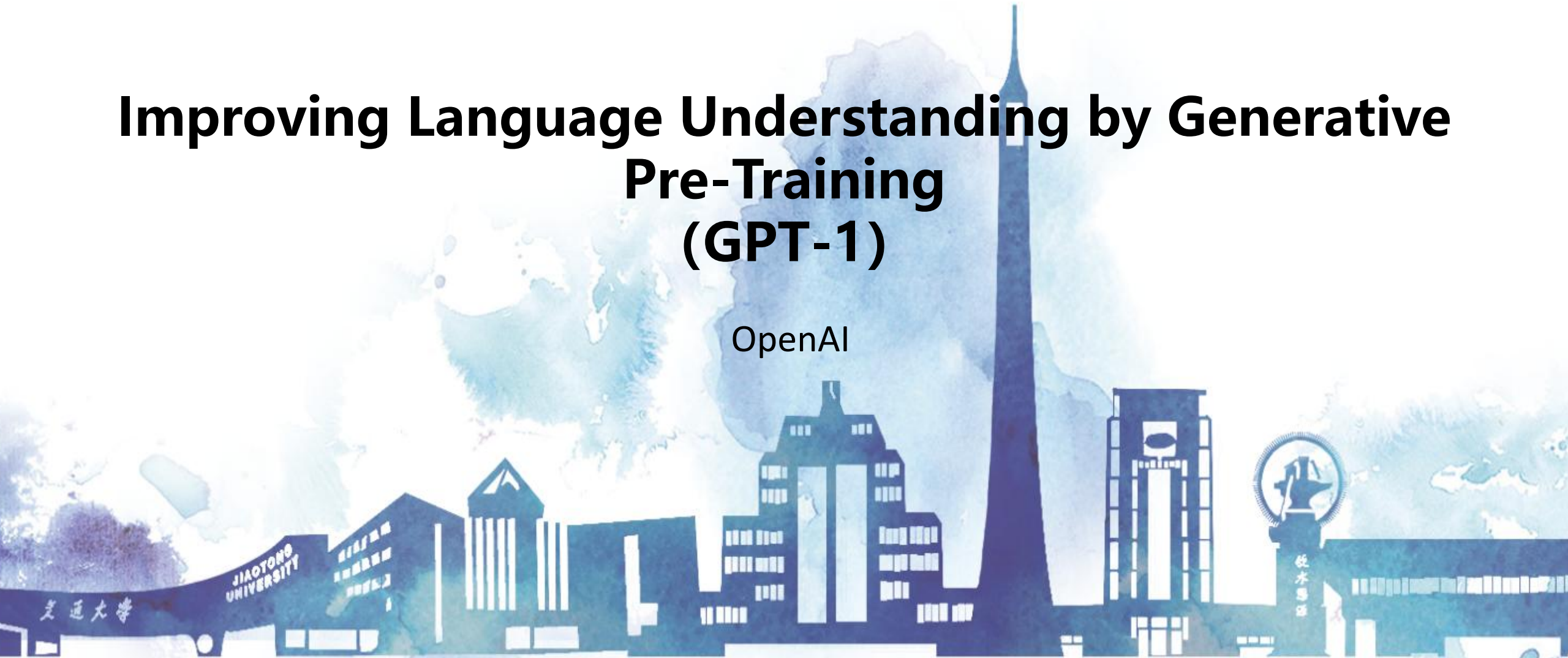
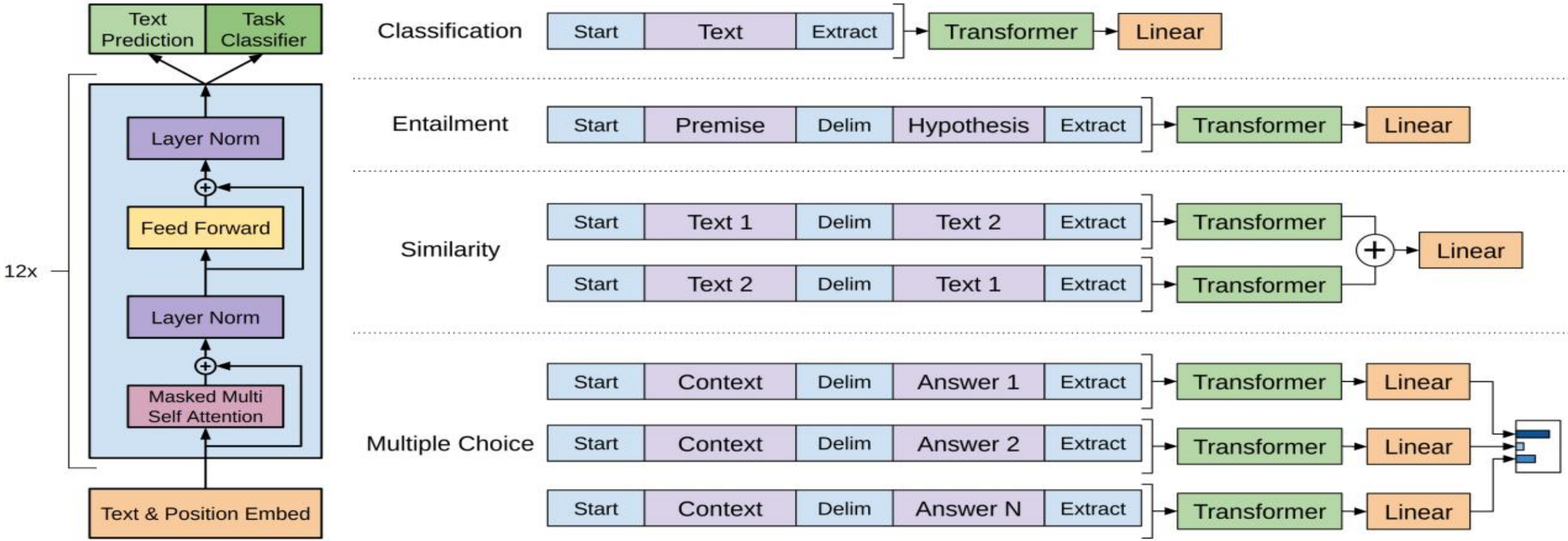
Improving Language Understanding by Generative Pre-Training (GPT-1)

OpenAI



概述

GPT之前大多数的深度学习方法都采用监督学习，需要使用大量的带**标签**的数据。如果现有的带标签数据不够，还需要对没有标签的数据大量的手工标记，人工成本很高，所以这种方法难以适用于那些带标记数据不足的领域。所以GPT采用**无监督学习**，即使是在那些有很多带标记数据的情况下，使用无监督的方式学习对文本的表示也可以更好地捕捉数据的特征，从而显著提高性能。所有语言模型，都是用来执行特定的或是多种综合的NLP任务。



预训练数据集

名称	简介	规模	特点	地址
BooksCorpus[1]	是一个由未出版作者撰写的免费小说集。	其中包含 11038 本书(约 7400 万个句子，1G 个单词)，共有 16 个不 同的子类型(如浪漫、历史、冒险等)。	该数据集的文本具有较高的语言复杂性和多样性，适合用于训练和评估各种语言模型，尤其是在需要处理长文本和复杂语境的任务中表现尤为突出。	https://hyper.ai/datasets/13642
1B Word Benchmark	一个基准语料库，用于衡量统计语言建模的进展。	该语料库在训练数据中有近十亿个单词。	该数据集的主要特点在于其庞大的规模和纯粹的文本特性。训练集包含超过三千万条文本实例，总字数接近十亿，为语言模型提供了丰富的训练材料。此外，数据集的结构简单，仅包含一个字段，即 'text' ，便于直接用于各种语言建模任务。其测试集相对较小，但足以用于模型性能的验证。	https://huggingface.co/datasets/billion-word-benchmark/lm1b

1. Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.

监督微调数据集

任务	名称	简介	特点	地址
自然语言推理	RTE[1]	首次创建于2005年，旨在评估自然语言处理系统在文本蕴含识别任务中的表现。	其专注于文本蕴含关系的识别，涵盖了多种语言现象和复杂的语义关系。	https://opendatalab.org.cn/OpenDataLab/RTE
	SNLI[2]	它包含了由人类注释员生成的成对句子，并标注了这对句子之间的逻辑关系。	SNLI数据集以其大规模和多样性著称，涵盖了广泛的主题和语言风格。	https://opendatalab.org.cn/OpenDataLab/SNLI
	SciTail[3]	是从多项选择科学考试和网络句子创建的蕴涵数据集，每个问题和正确的答案选择都被转换为一个断言陈述以形成假设。	核心研究问题是如何在自然语言处理领域中，通过文本蕴含任务来评估和提升机器对科学文本的理解能力。	https://opendatalab.org.cn/OpenDataLab/SciTail
	Question NLI[4]	是一个用于自然语言推理（NLI）任务的数据集，主要用于判断给定的句子是否包含问题的答案。	涵盖了多种类型的问答对，包括事实性问题、推理性问题等，能够全面评估模型的理解能力。	https://gluebenchmark.com/tasks
	MultiNLI[5]	以 SNLI 语料库为模型，但不同之处在于涵盖了一系列口语和书面文本类型，并支持独特的跨类型泛化评估。	具有跨领域的文本多样性，这使得模型能够在不同语境下进行推理，增强了其泛化能力。	https://opendatalab.org.cn/OpenDataLab/MultiNLI

1. L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo. The fifth pascal recognizing textual entailment challenge. In TAC, 2009.
2. S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. EMNLP, 2015.
3. T. Khot, A. Sabharwal, and P. Clark. Scitail: A textual entailment dataset from science question answering. In Proceedings of AAAI, 2018.
4. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.
5. A. Warstadt, A. Singh, and S. R. Bowman. Corpus of linguistic acceptability. http://nyu-mll.github.io/cola, 2018.

监督微调数据集

任务	名称	简介	特点	地址
问答	RACE[1]	是一个大规模的机器阅读理解数据集，专门用于训练和评估机器的阅读理解能力。	该数据集来自中国12-18岁之间的初中和高中英语考试阅读理解，包含28,000个短文、接近100,000个问题。	http://www.cs.cmu.edu/~glai1/data/race/
	Story Cloze[2]	是一个用于评估故事理解和生成能力的基准数据集。它包含一系列四句话的故事，要求模型从两个可能的结尾中选择一个正确的结尾。	是一个用于评估故事理解和生成能力的基准数据集。它包含一系列四句话的故事，要求模型从两个可能的结尾中选择一个正确的结尾。	https://cs.rochester.edu/nlp/rocstories/

1. G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. Race: Large-scale reading comprehension dataset from examinations. EMNLP, 2017.
2. N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, and J. Allen. Lsdsem 2017 shared task: The story cloze test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 46–51, 2017.

监督微调数据集

任务	名称	简介	特点	地址
语句相似度	STS Benchmark[1]	包括 2012 年至 2017 年间在 SemEval 环境中组织的 STS 任务中使用的英语数据集的选择。数据集的选择包括来自图像标题、新闻标题和用户论坛的文本。	该数据集涵盖了多种语言风格和主题，能够有效评估模型在不同语境下的语义理解能力。	https://opendatalab.org.cn/OpenDataLab/STS_Benchmark
	Quora Question Pairs[2]	源自于Quora平台上的用户提问数据。该数据集的构建基于一个核心任务：判断两个问题是否表达相同的意思。	源自于Quora平台上的用户提问数据。该数据集的构建基于一个核心任务：判断两个问题是否表达相同的意思。	https://www.kaggle.com/c/quora-question-pairs
	MSR Paraphrase Corpus[3]	是一个用于句子对相似度评估的数据集。它包含5801对句子，每对句子都标注了是否为释义关系。	MRPC数据集的构建基于新闻文章，包含了超过5800对句子，每对句子都标注了是否为释义关系。这一数据集的发布极大地推动了文本相似度检测技术的发展。	https://www.microsoft.com/en-us/download/details.aspx?id=52398

1. A. Warstadt, A. Singh, and S. R. Bowman. Corpus of linguistic acceptability. http://nyu-mll.github.io/cola, 2018.
2. Z. Chen, H. Zhang, X. Zhang, and L. Zhao. Quora question pairs. https://data.quora.com/First-QuoraDataset-Release-Question-Pairs, 2018.
3. W. B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005.

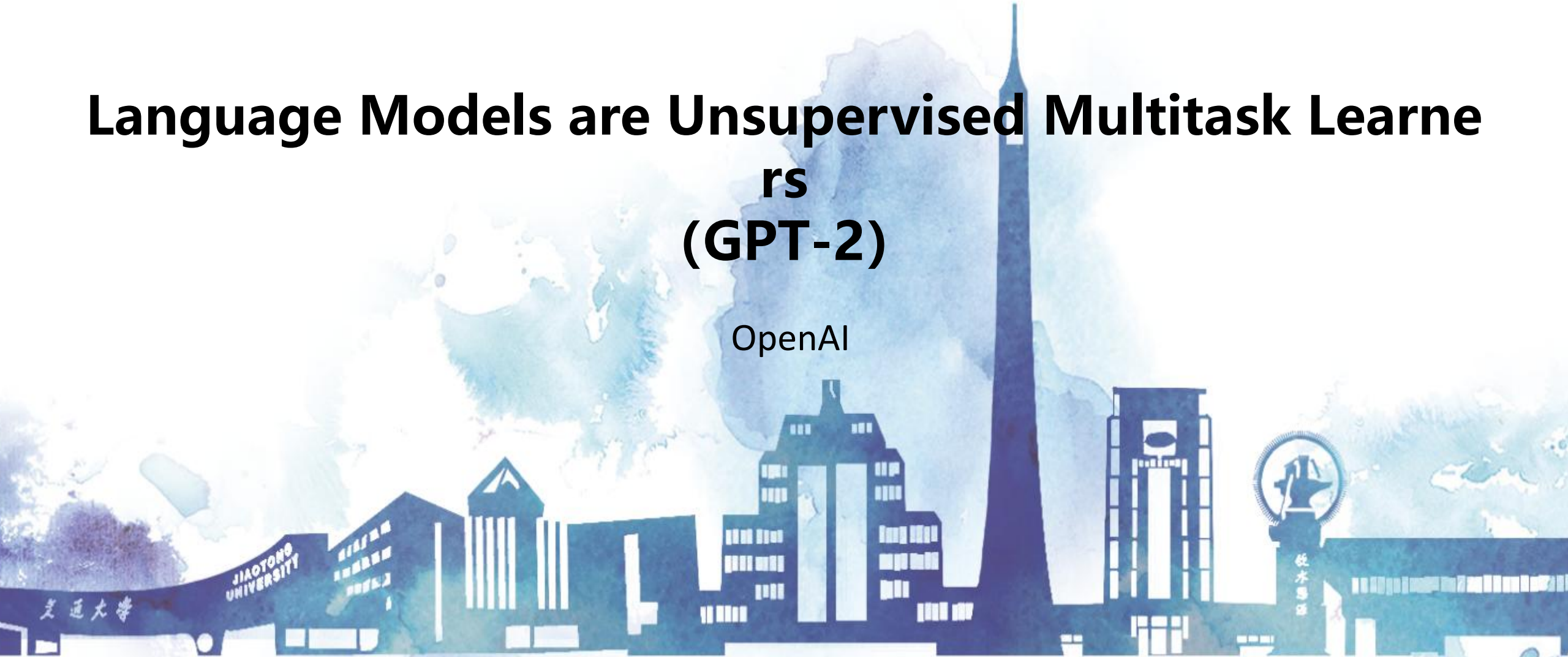
监督微调数据集

任务	名称	简介	特点	地址
分类	Stanford Sentiment Treebank-2 [1]	是一个情感分析数据集，包含电影评论的句子，每个句子都被标记为正面或负面情感。数据集由斯坦福大学发布，用于评估情感分类模型的性能。	以其简洁性和高实用性著称，特别适用于情感分析模型的训练和评估。其特点在于句子级别的情感标注，避免了复杂的短语或片段分析，使得模型能够更专注于整体情感的捕捉。	https://nlp.stanford.edu/sentiment/index.html
	COLA[2]	是一个用于评估句子语法合宜性的单句子分类任务数据集。它包含来自23本语言学出版物的10657个句子。	首次将语法正确性评估引入到模型训练和评估中，推动了语法相关任务的研究进展。	https://opendatalab.org.cn/OpenDataLab/CoLA

1. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.
2. A. Warstadt, A. Singh, and S. R. Bowman. Corpus of linguistic acceptability. http://nyu-mll.github.io/cola, 2018.

Language Models are Unsupervised Multitask Learners (GPT-2)

OpenAI



概述

GPT-1 和 BERT 都是先在大量无标签数据上预训练语言模型，然后在每个下游任务上进行有监督的微调，但是存在一些问题。

1.对于下游的每个任务，还是要重新训练模型。
2.需要收集有标签的数据。

基于监督学习训练的模型的泛化性不是很好，在一个任务上训练好的模型也很难迁移到下一个任务上。 该篇论文主要是探索在大规模无监督训练下，语言模型能否在没有标注数据的情况下完成多任务学习。同时传统有监督方法一般使用留出法等对数据集拆分成训练集和测试集分别进行训练和评估，这种训练方式假设数据集与真实情况的数据的分布情况相同。在字幕与阅读理解任务上，输入的变化情况很大，如果数据集带有某种非总体性的特征，则实际泛化能力可能不够好。提出无监督的语言模型直接进行下游任务。

GPT-2是一个拥有15亿参数的Transformer，它在零样本设置中在8个测试的语言建模数据集中的7个上取得了最先进的结果。

训练数据集

以往的大多数语言模型训练工作都基于单一领域的文本，比如新闻文章、维基百科或小说书籍。

GPT-2的做法则强调构建尽可能大且多样化的数据集，以收集来自尽可能多领域和语境下的自然语言任务示例。现有的一些数据集存在问题，比如网页抓取数据 Common Crawl，数据质量问题也非常严重，在研究常识推理任务时大量文档“内容几乎无法理解”。

GPT-2构建了一个新的网页抓取数据集，强调文档质量。为此， GPT-2仅抓取了经过人工筛选/推荐的网页。人工全面筛选网页代价极高，因此采用了 Reddit 作为起点 —— 抓取所有 Reddit 用户推荐（至少获得 3 karma）的链接。最终的数据集被称为 WebText，包含了 **4500 万**个链接中提取的文本子集。使用 Dragnet和 Newspaper1 两种内容抽取器，从 HTML 页面中提取正文文本。经过去重和启发式清洗后，最终得到约 **800 万**个文档，共计约 **40 GB** 文本数据。并且移除了所有维基百科文档，以避免和常见评估任务中的训练数据产生重叠。

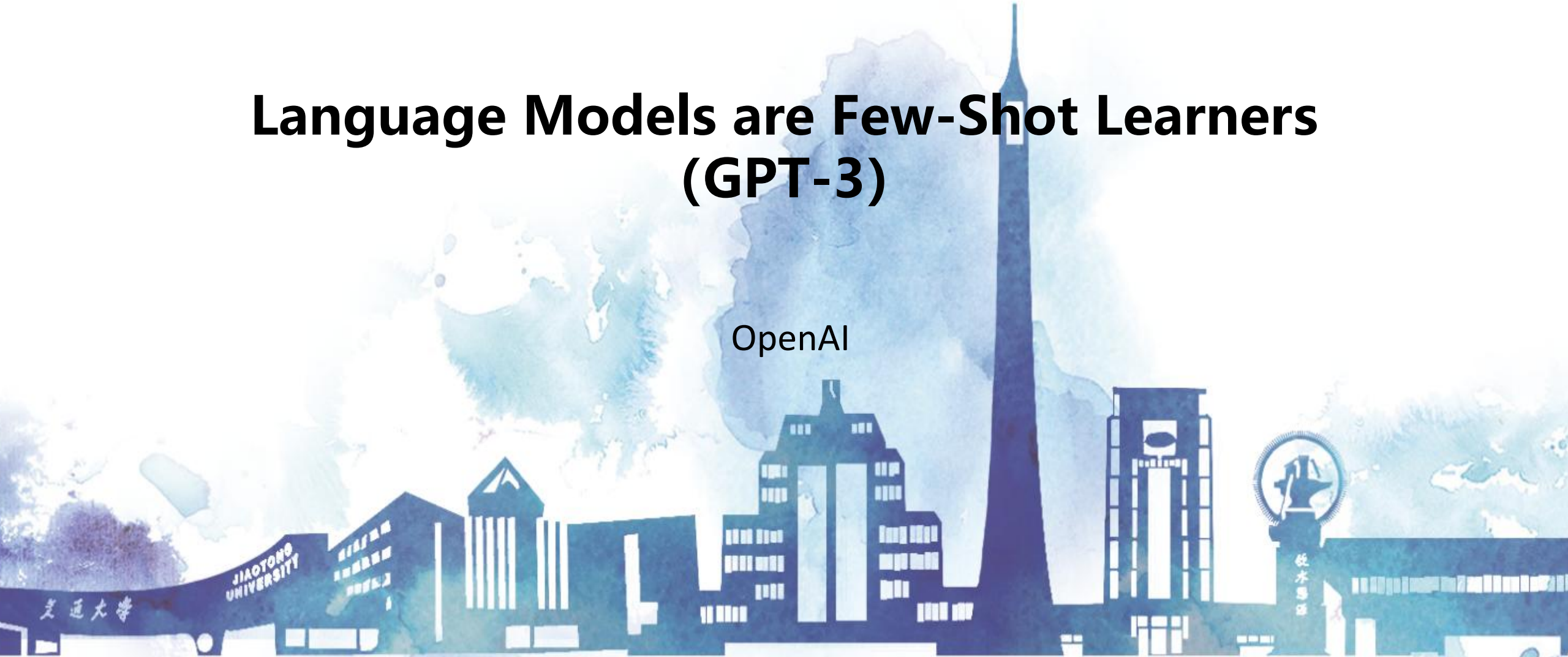
1. https://huggingface.co/datasets/Skylion007/openwebtext
2. https://github.com/CASIA-LM/ChineseWebText-2.0

测试数据集

任务	名称	简介	特点	地址
儿童图书测试	CBT	旨在直接衡量语言模型如何利用更广泛的上下文信息。该数据集基于免费提供的儿童书籍构建。	CBT报告的评估指标不是困惑度，而是在自动构建的完形填空测试中预测出一个被省略的词的10个可能选择中哪一个是正确的的准确性。	https://hf-mirror.com/datasets/cam-cst/cbt
文本理解	LAMBADA	用于评估计算模型在文本理解方面的能力，特别是通过单词预测任务来测试模型是否能够处理长距离依赖关系。	LAMBADA数据集的核心特点在于其对长程依赖关系的评估，要求模型不仅依赖局部上下文，还需理解更广泛的语境信息。	https://hf-mirror.com/datasets/cimec/lambada
常识推理	WinoGrande	是一个用于自然语言理解任务的数据集，特别设计来评估模型在解决Winograd Schema Challenge类型问题上的能力。	具有高度复杂的句子结构和丰富的上下文依赖性。每个句子都设计得极具歧义，要求模型不仅理解句子的表面意义，还需深入分析上下文以正确解析代词的指代对象。	https://winogrande.allenai.org/
阅读理解	CoQA	是一个大规模的对话式问答数据集，旨在帮助构建对话式问答系统。该数据集包含127,000个问题及其答案。	问题设计贴近实际对话场景，答案不仅包括提取式回答，还包括了自由形式的回答。	https://hf-mirror.com/datasets/stanfordnlp/coqa
摘要	CNN/Daily Mail	从 CNN 和每日邮报网站中的新闻故事中生成的问题（其中一个实体被隐藏），故事作为相应的段落，系统预计从中回答填空问题。	包含了超过30万篇新闻文章及其对应的摘要，涵盖了广泛的主题和领域。该数据集的显著特点是其摘要部分由新闻网站自动生成。	https://opendatalab.org.cn/OpenDataLab/CNN_Daily_Mail

Language Models are Few-Shot Learners (GPT-3)

OpenAI



概述

虽然 GPT-2 主推的 zero-shot 在创新度上有比较高的水平，但在效果上并没有很大的提升。为了解决效果上的问题提出了GPT-3。

GPT-3 不再去追求那种极致的不需要任何样本就可以表现很好的模型，而是考虑像人类的学习方式那样，仅仅使用极少数样本就可以掌握某一个任务。这里的 few-shot 不是像之前的方式那样，使用少量样本在下游任务上去做微调（因为在 GPT-3 那样的参数规模下，即使是参数微调的成本也是高到无法估计），而是作为示例（也就是in-context learning）进行效果的提升。

训练数据集

训练过程中使用了Common Crawl数据集，原始未处理的数据达到了 45TB，这个数据集的大小足以训练论文最大的模型（1500亿参数）。然而，Common Crawl的未过滤或轻度过滤版本的质量较低，因此论文采取了3个步骤来提高数据集的平均质量：

（1）根据与一系列高质量参考语料库的相似性下载并过滤了CommonCrawl版本。

（2）在文档级别、数据集内部和数据集之间执行了模糊重复数据删除，为了防止冗余并保持我们所提供的验证集的完整性，作为过度拟合的准确度量。

（3）论文还将已知的高质量参考语料库（GPT-2使用的WebText）添加到训练组合中，以增强CommonCrawl并增加其多样性。

测试数据集

任务	名称	简介	特点	地址
语言建模、文本生成、**完形填空**	LAMBADACBT、StoryCloze、**HellaSwag**	是一个用于常识自然语言推理的新数据集，旨在测试机器是否能完成句子。	具有高度复杂的语境设置和多样化的结尾选项，这使得模型在处理时需要具备较强的常识推理能力。	https://hf-mirror.com/datasets/Rowan/hellaswag
问答	TriviaQA	是一个现实的基于文本的问答数据集，其中包括来自维基百科和网络的 662K 文档中的 950K 问答对。	具有广泛的知识覆盖和复杂的问答结构著称。该数据集包含了超过65万个问答对，涉及多个领域，如科学、历史、文学等。	https://opendatalab.org.cn/OpenDataLab/TriviaQA
	ARC	是一个多项选择问答数据集，包含从 3 年级到 9 年级的科学考试的问题。	主要用于科学教育和研究领域，特别是用于评估和提升人工智能在科学问题解决中的能力。	https://modelscope.cn/datasets/OmniData/ARC
SuperGLUE基准	SuperGLUE	是一个用于评估自然语言处理模型性能的基准测试集，涵盖多种复杂任务，如 多轮对话推理、常识推理、阅读理解 等。	包含了多个子任务，每个子任务都有其独特的挑战和应用场景。数据集不仅涵盖了传统的自然语言处理任务，还引入了一些新兴的任务类型，如常识推理和情感分析，从而全面覆盖了语言理解的各个方面。	https://opendatalab.org.cn/OpenDataLab/SuperGLUE

总结

它是最早一批提出在 NLP 任务上使用 pre-train + fine-tuning 范式的工作。

GPT-1

在做下游任务时，不再进行微调，只进行简单的Zero-Shot，就能与同时期微调后的模型性能相差不大。

GPT-2

在被给定的几个任务示例或一个任务说明的情况下，模型应该能通过简单预测来补全任务中的其他示例。

GPT-3