NLP 中的生成式预训练模型

自我介绍：

马聪，2017 年获北京科技大学工学学士学位，专业智能科学与技术，现保送至中国科学院自动化研究所模式识别国家重点实验室，研究兴趣为自然语言处理、机器翻译、多模态信息处理等。曾任中国科学院大学人工智能技术学院首届学生会主席。研究生入学至现在，以第三作者的身份分别参与了一篇 EMNLP 会议论文和一篇 TKDE 期刊论文。

导读：

本次分享将主要关注 OpenAI 在自然语言处理领域的两个预训练的工作 GPT 和 GPT-2.0. 通过分析 GPT 的两个模型，重点探讨基于单向语言模型的 NLP 预训练过程对序列生成任务的作用以及利用预训练模型进行 NLP 多种任务无监督测试的方式和效果。GPT-2.0 在机器翻译、问答系统、文本摘要等复杂任务上的性能展示出 NLP 预训练模型的强大功能以及其在自然语言序列生成中性能。

正文：

2018 年 NAACL 会议的 Best paper 颁发给了预训练工作 ELMo，2019 年 NAACL 的 Best paper 颁发给了谷歌 AI 的预训练工作 Bert。从这连续两年的 NAACL 最佳论文的评选可以看出学界对自然语言处理中预训练的重视，同时预训练模型也没有辜负大家的期望，在一系列的任务上都得到了非常好的实验性能。在 Bert 工作的同期，OpenAI 研究机构在其官方博客上也发布了在自然语言处理中进行生成式模式的预训练模型 GPT 和 GPT-2. Bert 和 GPT 有相似也有不同，本文将重点针对 GPT 的思想和设计进行展开。

预训练模型在计算机视觉中已经有了非常广泛的应用，例如在ImageNet 上预训练 VGG、GoogLe Net、Res Net 等网络架构，在后续的下游任务中得到了非常好的效果，而在自然语言处理中预训练的工作之前主要是集中在文本表示的预训练中，例如 word2vec、sent2vec、doc2vec 这样对文本进行表示的工作中。近两年，有较多的研究工作关注在利用语言模型任务来进行预训练，以得到文本的更好表示。

在介绍预训练的模型之前，我们有必要先看看为什么要进行预训练，或者预训练到底能为后续的任务带来什么？无论是计算机视觉还是自然语言处理或者机器学习的其他任务，在进行模型设计前，特征的表达都是非常重要的工作，在之前传统方法中，主要的特征表示工作是由研究人员进行手工编撰、设计特征表达，进而将这些特征表达送到后续的模型中进行分类或回归等下游任务。对于计算机视觉中，最常见的数据形式为图像的像素值，自然语言处理中最常见的是字符串表征的文本，但无论是什么源数据形式，都需要将数据映射到特征空间，才能更好的进行模型的训练、学习。图 1 展示了机器学习过程中数据表示的流程图，现在大多数的机器学习任务都根据该流程进行数据的表征和模型的训练。在数据表示和训练的流程中，一个关键的问题是如何进行有效、合理的特征表达。在传统工作中，研究人员主要进行手工设计特征，而随着深度神经网络的发展和越来越大规模数据的获得，研究者们发现，深度学习框架可以从大规模的训练数据中自动的学习到比较好的数据特征表达，而不再需要手工编撰特征。显而易见，如果希望模型可以学习到比较好的数据表示，需要有大规模的数据支持，但是在很多具体的研究任务中，训练数据的规模还是比较有限，并不能支持模型学习到较好的特征表示，所以在特定的任务上进行预训练（例如计算机视觉中的 ImageNet 数据集上预训练，自然语言处理中的大规模语言模型预训练），再到具体任务上进行 fine-tune。

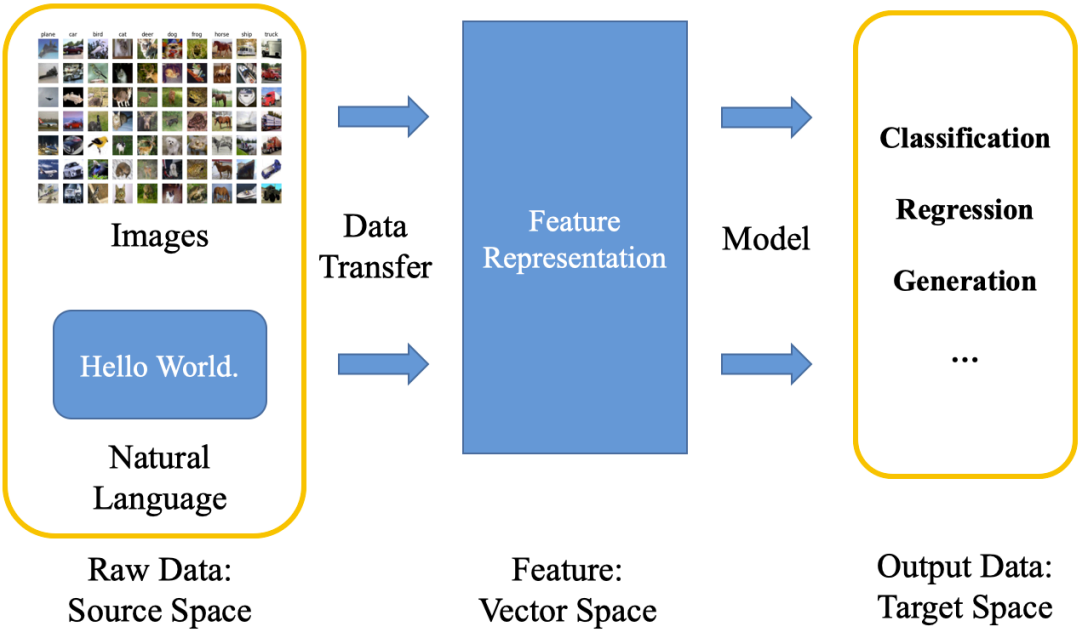

图 1 - 机器学习数据表示流程图

除了在数据表示工作的预训练，还有较多的工作关注在模型的预训练，该部分的工作主要是利用预训练得到模型的比较好的模型参数初始化范围，并在后续的 fine-tune 中在目标任务数据集上进行快速的收敛。

具体的在自然语言处理中利用语言模型进行预训练的形式化如图 2 所示。语言模型的任务是根据上下文的信息来预测后续将要出现的字符或单词的概率。前向语言模型是根据已经出现的若干词语来预测当前输出词的概率分布；而后向语言模型是根据未来的若干词语来预测当前输出词的概率分布。双向的语言模型则是根据历史信息和未来信息，共同作为条件来预测当前词的概率分布，从形式上来看，双向语言模型的定义非常类似“完形填空”的任务。常见的基于语言模型进行预训练的工作都是在前向、后向或双向的任务定义，例如 ELMo 是双向语言模型任务，Bert 也是双向语言模型任务，GPT 是单向语言模型任务。其中 ELMo 和 Bert 的区别在于，ELMo 的建模是利用前向语言模型和后向语言模型，两个模型的拼接来实现双向语言模型，而 Bert 是直接将历史信息和未来信息共同作为条件输入来对当前词的分布进行预测。GPT 则是只利用历史信息来预测接下来将要生成的词。所以 GPT 的作者在论文题目中称他们的工作为“Generative Pre-Training” - 生成式预训练。

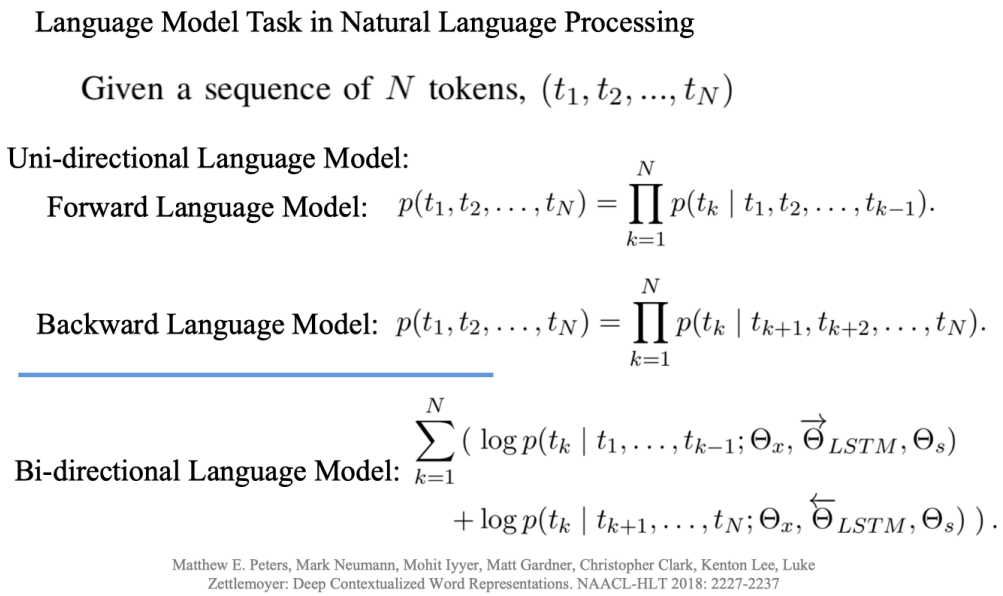

图 2 - 语言模型预训练形式化表示

基于语言模型的预训练框架中，之前的相关工作基本都是基于循环神经网络进行语言模型的建模。随着 2018 年《Attention is all you need》工作的提出（如图 3 所示），利用自注意力模型 Transformer 在自然语言处理中的各个任务中得到了非常好的效果，类似的，利用 Transformer 在语言模型中进行建模也都发现可以得到更好的效果。Bert 主要是利用了 Transformer 的编码器架构来设计双向语言模型，GPT 则是主要利用了 Transformer 的解码器部分来进行生成式的单向语言模型建模。

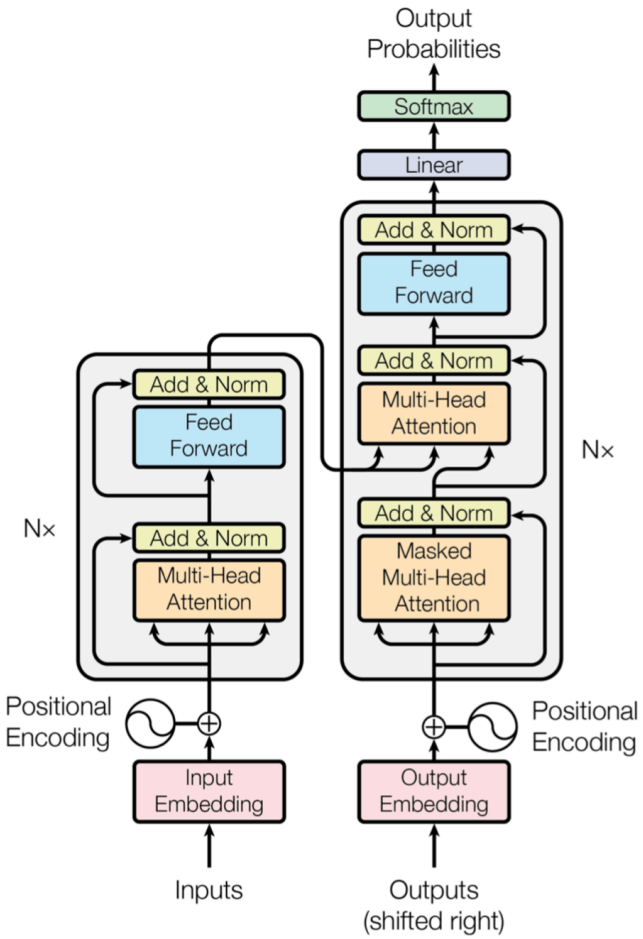

图 3 - Transformer 模型架构图

GPT 的建模思路如图 4 所示，其首先在语言模型上进行预训练，继而在特定的任务上进行 fine-tune。从图 4 中可以发现，其基础模型部分基本与 Transformer 的解码器架构一致，只是删去了 Transformer 中解码器与编码器之间的自注意力计算机制（因为在 GPT 的语言模型预训练中没有编码器，也没有源端语言的设定）。在预训练阶段，GPT 完全使用前向语言模型的优化目标进行语言模型的训练，而在特定任务的 fine-tune 部分，GPT 是通过对训练数据中添加相应的 token 来将各种任务转化为分类任务来进行训练。GPT 所实验的任务包含文本分类、文本蕴含、句子相似度预测和多项选择的阅读理解任务。需要特别注意的是 GPT 在 fine-tune 阶段不仅仅利用特定任务的训练目标，还将语言模型的任务同时作为辅助目标来进行统一训练，根据论文中的实验结果发现，具体任务搭配语言模型的辅助 fine-tune 在一些任务上可以得到性能的进一步提升，但在另外一些任务上没有达到更好的效果。个人认为在语言模型的预训练后的 fine-tune 中是否需要加入语言模型的辅助目标取决于下游任务的复杂度，有些下游任务相对比较简单，其对语言的条件生成的依赖比较小，此

时搭配语言模型进行辅助 fine-tune 可能并不能让实验性能进一步提升，而有一些较难的任务需要条件生成的更多信息来进行学习，此时在 fine-tune 的过程中搭配语言模型的目标进行 fine-tune 会得到更好的实验结果。

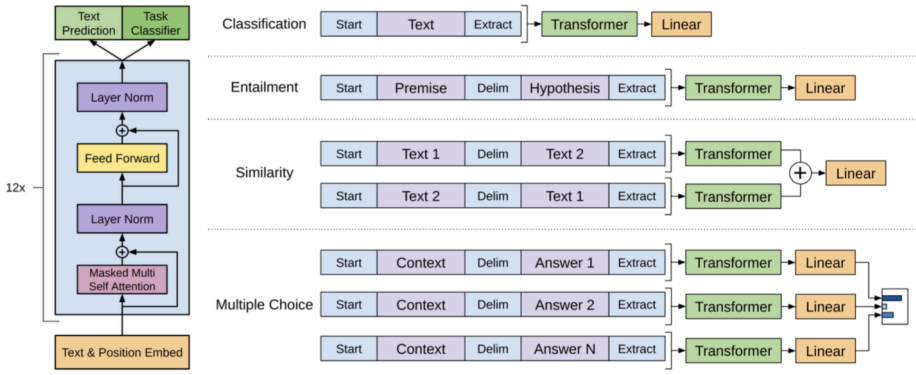

图 4 - GPT 模型架构图

在 GPT 发布 8 个月后，OpenAI 又发布了 GPT-2 版本，得到了更好的实验性能。与 GPT 不同的是，GPT-2 并没有在特定的下游任务上进行 fine-tune，而是全部采用了无监督测试的架构设置。在模型架构上，GPT-2 的总参数量是 GPT 的十倍还多，Transformer 的层数从 BPT 的 12 层增加到 GPT-2 的 48 层，隐层节点的维度也从 768 维增加到了 1600 维。在具体的模型细节上，GPT-2 与 GPT 的主要不同是在层归一化（layer norm）的放置和残差模块部分的初始化设置。更重要的一点，GPT-2 的训练数据集采用了 WebText，该数据集是 OpenAI 在网络资源上进行爬取、清理得到的，其总数据量约有 800 万文档，文本存储量有 40GB。更好的数据和更大的模型，让 GPT-2 在语言模型任务上得到了非常好的实验性能，GPT-2 虽然没有在各个特定的语言模型的训练集上进行训练（GPT-2 都是在 WebText 上进行训练的），但是在 8 个语言模型的任务上的 7 个得到了 state-of-the-art 的结果（如图 5 所示）。GPT-2 在 1BW 的测试集上没有得到最好的结果，作者分析是因为在 1BW 任务的训练集中，具有相当大比例的测试集数据存在（有 13.19%的测试数据存在于训练数据集中），而 WebText 与 1BW 的测试数据的交叉程度则较小，1BW 的测试数据中只有 3.75%的数据在 WebText 数据集中。

	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

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

图 5 - GPT-2 在语言模型上的实验效果

GPT-2 除了在语言模型的任务上进行了性能的分析，还在其他的语言生成任务上进行了性能的分析，展现了 GPT-2 在经过预训练后的强大语言生成能力。图 6 展示了 GPT-2 在部分自然语言处理任务上的实验性能。在所有的这些任务中，GPT-2 都是只在 WebText 的数据集上进行训练，而没有在下游任务上进行 fine-tune。没有进行 fine-tune 的工作，在测试阶段最重要的工作便是让模型知道是在什么任务上进行测试。

对于机器翻译工作，作者使用的方式是添加 token [English sentence = French sentence]来提示 GPT-2 进行翻译工作。虽然 WebText 中没有大规模的平行语料，甚至基本都是单语语料，GPT-2 还是展现出了不错的机器翻译实验性能。WebText 是针对英语构建的大规模训练语料，根据作者的统计，其中只检测到有 10MB 的法语资源（总语料的规模为 40GB），而在这样的数据中，GPT-2 的英法翻译的性能可以与基于平行词典的无监督翻译的工作性能可比。而法英的翻译，GPT-2 则是借助其强大的英语生成能力学习得到了较好的结果，其翻译性能超过了无监督的降噪机器翻译和词嵌入最近邻翻译，但是相比目前的无监督机器翻译的 state-of-the-art 工作还有一定的距离，具体结果如图 6 的第二张图所示。

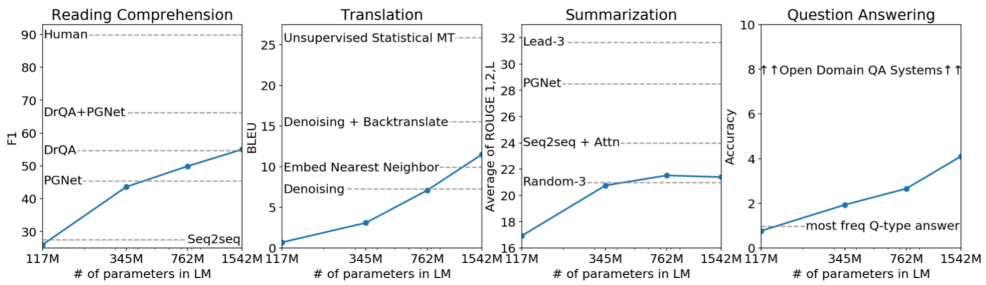

Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

图 6 - GPT-2 无监督实验结果

对于文本摘要的工作中，作者发现在网络爬取的数据中，存在一定的摘要的数据构造，在网络中，其存在形式为：［长文档。TL;DR:摘要文档]，其中的 TL；DR 所表示的含义为 Too Long， Don’t Read。基于这样的发现，作者在实验中加入了该标示，可以让 GPT-2 较好的展示出摘要的性能，如图 7 所示。当 GPT-2 没有添加任何标识时，模型倾向于在长文档的基础上继续进行后续的生成工作，而不是生成对应文本的摘要，所以在摘要的评测指标 ROUGE 上结果不好；而加入 TL；DR 的标识后，GPT-2 的实验性能可以达到与搭配注意力机制的序列模型可比的效果。

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 TL;DR:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

图 7 - GPT-2 在无监督摘要任务上的实验性能

GPT-2 除了在机器翻译、文本摘要的任务上进行无监督测试，还在语言生成的工作上进行了测试，在给定上文的基础上，GPT-2 展现了强大的生成能力（如图 8 所示，更多生成样例可参考 https://www.openai.com/blog/better-language-models/）。

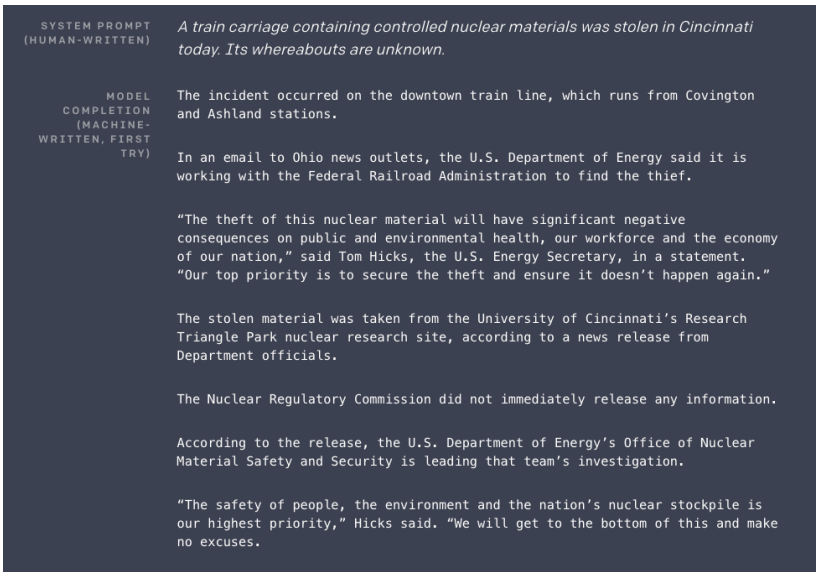

图 8 - GPT-2 语言生成样例

总结来说，GPT、GPT-2 使用前向语言模型进行预训练的工作，其在语言生成的模型设计上更加自然，但是只利用了一个方向的语言模型的建模确实会存在一定的局限性，作者也在论文中提到“It is unclear whether the additional training data and capacity of GPT-2 is sufficient to overcome the inefficiencies of uni-directional representations demonstrated by Bert.”可以看出 OpenAI 的研究工作是在扩大模型和训练数据量的基础上来弥补单方向语言模型和双向语言模型之间的性能差异。在可预见的未来，如何有效的利用双向语言模型进行语言生成任务上的工作，以及如何更加有效的利用大规模文本进行预训练来提高自然语言处理任务，仍将是自然语言处理领域研究的重点和热点。

参考文献：

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer: Deep Contextualized Word Representations. NAACL-HLT 2018: 2227-2237

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR abs/1810.04805 (2018)