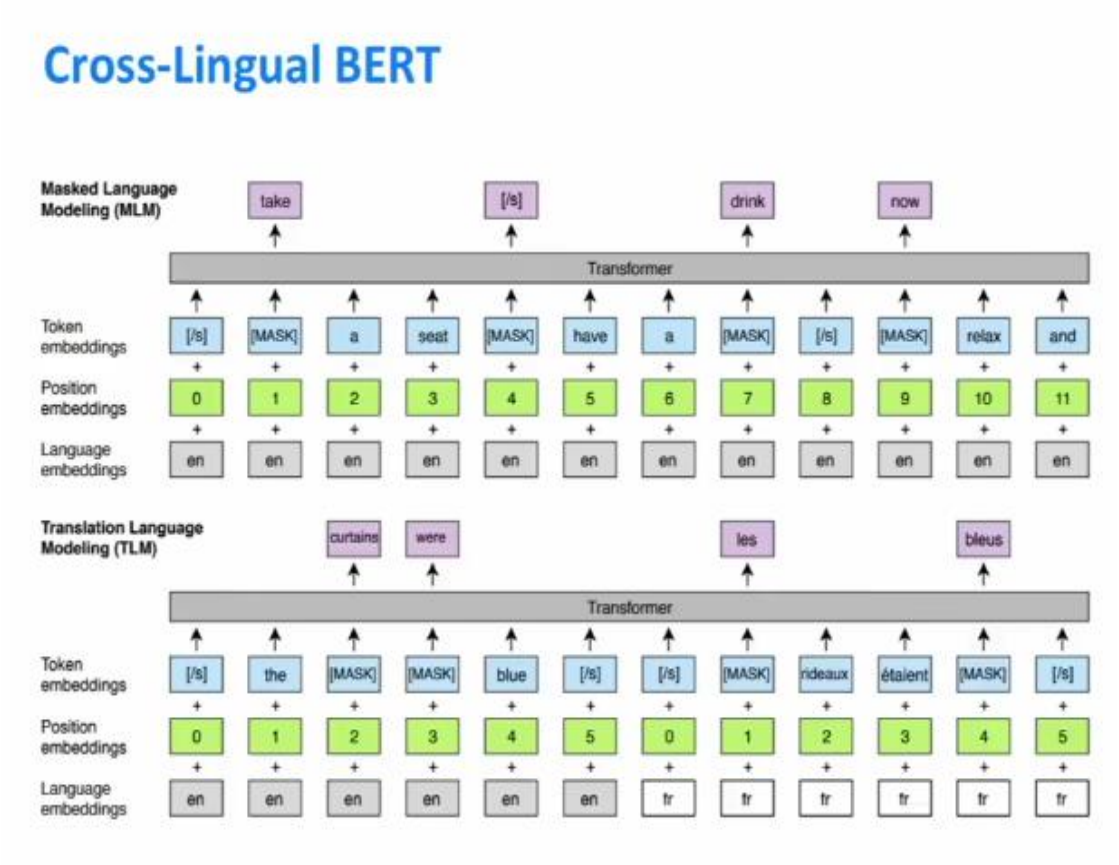
一 BERT



这是常规 BERT，使用英语一些句子（其中某些单词被删除），要求使用 BERT 模型来填补空白并预测这些单词。

谷歌实际上已经完成了多语言 BERT 训练。采用的是连接一大堆不同语言的语料库，然后训练一个模型在所有语言上使用模型。最近，由 Facebook 提出的这种新的扩展，实际上是将 LM 培训目标与翻译相结合。

在这种情况下，给这个模型一个英文序列和一个法语序列，删除一些单词，要求模型填写它，更好地使模型理解两种语言之间的关系。

Cross-Lingual BERT

Unsupervised MT Results

Model	En-Fr	En-De	En-Ro
UNMT	25.1	17.2	21.2
UNMT + Pre-Training	33.4	26.4	**33.3**
Current supervised State-of-the-art	**45.6**	**34.2**	29.9

因此像 BERT 用于 NLP 中的其他任务一样，基本上都采用这种跨语言 BERT，将其用作无监督机器翻译系统的初始化，并获得了大约 10 个 BLEU 点的增益，这样就可以实现无人监督的机器翻译。

二 Huge Models and GPT-2

Training Huge Models

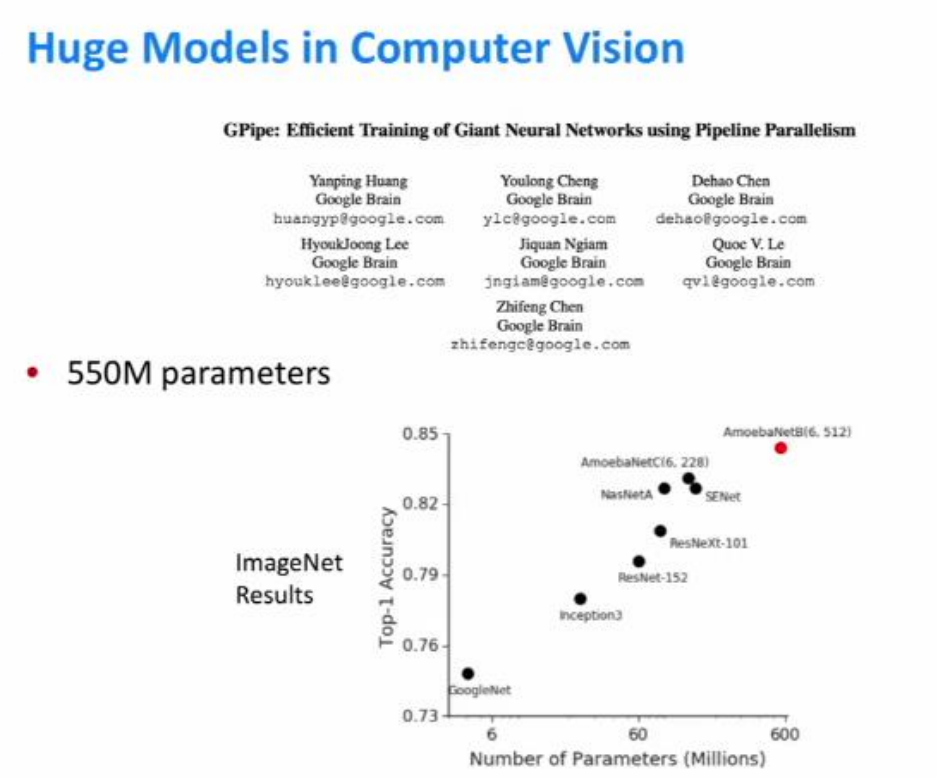
Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B
Honey Bee Brain	~1B synapses

首先，这是一些不同大小的 NLP 模型，也许几年前标准的 LSTM 中型模型大约有 1000 万个参数。在 OpenAI 论文之前，这个 GPT-2 大

约是它的 10 倍，大约相当于另一个数量级。当然，、神经网络中的突触和权重是完全不同的。

该图显示了 x 轴是时间，y 轴是用日志来衡量用于训练该模型的 petaFLOPS 的数量。这意味着至少目前的趋势是机器学习模型的计算能力呈现出指数增长。

这结果来自一个视觉的生成性对抗网络，它已经在很多数据上进行了训练，并且已经在大规模上进行了训练，这是 ELMo 和 BERT 之间的大型模型。如果你感兴趣的是 https://thispersondoesnotexist.com/。



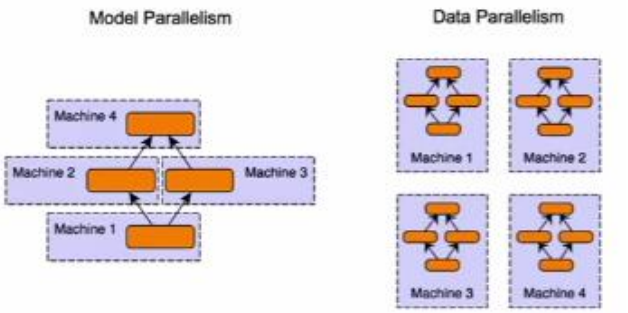
这是谷歌最近的工作，他们训练了一个有 5 亿个参数的图像网络模型。这里的图表显示 x 轴上的日志缩放参数数量，然后 ImageNet 在 y 轴上的准确性，这种大型模型表现得更好。并且似乎成为一种趋势，其精度随着模型尺寸的对数而增加。

Training Huge Models

- Better hardware
- Data and Model parallelism

Mesh-TensorFlow:
Deep Learning for Supercomputers

Noam Shazeer, Youlong Cheng, Niki Parmar,
Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee
Mingsheng Hong, Cliff Young, Ryan Sepassi, Blake Hechtman
Google Brain
{noam, ylc, nikip, trandustin, avaswani, penporn, phawkins,
hyouklee, hongm, cliffy, rsepassi, blakehechtman}@google.com



硬件在很大程度上扩展模型和训练模型。特别是，越来越多的公司正在开发深度学习的硬件。实际上另一种扩展模型的方法是利用并行性。

一种是数据并行性。在这种情况下，GPU 将拥有该模型的数据副本，将正在训练的数据分成小批量到这些模型中，这样就可以更快地训练模型。

另一种并行性是模型并行性。在这种情况下，您实际上需要将模型拆分为多个计算单元。

GPT-2

- Just a really big Transformer LM
- Trained on 40GB of text
  - Quite a bit of effort going into making sure the dataset is good quality
  - Take webpages from reddit links with high karma

So What Can GPT-2 Do?

- Obviously, language modeling (but very well)!
- Gets state-of-the-art perplexities on datasets it is not even trained on!

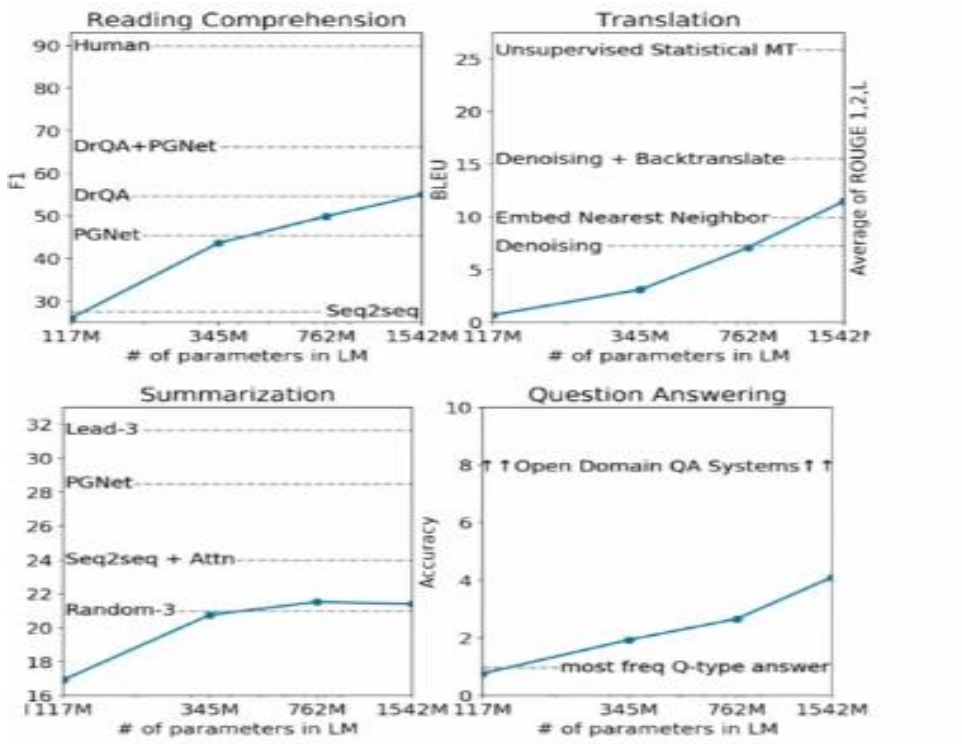
	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**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

拥有像 GPT-2 这样超级庞大的语言模型，可以用它进行语言建模。并在基准测试上运行这种语言模型。如果想评估语言模型，首先在 Penn Treebank 上训练，然后评估这个组合。在这种情况下，GPT-2 只是因为看过这么多的文字并且是如此大的模型，优于其他的先前，即使它没有那些数据也能在不同的语言基准中测试。

So What Can GPT-2 Do?

- **Zero-Shot Learning**: no supervised training data!
  - Ask LM to generate from a prompt
- **Reading Comprehension:** <context> <question> A:
- **Summarization:** <article> TL;DR:
- **Translation:**
  <English sentence1> = <French sentence1>
  <English sentence 2> = <French sentence 2>
  .....
  <Source sentence> =
- **Question Answering:** <question> A:

零射击学习只是尝试完成一项任务，而不需要对它进行训练。通过设计一个提示需要输入的语言模型，然后让它从那里生成，希望它生成与你想要解决的任务相关的语言。



x 轴是对数缩放的模型大小，y 轴是精确度，虚线基本上对应于这些任务的现有工作。

GPT-2 Question Answering

- Simple baseline: 1% accuracy
- GPT-2: ~4% accuracy

- Cherry-picked most confident results

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%

通常在 NLP 的历史中，如果想将一种世界知识带入一个 NLP 系统，你需要一个类似于事实的大数据库，它仍然可以通过阅读大量文本而无需明确地获取一些世界知识将这些知识付诸于模型。