变压器模型:简介和目录

泽维尔·阿马特里亚因 xavier@amatriain.net
阿南斯·桑卡尔 ansankar@linkedin.com
冰洁 jbing@linkedin.com
Praveen Kumar Bodigutla pbodigutla@linkedin.com
蒂莫西·J·黑森 thazen@linkedin.com
迈克尔·卡齐 mkazi@linkedin.com

2023 年 8 月 2 日

抽象的

这几年我们已经看到了几十个的横空出世
Transformer 系列的基础模型,所有这些模型都有令人难忘、有时甚至有趣但不言自
明的名称。目标是
本文旨在提供一个比较全面但简单的目录和
最流行的 Transformer 模型的分类。论文还
包括对最重要方面和创新的介绍
在变压器模型中。我们的目录将包括经过训练的模型
使用自我监督学习（例如 BERT 或 GPT3)以及那些
使用人机循环进一步训练（例如 InstructGPT 模型
由 ChatGPT 使用）。

内容

1 简介:什么是 Transformer

Transformer 是一类深度学习模型,由一些人定义
建筑特征。它们最初是在现在著名的“注意是
Google 研究人员于 2017 年发表的All you Need”论文（以及相关博客文章1）
（瓦斯瓦尼等人,2017）。该论文已累计被引用 38,000 次

只有5年。

1https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

最初的 Transformer 架构是编码器-解码器模型的一个特定实例（Cho 等人,2014 年)2 ,该模型在 2-3年前才开始流行。然而,直到那时,注意力只是这些模型使用的机制之一,这些模型主要基于 LSTM（长短期记忆）（Hochreiter & Schmidhuber,1997)和其他 RNN（循环神经网络）（Mikolov 等人,2017）。 ,2010)变化。正如标题所暗示的那样，Transformers 论文的主要见解是注意力可以用作导出输入和输出之间依赖关系的唯一机制。

Transformer 的输入是一系列标记。编码器的输出是每个标记的固定维度表示以及整个序列的单独嵌入。解码器将编码器的输出作为输入,并吐出一系列标记作为其输出。在自然语言处理（NLP)中,标记可以是单词或子单词。子词用于所有流行的 Transformer NLP 模型中,因为它们使我们能够解决基于单词的系统中固有的词汇外 (OOV) 问题。

为简单起见,我们将使用术语“令牌”来指代输入和输出序列中的项目,并理解这些令牌是 NLP系统的子词。当 Transformer 用于处理图像或视频时,令牌可以表示子图像或对象。

自论文发表以来,BERT 和 GPT 等流行模型仅使用原始架构的编码器或解码器方面。因此,这些模型的核心共性不是编码器-解码器方面,而是编码器和解码器中各个层的架构。

Transformers 的层架构基于自注意力机制和前馈层,其核心是每个输入令牌按照自己的路径流经各层,同时直接依赖于输入序列中的所有其他标记。这使得能够并行和直接计算上下文标记表示,这在以前使用 RNN 等顺序模型是不可能实现的。

深入探讨 Transformer 架构的所有细节超出了本文的范围。为此,我们将向您推荐原始论文（Vaswani 等人， 2017 年)或 The Illusterated Transformer3帖子。话虽这么说,我们将简要描述最重要的方面,因为我们将在下面的目录中提到它们。让我们从原始论文中的基本架构图开始,并描述一些组件。

1.1 编码器/解码器架构

通用编码器/解码器架构（见图 1)由两个模型组成。
编码器获取输入并将其编码为固定长度的向量。解码器获取该向量并将其解码为输出序列。编码器和解码器经过联合训练,以最大化给定输入的输出的条件对数似然。一旦经过训练，编码器/解码器就可以在给定输入序列的情况下生成输出,或者可以对一对输入/输出序列进行评分。

2https://machinelearningmastery.com/encoder-decoder-long-short-term-memory-networks/
3https://jalammar.github.io/illusterated-transformer/

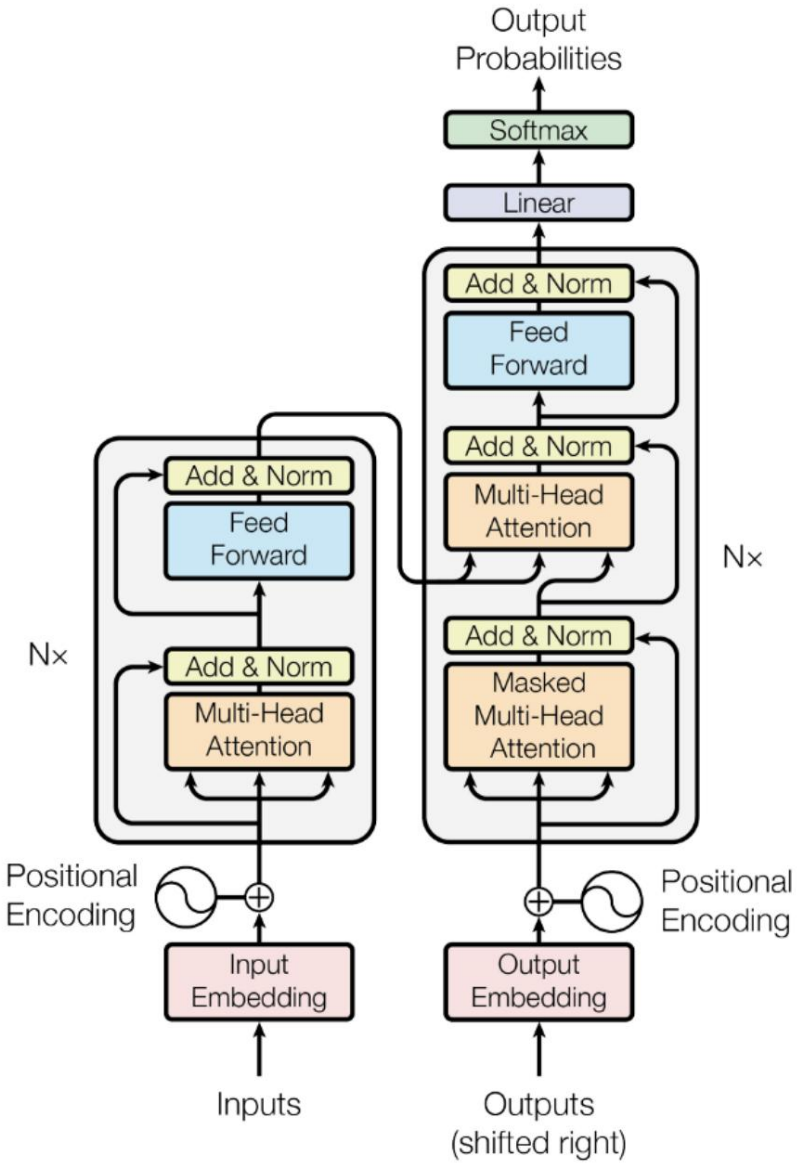

图 1:变压器架构（Vaswani 等人,2017 年）

在原始 Transformer 架构中,编码器和解码器都有 6 个相同的层。在这 6 层中,编码器的每一层都有两个子层:一个多头自注意力层和一个简单的前馈网络。自注意力层根据所有输入标记计算每个输入标记的输出表示。每个子层还具有残差连接和层归一化。编码器的输出表示大小为 512。

解码器中的多头自注意力层与编码器中的多头自注意力层略有不同。它屏蔽了正在计算表示的令牌右侧的所有令牌,以确保解码器只能处理位于它尝试预测的令牌之前的令牌。这在图 1 中显示为“屏蔽多头注意力”。解码器还添加了第三个子层,它是编码器所有输出上的另一个多头注意力层。请注意,所有这些具体细节已在我们将讨论的许多 Transformer 变体中进行了修改。例如,正如我们之前提到的,BERT和 GPT 等模型仅基于编码器或解码器。

1.2 注意事项

从上面的描述可以清楚地看出,模型架构中唯一的“奇异”元素是多头注意力层,但是,如上所述,这就是模型的全部力量所在!那么,注意力到底是什么?注意力函数是查询和一组键值对到输出之间的映射。使用三个相应的矩阵将输入到注意力层的每个标记转换为查询、键和值。每个令牌的输出表示被计算为所有令牌的值的加权和,其中分配给每个值的权重是通过其关联键和正在计算其表示的令牌的查询的兼容性函数来计算的。 Transformers 中使用的兼容性函数只是缩放后的点积。 Transformers 中这种注意力机制的一个关键方面是每个令牌都流经自己的计算路径,从而有助于并行计算输入序列中所有令牌的表示。现在我们了解了注意力的工作原理,那么什么是多头注意力呢?好吧,这只是多个注意力块独立计算每个标记的表示。然后聚合所有这些表示以给出令牌的最终表示。我们将再次向您推荐 The Illustrated Transformer[4]帖子,了解有关注意力机制如何工作的更多详细信息,但我们将重现图 2 中原始论文的图表,以便您了解主要想法。

与循环网络和卷积网络相比,注意力层有几个优点,最重要的两个优点是其较低的计算复杂性和较高的连接性,对于学习序列中的长期依赖性特别有用。

4https://jalammar.github.io/illusterated-transformer/

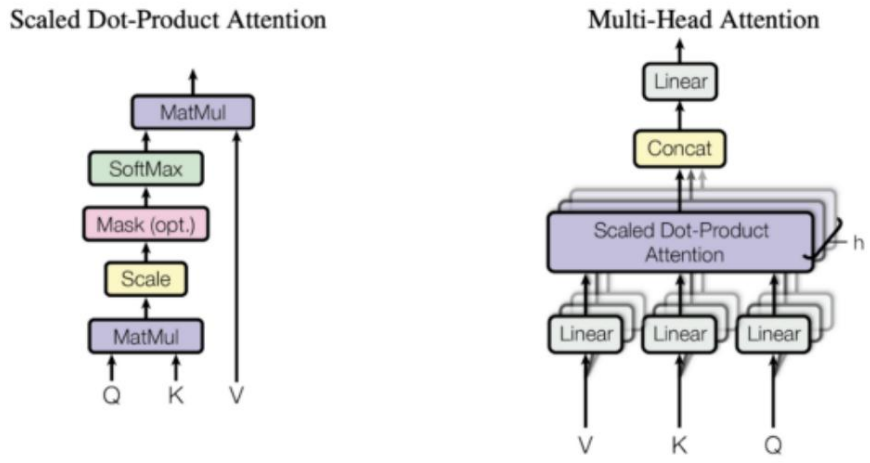

图 2:来自（Vaswani 等人,2017)的注意力机制。 （左)缩放点积注意力，（右)多头注意力

1.3 基础模型与微调模型

基础模型被定义为“任何经过广泛数据训练（通常使用大规模自我监督)的模型,可以适应（例如微调)广泛的下游任务”（Bommasani 等人,2021）). 当基础模型在少量特定于目标的数据上进一步训练时,它被称为微调模型5 ,因为它已经针对当前任务的具体情况进行了微调。

BERT 论文（Devlin 等人,2018)普及了这种自然语言处理预训练和微调的方法,导致许多研究人员使用这种方法来完成许多不同的任务。因此,任何与语言相关的机器学习 (ML) 任务的大多数排行榜都完全由某些版本的 Transformer 架构主导（例如,参见众所周知的问题解答 SQUAD 排行榜 6 或通用语言理解GLUE 排行榜 7 ） ,顶部的所有系统都采用基于 Transformer 的模型）。

在其最初的用法中，“微调”指的是针对特定任务调整基础模型,例如垃圾邮件分类或问题回答。 BERT 等模型会生成输入标记的表示,但它们本身并不能完成任何任务。因此，有必要通过在基础模型之上添加额外的神经层并端到端地训练模型来对它们进行微调。

对于像 GPT 这样的生成模型,情况有些不同。 GPT 是一个

5https://huggingface.co/docs/transformers/training
6https://rajpurkar.github.io/SQuAD-explorer
7https://gluebenchmark.com/leaderboard

解码器语言模型经过训练,可以在给定所有先前标记的情况下预测句子的下一个标记。通过对几乎涵盖人们能想到的任何主题的大量网络语料库进行训练,人们发现 GPT 实际上可以为输入查询或提示生成合理的输出。 GPT通过简单地预测给定输入提示序列和GPT 已经预测的输出序列的下一个标记来实现此目的。这种语言生成实际上在一些任务上完成了一些合理的工作,例如回答有关一般网络知识的问题、写诗等。尽管如此,GPT 的输出通常是不真实的,或者对用户来说确实没有多大帮助。为了解决这个问题,OpenAI 研究人员提出了训练 GPT 遵循人类指令的想法（Ouyang et al., 2022）。由此产生的模型称为 InstructGPT。

作者通过使用来自多种任务的少量人工标记数据来进一步训练 GPT 来实现这一目标。和以前一样,这是一个“微调”过程,但生成的 Instruct GPT 模型能够执行广泛的任务,并且实际上是流行的 ChatGPT 引擎使用的模型类别。由于这些模型可以完成无数任务,因此我们将它们称为基础模型。

这种额外的微调也已用于生成其他通用模型变体,专门为语言建模之外的用例（预测序列中的下一个标记)而设计。例如,有一个经过微调的模型子类,用于学习针对语义相关性优化的文本字符串嵌入,使它们直接可用于更高级别的语义任务（例如文本分类、聚类、搜索检索等）。示例包括 OpenAI 的文本嵌入模型8和 InstructOR10。 Transformer 编码器也已在多任务学习框架中成功进行了微调,能够使用单个共享Transformer 模型执行多个不同的语义任务（Liu 等人,2019 年;Aghajanyan 等人,2021 年）。

, E59,

因此,正如我们所看到的,虽然最初的基础模型是针对特定用户组的非常具体的目标任务进行微调的,但今天微调还用于创建可供大量用户使用的基础模型的进一步版本。 ChatGPT 和类似的对话代理（如 BlenderBot3 或 Sparrow)使用的过程相当简单:给定像 GPT 这样的预训练语言模型,我们用它来生成对输入提示（或指令)的不同响应,并让人类对结果进行排名。然后,我们使用这些排名（也称为偏好或反馈)来训练奖励模型。奖励模型为给定输入指令的每个输出附加一个分数。此后,使用人类反馈强化学习 (RLHF) 过程（Christiano 等人,2023)根据更多输入指令训练模型,但是,奖励模型不是使用人类来生成反馈,而是用于对模型的输出进行排序。您可以在 Huggingface11和 Ayush Thakur12的这两篇精彩文章中阅读更多内容。

8https://platform.openai.com/docs/guides/embeddings/what-are-embeddings
9https://huggingface.co/intfloat/e5-large
10https://huggingface.co/hkunlp/instructor-xl
11https://huggingface.co/blog/rlhf
12https://wandb.ai/ayush-thakur/RLHF/reports/Understanding-Reinforcement-Learning-from-Human-Feedback-RLHF-Par

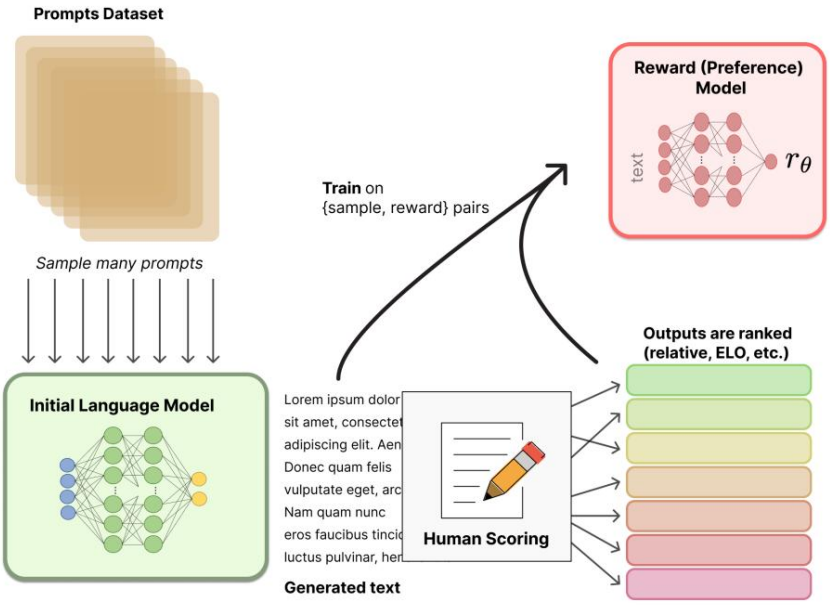

图 3:利用人类反馈进行强化学习。摘自 HuggingFace 的RLHF 博客文章： https://huggingface.co/blog/rlhf

1.4 变压器的影响

Transformer 原始论文（Vaswani 等人,2017)中演示的应用是语言翻译。这项开创性的工作还表明该架构可以很好地推广到其他语言任务。在接下来的几个月里,研究人员发现,通过对大量无监督文本进行预训练,变形金刚可以用来捕获大量关于语言的固有知识。然后,可以通过对少量标记数据进行训练,将这些模型中捕获的知识转移到目标任务。

虽然最初的 Transformer 是为语言任务而设计的,但相同的Transformer 架构已应用于许多其他应用程序,例如图像、音频、音乐甚至动作的生成。正因为如此,变形金刚被认为是所谓“生成人工智能”新浪潮的关键组成部分（如果不是关键的话）。生成式人工智能及其许多应用已经在社会的许多方面带来革命性的变化（Stokel-Walker & Noorden,2023; Baidoo- Anu & Owusu Ansah,2023）

当然,如果没有无数的工具,所有这些应用程序都是不可能实现的,任何人只要编写几行代码就可以轻松使用它们。 Transformer 不仅快速集成到主要的 AI框架（即 Pytorch13和 TensorFlow (TF)14)中,甚至还使得围绕它们创建了整个公司。 Huggingface15 是一家初创公司,迄今为止已筹集了超过 6000 万美元的资金,该公司几乎完全是围绕将其开源 Transformers 库商业化的想法而建立的16 。

随着商业参与者开发专用硬件以提高模型训练和推理速度， Transformer 模型的采用进一步加速。
NVIDIA 的 Hopper Tensor Cores17可以应用混合 FP8 和 FP16 精度来显着加速 Transformers 的 AI 计算。

最后但并非最不重要的一点是,如果我们不提及 ChatGPT 对变形金刚普及的影响,那就是我们的失职了。 ChatGPT 由 OpenAI 于 2022 年 11 月发布,并成为历史上增长最快的应用程序,不到一个月的时间就达到了 100 万用户,不到两个月的时间就达到了 1 亿（Dennean 等人, 2023）。 ChatGPT 最初是一个构建在 Instruct-GPT 模型（Ouyang et al., 2022)之上的聊天机器人应用程序,也称为 GPT-3.5。

不久之后,OpenAI 宣布发布更强大的 GPT-418 ,它在通过医生 USMLE 考试或律师律师 ，资格考试等任务中实现了人类能力（OpenAI,2023）。

1.5 关于扩散模型的注释

扩散模型已经成为图像生成领域最先进的技术,明显取代了之前的方法,例如 GAN（生成对抗模型）

13https://pytorch.org/tutorials/beginner/transformer_tutorial.html
14https://www.tensorflow.org/text/tutorials/transformer
15https://huggingface.co/docs
16https://github.com/huggingface/transformers
17https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet
18https://openai.com/research/gpt-4

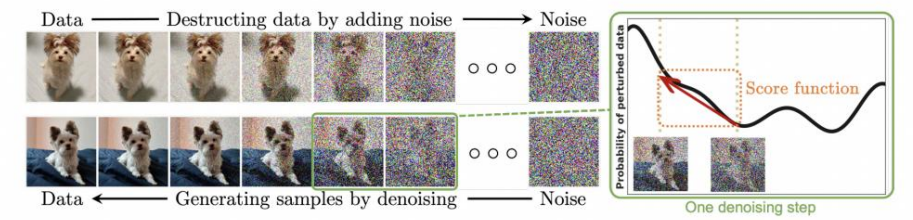

Fig. 2. Diffusion models smoothly perturb data by adding noise, then reverse this process to generate new data from noise. Each denoising step in the reverse process typically requires estimating the score function (see the illustrative figure on the right), which is a gradient pointing to the directions of data with higher likelihood and less noise.

图 4:来自“扩散模型:方法和应用的综合调查”的概率扩散模型架构,图 2（Yang 等人， 2022）

网络）。但值得注意的是,扩散机制并不依赖于 Transformer 架构。然而,大多数现代扩散方法确实包含 Transformer 主干（Esser 等人,2021）。

扩散模型是一类通过变分推理训练的潜变量模型。这在实践中意味着我们训练一个深度神经网络来对用某种噪声函数模糊的图像进行去噪。以这种方式训练的网络实际上正在学习这些图像所代表的潜在空间（见图 4）。

扩散模型与其他生成模型相关,例如去噪自动编码器和著名的生成对抗网络（GAN)19,它们在许多应用中大部分已被取代。一些作者20甚至会说扩散模型只是自动编码器的一个特定实例。

然而,他们也承认,微小的差异确实改变了他们的应用,从自动编码器的潜在表示到扩散模型的纯粹生成性质。

2 变形金刚目录

2.1 变压器的特点

在本节中,我们将介绍迄今为止已开发的最重要的 Transformer 模型的目录。我们将根据以下属性对每个模型进行分类:系列、预训练架构、预训练或微调任务、扩展、应用程序、日期（第一个已知出版物）、参数数量、语料库、许可证和实验室。有些相对容易理解:族代表特定模型扩展的原始基础模型,扩展描述模型向其本身添加的内容

19https://en.wikipedia.org/wiki/Generative_adversarial_network
20https://benanne.github.io/2022/01/31/diffusion.html

源自,日期是模型首次发布的时间,预训练模型的参数数量,语料库是模型预训练或微调的数据源,许可证描述了如何合法使用模型,实验室列表发布模型的机构。其余属性值得更多解释。我们在以下段落中这样做:

2.1.1 预训练架构

我们将 Transformer 架构描述为由编码器和解码器组成,对于原始 Transformer 来说也是如此。然而,从那时起,已经取得了不同的进展,表明在某些情况下仅使用编码器、仅使用解码器或两者都是有益的。

编码器预训练这些模型也称为双向或自动编码,仅在预训练期间使用编码器,这通常是通过屏蔽输入句子中的标记并训练模型以重建这些标记来完成。在预训练的每个阶段,自注意力层都可以访问其所有输入标记。该模型系列对于需要理解完整句子或段落的任务最有用,例如文本分类、蕴涵和提取式问答。

解码器预训练 解码器模型在预训练期间仅使用解码器。它们也称为自回归语言模型,因为它们被训练为根据先前的标记序列预测下一个标记。

自注意力层只能访问句子中给定标记之前的标记。它们最适合涉及文本生成的任务。

Transformer（编码器-解码器)预训练编码器-解码器模型,也称为序列到序列,使用 Transformer 架构的两个部分。

编码器的自注意力层可以访问所有输入令牌,而解码器的自注意力层只能访问位于给定令牌之前的令牌。如前所述,解码器中的附加关注层可以访问所有编码器令牌表示。

编码器-解码器模型可以通过优化去噪目标（Lewis et al., 2019)或去噪和因果语言建模目标的组合（Soltan et al., 2022)来进行预训练。与仅用于预训练编码器或仅解码器模型的目标函数相比,这些目标函数更为复杂。

编码器-解码器模型最适合根据给定输入生成新句子的任务,例如摘要、翻译或生成式问答。

2.1.2 预训练或微调任务

训练模型时,我们需要定义模型学习的目标或任务。上面已经提到了一些典型的任务,例如预测下一个标记或学习重建屏蔽标记。 “预训练模型

“自然语言处理:一项调查”（Qiu et al., 2020)包括一个相当全面的预训练任务分类,所有这些都可以被认为是自我监督的:

1.语言建模（LM）:预测下一个标记（在单向 LM 的情况下)或前一个和下一个标记（在双向 LM 的情况下）。

2.因果语言模型（Causality-masked LM）:自回归（一般从左到右)预测文本序列,类似于单向LM。

3.前缀语言建模（Prefix LM）:在此任务中,单独的“前缀”部分与主序列分开。在前缀内,任何标记都可以参与任何其他标记（非因果）。在前缀之外,解码以自回归方式进行。

4.屏蔽语言建模（MLM）:从输入句子中屏蔽掉一些标记,然后训练模型使用周围的上下文来预测屏蔽的标记。

5.排列语言建模 (PLM):与 LM 相同,但基于输入序列的随机排列。排列是从所有可能的排列中随机采样的。然后选择一些标记作为目标,并训练模型来预测这些目标。

6.去噪自动编码器 (DAE):采用部分损坏的输入,旨在恢复原始的、未失真的输入。损坏输入的示例包括从输入中随机采样标记并将其替换为“[MASK]”元素、从输入中随机删除标记或以随机顺序打乱句子。

7.替换标记检测（RTD）:使用“生成器”模型,随机替换文本中的某些标记。 “鉴别器”的任务是预测标记是来自原始文本还是生成器模型。

8.下一句预测（NSP）:训练模型以区分两个输入句子是否是训练中的连续片段
语料库。

请注意,在微调模型的情况下,此属性用于描述
模型微调的任务,而不是预训练的方式。

2.1.3 应用

这里我们要注意Transformer模型的主要实际应用有哪些。大多数这些应用程序将在语言领域（例如问答、情感分析或实体识别）。然而,如前所述,一些 Transformer 模型也发现了远远超出 NLP 的应用,并且也包含在目录中。

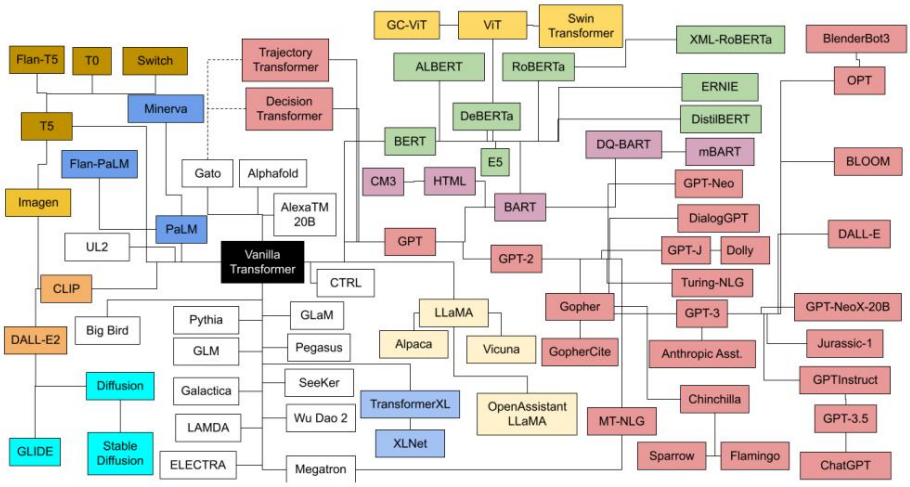

图 5:变形金刚家谱

2.2 目录表

您可以在http://bit.ly/3YFqRn9访问目录的表格格式以便更轻松地浏览不同型号的功能。

2.3 家谱

图 5 中的图表是一个简单视图,突出显示了 Transformer 的不同系列以及它们之间的相互关系。

2.4 时间线

目录的另一个有趣的视角是将其视为按时间顺序排列的时间线。在图 6 中,您将发现目录中的所有 Transformer 按发布日期排序。在第一个可视化中,Y 轴仅用于对相关遗产/家族的变形金刚进行聚类。

在图 7 中,Y 轴代表模型大小（以百万个参数为单位）。您将无法看到目录中的所有型号,因为许多型号都具有相同的时间和尺寸,因此请参阅上一张图片。

自从引入 chatGPT 以来,LLM 开源社区的活动量大幅增加。每过一周,我们都会观察到使用最新技术进行微调的精致模型的激增。

因此,这些模型不断改进,变得更加稳健和强大。图 8 显示了自 2023 年 2 月以来最新出现的模型。

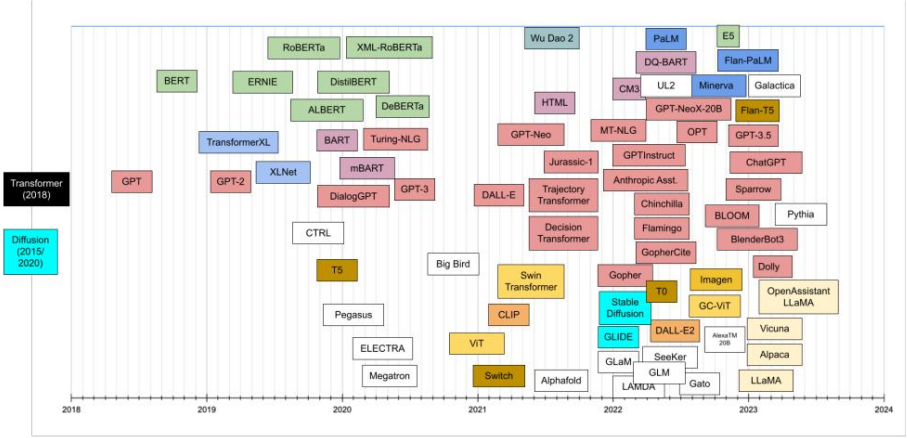

图 6:Transformer 时间轴。颜色描述了 Transformer 系列。

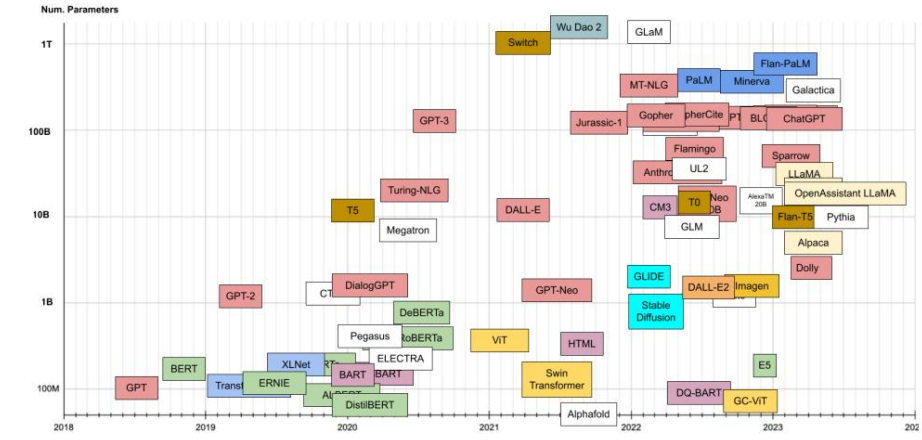

图 7:Transformer 时间线。纵轴为参数数量。
颜色描述了 Transformer 系列。

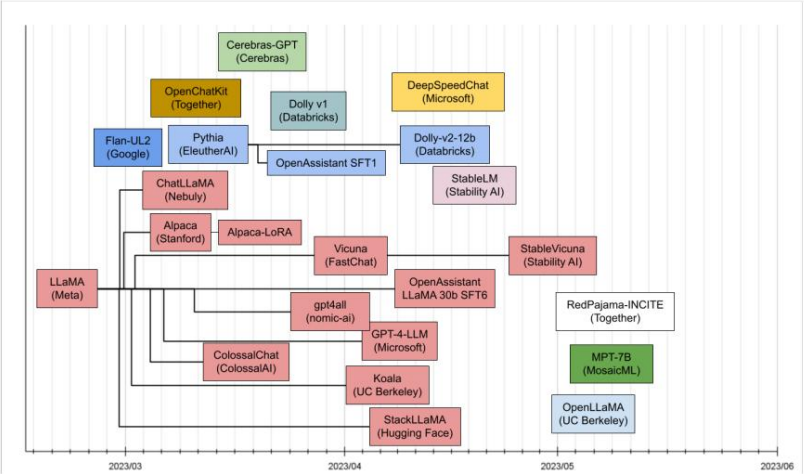

图 8:最近发表的法学硕士

目录列表

最后,这是完整的列表视图,在某些方面可能更容易理解
案例:

A.1 阿尔伯特

- 参考文献： （Lan 等人,2019）
- 链接： https: //huggingface.co/docs/transformers/model_doc/albert
- 家族:BERT
- 预训练架构:编码器
- 预训练任务:MLM/NSP
- 扩展:使用参数共享的 BERT 压缩版本，
  给定相同数量的参数,效率要高得多
- 应用:与BERT相同
- 日期（首次已知出版物） :09/2019
- 数量。参数:基本 = 12M,大 = 18M,XLarge = 60M*
- 语料库:与 BERT 相同
- 许可证:开放、Apache-2.0
- 实验室:谷歌

A.2 AlexaTM 20B

- 参考文献：（Soltan 等人,2022）
- 链接:https: //github.com/amazon-science/alexa-teacher-models
- 系列:变压器
- 预训练架构:编码器/解码器 · 预训练任务:优化去噪 (80%) 和 Prefix LM (20%)
- 扩展:源自BART 和恰好位于每层开始处的层规范。使用内部 10B 预训练编码器初始化编码器。
- 应用:摘要、多语言机器翻译和 NLU 任务
- 日期（首次已知出版物）:08/2022
- 数量。参数:20B
- 语料库:12 种语言的维基百科和mC4 数据集。
- 许可:有限、非商业性
- 实验室:亚马逊

A.3 羊驼毛

- 参考文献：（Taori 等人,2023）
- 链接:https: //github.com/tatsu-lab/stanford_alpaca
- 家庭： LLaMA
- 预训练架构:解码器
- 微调任务:人工指令
- 扩展:羊驼毛是在7B LLaMA 模型的基础上进行微调的。
- 应用:评估各种文本生成和分类 任务。
- （第一个已知出版物的日期）:03/2023
- 数量。参数:7B
- 语料库:使用自指令机制从175 个人工编写的指令输出对生成的52K 指令跟踪数据。
- 许可证:有限、非商业定制许可证
- 实验室:斯坦福大学

A.4 AlphaFold · 参

考：（Jumper 等人，2021）

- 链接：https: //github.com/deepmind/alphafold
- 系列：SE(3) Transformer（Fuchs 等人，2020）
- 预训练架构：编码器
- 预训练任务：使用参数共享对 BERT 进行蛋白质折叠预测，在参数数量相同的情况下效率更高
- 扩展：最初的Alphafold 使用了BERT 风格的Transformer。 Alphafold 的 Transformer 的细节尚不清楚，但据信它是SE(3)-Transformer（3-D 等变 Transformer)的扩展（请参阅此博客文章21）
- 应用：蛋白质折叠
- 日期（首次已知出版物）：09/2019
- 数量。参数：b12M，大= 18M，XLarge = 60M*
- 语料库：与 BERT 相同
- 许可证：代码开源，采用Apache-2.0
- 实验室：Deepmind

A.5 人择助手

- 参考文献：（Bai 等人，2022a；Askel 等人，2021；Bai 等人，2022b）
- 链接：不适用
- 系列：变压器
- 预训练架构：解码器
- 预训练任务：LM
- 扩展：这些模型没有在架构/预训练层面引入新颖性，它们与 GPT-3 类似，但它们专注于如何通过微调和提示来改进对齐。请注意， Anthropic Assistant 包括针对不同任务优化的多个模型。这些工作通常关注 RLHF 的好处。这项工作研究的最新版本使用法学硕士来批评模型输出的无害性，并以这种方式为强化学习提供反馈数据（RLHF -> RLAIF）。

21https://fabianfuchsml.github.io/alphafold2/

·应用程序:不同型号有不同的应用程序,从一般对话框到代码助手。

· 日期（首次已知出版物）:12/2021

· 数量。参数:10M 至 52B

·语料库:来自过滤后的Common Crawl 和Books 的400B 令牌,以及10% python 代码。他们还为RLHF
训练创建了多个对话偏好数据集。

· 许可证:不适用

· 实验室:人择

A.6 捷运

· 参考文献：（Lewis 等人,2019）

· 链接:https: //huggingface.co/docs/transformers/model_doc/bart

· 系列:用于编码器的BERT,用于解码器的GPT

· 预训练架构:编码器/解码器

· 预训练任务:DAE

·扩展:可以看作是 BERT 和 GPT 的推广：
它结合了编码器和解码器的想法

·应用:主要是文本生成,但也有一些文本理解
任务*

· 日期（首次已知出版物）:10/2019*

·数量。参数:基本 = 140M,大 = 400M。一般来说,对于同等架构,大约比 BART 大 10%。

· 语料库:与 RoBERTa 相同（160Gb 的新闻、书籍、故事）

· 许可证:开放、Apache-2.0

· 实验室:Facebook

A.7 BERT

· 参考文献：（Devlin 等人,2018）

· 链接:https: //huggingface.co/docs/transformers/model_doc/bert

· 家族:BERT

· 预训练架构:编码器

· 预训练任务:MLM/NSP

·扩展:它可以被视为 BERT 和 GPT 的推广,因为它结合了编码器和解码器的思想

·应用:一般语言理解和问答。
随后出现了许多其他语言的应用程序

· 日期（首次已知出版物）:10/2018

· 数量。参数:基础 = 110M,大 = 340MT

· 语料库:多伦多图书语料库和维基百科（3.3B 代币）

· 许可证:开放、Apache-2.0

· 实验室:谷歌

A.8 大鸟

· 参考文献：（Zaheer 等人,2020）

·链接： https: //huggingface.co/docs/transformers/model_doc/big_
鸟

· 家族:BERT

· 预训练架构:编码器

· 预训练任务:传销

·扩展:Big Bird 可以扩展其他架构,例如 BERT、
Pegasus 或 RoBERTa 通过使用稀疏注意机制消除二次依赖性,从而使其更适合更长的序列

·应用:特别适合较长的序列,不仅限于文本
但也包括在基因组学中

· 日期（首次已知出版物）:07/2020

· 数量。参数:取决于整体架构

· 语料库:书籍、CC 新闻、故事和维基百科）

· 许可证:开放、Apache-2.0

· 实验室:谷歌

A.9 BlenderBot3

- 参考文献：（Shuster 等人,2022b）
- 链接:https: //parl.ai/projects/bb3/
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:BlenderBot 3 基于预训练的OPT。它增加了对话代理所需的功能,例如长期记忆或搜索互联网的能力。根据人类反馈,它还针对某些特定任务进行了微调。
- 应用:与GPT-3相同
- 日期（首次已知出版物）:08/2022
- 数量。参数:3B、30B 和 175B
- 语料库:180B 代币 = RoBERTa + Pile + PushShift.io Reddit 项目
  许可:有限、非商业、仅限研究
- 实验室:Facebook

A.10 布卢姆

- 参考:参见博客文章22
- 链接： https: //huggingface.co/docs/transformers/model_doc/bloom
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:与 GPT-3 的主要区别在于它使用完全注意力
  稀疏关注
- 应用:与GPT-3相同
- 日期（首次已知出版物）:07/2022
- 数量。参数:176B
- 语料库:366B 个标记（1.5 TB 文本数据)多语言数据集 · 实验室:Big Science/Huggingface
- 许可证:开放,但需要遵循附件 A、BigScience 中的限制
  铁路许可证 v1.0

22https://huggingface.co/blog/bloom-inference-optimization

A.11 聊天GPT

- 参考:参见博客文章23
- 链接:https: //chat.openai.com
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:ChatGPT 采用 GPT3.5（又名 GPT3 Davinci-003)预训练模型,并使用 RLHF 来微调模型,大部分与InstructGPT 中描述的类似,但数据收集略有不同。 ChatGPT不仅仅是一个模型,因为它包含类似于 BlenderBot3 的内存存储和检索扩展
- 应用程序:对话代理
- 日期（首次已知出版物）:10/2022
- 数量。参数:与GPT3相同
- 语料库:与 GPT3 相同 + 为 RLHF 生成的数据集
- 许可证:闭源,可通过 API 访问
- 实验室:OpenAI

A.12 龙猫

- 参考文献: （Hoffmann 等人,2022） ·

链接:不适用

- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:与 Gopher 相同,但进行了优化以减少模型大小,从而减少训练/推理时间,同时具有相同或更好的性能
- 应用:与 Gopher/GPT3 相同 · 日期（首次已知发布）:03/2022
- 数量。参数:70B
- 语料库:海量文本
- 许可证:闭源。
- 实验室:Deepmind

23 https://openai.com/blog/chatgpt/

A.13 剪辑

- 参考文献：（Radford 等人，2021）
- 链接：https: //huggingface.co/docs/transformers/model_doc/clip
- 系列：CLIP（还使用 Resnet、ViT 和 vanilla Transformer 来处理文本）
- 预训练架构：编码器
- 预训练任务：预测 N × N 中哪一个可能（图像、文本）
  批次间的配对实际发生
- 扩展：结合 Resnet 和 ViT 进行视觉编码
  文本编码器的变压器
- 应用：图像/物体分类
- 日期（首次已知出版物）：02/2021
- 数量。参数：不适用
- 语料库：WIT (WebImageText) - 4 亿个文本、图像对
- 许可证：开放、MIT 许可证
- 实验室：OpenAI

A.14 CM3

- 参考文献： （Aghajanyan 等人，2022）
- 链接：不适用
- 系列：HTML
- 预训练架构：解码器
- 预训练任务：因果屏蔽 LM
- 扩展：这在结构化训练数据的使用方面有点类似于HTML 。然而，它是一种不同的架构，并使用因果屏蔽，这使得模型在序列末尾预测整个缺失的文本范围。它还包括通过矢量量化变分自动编码 (VQ-VAE) 令牌输入的图像。
- 应用：多模态语言模型，能够进行结构化提示、零镜头字幕、图像生成和实体链接（通过超链接的目标文本预测）
- 日期（首次已知出版物）：01/2022
- 数量。参数：13B（最大）

- 语料库:CC-News、英语维基百科
- 许可证:不适用
- 实验室:Facebook

A.15 控制

- 参考文献：（Keskar 等人,2019）
- 链接:https: //huggingface.co/docs/transformers/model_doc/ctrl
- 家庭：
- 预训练架构:解码器
- 预训练任务：
- 扩展:模型可以生成以控制代码为条件的文本,这些控制代码指定域、样式、主题、日期、实体、实体之间的关系、情节点和任务相关行为
- 应用:可控文本生成
- 日期（首次已知出版物）:09/2019
- 数量。参数:1.63B
- 语料库:140 GB 文本,包括:维基百科（En、De、Es、Fr）、项目古腾堡、45 个 subreddits、OpenWebText2、亚马逊评论、Europarl 和来自 WMT 的联合国数据、来自 ELI5 的问答对以及 MRQA 共享任务3,其中包括斯坦福问答数据集，NewsQA、TriviaQA、SearchQA、HotpotQA 和 Natural Questions
- 许可证:开放、BSD-3 条款许可证
- 实验室:Salesforce

A.16 达尔-E

- 参考文献：（Ramesh 等人,2021）
- 链接:https: //openai.com/blog/dall-e
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:字幕预测
- 扩展:差分变分自动编码器用于学习视觉密码本。 Transformer 是 GPT-3 的变体

- 应用:文本转图像
- 日期（首次已知出版物）:01/2021
- 数量。参数:12B
- 语料库:来自互联网的 2.5 亿文本-图像对
- 许可证:不适用
- 实验室:OpenAI

A.17 达尔-E 2

- 参考文献：（Ramesh 等人,2022）
- 链接:https: //openai.com/dall-e-2
- 系列:CLIP、GLIDE
- 预训练架构:编码器/解码器
- 预训练任务:字幕预测
- 扩展:结合 CLIP 编码器和 Diffusion 解码器,类似于滑行
- 应用:文本转图像
- 日期（首次已知出版物）:04/2022
- 数量。参数:3.5B
- 语料库:DALL-E 和 CLIP 数据集的组合
- 许可证:闭源,可通过 API 访问
- 实验室:OpenAI

A.18 德贝尔塔

- 参考文献：（He et al., 2021）
- 链接:https: //huggingface.co/microsoft/deberta-large
- 家族:BERT
- 预训练架构:编码器
- 预训练任务:传销
- 扩展:使用内容和相对位置的解缠注意力矩阵,独立于内容嵌入的独立位置嵌入向量

- 应用:与BERT相同
- 日期（首次已知出版物）:06/2020
- 数量。参数:134M（基本）、384M（大）、750M（超大）
- 语料库:英语维基百科、BookCorpus、OPENWEBTEXT 和 STO-里斯
- 许可证:开放、MIT 许可证
- 实验室:微软

A.19 决策转换器

- 参考文献： （Chen 等人,2021）
- 链接:https: //github.com/kzl/decision-transformer
- 系列:GPT,控制 Transformers”（本身不是一个系列,而是将那些尝试对更通用的控制、类 RL任务进行建模的 Transformer 分组）
- 预训练架构:解码器
- 预训练任务:下一步动作预测
- 扩展:决策转换器使用 GPT 架构,并通过以可通过自回归任务学习的方式对轨迹进行编码来扩展它
- 应用:通用强化学习（强化学习任务）
- 日期（首次已知出版物）:06/2021
- 数量。参数:与GPT相同
- 语料库:不同实验的不同语料库
- 许可证:开放、MIT 许可证
- 实验室:Google/加州大学伯克利分校/Facebook

A.20 DialoGPT

- 参考文献:（Zhang 等人,2019a）
- 链接： https: //huggingface.co/docs/transformers/model_doc/dialogpt
- 系列:GPT
- 预训练架构:解码器

- 预训练任务:LM
- 扩展:基于对话数据训练的 GPT-2 架构
- 应用:对话框设置中的文本生成
- 日期（首次已知出版物）:10/2019
- 数量。参数:1.5B
- 语料库:1.4 亿 Reddit 对话
- 许可证:开放、MIT 许可证
- 实验室:微软

A.21 DistilBERT

- 参考文献：（Sanh 等人,2019）
- 链接： https: //huggingface.co/docs/transformers/model_doc/distilbert
- 家族:BERT
- 预训练架构:编码器
- 预训练任务:MLM/NSP
- 扩展:使用蒸馏的 BERT 压缩版本,这比
  给定相同数量的参数,效率更高
- 应用:与BERT相同
- 日期（首次已知出版物）:10/2019
- 数量。参数:66M
- 语料库:与 BERT 相同
- 许可证:开放、Apache-2.0
- 实验室:Hugingface

A.22 DQ-BART

- 参考文献：（Li 等人,2022）
- 链接:https: //github.com/amazon-science/dq-bart
- 家庭:BART
- 预训练架构:编码器/解码器
- 预训练任务:DAE

·扩展:向 BART 模型添加量化和蒸馏,以提高性能和模型大小

· 应用:文本生成和理解

· 日期（首次已知出版物）:03/2022

·数量。参数:与标准相比,参数减少高达 30 倍
捷运

· 语料库:CNN/DM、XSUM、ELI5、WMT16 En-Ro（100 万个代币）

· 许可证:开放、Apache-2.0

· 实验室:亚马逊

A.23 多莉

· 参考:参见博客文章24

· 链接:https: //huggingface.co/databricks/dolly-v1-6b

· 系列:GPT

· 预训练架构:解码器

· 微调任务:人工指令

·扩展:基于GPT-J-6B（V1)和Pythia模型进行微调
(V2)

· 应用:类似于羊驼毛

·（第一个已知出版物的日期）:03/2023

· 数量。参数:V1:6B,V2:12B

·语料库:V1:与 Alpaca 相同的指令语料库,V2:databricks 自己的语料库
数据集。

· 许可证:开放

· 实验室:Databricks, Inc

24https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

A.24 E5

- 参考文献：（Wang 等人，2022）
- 链接：https: //huggingface.co/intfloat/e5-large
- 家族：BERT
- 预训练架构：编码器
- 微调任务：使用对比损失的语义相似度
- 扩展：微调基于 BERT 的模型以创建文本字符串嵌入
  针对语义相关性进行了优化。
- 应用：用于语义相关任务（例如文本）的文本嵌入
  聚类或搜索检索。
- 日期（首次已知出版物）：12/2022
- 数量。参数：300M（大版）
- 语料库：MS-MARCO、NQ、NLI
- 许可证：开放、MIT 许可证
- 实验室：微软

A.25 伊莱克特拉

- 参考文献：（Clark 等人，2020）
- 链接： https: //huggingface.co/docs/transformers/model_doc/electra
- 家族：BERT
- 预训练架构：编码器
- 预训练任务：RTD
- 扩展：应用新的训练技术，包括替换令牌
  检测
- 申请：03/2020
- （首次已知出版物的日期）日期：2020 年
- 数量。参数：基础 = 110M，大 = 330M
- Corpus：与 BERT 相同，但 Large 与 XLNet 相同
- 许可证：开放、Apache-2.0
- 实验室：斯坦福/谷歌

A.26 厄尼

- 参考文献：（Zhang et al., 2019b）
- 链接：不适用
- 家族：BERT
- 预训练架构：编码器
- 预训练任务：传销
- 扩展：使用BERT 进行编码器架构，但对其中的两个文本和实体进行堆栈和聚合。这个架构可以理解为文本+知识图谱的BERT
- 应用：可能受益于知识图或实体（例如实体识别）的知识密集型相关任务
- 日期（首次已知出版物）：05/2019
- 数量。参数：Ernie-ViLG 2.0 = 10B，Ernie 3.0 Titan = 260B
- 语料库：英文维基百科 + 实体的维基数据（请注意，它们将模型初始化为原始 BERT 参数值
- 许可证：闭源
- 实验室：百度、鹏程实验室

A.27 火烈鸟

- 参考文献：

（Alayrac 等人，2022）

- 链接：不适用
- 家族：龙猫
- 预训练架构：解码器
- 预训练任务：记录给定一些视觉输入的文本的可能性
- 扩展：它使用以视觉表示为条件的冻结文本语言模型（如 Chinchilla），该模型是从规范化器编码的免费 ResNet
- 应用：文本转图像
- 日期（首次已知出版物）：04/2022
- 数量。参数：80B（最大）

·语料库:MultiModal MassiveWeb (M3W):1.85 亿张图像和 182
GB文本+与图像数据集配对的多个文本:ALIGN + LTIP
（长文本和图像对)= 3.12 亿张图像,以及 VTP（视频和文本）
对） = 2700 万个短视频（平均约 22 秒）

· 许可证:闭源

· 实验室:Deepmind

A.28 果馅饼-T5

· 参考文献： （Chung 等人,2022）

·链接： https://huggingface.co/docs/transformers/model_doc/flan-t5

· 家族:T5

· 预训练架构:编码器/解码器

· 微调任务:零样本和少样本任务的说明

·扩展:Flan-T5 是通过“Flan Finetuning”T5 模型生成的： (1) 将任务数量扩展到1,836,(2) 缩放模型大小,以及(3)对思维链数据进行微调。

·应用:主要用途是了解如何通过正确的指令微调来改进大型语言模型。重点研究零样本和上下文中的小样本学习 NLP 任务,例如推理和问答;推进公平性和安全性研究,并了解当前大型语言模型的局限性

· 日期（首次已知出版物）:11/2022

·数量。参数:80M(小)、250M(底座)、780M(大)、3B(XL)、11B(XXL)

· 语料库:Flan 针对Muffin、T0-SF、NIV2 和CoT 中的任务进行了微调。

· 许可证:开放、Apache-2.0

· 实验室:谷歌

A.29 弗兰-帕尔姆

· 参考文献： （Chung 等人,2022）

· 链接:不适用

· 家庭:PaLM

· 预训练架构:解码器

· 微调任务:零样本和少样本任务的说明

·扩展:Flan-PaLM 是通过“Flan Finetuning”PaLM 模型生成的:(1) 将任务数量扩展到 1,836,(2) 缩放模型大小,以及 (3) 对思维链数据进行微调。

·应用:与Flan-T5 相同。目标是展示 Flan 微调甚至可以改进最大的 Google LM（跨任务平均改进 9.4%）,并改进思维链、自我一致性、多语言任务、算术推理

· 日期（首次已知出版物）:11/2022

· 数量。参数:8B、62B、540B

· 语料库:Flan 针对Muffin、T0-SF、NIV2 和CoT 中的任务进行了微调。

· 许可证:闭源

· 实验室:谷歌

A.30 卡拉狄加

· 参考文献： （Taylor 等人,2022）

· 链接:https: //galacta.org

· 系列:变压器

· 预训练架构:解码器

· 预训练任务:科学领域的LM

·扩展:仅解码器设置中基于变压器的架构,经过一些修改。数据扩展包括工作记忆、引文、遗传数据和其他一些生物学相关任务的特殊标记。

·应用:模型旨在执行科学任务,包括但不限于引文预测、科学质量保证、数学推理、摘要、文档生成、分子属性预测和实体提取。

· 日期（首次已知出版物）:11/2022

·数量。参数:迷你:125M,底座:1.3B,标准:6.7B,大号:30B,
巨大:120B

·语料库:使用 1060 亿个开放获取科学文本和数据进行训练。这包括论文、教科书、科学网站、百科全书、参考资料、知识库等

· 许可证:有限、非商业 CC BY-NC 4.0 许可证

· 实验室:元

A.31 加托

- 参考文献：（Reed 等人,2022）
- 链接:https: //www.deepmind.com/blog/a-generalist-agent
- 系列：“控制变压器”（本身不是一个系列,而是将那些尝试对更通用的控制、类似 RL 的任务进行建模的变压器分组）
- 预训练架构:解码器
- 预训练任务:MLM（其中标记是文本或代理操作）
- 扩展:标准的仅解码器 Transformer 架构前面是一个嵌入层,可以嵌入文本和图像,并在适用时添加位置编码以添加空间信息。
- 应用:Gato 提出了一种通用代理,除了文本之外,还可以用于玩 Atari 或控制机器人手臂等任务。
- 日期（首次已知出版物）:05/2022
- 数量。参数:1.2B
- 语料库:1.5T 令牌,包括标准文本（例如 MassiveText）、视觉（例如 ALIGN)和模拟环境（例如 ALE Atari 或 RGB Stacking Real Robot）
- 许可证:闭源
- 实验室:Deepmind

A.32 华丽

- 参考文献：（Du 等人,2022a）
- 链接:参见博客文章25
- 系列:变压器
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:GLaM 引入了 64 个专家的混合,以增加某种标准解码器中的参数数量和泛化属性。变压器架构。每个令牌一次只有两名专家被激活,这使得模型在训练和推理方面也更加高效。

25https://ai.googleblog.com/2021/12/more-efficient-in-context-learning-with。
html

- 应用:通用语言建模
- 日期（首次已知出版物）:12/2021
- 数量。参数:64位专家1.2T,但只有96B被激活
  推理
- Corpus:1.6T 代币,包括维基百科过滤的网页和书籍
  为了质量
- 许可证:闭源
- 实验室:谷歌

A.33 滑翔

- 参考文献： （Nichol 等人,2021）
- 链接:https: //github.com/openai/glide-text2im
- 系列:扩散模型
- 预训练架构:编码器
- 预训练任务:字幕预测
- 扩展:GLIDE 可以被视为同一作者对 ADM（消融扩散模型)的扩展。然而,ADM 本身并不是一种
  Transformer 架构,尽管它确实类似于作者使用的一些配置。鉴于 ADM 是由相同的作者编写的,并且
  很快就被 GLIDE 跟进,我们认为将GLIDE 视为同类中的第一个是公平的。

- 应用:文本转图像
- 日期（首次已知出版物）:12/2021
- 数量。参数:3.5B扩散模型（2.3B用于视觉编码,1.2B用于
  文本)+ 1.5B 用于上采样模型
- 语料库:与 DALL-E 相同
- 许可证:开放、MIT 许可证
- 实验室:OpenAI

A.34 GLM

- 参考文献： （Du 等人,2022b）
- 链接:https: //github.com/THUDM/GLM-130B
- 系列:GLM（通用语言模型）
- 预训练架构:编码器和解码器
- 预训练任务:自回归空白填充
- 扩展:GLM 有一个双向编码器和一个单向解码器
  在一个统一的模型中。
- 应用:使用自回归填空目标进行预训练的通用语言模型,可以针对各种自然语言理解和生成任务进行微调。
- 日期（首次已知出版物）:03/2022
- 数量。参数:Base = 110M,Large = 335M,还有 2B、10B、130B
- 语料库:Pile、GLM-130B 中文语料库、P3、DeepStruct 微调
  数据集
- 许可证:开放、MIT 许可证
- 实验室:清华大学

A.35 全局上下文 ViT

- 参考文献： （Hatamizadeh 等人,2022）
- 链接:https: //github.com/NVlabs/GCVit
- 家庭:ViT
- 预训练架构:编码器
- 预训练任务:图像分类
- 扩展:由局部和全局自注意力模块组成的分层ViT架构
- 应用:图像生成
- 日期（首次已知出版物）:06/2022
- 数量。参数:90M
- 语料库:Imagenet-1K 和其他任务相关数据集
- 许可证:有限的非商业许可证CC-BY-NC-SA-4.0
- 实验室:NVidia

A.36 地鼠

- 参考文献：（Rae 等人,2021）
- 链接:参见博客文章26
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:与 GPT-2 相同,但使用 RSNorm 而不是 LayerNorm,使用相对位置编码而不是绝对位置编码
- 应用程序:主要是语言建模和 NLU,但也可扩展
  像 GPT
- 日期（首次已知出版物）:12/2021
- 数量。参数:280B
- 语料库:海量文本（23.5 亿文档,或约 10.5 TB 文本,包括海量网络、书籍、Github、新闻、C4 和维基百科。
- 许可证:闭源
- 实验室:Deepmind

A.37 地鼠引用

- 参考文献： （Menick 等人,2022）
- 链接:参见博客文章27
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:GopherCite 基于 Gopher,但添加了使用 RLHP 的步骤
  （根据人类偏好进行强化学习）了解响应是否合理且得到支持
- 应用:对话系统、问答、通用语言生成任务
- 日期（首次已知出版物）:03/2022
- 数量。参数:280B

26https://www.deepmind.com/blog/language-modelling-at-scale-gopher-ethical-considerations-and-retrieval
27https://www.deepmind.com/blog/gophercite-teaching-language-models-to-support-answers-with-verified-quotes

·语料库:与 Gopher 相同加上 RLHP 中生成的特定数据集
过程

· 许可证:闭源

· 实验室:Deepmind

A.38 GPT

· 参考文献：（Radford 等人,2018）

·链接： https ://huggingface.co/docs/transformers/model_doc/openai-gpt

· 系列:GPT

· 预训练架构:解码器

· 预训练任务:LM

· 扩大：

·应用:文本生成,但经过微调后可适用于许多其他 NLP 任务。

· 日期（首次已知出版物）:06/2018

· 数量。参数:117M

· Corpus:BookCorpus 数据集上的无监督预训练。对多个特定任务数据集（包括 SNLI、RACE、Quora）
进行监督微调。 。 。

· 许可证:不适用

· 实验室:OpenAI

A.39 GPT-2

· 参考文献：（Radford 等人,2019）

· 链接： https ://huggingface.co/docs/transformers/model_doc/gpt2

· 系列:GPT

· 预训练架构:解码器

· 预训练任务:LM

·扩展:GPT 架构的小扩展（例如,层标准化移至每个子层的输入,或将上下文大小从 512 增加到 1024）

·应用:文本生成,但经过微调后可适用于许多其他 NLP 任务。

· 日期（首次已知出版物）:02/2019

· 数量。参数:124M、355M、774M、1.5B

·语料库:800 万个网页(40 GB)。 10 倍 GPT 。 WebText 数据集是通过抓取 Reddit 上至少有 3 个 Karma 点的所有链接来创建的。

· 许可证:开放、修改的 MIT 许可证

· 实验室:OpenAI

A.40 GPT-3

· 参考文献:（Brown 等人,2020）

·链接:https: //github.com/openai/gpt-3

· 系列:GPT

· 预训练架构:解码器

· 预训练任务:LM

·扩展:与 GPT-2 相同,只是增加了交替的密集和局部带状稀疏注意力模式,灵感来自于稀疏变换器

·应用程序:最初用于文本生成,但随着时间的推移,已用于代码生成等领域的广泛应用,还包括图像和音频生成

· 日期（首次已知出版物）:05/2020

· 数量。参数:175 B

·语料库:500B 代币,包括 CommonCrawl (410B)、WebText2 (19B)、书籍 1 (12B)、书籍 2 (55B) 和维基百科 (3B)

· 许可证:闭源

· 实验室:OpenAI

A.41 GPT-3.5

- 参考:不适用
- 链接： https : //platform.openai.com/docs/model-index-for-researchers/models-referred-to-as-gpt-3-5
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:GPT3.5 系列包括许多模型,如Davinci- 003。它们基本上是 InstructGPT 模型的版本。有关与旧版 GPT3 模型的性能比较的详细信息，请参阅博客文章 28 。
- 应用程序:对话和通用语言,但有特定的代码 模型 - 法典
- 日期（首次已知出版物）:10/2022
- 数量。参数:175B
- 语料库:与 InstructGPT 相同
- 许可证:闭源,可通过 API 访问
- 实验室:OpenAI

A.42 GPT-J

- 参考文献：（Wang 和 Komatsuzaki,2021）
- 链接:https: //huggingface.co/EleutherAI/gpt-j-6B
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:GPT-J 6B 是使用 Mesh Trans- 训练的 Transformer 模型 前 JAX 和与 GPT2/3 相同的标记器
- 应用:与GPT-3相同
- 日期（首次已知出版物）:05/2021
- 数量。参数:6B

28https://scale.com/blog/gpt-3-davinci-003-comparison

- Corpus:Pile corpus,由EleutherAI 创建的大规模精选数据集。
- 许可证:开放、Apache-2.0
- 实验室:EleutherAI

A.43 GPT-Neo

- 参考： https: //huggingface.co/docs/transformers/model_doc/gpt_neo
- 链接:https: //github.com/EleutherAI/gpt-neo
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:与 GPT-2 类似,但在每个其他层中使用局部注意力,窗口大小为 256 个令牌
- 应用:文本生成,但经过微调后可适用于许多其他 NLP 任务
- 日期（首次已知出版物）:03/2021
- 数量。参数: 5B, 2.7B (XL)
- Corpus:Pile 840 GB 开源文本数据集,结合了 22 个现有数据集
- 许可证:开放、MIT 许可证
- 实验室:EleutherAI

A.44 GPT-NeoX-20B

- 参考文献： （Black 等人,2022）
- 链接： https://huggingface.co/EleutherAI/gpt-neox-20b
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:类似于 GPT-3,使用旋转编码器代替位置、并行注意和前馈层、不同的初始化以及所有密集层而不是交替密集/稀疏层
- 应用:与GPT-3相同

- 日期（首次已知出版物）:04/2022
- 数量。参数:20B
- Corpus:Pile 840 GB 开源文本数据集,结合了 22 个现有数据集
- 许可证:开放、Apache-2.0
- 实验室:EleutherAI

A.45 HTLM

- 参考文献： （Aghajanyan 等人,2021）
- 链接:不适用
- 家庭:BART
- 预训练架构:编码器/解码器
- 预训练任务:DAE
- 扩展:与 BART 不同,他们不进行句子改组
- 应用程序:通用语言模型,允许结构化 HTML提示
- 日期（首次已知出版物）:07/2021
- 数量。参数:400M
- 语料库:从 CommonCrawl 中提取的 23TB 简化 HTML
- 许可证:不适用
- 实验室:Facebook

A.46 图像

- 参考文献：（Saharia 等人,2022）
- 链接:https: //imagen.research.google
- 系列:T5、CLIP、Diffusion 型号
- 预训练架构:用于冻结文本编码器的 T5（或 CLIP 或 BERT)+ 用于文本到图像的级联扩散模型的 U-net 架构
- 预训练任务:图像/文本对预测

·扩展:Imagen 为 U-net 扩散架构添加了一些扩展（池化嵌入向量、文本嵌入的交叉关注,以及
层标准化）

· 应用:文本转图像

· 日期（首次已知出版物）:06/2022

· 数量。参数:2B

·语料库:内部数据集（包含 4.6 亿图像文本对)和公开可用的 Laion 数据集（包含 4 亿图像文本对)的组合

· 许可证:闭源

· 实验室:谷歌

A.47 指令GPT

· 参考文献：（欧阳等人,2022）

·链接： https: //github.com/openai/following-instructions- human-feedback

· 系列:GPT

· 预训练架构:解码器

· 预训练任务:LM

·扩展:GPTInstruct 从预训练的 GPT3 模型开始,并在监督微调后通过强化学习添加奖励建模

· 应用:知识密集型对话或语言任务

· 日期（首次已知出版物）:01/2022

· 数量。参数:与GPT3相同

·语料库:与 GPT3 的预训练相同,但使用标签数据和提示进行微调和优化

· 许可证:闭源,可通过 API 访问

· 实验室:OpenAI

A.48 指导者或

- 参考文献：（Su 等人,2022）
- 链接:https: //huggingface.co/hkunlp/instructor-xl
- 家族:T5
- 预训练架构:编码器/解码器
- 微调任务:各种基于指令的文本到文本任务
- 扩展:显式微调T5 以优化编码器,生成对许多NLU 任务有用的通用文本字符串嵌入。
- 应用程序:任何需要单个文本字符串嵌入的NLU 任务。截至2023 年 4 月,InstructOR 是大规模文本嵌入基准 (MTEB)上排名第一的系统。29
- 日期（首次已知出版物）:12/2022
- 数量。参数:330M
- 语料库:在 MEDI 上进行微调
- 许可证:开放、Apache-2.0
- 实验室:香港大学、华盛顿大学、META AI

A.49侏罗纪-1

- 参考文献： （Lieber 等人,2021）
- 链接:https: //github.com/ai21labs/lm-evaluation
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:与GPT-3 非常相似,但更多的参数和更高的训练效率主要是因为改进了分词器。另外,深度与宽度的比例也不同
- 应用:类似于GPT-3
- 日期（首次已知出版物）:09/2021 · 编号。参数:178B (Jumbo)、

17B (Grande)、7.5B (Large) · 语料库:300B 代币（与 GPT-3 相同）

- 许可证:闭源,可通过 API 访问
- 实验室:AI21

29https://huggingface.co/spaces/mteb/leaderboard

A.50 拉姆达

- 参考文献：（Thoppilan 等人，2022）
- 链接：参见博客文章30
- 系列：变压器
- 预训练架构：解码器
- 预训练任务：LM
- 延伸：LAMDA 关注如何使用不同的微调策略来提高安全性、质量和接地性
- 应用：通用语言建模，例如翻译、摘要化，问题和答案。
- 日期（首次已知出版物）：01/2022
- 数量。参数：137B
- 语料库：来自公共对话数据和其他公共网络的 1.56T 单词文件
- 许可证：闭源
- 实验室：谷歌

A.51 美洲驼

- 参考文献：（Touvron 等人，2023）
- 链接： https: //huggingface.co/docs/transformers/main/model_doc/骆驼
- 系列：变压器
- 预训练架构：解码器
- 预训练任务：LM
- 扩展：LLaMA 使用 Transformer 架构，并具有扩展功能：预归一化、SwiGLU 激活、RoPE 嵌入、通过有效实现因果多头注意力来减少内存使用和运行时间、检查点以减少激活量在后向传递、模型和序列并行期间重新计算，以减少模型的内存使用，并在标记化后使用 1.4T BPE 标记。

30https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html

·应用:零次和少量的常识推理、问答、代码生成和阅读理解。

· 日期（首次已知出版物）:02/2023

· 数量。参数:7B、13B、33B 和 65B

·语料库:英语 CommonCrawl + C4 + Github + 维基百科 + 古腾堡
和 Books3 + ArXiv + Stack Exchange

· 许可证:有限、非商业定制许可证

· 实验室:元

A.52 mBART

· 参考文献：（Liu 等人,2020）

·链接： https: //huggingface.co/docs/transformers/model_doc/mbart

· 家庭:BART

· 预训练架构:编码器/解码器

· 预训练任务:DAE

· 扩展:将 BART 扩展到多语言功能

· 应用:翻译

·（首次已知出版物的日期）:01/2020

· 数量。参数:与 BART 相同

·语料库:CC25 语料库包括25 个不同语言的单语语料库。最大的语料库是英语 (300 GB) 和俄语 (280 GB)

· 许可证:开放、MIT 许可证

· 实验室:脸书

A.53 威震天

· 参考文献：（Shoeybi 等人,2019）

· 链接:https: //github.com/NVIDIA/Megatron-LM

· 系列:GPT/BERT/T5

·预训练架构:编码器或解码器,取决于基础
模型

· 预训练任务:与基础模型相同

·扩展:Megatron 是一系列模型,通过引入模型并行原语来扩展先前已知的架构（即最初的 GPT-2 和 BERT,以及最近的 T5）。就 BERT 而言,作者还用句子顺序预测替换了下一个句子预测头,并使用全词 n-gram 掩码。

· 应用:与基本型号相同

· 日期（首次已知出版物）:03/2020

· 数量。参数:8.3B（类似 GPT）、3.9B（类似 BERT）

·语料库:原始论文使用由维基百科组成的聚合数据集）,
CC-Stories)、RealNews 和 OpenWebtext

· 许可证:有限、非商业用途

· 实验室:NVidia

A.54 密涅瓦

· 参考文献:（Lewkowycz 等人,2022）

· 链接:参见博客文章31

· 家庭:PaLM

· 预训练架构:解码器

· 预训练任务:LM

· 扩展:通过对数学数据集进行微调来扩展PaLM

· 应用:数学推理

· 日期（首次已知出版物）:06/2022

· 数量。参数:540B

·语料库:与 PaLM 相同 + 来自 arXiv 预印本服务器的 118GB 科学论文数据集以及包含使用 LaTeX、MathJax 或其他数学排版格式的数学表达式的网页

· 许可证:闭源

· 实验室:谷歌

31https://ai.googleblog.com/2022/06/minerva-solving-quantitative-reasoning.html

A.55 MT-NLG（威震天图灵NLG）

· 参考资料：（Smith 等人，2022） ·

链接：参见博客文章32

· 系列：GPT

· 预训练架构：解码器

· 预训练任务：LM

·扩展：使用类似于威震天的并行化来训练 LM 双精度
GPT-3的大小

· 应用：语言生成及其他（类似于 GPT-3） · （第一个已知出版物的日期）：10/2021

· 数量。参数：530B

· 语料库：Pile33 （800GB 数据集）+ 2 个常见爬网快照

· 许可证：有限、非商业用途

· 实验室：NVidia

A.56 OpenAssistant LLaMa

· 参考：不适用

· 链接：https: //open-assistant.io/

· 家庭： LLaMA

· 预训练架构：解码器

·扩展：对众包对话/助理进行监督微调
数据。

·应用程序：与ChatGPT 相同，但开源。与替代方案相比，它使用人类生成的对话数据

· 日期（首次已知出版物）：04/2023

· 数量。参数：LLaMa 为 30B

·语料库：志愿者收集的对话（Köpf 等人，2023）可在https://huggingface.co/datasets/OpenAssistant/
oasst1获取

·许可证：有限、非商业定制许可证。还有一个版本
基于 Apache 许可的 Pythia。

· 实验室：各种开源贡献者

32https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-
33https://arxiv.org/abs/2101.00027

A.57 选择

- 参考文献：（Zhang 等人,2022）
- 链接:参见博客文章34
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:与 GPT-3 基本相同的架构,但进行了一些训练 Megatron-LM 中引入的改进
- 应用:与GPT-3相同
- 日期（首次已知出版物）:05/2022
- 数量。参数:175B（和其他较小的版本）
- 语料库:180B 代币 = RoBERTa + Pile + PushShift.io Reddit
- 许可证:有限的非商业许可证
- 实验室:Facebook

A.58 手掌

- 参考文献：（Chowdhery 等人,2022）
- 链接:参见博客文章35
- 系列:变压器
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:Palm 使用典型的仅解码器 Transformer 架构,但添加了相当多的扩展:SwiGLU 激活、并行层、多查询注意力、RoPE 嵌入、共享输入输出嵌入、无偏差以及从生成的 256k SentencePiece 词汇表训练数据。
- 应用:PalM 被设计为通用语言模型，适用于数百种不同的语言任务
- 日期（首次已知出版物）:04/2022

34https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/
35https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to。
html

- 数量。参数:540B
- 语料库:来自过滤网页、书籍、维基百科、新闻文章、源代码和社交媒体对话的 780B 代币。代码包括 24 种编程语言。
- 许可证:闭源,可通过 API 访问
- 实验室:谷歌

A.59 飞马座

- 参考文献:(Zhang 等人,2020)
- 链接: https: //huggingface.co/docs/transformers/model_doc/pegasus
- 系列:变压器
- 预训练架构:编码器/解码器
- 预训练任务:DAE(更具体地说是 GSG)和 MLM
- 扩展:通过使用更适合摘要的不同预训练任务(GSG:间隙句子生成)来扩展vanilla Transformer
- 应用:总结
- 日期(首次已知出版物):12/2019
- 数量。参数:基础 = 223M,大 = 568M
- 语料库:C4 (750GB) + HugeNews (3.8 TB)
- 许可证:不适用
- 实验室:伦敦大学学院/谷歌

A.60 皮提亚

- 参考文献: (Biderman 等人,2023)
- 链接:https: //github.com/EleutherAI/pythia
- 家族:皮提亚
- 预训练架构:解码器
- 扩展:使用 GPT-NeoX 库进行训练
- 应用:研究语言模型的行为、功能和限制。

- 日期（首次已知出版物）:04/2023
- 数量。参数:70M、160M、410M、1B、1.4B、2.8B、6.9B、12B
- 语料库:桩
- 许可证:开放、Apache-2.0
- 实验室:Eleuther AI

A.61 罗伯塔

- 参考文献：（Liu 等人,2019）
- 链接： https: //huggingface.co/docs/transformers/model_doc/roberta
- 家族:BERT
- 预训练架构:编码器
- 预训练任务:MLM（动态）
- 扩展:通过优化训练程序来扩展 BERT
  更多数据
- 应用:与BERT相同
- 日期（首次已知出版物）:07/2019
- 数量。参数:356M
- 语料库:与 BERT + CC News + OpenWebText + Stories 相同（33B
  代币）
- 许可证:不适用
- 实验室:华盛顿大学/谷歌

A.62 搜寻者

- 参考文献：（Shuster 等人,2022a）
- 链接:https: //parl.ai/projects/seeker
- 家庭:GPT（但可以扩展任何家庭）
- 预训练架构:编码器/解码器或仅解码器,具体取决于其扩展的基本模型
- 预训练任务:LM 训练、对话训练
- 扩展:SeeKer 是一个扩展,通过引入预训练期间引入的“搜索”、“知识”和“响应”模块,可以应用于任何
  Transformer架构

- 应用:与基本型号相同
- 日期（首次已知出版物）:03/2022
- 数量。参数:SeeKeR 对话:400M、3B； SeeKeR LM:365M、762M、 1.5B、R2C2 BlenderBot:400M、3B
- 语料库:互联网向导/维基百科、PersonaChat、混合技能谈话、同理心对话、多会话聊天、MS MARCO、自然问题、SQuAD、TriviaQA
- 许可证:代码是开源的。
- 实验室:Facebook

A.63麻雀

- 参考文献： （Glaese 等人,2022）
- 链接:不适用
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:从Chinchilla 70B 模型开始,但添加了RLHF（带有人类反馈的强化学习）。它还添加了GopherCite 的内联证据
- 应用程序:对话代理和通用语言生成应用程序喜欢问答
- 日期（首次已知出版物）:09/2022
- 数量。参数:70B
- 语料库:与 Chinchilla 相同 + 与人类交互数据收集 RLHF 过程中的注释者
- 许可证:闭源
- 实验室:Deepmind

A.64 稳定扩散

- 参考文献：（Rombach 等人,2022）
- 链接:https: //huggingface.co/CompVis/stable-diffusion
- 家族:扩散
- 预训练架构:编码器/解码器
- 预训练任务:字幕预测
- 扩展:稳定扩散基本上是由慕尼黑大学研究人员开发的潜在扩散模型+来自 DALL-e 和 Imagen 的一些关于条件扩散的学习
- 应用:文本转图像
- 日期（首次已知出版物）:12/2021
- 数量。参数:890M（尽管有不同的、更小的变体）
- 语料库:LAION-5B,源自 Common 的公开数据集
  爬行
- 许可证:开放、CreativeML Open RAIL++-M 许可证
- 实验室:慕尼黑大学 + Stability.ai + Eleuther.ai

A.65 Swin 变压器

- 参考文献：（Liu 等人,2021）
- 链接:https: //github.com/microsoft/Swin-Transformer
- 家庭:ViT
- 预训练架构:编码器
- 预训练任务:与ViT相同
- 扩展:通过使用基于移位窗口 (Swin) 的模块替换标准多头自注意力 (MSA) 模块来扩展 ViT,从而允许

  类似 ViT 的架构可推广到更高分辨率的图像
- 应用:图像（对象检测、图像分类..） · 日期（第一个已知出版物）:03/2021
- 数量。参数:29M-197M
- 语料库:Imagenet 和 Imagenet-22k
- 许可证:代码开源,具有 MIT 许可证
- 实验室:微软

A.66 开关

- 参考文献： （Fedus 等人,2021）
- 链接:https: //github.com/google-research/t5x
- 家族:T5
- 预训练架构:编码器/解码器
- 预训练任务:DAE
- 扩展:目标是通过使用 MoE（专家混合)的高效路由来增加参数数量,同时保持 FLOP 操作恒定
- 应用:一般语言任务（例如回答问题）
- 日期（首次已知出版物）:01/2021
- 数量。参数:1T
- 语料库:庞大的干净爬行语料库
- 许可证:开放、Apache-2.0
- 实验室:谷歌

A.67 T0

- 参考文献： （Sanh 等人,2021）
- 链接:https: //huggingface.co/bigscience/T0
- 家族:T5
- 预训练架构:编码器/解码器
- 微调任务:自然语言提示
- 扩展:T0 代表“T5 for Zero Shot”,是通过在涵盖许多不同 NLP 任务的多任务混合上微调 T5 模型而获得的。
  与T0相比,T0p和T0pp用更多的数据集进行了微调。
  推荐使用 T0pp,因为它（平均)在各种 NLP 任务上都能带来最佳性能。
- 应用:通过用自然语言指定查询来执行零样本推理任务,模型将生成预测。
- 日期（首次已知出版物）:03/2022
- 数量。参数:T0-3B:30亿,T0、T0p、T0pp:110亿

·语料库:T0（多项选择 QA、抽取式 QA、闭卷 QA、结构到文本、情感、总结、主题分类、短语识别。T0p（与 T0 相同,带有来自GPT-3 评估的附加数据集） suite). T0pp（与 T0p 相同,带有来自 SuperGLUE 的附加数据集,不包括 NLI 集）

· 许可证:开放、Apache-2.0

· 实验室:BigScience

A.68 T5

· 参考文献：（Raffel 等人,2020）

· 链接:https: //huggingface.co/docs/transformers/model_doc/t5

· 系列:变压器

· 预训练架构:编码器/解码器

· 预训练任务:DAE

·扩展:与原始 Transformer 相同,但添加了一些内容,例如
相对位置嵌入,例如 Transformer XL

·应用:一般语言任务,包括机器翻译、问答、抽象概括和文本分类

· 日期（首次已知出版物）:10/2019

· 数量。参数:11 B（最多）

·语料库:巨大的干净爬行语料库 (C4) 清理版本
Common Crawl 数据集 750 GB

· 许可证:开放、Apache-2.0

· 实验室:谷歌

A.69 轨迹变压器

· 参考文献： （Janner 等人,2021）

· 链接:https: //trajectory-transformer.github.io

·系列:GPT,控制 Transformers”（本身不是一个系列,而是将那些尝试对更通用的控制、类 RL任务进行建模的 Transformer 分组）

· 预训练架构:解码器

· 预训练任务:预测最可能的序列

·扩展:与决策变压器类似,轨迹变压器引入的主要扩展是一种对轨迹（状态、动作、奖励)进行编码的方法

· 应用:通用强化学习（强化学习任务）

· 日期（首次已知出版物）:06/2021

· 数量。 Params:比 GPT 更小的架构

·语料库:D4RL 数据集和其他 RL 数据集,具体取决于任务
手

· 许可证:开放、MIT 许可证

· 实验室:加州大学伯克利分校

A.70 变压器 XL

· 参考文献：（Dai 等人,2019）

·链接： https://huggingface.co/docs/transformers/model_doc/transfo-xl

· 家庭:

· 预训练架构:解码器

· 预训练任务:LM

· 扩展:相对定位的嵌入可以实现更长的上下文关注
与普通 Transformer 模型相比

· 应用:一般语言任务

· 日期（首次已知出版物）:01/2019

· 数量。参数:151M

·语料库:根据实验不同的训练数据集,但基础
行是 Wikitext-103

· 许可证:不适用

· 实验室:CMU/Google

A.71 图灵-NLG

- 参考文献： （Rosset,2020）
- 链接:不适用
- 系列:GPT
- 预训练架构:解码器
- 预训练任务:LM
- 扩展:具有最佳超参数的 GPT2 优化版本
  和软件/硬件平台来改进培训
- 应用:与GPT-2/3相同
- 日期（首次已知出版物）:02/2020
- 数量。参数:最初为 17B,最近提高到 530B
- 语料库:来自 The Pile 的最高质量子集 + 2 个 CC 快照 (339B
  代币）
- 许可证:不适用
- 实验室:微软

A.72 UL2

- 参考文献： （Tay 等人,2022）
- 链接： https: //github.com/google-research/google-research/tree/
  主控/ul2
- 系列:变压器
- 预训练架构:编码器/解码器
- 预训练任务:混合降噪器,结合了不同的预训练任务
  一起训练范式
- 扩展:UL2-20B（统一语言学习)可以解释为与T5 非常相似的模型,但使用不同的目标和略有不同的缩放旋
  钮进行训练。
- 应用:预训练模型的统一框架
  跨数据集和设置普遍有效。
- 日期（首次已知出版物）:05/2022
- 数量。参数:20B

- Corpus:C4 上有 1 万亿个代币
- 许可证:开放、Apache-2.0
- 实验室:谷歌

A.73 骆马毛

- 参考:不适用
- 链接:https: //vicuna.lmsys.org
- 家庭： LLaMA
- 预训练架构:解码器
- 微调任务:人工指令
- 扩展:LLaMA 对从ShareGPT 收集的用户共享对话进行了微调。
- 应用程序:与ChatGPT 相同
- （第一个已知出版物的日期）:03/2023
- 数量。参数:13B
- 语料库:从 ShareGPT 收集的对话
- 许可证:有限、非商业定制许可证
- 实验室:加州大学伯克利分校、卡内基梅隆大学、斯坦福大学、加州大学圣地亚哥分校和 MBZUAI

A.74 维特

- 参考文献：（Dosovitskiy 等人,2020）
- 链接:https: //huggingface.co/docs/transformers/model_doc/vit
- 家族:BERT
- 预训练架构:编码器
- 预训练任务:图像分类
- 扩展:BERT 架构的扩展以训练图像补丁
- 应用:图像分类
- 日期（首次已知出版物）:10/2020 · 编号。参数:86M（基本)到 632M（巨大）
- 语料库:从标准 Imagenet 到 JFT-300M（大型内部数据集）
- 许可证:不适用
- 实验室:谷歌

A.75五刀2.0

- 参考:参见维基百科第36页
- 链接:参见博客文章37
- 系列:GLM（通用语言模型）
- 预训练架构:解码器
- 预训练任务:自回归空白填充
- 扩展:与GPT 类似,它使用解码器/自回归架构,但应用GLM模型系列中提出的不同预训练任务。此外,Wu Dao 使用了专家的快速混合（参见https://github.com/laekov/fastmoe)数万亿参数的规模训练方法
- 应用:语言和多模态（特别是图像）
- 日期（首次已知出版物）:06/2021
- 数量。参数:1.75T
- 语料库:4.9 TB 的高质量图像和文本（英语和英语）中国人
- 许可证:闭源
- 实验室:北京人工智能研究院

A.76 XLM-罗伯塔

- 参考文献：（Conneau 等人,2019）
- 链接： https: //huggingface.co/docs/transformers/model_doc/xlm-roberta
- 家人:罗伯塔
- 预训练架构:编码器
- 预训练任务:MLM（动态）
- 扩展:RoBERTa 的扩展,引入了参数调整多语言应用背景下的见解
- 应用:翻译和其他跨语言的语言任务
- 日期（首次已知出版物）:10/2019
- 数量。参数:基础 = 270M,大 = 550M

36https://en.wikipedia.org/wiki/Wu_Dao
37https://mp.weixin.qq.com/s/BUQWZ5EdR19i40GuFofpBg

- 语料库:清理了 100 种语言的 Common Crawl
- 许可证:开放、MIT 许可证
- 实验室:Facebook

A.77 XLNet

- 参考文献:(Yang 等人,2019)
- 链接: https: //huggingface.co/docs/transformers/model_doc/xlnet
- 系列:Transformer XL
- 预训练架构:解码器
- 预训练任务:PLM
- 扩展:该模型基本上将 Transformer XL 架构适应于基于排列的 LM
- 应用:一般语言任务
- 日期(首次已知出版物):05/2019
- 数量。参数:Base=117M,Large=360M
- 语料库:与 BERT + Giga5(16GB 文本)相同 + 并进行积极过滤 ClueWeb 2012-B (19GB)、普通爬网 (110GB)
- 许可证:开放、MIT 许可证
- 实验室:CMU/Google

参考

Aghajanyan, A.、Gupta, A.、Shrivastava, A.、Chen, X.、Zettlemoyer, L. 和 Gupta, S. (2021)。 Muppet:具有预微调的大规模多任务表示。 https://arxiv.org/abs/2101.11038。

Aghajanyan, A.、黄 B.、罗斯 C.、卡普欣 V.、徐 H.、戈亚尔 N.、奥洪科D.、乔希 M.、戈什 G.、刘易斯 M.、等人。 (2022)。 Cm3:互联网的因果屏蔽多模式模型。 https://arxiv.org/abs/2201.07520。

Aghajanyan, A.、Okhonko, D.、Lewis, M.、Joshi, M.、Xu, H.、Ghosh, G. 和Zettlemoyer, L. (2021)。 Htlm:语言模型的超文本预训练和提示。 https://arxiv.org/abs/2107.06955。

Alayrac, J.-B.、Donahue, J.、Luc, P.、Miech, A.、Barr, I.、Hasson, Y.、Lenc, K.、Mensch, A.、Millican, K.、Reynolds,金属。 (2022)。 Flamingo:用于小样本学习的视觉语言模型。 https:// arxiv.org/abs/ 2204.14198。

Askell, A.、Bai, Y.、Chen, A.、Drain, D.、Ganguli, D.、Henighan, T.、Jones, A.、Joseph, N.、Mann, B.、DasSarma, N. ,等人。 （2021）。通用语言助手作为对齐实验室。 https://arxiv.org/abs/2112。 00861。

Bai, Y.、Jones, A.、Ndousse, K.、Askel, A.、Chen, A.、DasSarma, N.、Drain, D.、Fort, S.、Ganguli, D.、Henighan, T.,约瑟夫,N.、卡达瓦斯,S.、凯尼恩， J.、康纳利,T.、埃尔-肖克，S.、埃尔哈格,N.、哈特菲尔德-多兹,Z.、埃尔南德斯,D.、休姆,T.、约翰斯顿、S.、Kravec, S.、Lovitt, L.、Nanda, N.、Olsson, C.、Amodei, D.、Brown, T.、Clark, J.、McCandlish, S.、Olah, C.、Mann ， B.,&卡普兰,J.（2022a）。通过人类反馈的强化学习来训练一个有用且无害的助手。 https://arxiv.org/abs/2204.05862 。

Bai, Y.、Kadavath, S.、Kundu, S.、Askel, A.、Kernion, J.、Jones, A.、Chen, A.、Goldie, A.、Mirhoseini, A.、McKinnon, C.、 Chen, C.、Olsson, C.、Olah, C.、 Hernandez, D.、Drain, D.、Ganguli, D.、Li, D.、Tran-Johnson, E.、Perez, E.、 Kerr, J .、Mueller, J.、Ladish, J.、Landau, J.、Ndousse, K.、Lukosuite, K.、 Lovitt, L.、Sellitto, M.、Elhage, N.、Schiefer, N.、Mercado, N. .、DasSarma, N.、Lasenby, R.、Larson, R.、Ringer, S.、Johnston, S.、Kravec, S.、Showk, SE、Fort, S.、Lanham, T.、Telleen-Lawton, T.、Conerly, T.、Henighan, T.、Hume, T.、Bowman, SR、Hatfield-Dodds, Z.、Mann, B.、Amodei, D.、Joseph, N.、 McCandlish, S.、Brown 、T. 和 Kaplan, J. (2022b)。宪法人工智能：人工智能反馈的无害性。 https://arxiv.org/abs/2212.08073。

Baidoo-Anu, D. 和 Owusu Ansah, L. (2023)。生成人工智能 (AI)时代的教育：了解 ChatGPT 在促进教学方面的潜在优势。 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4337484 。

Biderman, S.、Schoelkopf, H. 和 QA (2023)。 Pythia：用于分析跨训练和扩展的大型语言模型的套件。 https://arxiv.org/abs/2304.01373 。

Black, S.、Biderman, S.、Hallahan, E.、Anthony, Q.、Gao, L.、Golding, L.、He, H.、Leahy, C.、McDonell, K.、Phang, J.、等人。 （2022）。 Gpt-neox-20b：一种开源自回归语言模型。 https://arxiv.org/abs/2204.06745 。

Bommasani, R.、Hudson, DA、Adeli, E.、Altman, R.、Arora, S.、von Arx, S.、 Bernstein, MS、Bohg, J.、Bosselut, A.、Brunskill, E. 等等人。 （2021）。论基础模型的机遇与风险。 https://arxiv.org/abs/2108.07258 。

Brown, T.、Mann, B.、Ryder, N.、Subbiah, M.、Kaplan, JD、Dhariwal, P.、Neelakantan, A.、 Shyam, P.、Sastry, G.、Askel, A. 等等人。 （2020）。语言模型是小样本学习者。神经信息处理系统的进展,33,1877-1901。

Chen, L.、Lu, K.、Rajeswaran, A.、Lee, K.、Grover, A.、Laskin, M.、Abbeel, P.、 Srinivas, A. 和 Mordatch, I. (2021)。决策转换器：通过序列建模强化学习。神经信息处理系统的进展，34，15084–15097。

Cho, K.、Van Merriënboer, B.、Bahdanau, D. 和 Bengio, Y. (2014)。关于神经机器翻译的特性：编码器-解码器方法。 https://arxiv.org/abs/1409.1259。

Chowdhery, A.、Narang, S.、Devlin, J.、Bosma, M.、Mishra, G.、Roberts, A.、Barham, P.、Chung, HW、Sutton, C.、Gehrmann, S. 等等人。 （2022）。 Palm：通过路径扩展语言建模。 https://arxiv.org/abs/2204。 02311。

Christiano, P.、Leike, J.、Brown, TB、Martic, M.、Legg, S. 和 Amodei, D.（2023）。根据人类偏好进行深度强化学习。 https://arxiv.org/abs/1706.03741 。

Chung, HW, Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson , A., Gu, SS, Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D. , Narang, S.、Mishra, G.、Yu, A.、Zhao, V.、Huang, Y.、Dai, A.、 Yu, H.、Petrov, S.、Chi, EH、Dean, J.、Devlin ，J.，罗伯茨，A.，周， D.，乐，QV，&魏，J.（2022）。扩展指令微调语言模型。 https://arxiv.org/abs/2210.11416。

Clark, K.、Luong, M.-T.、Le, QV 和 Manning, CD (2020)。 Electra：将文本编码器预训练为判别器而不是生成器。 https://arxiv.org/abs/2003.10555 。

Conneau, A.、Khandelwal, K.、Goyal, N.、Chaudhary, V.、Wenzek, G.、Guzmán, F.、 Grave, E.、Ott, M.、Zettlemoyer, L. 和 Stoyanov, V. （2019）。大规模无监督跨语言表征学习。 https://arxiv.org/abs/1911.02116 。

Dai, Z.、Yang, Z.、Yang, Y.、Carbonell, J.、Le, QV 和 Salakhutdinov, R. (2019)。Transformer-xl：超越固定长度上下文的细心语言模型。 https://arxiv.org/abs/1901.02860。

Dennean, K.、Gantori, S.、Limas, DK 和 Allen Pu, a。 RG（2023）。 https://www.ubs.com/global/en/我们来聊聊ChatGPT。财富管理/our-approach/marketnews/article.1585717。 html。

Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. (2018)。 Bert：用于语言理解的深度双向转换器的预训练。 https://arxiv.org/abs/1810.04805 。

Dosovitskiy, A.、Beyer, L.、Kolesnikov, A.、Weissenborn, D.、Zhai, X.、Unterthiner, T.、Dehghani, M.、Minderer, M.、Heigold, G.、Gelly, S.,等人。 （2020）。一张图像相当于 16x16 个单词:用于大规模图像识别的 Transformer。 https://arxiv.org/abs/2010.11929。

杜娜、黄 Y.、戴 AM、童 S.、Lepikhin, D.、徐 Y.、Krikun, M.、周 Y.、于 AW、Firat, O. 等。（2022a）。 Glam:专家混合的语言模型的有效扩展。国际机器学习会议,第 5547-5569 页。 PMLR。

杜Z.、钱Y.、刘X.、丁明、邱J.、杨Z.和唐J. (2022b)。 GLM:具有自回归空白填充的通用语言模型预训练。 https://arxiv.org/abs/2103.10360。

Esser, P.、Rombach, R. 和 Ommer, B. (2021)。驯服变压器以进行高分辨率图像合成。 IEEE/CVF 计算机视觉和模式识别会议记录,第 12873–12883 页。

Fedus, W.、Zoph, B. 和 Shazeer, N. (2021)。开关变压器:通过简单高效的稀疏性扩展到万亿参数模型。机器学习研究杂志,23,1-40。

Fuchs, FB、Worrall, DE、Fischer, V. 和 Welling, M. (2020)。 SE(3)-变形金刚:3D 旋转翻译等变注意网络。 https://arxiv.org/abs/2006.10503 。

Glaese, A.、McAleese, N.、Tr bacz, M.、Aslanides, J.、Firoiu, V.、Ewalds, T.、Rauh, M.、Weidinger, L.、Chadwick, M.、Thacker, P.,等人。 （2022）。通过有针对性的人类判断来改善对话代理的一致性。 https://arxiv.org/abs/2209.14375 。

Hatamizadeh, A.、Yin, H.、Kautz, J. 和 Molchanov, P. (2022)。全球背景视觉转换器。 https://arxiv.org/abs/2206.09959。

何平、刘X、高静、陈文（2021）。 DeBERTa:具有解纠缠注意力的解码增强型BERT。在国际学习表征会议上。

Hochreiter, S. 和 Schmidhuber, J. (1997)。长短期记忆。神经计算,9 (8),1735–1780。

霍夫曼,J.,博尔若,S.,门施,A.,布哈茨卡亚,E.,蔡,T.,卢瑟福， E.,卡萨斯,D.d. L.、Hendricks, LA、Welbl, J.、Clark, A. 等。 （2022）。
训练计算优化的大型语言模型。 https://arxiv.org/abs/2203.15556 。

Janner, M.、Li, Q. 和 Levine, S. (2021)。离线强化学习是一个大的序列建模问题。神经信息处理系统的进展,34,1273–1286。

Jumper, J.、Evans, R.、Pritzel, A.、Green, T.、Figurnov, M.、Ronneberger, O.、Tunyasuvunakool, K.、Bates, R.、 Zidek, A.、Potapenko, A.,等人。 （2021）。使用 AlphaFold 进行高度准确的蛋白质结构预测。自然， 596（7873）,583-589。

Keskar, NS、McCann, B.、Varshney, LR、Xiong, C. 和 Socher, R. (2019)。Ctrl：用于可控生成的条件转换器语言模型。 https://arxiv.org/abs/1909.05858。

Köpf, A.、Kilcher, Y.、von Rütte, D.、Anagnostidis, S.、Tam, Z.-R.、Stevens, K.、Barhoum, A.、Duc, NM、Stanley, O.、Nagyfi, R.、ES, S.、Suri, S.、 Glushkov, D.、Dantuluri, A.、Maguire, A.、Schuhmann, C.、Nguyen, H. 和Mattick, A. (2023)。 OpenAssistant Conversations – 大众化大语言模型对齐。 https://arxiv.org/abs/2304.07327。

Lan, Z.、Chen, M.、Goodman, S.、Gimpel, K.、Sharma, P. 和 Soricut, R. (2019)。ALBERT：用于语言表示自我监督学习的精简版 BERT 。 https://arxiv.org/abs/1909.11942。

Lewis, M.、Liu, Y.、Goyal, N.、Ghazvininejad, M.、Mohamed, A.、Levy, O.、Stoyanov, V. 和 Zettlemoyer, L. (2019)。 Bart：用于自然语言生成、翻译和理解的序列到序列去噪预训练。 https://arxiv.org/abs/1910.13461。

Lewkowycz, A.、Andreassen, A.、Dohan, D.、Dyer, E.、Michalewski, H.、Ramasesh, V.、Slone, A.、Anil, C.、Schlag, I.、Gutman-Solo, T .,等人。 （2022）。使用语言模型解决定量推理问题。 https://arxiv. org/abs/2206.14858。

Li, Z.、Wang, Z.、Tan, M.、Nallapati, R.、Bhatia, P.、Arnold, A.、Xiang, B. 和 Roth, D. (2022)。 DQ-BART：通过联合蒸馏和量化的高效序列到序列模型。 https://arxiv.org/abs/2203。 11239。

Lieber, O.、Sharir, O.、Lenz, B. 和 Shoham, Y. (2021)。 Jurassic-1：技术细节和评估。https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf 。–

刘X.、何平、陈文、高J. (2019)。用于自然语言理解的多任务深度神经网络。计算语言学协会第 57 届年会论文集，第 4487-4496 页，意大利佛罗伦萨。计算语言学协会。

Liu, Y.、Gu, J.、Goyal, N.、Li, X.、Edunov, S.、Ghazvininejad, M.、Lewis, M. 和Zettlemoyer, L. (2020)。神经机器翻译的多语言去噪预训练。计算语言学协会汇刊，8, 726–742。

Liu, Y.、Ott, M.、Goyal, N.、Du, J.、Joshi, M.、Chen, D.、Levy, O.、Lewis, M.、Zettlemoyer, L. 和 Stoyanov, V. （2019）。 Roberta：一种稳健优化的bert 预训练方法。 https://arxiv.org/abs/1907.11692。

刘Z.、林Y.、曹Y.、胡H.、魏Y.、张Z.、林S.和郭B. (2021)。 Swin 变压器：使用移动窗口的分层视觉变压器。 IEEE/CVF 国际计算机视觉会议记录，第 10012–10022 页。

Menick, J.、Trebacz, M.、Mikulik, V.、Aslanides, J.、Song, F.、Chadwick, M.、 Glaese, M.、Young, S.、Campbell-Gillingham, L.、Irving, G .，等人。 （2022）。 教授语言模型以支持带有经过验证的引用的答案。 https://arxiv.org/abs/2203.11147 。

Mikolov, T.、Karafiat, M.、Burget, L.、Cernocky, J. 和 Khudanpur, S. (2010)。 基于循环神经网络的语言模型。在语音间，卷。 2，第 1045–1048 页。

Nichol, A.、Dhariwal, P.、Ramesh, A.、Shyam, P.、Mishkin, P.、McGrew, B.、Sutskever, I. 和 Chen, M. (2021)。 Glide：使用文本引导的扩散模型生成和编辑逼真的图像。 https://arxiv.org/abs/2112.10741。

开放人工智能（2023）。 GPT-4 技术报告。 https://arxiv.org/abs/2303。 08774。

欧阳 L.、吴 J.、蒋 X.、阿尔梅达 D.、温赖特 CL、米什金 P.、张 C.、阿加瓦尔 S.、斯拉马 K.、雷 A. 等等人。 （2022）。训练语言模型遵循人类反馈的指令。 https://arxiv.org/abs/2203.02155 。

邱X.、孙T.、徐Y.、邵Y.、戴宁和黄X. (2020)。用于自然语言处理的预训练模型：一项调查。科学中国技术科学，63（10），1872-1897。

Radford, A.、Kim, JW、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askel, A.、Mishkin , P.、Clark, J. 等等人。 （2021）。从自然语言监督中学习可迁移的视觉模型。国际机器学习会议，第 8748–8763 页。 PMLR。

Radford, A.、Narasimhan, K.、Salimans, T.、Sutskever, I. 等。 （2018）。通过生成预训练提高语言理解。 https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf 。

Radford, A.、Wu, J.、Child, R.、Luan, D.、Amodei, D.、Sutskever，语言模型是无监督的https://I.，等人。 （2019）。多任务 paperswithcode.com/paper/学习者。

语言模型是无监督的多任务。

Rae, JW、Borgeaud, S.、Cai, T.、Millican, K.、Hoffmann, J.、Song, F.、Aslanides, J.、Henderson, S.、Ring, R.、Young, S. 等等人。 （2021）。扩展语言模型：来自训练 gopher 的方法、分析和见解。 https://arxiv.org/abs/2112.11446 。

Raffel, C.、Shazeer, N.、Roberts, A.、Lee, K.、Narang, S.、Matena, M.、Zhou, Y.、Li, W. 和 Liu, PJ (2020)。使用统一的文本到文本转换器探索迁移学习的局限性。机器学习研究杂志，21 (1), 5485–5551。

Ramesh, A.、Dhariwal, P.、Nichol, A.、Chu, C. 和 Chen, M. (2022)。具有剪辑潜在特征的分层文本条件图像生成。 https://arxiv.org/abs/2204.06125 。

Ramesh, A.、Pavlov, M.、Goh, G.、Gray, S.、Voss, C.、Radford, A.、Chen, M. 和Sutskever, I. (2021)。零样本文本到图像生成。国际机器学习会议，第 8821–8831 页。 PMLR。

里德，S.，佐尔纳，K.，帕里索托，E.，科尔梅纳雷霍，SG，诺维科夫，A.，巴特-马龙，G.，希门尼斯，M.，苏尔斯基，Y.，凯，J.，斯普林伯格，JT，等人。 （2022）。多才多艺的经纪人。 https://arxiv.org/abs/2205.06175。

Rombach, R.、Blattmann, A.、Lorenz, D.、Esser, P. 和 Ommer, B. (2022)。使用潜在扩散模型进行高分辨率图像合成。 IEEE/CVF 计算机视觉和模式识别会议论文集，第 10684–10695 页。

罗塞特，C.（2020）。 Turing-NLG：Microsoft 的 170 亿参数语言模型。 https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/ 。

Saharia, C.、Chan, W.、Saxena, S.、Li, L.、Whang, J.、Denton, E.、Ghasemipour, S. KS、Ayan、BK、Mahdavi、SS、洛佩斯、RG 等。 （2022）。具有深入语言理解的真实感文本到图像扩散模型。 https://arxiv.org/abs/2205.11487 。

Sanh, V.、Debut, L.、Chaumond, J. 和 Wolf, T. (2019)。 DistilBERT，BERT 的精炼版：更小、更快、更便宜、更轻。 https://arxiv. org/abs/1910.01108。

Sanh, V.、Webson, A.、Raffel, C.、Bach, SH 和 Lintang Sutawika, ea (2021)。多任务提示训练可实现零样本任务泛化。 https://arxiv.org/abs/2110.08207。

Shoeybi, M.、Patwary, M.、Puri, R.、LeGresley, P.、Casper, J. 和 Catanzaro, B.（2019）。 Megatron-lm：使用模型并行性训练数十亿参数语言模型。 https://arxiv.org/abs/1909.08053。

Shuster, K.、Komeili, M.、Adolphs, L.、Roller, S.、Szlam, A. 和 Weston, J. (2022a)。
寻求知识的语言模型:对话和提示完成的模块化搜索和生成。 https://arxiv.org/abs/
2203.13224。

Shuster, K.、Xu, J.、Komeili, M.、Ju, D.、Smith, EM、Roller, S.、Ung, M.、Chen, M.、Arora, K.、
Lane, J. 等等人。 （2022b）。 Blenderbot 3:一个部署的对话代理,不断学习负责任
地参与。 https://arxiv.org/abs/2208.03188 。

史密斯,S.,帕特瓦里,M.,诺里克,B.,勒格雷斯利,P.,拉杰班达里,S.,卡斯珀,J.,刘,Z.,普拉布胡莫耶,S.,泽维斯,G.,科
蒂坎蒂,V.,等人。 （2022）。使用deepspeed 和 megatron 训练大型生成语言模型 megatron-turing
nlg 530b。 https://arxiv.org/abs/2201.11990。

Soltan, S.、Ananthakrishnan, S.、FitzGerald, J.、Gupta, R.、Hamza, W.、Khan, H.、Peris, C.、
Rawls, S.、Rosenbaum, A.、Rumshisky, A., Prakash, CS、Sridhar, M.、Triefenbach,
F.、Verma, A.、Tur, G. 和 Natarajan, P. (2022)。 AlexaTM 20B:使用大规模多语言
Seq2Seq 模型的少样本学习。 https://arxiv.org/abs/2208.01448。

Stokel-Walker, C. 和 Noorden, RV (2023)。 ChatGPT 和生成人工智能对科学意味着什
么。 https://www.nature.com/articles/
d41586-023-00340-6。

Su, HS, Shi, WS, Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, NA,
Zettlemoyer, L., & Yu, T. （2022）。一台嵌入器,任何任务:指令微调文本嵌入。
https://arxiv.org/abs/2212。 09741。

Taori, R.、Gulrajani, I.、Zhang, T.、Dubois, Y.、Li, X.、Guestrin, C.、Liang, P.和 Hashimoto, TB
(2023)。斯坦福羊驼:遵循指令的LLaMA 模型。 https://github.com/tatsu-lab/
stanford_alpaca。

Tay, Y.、Dehghani, M.、Tran, VQ、Garcia, X.、Bahri, D.、Schuster, T.、Zheng, HS、Houlsby,
N. 和 Metzler, D. (2022)。统一语言学习范式。 https://arxiv.org/abs/2205.05131。

Taylor, R.、Kardas, M.、Cucurull, G.、Scialom, T.、Hartshorn, A.、Saravia, E.、Poulton, A.、
Kerkez, V. 和 Stojnic, R. (2022)。 GALACTICA:大型科学语言模型。 https://
arxiv.org/abs/2211.09085。

Thoppilan, R.、De Freitas, D.、Hall, J.、Shazeer, N.、Kulshreshtha, A.、Cheng, H.-T.、Jin,
A.、Bos, T.、Baker, L.、Du ,Y.,等人。 （2022）。 Lambda:对话应用程序的语言模
型。 https://arxiv.org/abs/2201.08239。

Touvron, H.、Lavril, T.、Izacard, G.、Martinet, X.、Lachaux, M.-A.、Lacroix, T.、 Rozière, B.、
Goyal, N.、Hambro, E.、Azhar, F.、Rodriguez, A.、Joulin, A.、 Grave, E. 和 Lample,
G. (2023)。 LLaMA:开放且高效的基础语言模型。 https://arxiv.org/abs/2302.13971。

Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, AN、 Kaiser, . 和 Polosukhin, I. (2017)。您所需要的就是关注。神经信息处理系统的进展,30。

王 B. 和小松崎 A. (2021)。 GPT-J-6B:60 亿参数的自回归语言模型。 https://github.com/kingoflolz/mesh-transformer-jax 。

王 L.、杨 N.、黄 X.、焦 B.、杨 L.、江 D.、Majumder, R. 和魏 F. (2022)。通过弱监督对比预训练进行文本嵌入。 https://arxiv.org/abs/2212.03533。

杨丽、张志、宋勇、洪胜、徐荣、赵勇、邵勇、张文、崔斌、杨明。 -H。 （2022）。扩散模型:方法和应用的全面调查。 https://arxiv.org/abs/2209.00796。

Yang, Z.、Dai, Z.、Yang, Y.、Carbonell, J.、Salakhutdinov, RR 和 Le, QV （2019）。 Xlnet:语言理解的广义自回归预训练。神经信息处理系统的进展,32。

Zaheer, M.、Guruganesh, G.、Dubey, KA、Ainslie, J.、Alberti, C.、Ontanon, S.、Pham, P.、Ravula, A.、Wang, Q.、Yang , L. 等等人。 （2020）。大鸟:用于较长序列的变形金刚。神经信息处理系统的进展,33,17283–17297。

张静、赵云、萨利赫 M. 和刘 P. (2020)。 Pegasus:使用提取的间隙句子进行预训练以进行抽象总结。国际机器学习会议,第 11328–11339 页。 PMLR。

张,S.,罗勒,S.,戈亚尔,N.,Artetxe,M.,陈,M.,陈,S.,德万,C.,迪亚布,M.,李,X.,林,XV,等等人。 （2022）。 Opt:打开预先训练的Transformer 语言模型。 https://arxiv.org/abs/2205.01068。

张 Y.、孙 S.、Galley, M.、陈 Y.-C.、Brockett, C.、高 X.、高 J.、刘 J. 和 Dolan, B. (2019a ）。 Dialogpt:用于对话响应生成的大规模生成预训练。 https://arxiv.org/abs/1911。 00536。

张Z.,韩X.,刘Z.,蒋X.,孙明,&刘Q.（2019b）。 ERNIE:通过信息实体增强语言表示。 https://arxiv.org/abs/1905.07129 。