「GPT2」Language Models are Unsupervised Multitask Learners

GPT2 的文章看似写了但又似乎没写，文章中对于训练的细节并没有详细介绍，更多的是向我们展示使用了 GPT2 对于不同 task 的效果。首先作者表明了这次的工作，相对于前一次的无监督+fine-tune，作者这次会在更大的数据集上仅使用无监督方式进行训练，并且会在不进行任何微调的情况下测试多种 task 的表现。作者实验发现模型的大小也关系着性能的优良，于是 GPT2 相比于原先，参数量达到了1.5b。这次主要优化点：

1、只使用无监督方式进行训练

2、数据集相比原来更多，达到了40GB

3、模型侧将 layer-norm 放到了每个 transformer block 的前面，并在最后一个注意力之后添加一个额外的 layer-norm，以及初始化将每个层的权重按照1/√N进行缩放，N为当前层数

4、字符扩展到了 50257 个，max_seq_len 也从之前的512扩展到了1024

作者在 token 编码这块仍然采用 BPE 形式进行字节级编码，但是作者也说明了效果不如 One Billion Word Benchmark 这种更大单词级的，更大的 vocab 意味着更多的参数以及训练难度的加大，因此采用字节级尽可能的在缩小 vocab 的情况下能够组合出更多的文本也是一种无奈的妥协。

随后作者花了大量篇幅介绍了不同参数即不同模型大小下，在不进行 fine-tune 情况下生成模型在不同任务上的表现，首先展示了一张多任务效果图

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

然后再单独展示了各个任务上的效果，第一个 ChildBook Test

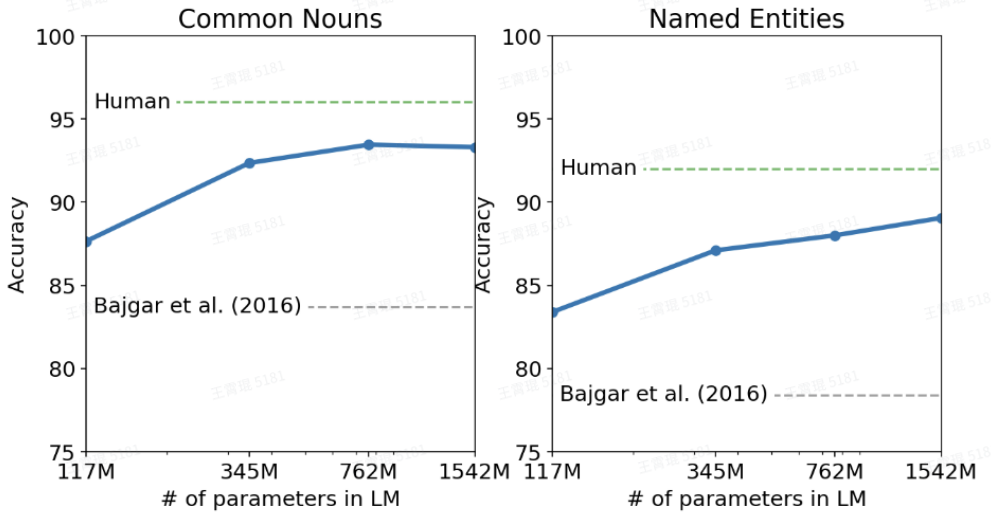

Figure 2. Performance on the Children's Book Test as a function of model capacity. Human performance are from Bajgar et al. (2016), instead of the much lower estimates from the original paper.

第二个 LAMBADA没有图，随后是 Winograd Schema Challenge

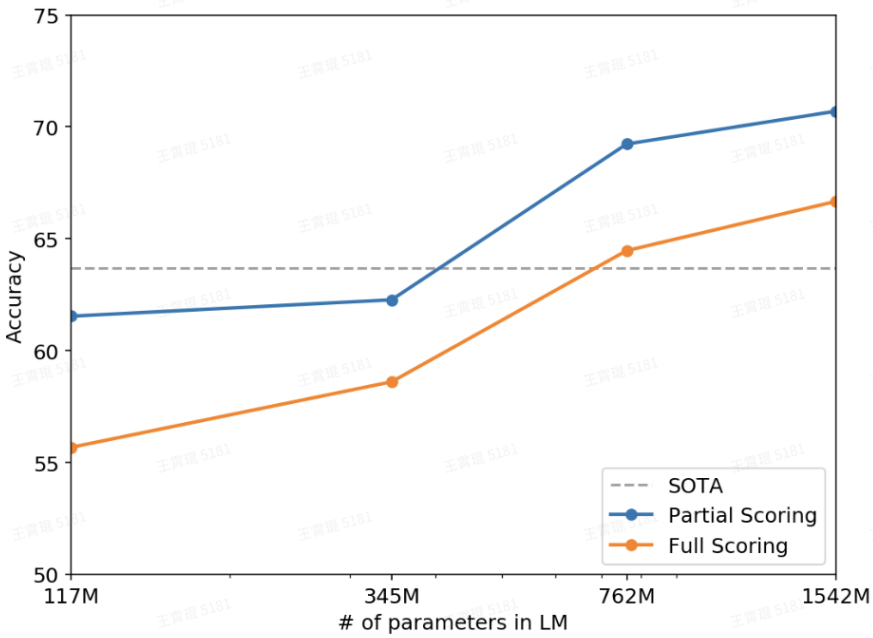

Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

然后是阅读理解能力

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 TL;DR:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

之后介绍了摘要任务，以及翻译任务等，就不展开说明了，总之就是8个任务中有七个任务效果非常好。之后作者介绍了数据集越大，重复数据的概率就越大，但占比不是很高，不过还是需要好好的清洗数据集，最后作者只是告诉我们模型越大效果越好，于是开启大模型之路

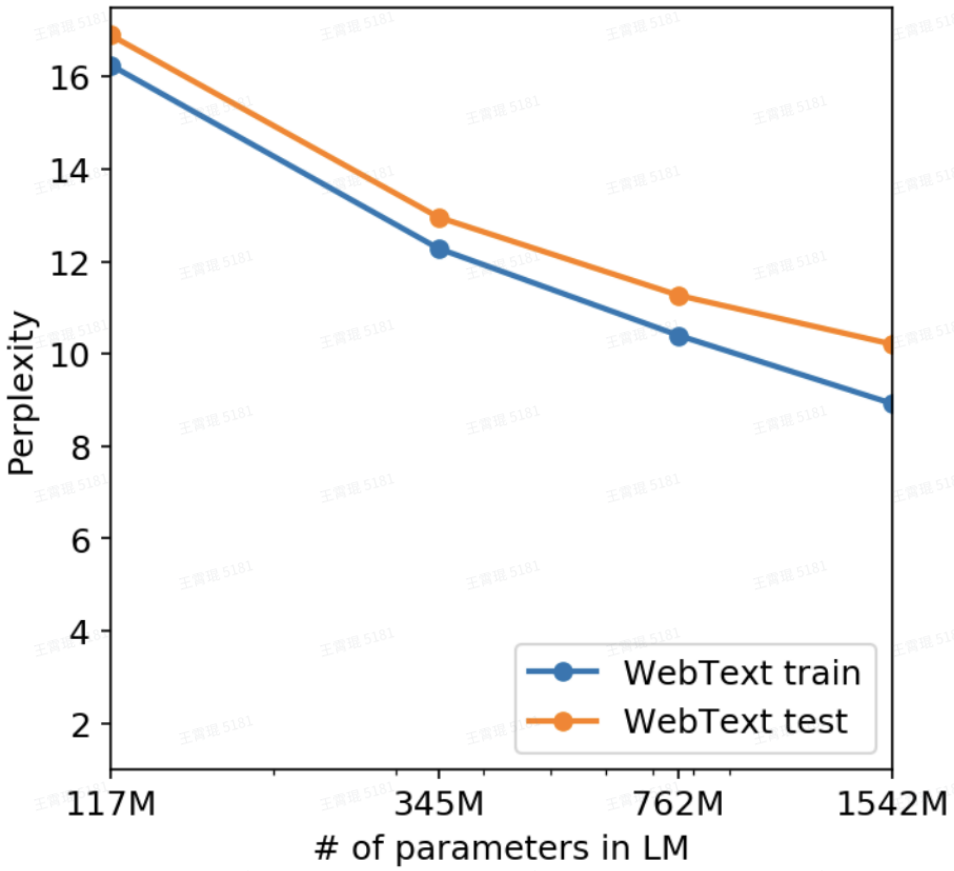

Figure 4. The performance of LMs trained on WebText as a function of model size.

然后后面的附录大致都是挑选出了不同任务的一些优质 case展示，体现出 GPT2 模型的好，没什么可看的，总之GPT2 这篇文章给我感觉是很水，说了一堆东西但又不告诉你细节。