3.1 **Unsupervised pre-training**

3.2

Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \dots, u_n\}$, we use a standard language modeling objective to maximize the following likelihood: (1)

$$L_1(\mathcal{U}) = \sum_i \log P(u_i|u_{i-k}, \dots, u_{i-1}; \Theta) \tag{1}$$
 where k is the size of the context window, and the conditional probability P is modeled using a neural

network with parameters Θ . These parameters are trained using stochastic gradient descent [51]. In our experiments, we use a multi-layer *Transformer decoder* [34] for the language model, which is

a variant of the transformer [62]. This model applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens: $h_0 = UW_e + W_p$ $h_l = \texttt{transformer_block}(h_{l-1}) orall i \in [1, n]$ (2)

 $P(u) = \mathtt{softmax}(h_n W_e^T)$

where $U = (u_{-k}, \dots, u_{-1})$ is the context vector of tokens, n is the number of layers, W_e is the token embedding matrix, and W_p is the position embedding matrix. **Supervised fine-tuning**

After training the model with the objective in Eq. [1], we adapt the parameters to the supervised target task. We assume a labeled dataset \mathcal{C} , where each instance consists of a sequence of input tokens, x^1, \ldots, x^m , along with a label y. The inputs are passed through our pre-trained model to obtain the final transformer block's activation h_l^m , which is then fed into an added linear output layer with parameters W_y to predict y:

 $P(y|x^1,\ldots,x^m) = \operatorname{softmax}(h_l^m W_y).$

$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1,\dots,x^m).$ (4)

(3)

We additionally found that including language modeling as an auxiliary objective to the fine-tuning helped learning by (a) improving generalization of the supervised model, and (b) accelerating

Layer Norm

Feed Forward

This gives us the following objective to maximize:

convergence. This is in line with prior work [50, 43], who also observed improved performance with such an auxiliary objective. Specifically, we optimize the following objective (with weight
$$\lambda$$
):
$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C}) \tag{5}$$

Transformer

+)→ Linear

Text Task Classification Start Text Extract Transformer Prediction Entailment Start Premise Hypothesis Extract ➤ Transformer

Text 1

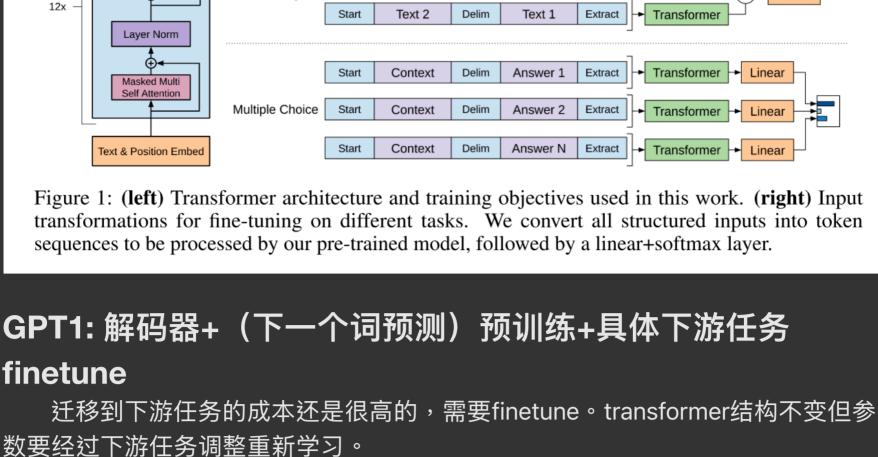
Start

Similarity

Delim

Text 2

Extract



GPT2: Language Models are Unsupervised Multitask

Learners Large dataset + zero shot (无需重新针对下游任务训练,直接用:zeroshot)

1. 去掉下游任务没见过的分隔符之类的特殊符号。引入prompt方法:

• 阅读理解任务prompt: (answer the question, document, question, answer)

• 翻译任务prompt: (translate to france, English text, French text)

2. 模型初始改变(scale the weights of residual layers at initial- ization by a factor of $1/\sqrt{N}$ where N is the number of residual layers) 3. Add sparse transformer 的改进

采用Reddit,kamma(3个以上的问答数据集)。4500w个link,800w文 本,40GB文字。

4. pre normalization, reversible tokenization

没用Common Crawl 信噪比很低。

Using prompt

GPT2改进之处:

5. 训练数据集构建:

https://github.com/milmor/GPT

• 4个大小不同的模型

GPT3: Language Models are Few-Shot Learners

。 Few shot learning每个sample提供10-100个例子。

task description

task description

task description examples

example

prompt

shot setting. GPT-3 is applied without any gradient updates or fine-tuning. GPT-

2. in-context learning

з. Evaluation of GPT-3

One shot learning

GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-

distinguishing from articles written by humans.

性能提升: 1. weights 训练好之后不会再更新了

3 can generate samples of news articles which human evaluators have difficulty

Zero shot learning The three settings we explore for in-context learning Traditional fine-tuning (not used for GPT-3) Zero-shot Fine-tuning

The model is trained via repeated gradient updates using a

example #1

example #2

example #N

prompt

large corpus of example tasks.

cheese =>

sea otter => loutre de mer

peppermint => menthe poivrée

One-shot

Translate English to French:

Translate English to French:

The model predicts the answer given only a natural language

description of the task. No gradient updates are performed.

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

In addition to the task description, the model sees a few

Translate English to French:

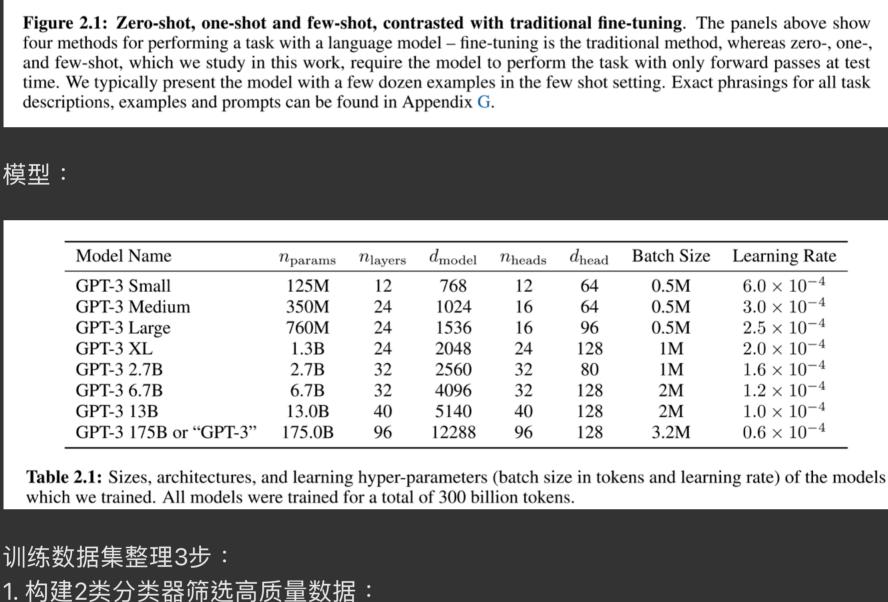
sea otter => loutre de mer peppermint => menthe poivrée plush girafe => girafe peluche

sea otter => loutre de mer cheese =>

cheese =>

cheese =>

plush giraffe => girafe peluche prompt Few-shot



• 正类:GPT2训练数据。负类:Common Crawl 数据。

2. 用LSH方法去重文章(去掉相似度很高的文章)

Statistically-Trained Natural Language Processing System (STNLP) developed by Terry Winograd at MIT. STNLP was a language model that could

> Nvidia introduces first Graphics Processing Unit (GPU), the

generate text from statistical rules.

NVIDIA GPU

Geforce 256.

Facebook AI Research (FAIR) is created, focused on advancing the field of AI through cutting-edge research.

> Google releases the Google Brain project, a deep learning artificial intelligence research project.

OpenAl Founded

general Al.

Introduction of Transformer Models

Tensor Flow Framework

and train their own AI models.

Transformer Models are introduced through papers like Googles Transformer: A Novel Neural Network Architecture for Language Understanding

and Attention Is All You Need, Vaswani et al., 2017.

November: Google releases its TensorFlow framework, which allows developers to build

December: OpenAI founded to pursue the develoment of

FAIR Facebook AI Research

Google Brain

此后的数据构成如下:

3. 加入已有高质量文本数据(BERT, GPT2等训练数据)

• 分类器判断偏正类的common crawl网页被留作GPT3训练样本。

Quantity Weight in Epochs elapsed when **Dataset** (tokens) training mix training for 300B tokens Common Crawl (filtered) 410 billion 60% 0.44 2.9 WebText2 19 billion 22% 8% 1.9 Books1 12 billion 55 billion Books2 8% 0.43 3 billion Wikipedia 3% 3.4

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets

HAT SUGGET TO YOU)

STRONGLY ABOUT DISCUSSING SUCH THENGS)
) (YOU CON'T SEEM GUITE CERTAIN)
E MORE POSITIVE)
E MORE POSITIVE) NO. ARE YOU 03 100 YOU BELIEVE YOU ARE 43 NAME TO BE 43 (YOU WISH I WOULD TELL YOU YOU ARE AN IF YOU WERE 4)) SAY "AR") (I DON'T UNDERSTAND THAT!))

1966

1997

1999

2000

2006

2013

2014

2015

2016

2018

ELIZA ELIZA developed by

with a human.

Joseph Weizenbaum at MIT to simulate limited conversations

IBM Model 1

internaly.

Stanford SQuAD

Stanford University's NLP Group releases the Stanford Question Answering Dataset (SQuAD), a dataset for NLP research.

Universal Sentence Encoder

pairs, is introduced by Google.

December: The Universal Sentence Encoder, a pre-trained model for encoding sentence

IBM releases IBM Model 1, the first version of its statistical machine translation system.

IBM Tangora

20,000 spoken words.

Google Tensor Processing Units Google begins using Tensor Processing Units (TPUs)

Elmo

fine-tuned GPT-3 model for web browsers that can answer open-ended questions with

citations and links to sources.

Bing + OpenAl

Al21 lak

Prometheus Model.

Microsoft Introduces GPT-4

March: Satya Nadella and Microsoft debut OpenAl's GPT-4

February: Microsoft debuts the New Bing powered by OpenAi and its own

AI21 Labs Jurassic-2

March 2023 - AI21 Labs

releases Jurassic-2, with three different sizes. Also introduces

<u>Synthedia</u>

five new generative AI APIs.

Google PaLM

Model (PaLM).

April: Google introduce Pathways Language

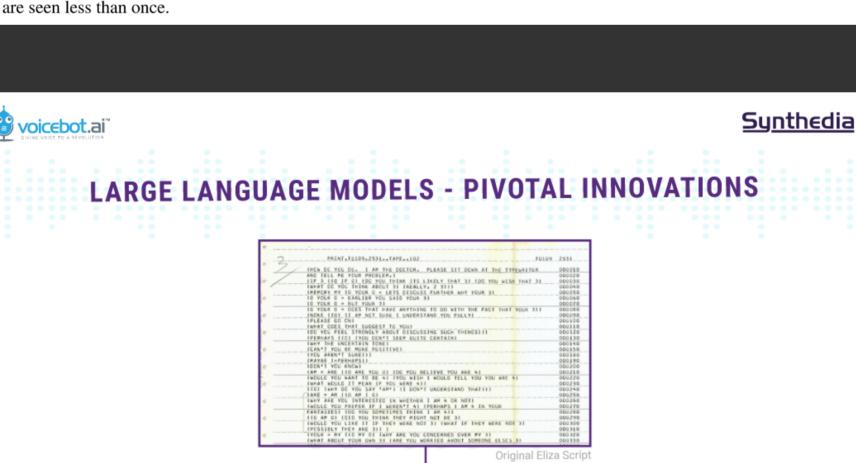
October: Allen Institute for Al

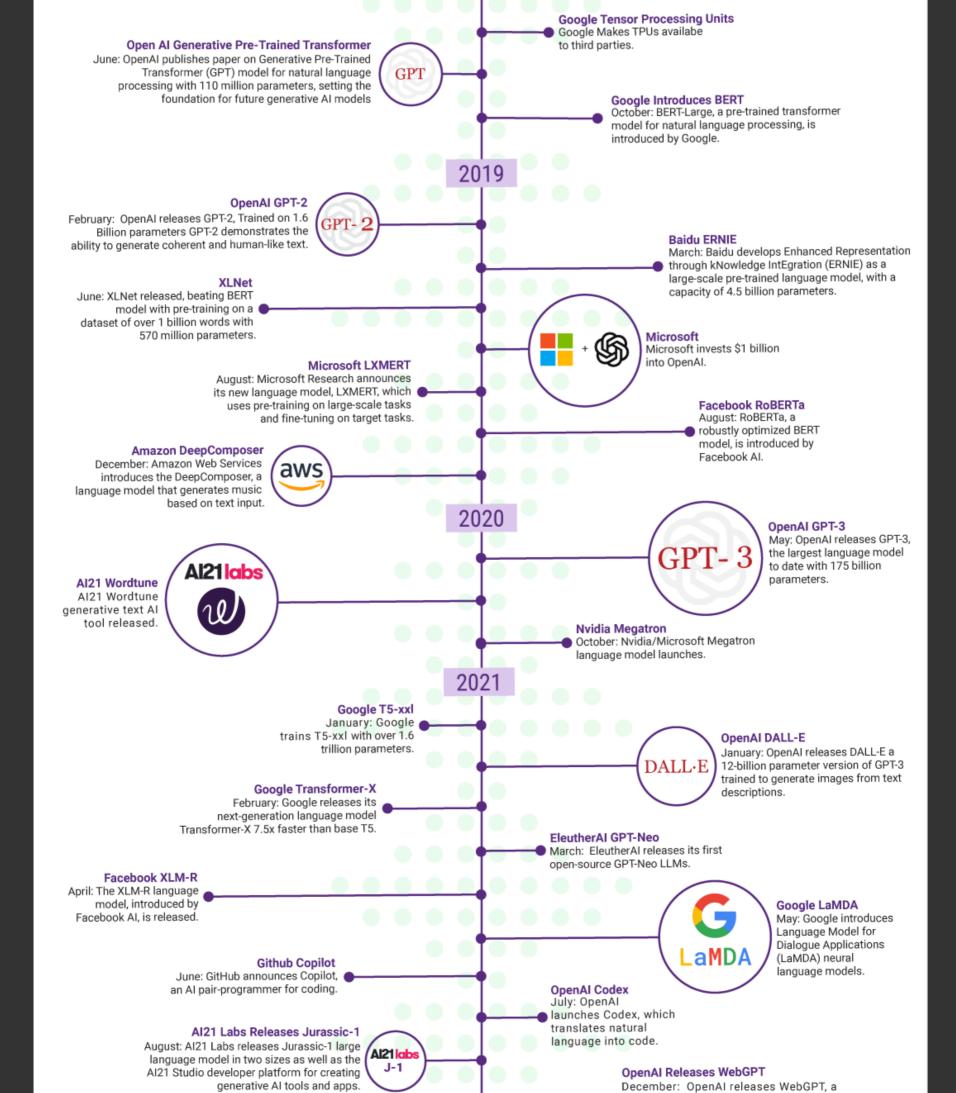
5.5B words and 1B parameters.

researchers develop Embeddings from Language Models (Elmo) trained on

IBM Tangora, is able to recognize

Long Short-Term Memory (LSTM) network developed by Sepp Hochreiter and Jürgen Schmidhuber as recurrent neural network able to learn from data and generate text.





Google DeepMind Model trained on open-source databases. October: Google DeepMind Sparrow dialogue agent introduced **Google Wordcraft** November: Google Wordcraft text generating tool launches. OpenAl Chat GPT November 30: OpenAI debuts generative AI chatbot ChatGPT to explosive interest. 2023 Microsoft Invests in OpenAI January: MIcrosoft invests a rumored \$10 billion more into OpenAl. **Baidu ERNIE Bot** February: Baidu previews Ernie Bot. Google BARD February: Google

OpenAl Introduces InstructGPT January: InstructGPT OpenAI introduces InstructGPT model as new default model, one better at following instructions and

> announces BARD, a ChatGPT rival.

Meta Introduces LLaMA

February: Meta LLaMA (Large Language Model Meta AI) a competitive multiple sized model LLaMA 65B and LLaMA 33B on 1.4 trillion tokens. Our smallest

model, LLaMA 7B, is trained on one trillion tokens.

directly into their apps

March: OpenAl introduces API's. Developers can now integrate ChatGPT and Whisper models

OpenAl APIs

less likely to hallucinate wrong answers.

Hugging Face BLOOM

July: Machine learning developer Hugging Face launches BigScience Large Open-science Open-access Multilingual (BLOOM) Language

> **Baidu Releases ERNIE Bot** likely a multimodal trillion March 2023: Baidu to release ERNIE parameter version of GPT-3 chatbot based on third-generation of the large language model and designed to be better understand Chinese culture than existing generative AI chatbots.

> > © Voicebot.ai All Rights Reserved 2023