学习 ChatGPT 必看的 10 篇论文

1.Transformer

ChatGPT 使用的预训练模型 GPT，是在 Transformer 中的 decoder 基础上进行改造的。

论文标题：Attention Is All You Need

论文链接：https://arxiv.org/pdf/1706.03762.pdf

摘要：占主导地位的序列转导模型是基于复杂的递归或卷积神经网络，包括一个编码器和一个解码器。性能最好的模型还通过注意机制将编码器和解码器连接起来。我们提出了一个新的简单的网络结构–Transformer，它只基于注意力机制，完全不需要递归和卷积。在两个机器翻译任务上的实验表明，这些模型在质量上更胜一筹，同时也更容易并行化，需要的训练时间也大大减少。我们的模型在WMT 2014英德翻译任务中达到了28.4 BLEU，比现有的最佳结果（包括合集）提高了2 BLEU以上。在WMT 2014英法翻译任务中，我们的模型在8个GPU上训练了3.5天后，建立了新的单模型最先进的BLEU得分，即41.0分，这只是文献中最佳模型的训练成本的一小部分。

2.GPT-3

GPT 家族与 BERT 模型都是知名的 NLP 预训练模型，都基于 Transformer 技术。GPT-1 只有12个 Transformer 层，而到了 GPT-3，则增加到 96 层。

论文标题：Language Models are Few-Shot Learners

论文链接：https://arxiv.org/pdf/2005.14165.pdf

摘要：最近的工作表明，在许多NLP任务和基准上，通过对大型文本语料库进行预训练，然后对特定的任务进行微调，可以获得巨大的收益。虽然在结构上通常是任务无关的，但这种方法仍然需要特定任务的微调数据集，包括几千或几万个例子。相比之下，人类通常只需通过几个例子或简单的指令就能完成一项新的语言任务–而目前的NLP系统在很大程度上仍难以做到这一点。在这里，我们展示了扩大语言模型的规模，大大改善了与任务无关的、少量的性能，有时甚至达到了与之前最先进的微调方法的竞争力。具体来说，我们训练了GPT-3，一个具有1750亿个参数的自回归语言模型，比以前的任何非稀疏语言模型多10倍，并测试了它在少数情况下的性能。对于所有的任务，GPT-3的应用没有任何梯度更新或微调，纯粹通过与模型的文本互动来指定任务和少量演示。GPT-3在许多NLP

数据集上取得了强大的性能，包括翻译、回答问题和cloze任务，以及一些需要即时推理或领域适应的任务，如解读单词、在句子中使用一个新词或进行3位数的算术。同时，我们也发现了一些数据集，在这些数据集中，GPT-3的几率学习仍然很困难，还有一些数据集，GPT-3面临着与大型网络语料库训练有关的方法学问题。最后，我们发现，GPT-3可以生成人类评价者难以区分的新闻文章样本。我们讨论了这一发现和GPT-3总体上的更广泛的社会影响。

3.InstructGPT

ChatGPT 的训练流程，主要参考自 instructGPT ，ChatGPT 是改进的 instructGPT。

论文标题：Training language models to follow instructions with human feedback

论文链接：https://arxiv.org/pdf/2203.02155.pdf

摘要：让语言模型变得更大并不意味着它们能更好地遵循用户的意图。例如，大型语言模型可以产生不真实的、有毒的或根本对用户没有帮助的输出。换句话说，这些模型没有与用户保持一致。在本文中，我们展示了一个途径，通过人类反馈的微调，在广泛的任务中使语言模型与用户的意图保持一致。从一组标签员写的提示语和通过OpenAI API提交的提示语开始，我们收集了一组标签员演示的所需模型行为的数据集，我们利用监督学习对GPT-3进行微调。然后，我们收集模型输出的排名数据集，我们利用人类反馈的强化学习来进一步微调这个监督模型。我们把产生的模型称为InstructGPT。在人类对我们的提示分布的评估中，尽管参数少了100倍，但1.3B参数的InstructGPT模型的输出比175B的GPT-3的输出更受欢迎。此外，InstructGPT模型显示了真实性的改善和有毒输出生成的减少，同时在公共NLP数据集上的性能回归最小。尽管InstructGPT仍然会犯一些简单的错误，但我们的结果表明，利用人类反馈进行微调是使语言模型与人类意图相一致的一个有希望的方向。

4.Sparrow

DeepMind 的 Sparrow，这个工作发表时间稍晚于 instructGPT，其大致的技术思路和框架与 instructGPT 的三阶段基本类似，不过明显Sparrow 在人工标注方面的质量和工作量是不如instructGPT的。反过来，Sparrow 中把奖励模型分为两个不同 RM 的思路，理论上是优于instructGPT 的。

论文标题：Improving alignment of dialogue agents via targeted human judgements

论文链接：https://arxiv.org/pdf/2209.14375.pdf

摘要：我们提出了Sparrow，一个寻求信息的对话代理，与提示的语言模型基线相比，它被训练得更有帮助，更正确，更无害。我们使用来自人类反馈的强化学习来训练我们的模

型，并增加了两个新的内容来帮助人类评分者判断代理行为。首先，为了使我们的代理更有帮助和无害，我们将良好对话的要求分解为代理应该遵循的自然语言规则，并分别询问评分者每条规则。我们证明，这种分解使我们能够收集更有针对性的人类对代理行为的判断，并允许更有效的规则条件的奖励模型。第二，我们的代理在收集对模型声明的偏好判断时，提供支持事实性要求的来源的证据。对于事实问题，麻雀提供的证据在78%的情况下支持采样的反应。Sparrow比基线更经常受到青睐，同时对人类的对抗性探测更有弹性，在被探测时只有8%的时间违反了我们的规则。最后，我们进行了广泛的分析，表明尽管我们的模型学会了遵循我们的规则，但它会表现出分布性的偏差。

5.RLHF

InstructGPT/GPT3.5（ChatGPT的前身）与 GPT-3 的主要区别在于，新加入了被称为 RLHF（Reinforcement Learning from Human Feedback，人类反馈强化学习）。这一训练范式增强了人类对模型输出结果的调节，并且对结果进行了更具理解性的排序。

论文标题：Augmenting Reinforcement Learning with Human Feedback

论文链接：https://www.cs.utexas.edu/~ai-lab/pubs/ICML_IL11-knox.pdf

摘要：随着计算代理越来越多地被用于研究实验室之外，它们的成功将取决于它们学习新技能和适应其动态、复杂环境的能力。如果人类用户–没有编程技能–能够将他们的任务知识转移给代理，那么学习就会大大加快，减少昂贵的试验。TAMER框架指导代理人的设计，其行为可以通过批准和不批准的信号来塑造，这是人类反馈的一种自然形式。最近，TAMER+RL被引入，使人类反馈能够增强传统的强化学习（RL）代理，该代理从马尔科夫决策过程（MDP）的奖励信号中学习。通过对TAMER和TAMER+RL的重新实现，我们解决了先前工作的局限性，在两个关键方向上做出了贡献。首先，我们在第二个任务上测试了先前TAMER+RL工作中结合人类强化和RL的四种成功技术，并分析了这些技术对参数变化的敏感性。这些检查共同产生了更多的一般性和规范性的结论，以指导那些希望将人类知识纳入RL算法的其他人。第二，TAMER+RL到目前为止仅限于顺序设置，即在从MDP奖励中学习之前发生训练。我们对顺序算法进行了修改，使其能够同时从两个来源进行学习，从而使人类的反馈能够在强化学习过程中的任何时候出现。为了实现同步学习，我们引入了一种新的技术，适当地确定人类模型在整个时间和状态动作空间对RL算法的影响程度。

6.TAMER

ChatGPT 中的 TAMER（Training an Agent Manually via Evaluative Reinforcement，评估式强化人工训练代理）框架，将人类标记者引入到 Agents 的学习循环中，可以通过人类向 Agents 提供奖励反馈（即指导 Agents 进行训练），从而快速达到训练任务目标。

论文标题：Interactively Shaping Agents via Human Reinforcement

论文链接：https://www.cs.utexas.edu/~bradknox/papers/kcap09-knox.pdf

摘要：随着计算学习代理进入产生实际成本的领域（例如，自动驾驶或金融投资），有必要在没有大量高成本学习试验的情况下学习好的政策。减少学习任务的样本复杂性的一个有希望的方法是将知识从人类转移到代理人。理想情况下，转移的方法应该是任何拥有任务知识的人都可以使用的，不管这个人在编程和人工智能方面的专业知识如何。本文的重点是允许人类培训师通过强化信号互动地塑造一个代理的政策。具体来说，本文介绍了“通过评估性强化训练代理”，即tamer，一个能够实现这种塑造的框架。与以前的交互式塑造方法不同，tamer代理对人类的强化进行建模，并通过选择预期会得到最多强化的行动来利用其模型。来自两个领域的结果表明，非专业人员可以在不定义环境奖励函数（如MDP）的情况下训练驯兽师代理，并表明在驯兽师框架内的人类训练可以比自主学习算法降低样本的复杂性。

7.PPO

PPO（Proximal Policy Optimization，近端策略优化）强化学习模型，是 ChatGPT 训练的第三阶段。

论文标题：Proximal Policy Optimization Algorithms

论文链接：https://arxiv.org/pdf/1707.06347.pdf

摘要：我们为强化学习提出了一个新的策略梯度方法系列，它通过与环境的交互作用在数据采样和使用随机梯度上升优化一个 "代理 "目标函数之间交替进行。标准的策略梯度方法对每个数据样本进行一次梯度更新，而我们提出了一个新的目标函数，可以进行多次的小批量更新。我们称之为近似策略优化（PPO）的新方法具有信任区域策略优化（TRPO）的一些优点，但它们的实现要简单得多，更通用，并且具有更好的样本复杂性（经验上）。我们的实验在一系列基准任务上测试了PPO，包括模拟机器人运动和Atari游戏，我们表明PPO优于其他在线策略梯度方法，并且总体上在样本复杂性、简单性和壁垒时间之间取得了有利的平衡。

8.In-Context Learning

ChatGPT 的认知能力不全是从语料的统计中习得的，他还有临场学习的能力，这种能力称作 In-Context Learning，学术界本身对这种能力也还没有充分理解。

8.1 Why Can GPT Learn In-Context

论文标题：Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers

论文链接：https://arxiv.org/pdf/2212.10559.pdf

摘要：大型预训练的语言模型显示了令人惊讶的语境学习（In-Context Learning，ICL）能力。通过一些示范性的输入-标签对，它们可以预测未见过的输入的标签，而无需额外的参数更新。尽管在性能上取得了巨大的成功，但ICL的工作机制仍然是一个开放的问题。为了更好地理解ICL的工作原理，本文将语言模型解释为元优化器，并将ICL理解为一种隐性的微调。从理论上讲，我们弄清楚了Transformer注意力有一个基于梯度下降的优化的双重形式。在此基础上，我们对ICL的理解如下。GPT首先根据示范实例产生元梯度，然后将这些元梯度应用于原始的GPT，建立ICL模型。在实验中，我们综合比较了ICL和基于真实任务的显式微调的行为，以提供支持我们理解的经验证据。结果证明，ICL在预测层面、表征层面和注意行为层面的表现与显式微调类似。此外，受我们对元优化的理解启发，我们通过与基于动量的梯度下降算法的类比，设计了基于动量的注意力。它比香草式注意力持续更好的表现从另一个方面再次支持了我们的理解，更重要的是，它显示了利用我们的理解进行未来模型设计的潜力。

8.2 What learning algorithm is in-context learning

论文标题：What learning algorithm is in-context learning? Investigations with linear models

论文链接：https://arxiv.org/pdf/2211.15661.pdf

摘要：神经序列模型，特别是转化器，表现出显著的语境中学习的能力。它们可以从输入的标记例子序列（x,f(x)）中构建新的预测器，而无需进一步更新参数。我们研究了这样一个假设：基于转化器的语境中学习者通过在其激活中编码较小的模型，并在语境中出现新的例子时更新这些隐性模型，从而隐性地实施标准的学习算法。使用线性回归作为一个原型问题，我们为这个假设提供了三个证据。首先，我们通过构造证明，转化器可以实现基于梯度下降和闭式脊回归的线性模型的学习算法。第二，我们表明，经过训练的语境中的学习者与梯度下降、山脊回归和精确最小二乘回归计算的预测者密切匹配，随着转化器深度和数据集噪声的变化，在不同的预测者之间过渡，并在大宽度和大深度下收敛到贝叶斯估计者。第三，我们提出了初步证据，证明in-context学习者与这些预测者共享算法特征：学习者的后期层非线性地编码权重向量和矩矩阵。这些结果表明，上下文学习在算法方面是可以理解的，而且（至少在线性情况下）学习者可以重新发现标准的估计算法。

9.Prompt

ChatGPT 训练时的输入使用的是 Prompt，Prompt 是研究者们为了下游任务设计出来的一种输入形式或模板，它能够帮助预训练模型“回忆”起自己在预训练时“学习”到的东西。

论文标题：Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

论文链接：https://dl.acm.org/doi/pdf/10.1145/3560815

摘要：本文调查并组织了自然语言处理中的一个新范式的研究工作，我们称之为 “基于 prompt 的学习”。与传统的监督学习不同的是，传统的监督学习是训练一个模型来接受一个输入x并预测一个输出y作为P(y|x)，而基于提示的学习是基于语言模型，直接对文本的概率进行建模。为了使用这些模型来执行预测任务，原始输入x被使用模板修改成一个文本字符串 prompt x’，其中有一些未填充的槽，然后语言模型被用来概率性地填充未填充的信息，得到最终的字符串x，从中可以得出最终的输出y。由于一些原因，这个框架是强大和有吸引力的：它允许语言模型在大量的原始文本上进行预训练，并且通过定义一个新的 prompting 函数，模型能够进行少次甚至零次的学习，适应只有很少或没有标记数据的新场景。在本文中，我们介绍了这种有前途的范式的基本原理，描述了一套统一的数学符号，可以涵盖各种现有的工作，并沿着几个维度组织现有的工作，例如选择预训练的模型、prompts 和调整策略。为了让感兴趣的初学者更容易了解这个领域，我们不仅对现有的工作进行了系统的回顾，并对基于 prompt 的概念进行了高度结构化的分类，而且还发布了其他资源，例如，一个包括不断更新的调查的网站http URL，以及论文清单。