# Title：

Language Models are Unsupervised Multitask Learners

# Abstract：

做了一个新的数据集叫WebText，有百万级别的文本。有了更大的数据集就能训练更大的模型，故作者训练了一个1.5B参数的模型。相比BERT优势不大，但是zero-shot能力很好。

# Introduction：

现在主流的方法是对一个任务收集一个数据集，然后在上面训练模型做预测，为什么这件事情很流行，因为【PROBLEM】现在模型的泛化性不是很好。

多任务学习：在训练的时候设置多个损失函数，在多个数据集上进行训练，为了让这个模型在多个任务上都能够使用，但是现在用的不多。

在本篇文章中，还是在做GPT模型，但是在做下游任务的时候使用了一个叫zero-shot的设定（在做下游任务的时候不需要任何下游任务的标注信息），得到了有一定竞争力的结果。

# Approach

之前的GPT-1，在做下游任务的时候会有线性层，开始符，间隔符，结束符等等，但是现在要做zero-shot就不能这么做了。**输入的形式应该更像一个自然语言。（prompt）**

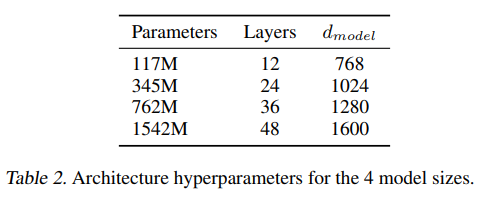
## Training Dataset

一个公开的项目Common Crawl.收集网页：一群人写的爬虫公开抓取网页放在aws的s3上供免费下载，但是信噪比比较低，清理需要花很多时间。

如何用大家已经过滤好的网页呢？

用reddit上karma>3的网页爬下来，得到45M的链接，然后再把文字信息抽取出来得到了一个数据集。大概8M的文本，40GB的文字。

## Model



作者设计了四个模型

第一个模型有12层，宽度为768，可学习参数为117M

# Experiments

