.6%的重合。更让人担忧的是，1BW的训练集和测试集有13.2%的重复。

For the Winograd Schema Challenge, we found only 10 schemata which had any 8-gram overlaps with the WebTexttraining set. Of these, 2 were spurious matches. Of the remaining 8, only 1 schema appeared in any contexts that gave away the answer。
对Winograd Schema Challenge来说，我们发现也就只是和WebText匹配上了10个8-gram。其中2个是无关的。剩下的8个中，只有1个是可以WebText看到答案的。
For CoQA, about 15% of documents in the news domain are already in WebText and the model performs about 3 F1 better on these. CoQA’s development set metric reports the average performance over 5 different domains and we measure a gain of about 0.5-1.0 F1 due to overlap across the various domains. However, no actual training questions or answers are in WebText since CoQA was released after the cutoff date for links in WebText.
对于CoQA，大约15%的新闻领域的文档存大于WebText中，这部分重合部分，模型有3个F1值的超越。CoQA的开发集报告了5个不同领域的表现，我们模型在重合部分也有0.5到1个F1值提升。然后实际的问答和答案是没有重复的，因为CoQA发布的时候，是在WebText数据集之后的。
On LAMBADA, the average overlap is 1.2%. GPT-2 performs about 2 perplexity better on examples with greater than 15% overlap. Recalculating metrics when excluding all examples with any overlap shifts results from 8.6 to 8.7 perplexity and reduces accuracy from 63.2% to 62.9%. This very small change in overall results is likely due to only in 200 examples having significant overlap.
在LAMBADA数据集上，它与WebText之间的平均重叠率为1.2%。GPT-2模型在15%的重合率上降低了2个困惑度。把重合的样本去掉重新评估，困惑度从8.6升到了8.7，准确率从63.2%降到了62.9%。变化这么小可能是因为只有200个样本有重大重叠。
Overall, our analysis suggests that data overlap between WebText training data and specific evaluation datasets provides a small but consistent benefit to reported results. However, for most datasets we do not notice significantly larger overlaps than those already existing between standard train ing and test sets, as Table 6 highlights.
总之，我们分析显示，WebText训练集和特殊任务的评估集的重复，可以使评估效果有一点提升。同时，对于大多数据集来说，我们并没有发现有比表6说得有更大的重复。
Understanding and quantifying how highly similar text impacts performance is an important research question. Better de-duplication techniques such as scalable fuzzy matching could also help better answer these questions. For now, we recommend the use of n-gram overlap based de-duplicationas an important verification step and sanity check during thec reation of training and test splits for new NLP datasets.
理解和量化文本相似对性能的影响是一个重要的研究课题。好的去重技术例如可扩展的模糊匹配可以更好地应对这个问题。但是现在，我们推荐n-gram重叠检验技术，在训练和测试集上去重。
Another potential way of determining whether the performance of WebText LMs is attributable to memorization is inspecting their performance on their own held-out set. As shown in Figure 4, performance on both the training and test sets of WebText are similar and

improve together as model size is increased. This suggests even GPT-2 is still underfitting on WebText in many ways.
还有一个不错的方法可以检查WebText语言模型的泛化能力，就是看它们在测试集上的表现。如图4所示，训练和测试集上的表现相近，而且随着模型变大，效果一起变好。这表明即使像GPT-2这样大的模型依然在WebText表现了出色地泛化能力。

图 4. WebText上模型训练和测试表现在不同模型大小上的表现

GPT-2 is also able to write news articles about the discovery of talking unicorns. An example is provided in Table 13.
GPT-2也可以写关于发现会说话的独角兽的新闻文章。表13是一个例子。

表13 GPT-2在超出分布的情境下生成的内容。在k=40的情况下的10个较好的样本。

5 相关工作

A significant portion of this work measured the performance of larger language models trained on larger datasets. This is similar to the work of Jozefowicz et al. (2016) which scaled RNN based language models on the 1 Billion Word Benchmark. Bajgar et al. (2016) also previously improved results on the Children’s Book Test by creating a much largert raining dataset out of Project Gutenberg to supplement the standard training dataset. Hestness et al. (2017) conducted a thorough analysis of how the performance of various deep learning models changes as a function of both model capacity and dataset size. Our experiments, while much noisier across tasks, suggest similar trends hold for sub-tasks of anobjective and continue into the 1B+ parameter regime.
我们工作的主要部分就是在更大的数据集、更大的语言模型上评估了模型表现。前人出在这方面做过研究，比如 Jozefowicz et al. (2016)在1 Billion Word Benchmark上扩大了语言模型的结构。Bajgar et al. (2016) 也通过创建了一个更大的数据集（来自Gutenberg）来补充标准数据集，这样提升了Children’s Book的测试结果。Hestness et al. (2017)对模型数据和容量如何影响模型表现进行了全面的分析。

Interesting learned functionality in generative models has been documented before such as the cells in an RNN language model performing line-width tracking an dquote/comment detection Karpathy et al. (2015). More inspirational to our work was the observation of Liu et al.(2018) that a model trained to generate Wikipedia articles also learned to translate names between languages。
生成模型的相关文献也记录了一些有趣的能力。比如RNN模型在生成单元时可以自动调整行宽和识别符号Karpathy et al. (2015)。更有启发地是 Liu et al.(2018) 训练了的用于生成Wikipedia文章的模型可以在不同语言之间翻译名字。

Previous work has explored alternative approaches to filtering and constructing a large text corpus of web pages, such as the iWeb Corpus (Davies, 2018).
前人也有探索过其它语料库，如iWeb语料(Davies, 2018)。

There has been extensive work on pre-training methods for language tasks. In addition to those mentioned in the introduction, GloVe (Pennington et al., 2014) scaled word vector representation learning to all of Common Crawl. An influential early work on deep representation learning for text was Skip-thought Vectors (Kiros et al., 2015). McCannet al. (2017) explored the use of representations derived from machine translation models and Howard & Ruder (2018) improved the RNN based fine-tuning approaches of (Dai& Le, 2015). (Conneau et al., 2017a) studied the transfer performance of representations learned by natural language inference models and (Subramanian et al., 2018) explored large-scale multitask training.

关于语言模型的预训练有大量的相关工作。除了引言中提到的这些， GloVe (Pennington et al., 2014) 在Common Crawl研究了词向量生成方法。一个早期的影响力的深度学习文本表示的方法是Skip-thought Vectors (Kiros et al., 2015)。McCannet al. (2017)探索了从机器翻译而来的表示方法。 Howard & Ruder (2018)提升了基于RNN模型的微调方法。(Conneau et al., 2017a)研究了通过自然语言推理而来的表示进行迁移的方法。(Subramanian et al., 2018) 探索了大规模的多任务训练。

Ramachandran et al., 2016) demonstrated that seq2seq models benefit from being initialized with pre-trained language models as encoders and decoders. More recent work has shown that LM pre-training is helpful when fine-tuned for difficult generation tasks like chit-chat dialog and dialog based question answering systems as well (Wolf et al., 2019)(Dinan et al., 2018).

Ramachandran et al., 2016) 证实了将预训练模型来初始化encoders和decoders可以为seq2seq模型带来好处。更多地工作也表明预训练语言模型对于较困难的生成任务也是有帮助的，比如闲聊对话和基于对话的问答系统(Wolf et al., 2019)(Dinan et al., 2018)。

6 讨论

Much research has been dedicated to learning (Hill et al.,2016), understanding (Levy & Goldberg, 2014), and critically evaluating (Wieting & Kiela, 2019) the representations of both supervised and unsupervised pre-training methods.Our results suggest that unsupervised task learning is an additional promising area of research to explore. These findings potentially help explain the widespread success of pre-training techniques for down-stream NLP tasks as we show that, in the limit, one of these pre-training techniques begins to learn to perform tasks directly without the needfor supervised adaption or modification.

许多研究致力于学习、理解和评估有监督和无监督预训练方法的向量表示。我们的方法表明无监督学习是一个有潜力的研究领域。这些发现帮助解释了预训练模型迁移在NLP广泛的下游任务上的成功。极限情况下，预训练任务可以直接在下游任务使用，而不需要微调。

On reading comprehension the performance of GPT-2 is competitive with supervised baselines in a zero-shot setting.However, on other tasks such as summarization, while it is qualitatively performing the task, its performance is still only rudimentary according to

quantitative metrics. While suggestive as a research result, in terms of practical applications, the zero-shot performance of GPT-2 is still far from use-able.
在阅读理解上，GPT-2在零样本上的表现与基模型相当。然而，在其它任务比如文文摘要，虽然从质量上来说，可以完成任务，但是从量化指标上来看，还是很初级的。从结果上看，虽说是有借鉴意义，但是，从实践来说，零样本表现离实用还有很长的路要走。
We have studied the zero-shot performance of WebText LMs on many canonical NLP tasks, but there are many additional tasks that could be evaluated. There are undoubtedly many practical tasks where the performance of GPT-2 is still no better than random. Even on common tasks that we evaluated on, such as question answering and translation,language models only begin to outperform trivial baselines when they have sufficient capacity.
我们研究了WebText语言模型在许多经典的NLP任务上的表现，当然也有许多任务没有评估。毫无疑问的是，许多实际的任务在GPT-2的表现还不如随机猜测。即使在普通的任务上，比如问答、翻译，语言模型也只有在足够的容量上才比一般的基线好一点点。
While zero-shot performance establishes a baseline of the potential performance of GPT-2 on many tasks, it is notclear where the ceiling is with finetuning. On some tasks,GPT-2’ s fully abstractive output is a significant departure from the extractive pointer network (Vinyals et al., 2015)based outputs which are currently state of the art on many question answering and reading comprehension datasets.Given the prior success of fine-tuning GPT, we plan to investigate fine-tuning on benchmarks such as decaNLP and GLUE, especially since it is unclear whether the additional training data and capacity of GPT-2 is sufficient to overcome the inefficiencies of uni-directional representations demonstrated by BERT (Devlin et al., 2018).
虽然，零样本表现奠定了GPT-2在许多任务上的基线，但是，微调的上限尚不得而知。在许多任务上，GPT-2完全抽象的输出和基于指针的网络不一样。鉴于前期的GPT在微调上的成功，我们计划在decaNLP和GLUE上进行微调，目前，还不清楚GPT-2增加了数据容量后是否可以克服BERT论文中所提到的单向表示的不足。
7 总结
When a large language model is trained on a sufficiently large and diverse dataset it is able to perform well across many domains and datasets. GPT-2 zero-shots to state of the art performance on 7 out of 8 tested language modeling datasets. The diversity of tasks the model is able to perform in a zero-shot setting suggests that high-capacity models trained to maximize the likelihood of a sufficiently varied text corpus begin to learn how to perform a surprising amount of tasks without the need for explicit supervision.
当一个很大的语言模型在足够大且多样的数据上训练后，它有可能在许多领域和数据上表现良好。GPT-2在8个零样本测试上有7个表现良好。模型在零样本设置下在很多任务上的表现，证明了高容量模型进行训练，在足够多得语料上最大化似然函数进行学习，在许多任务上得到了惊艳的效果，同时还不需要显示地监督。

致谢

Thanks to everyone who wrote the text, shared the links,and upvoted the content in WebText. Many millions ofpeople were involved in creating the data that GPT-2 wastrained on. Also thanks to all the Googlers who helped uswith training infrastructure, including Zak Stone, JS Riehl,Jonathan Hseu, Russell Power, Youlong Cheng, NoamShazeer, Solomon Boulos, Michael Banfield, Aman Gupta,Daniel Sohn, and many more. Finally thanks to the peoplewho gave feedback on drafts of the paper: Jacob Steinhardt,Sam Bowman, Geoffrey Irving, and Madison May.

感谢所有参与撰写文本、分享链接和点赞内容的WebText的人。有数百万人参与了创建GPT-2训练所用的数据。也感谢所有帮助我们提供训练基础设施的谷歌员工，包括Zak Stone, JS Riehl, Jonathan Hseu, Russell Power, Youlong Cheng, Noam Shazeer, Solomon Boulos, Michael Banfield, Aman Gupta, Daniel Sohn等等。最后感谢给我们提供论文草稿反馈的人：Jacob Steinhardt, Sam Bowman, Geoffrey Irving和Madison May。