同济大学

TONGJI UNIVERSITY

人工智能期中论文

课题名称	大语言模型的产生与发展
副 标 题	ChatGpt 的前世今生
学　　院	同济大学
专　　业	计算机科学与技术专业
学生姓名	吕博文
学　　号	2151769
指导教师	王俊丽
日　　期	2023 年 04 月 30 日

大语言模型的产生与发展

摘要

这篇论文旨在研究人工智能大语言模型的基本原理，发展历程以及未来前景。通过对大语言模型的发展历程和相关领域前沿技术的探讨，总结已有的人工智能大语言模型研究。研究方法包括模型结构和原理、训练数据和预处理、模型训练和微调。在实验结果和分析部分，参考了 GPT 大语言模型的相关测试结果进行分析。在应用与展望部分，讨论了大语言模型在自然语言处理中的应用和未来发展趋势。最后总结了研究结论，并展望了未来工作的方向。本论文旨在全面深入地探讨人工智能大语言模型，并为相关领域的研究提供一定的参考和指导。

关键词：人工智能，大语言模型，ChatGpt

装 订 线

目 录

1 大语言模型的提出以及历史发展

1.1 大语言模型的提出背景

自问世以来，人工智能的发展便一波三折，从刚刚问世的蓬勃发展期（规则系统与专家系统的广泛应用），到 20 世纪 90 年代的低谷期（计算机硬件软件技术限制，专家系统的失败），再到如今的高潮期（大数据，机器学习等），大语言模型的提出可谓是人工智能由衰转盛的一个转折点，它标志着人们不再满足于人工智能在单一领域的精通，而是希望其尽可能的接近“全能”，来帮助人们解决更多现实生活中的问题。

1.2 大语言模型的提出历史

经典语言模型（Classic Language Model，CLM）：经典语言模型是最早的语言模型之一，由 Jelinek 等人于 1991 年提出。该模型基于 N-gram 统计语言模型，使用 n-1 个词作为上下文来预测下一个词。虽然经典语言模型在短语和句子级别上表现良好，但是当面对长文本时，其准确性会受到限制。

循环神经网络语言模型（Recurrent Neural Network Language Model，RNNLM）：由 Mikolov 等人于 2010 年提出。RNNLM 使用循环神经网络来解决经典语言模型的问题，并将前面的词作为输入来预测下一个词。与经典语言模型相比，RNNLM 在长文本的建模方面表现更好。但是，RNNLM 的训练速度较慢，并且存在梯度消失和爆炸的问题。

序列到序列语言模型（Sequence-to-Sequence Language Model，Seq2SeqLM）：Seq2SeqLM 是由 Sutskever 等人于 2014 年提出的，基于编码器-解码器框架的语言模型。Seq2SeqLM 使用编码器将输入序列转换为向量表示，并使用解码器从该向量表示中生成输出序列。Seq2SeqLM 在机器翻译和对话系统等任务中表现出色，但在生成长序列时也存在限制。

改进的 Transformer 语言模型（Transformer-based Language Model）：Transformer 是由谷歌团队于 2017 年提出的一种深度神经网络结构，用于序列到序列的任务。基于 Transformer 的语言模型使用自注意力机制（self-attention）来捕获序列中的长程依赖关系，大大提高了语言模型的建模能力。由此，出现了一系列基于 Transformer 的语言模型，如 BERT、GPT、XLNet 等，这些模型在各种自然语言处理任务中表现出了显著的性能优势。

1.3 大语言模型的广为人知

虽说人工智能大语言模型经历了如此之多的发展阶段，但以 ChatGpt 为代表的大语言模型真正做到广为人知也是在最近才发生的，得益于 ChatGpt 采用的先进的 Transformer 模型以及大量的数据预处理，也得益于人们对人工智能技术接受度的提高，ChatGpt 的发布在全球各个领域都掀起了一场不小的风波，人们惊叹于它回复的准确性的同时，也有越来越多的科研工作者投身于人工智能领域。有人声称，ChatGpt 在 21 世纪出现的意义不亚于 20 世纪互联网的出现，这样的评价明显彰显了人们对其未来发展前景的远大期待。

2 Transformer 语言模型深究

2.1 模型介绍

Transformer 是一种基于注意力机制的序列到序列（Sequence-to-Sequence，简称 Seq2Seq）模型，由 Google 公司提出，主要用于自然语言处理（NLP）领域中的文本序列转换任务，例如机器翻译、文本摘要等。Transformer 最初是被用于机器翻译任务的，但后来逐渐被应用于文本生成、问答系统、语音识别等各种任务。

相较于之前的基于循环神经网络（Recurrent Neural Network，RNN）的 Seq2Seq 模型，Transformer 最大的特点是采用了自注意力机制（Self-Attention Mechanism），这种机制能够有效地处理输入序列中各个位置之间的关系，提高了模型的性能和训练速度。

Transformer 由多个相同的模块（Transformer Block）组成，每个模块由两个子层组成，即多头自注意力机制（Multi-Head Self-Attention）和前馈神经网络（Feed-Forward Neural Network）。其中，多头自注意力机制用于学习输入序列中各个位置之间的依赖关系，而前馈神经网络用于对注意力机制得到的向量进行非线性变换，从而得到最终的输出向量。模块之间还通过残差连接（Residual Connection）和层归一化（Layer Normalization）进行连接和规范化，进一步提高了模型的训练效率和准确性。

Transformer 的优点是可以并行计算，不像 RNN 的计算是串行的。另外，Transformer 还引入了“遮挡掩码”（Masking）的概念，用于在训练时避免模型在预测下一个词时使用到未来的信息。这些创新性的设计和优化，使得 Transformer 成为了当前最先进的自然语言处理模型之一，大大提高了文本序列转换任务的效果和速度。

2.2 模型基本结构与原理

首先我们要知道，Transformer 大语言模型本质上仍然是一个神经网络结构模型，它整体的功能实现依赖于整个网络神经元的个数，神经网络的层数以及神经网络函数设计的优劣，具体了解 Transfomer 模型之前，我们首先要了解一些神经网络的基础知识。

神经网络是一种模仿人类神经系统构造的计算模型，它主要由若干个神经元（neuron）组成，每个神经元接收多个输入信号，经过一定的加权和非线性变换后产生一个输出信号，最终输出给下一个神经元或输出层。神经网络的基本结构通常由输入层、隐藏层和输出层组成。输入层接收外界输入的数据，隐藏层对输入层的信息进行处理，输出层将处理结果输出给外界。

神经网络运行时，会先将输入数据喂入网络中，通过神经元之间的连接传递信息，不断进行加权和非线性变换，最终得到输出结果。整个过程可以被描述为一种前向传播的过程，也就是从输入层开始，逐层向前传播，直到输出层。

在训练神经网络时，通常会先给定一个损失函数（loss function），用来度量神经网络在当前参数下预测结果与真实结果之间的差距。然后利用反向传播算法（Back Propagation，BP）对损失函数

进行优化，更新神经网络中的参数，不断降低损失函数的值，使神经网络的输出结果更加准确。

总的来说，神经网络是一种基于加权和非线性变换的模型，通过前向传播和反向传播等算法进行运算和优化，能够对输入数据进行复杂的处理和分析，并输出相应的结果。

而相比于普通的神经网络模型，大语言模型通常是基于深度神经网络（DNN）的架构设计，使用反向传播算法进行训练优化，并在不断的更新迭代中，从原来的单向递归神经网络升级为 Transformer 中引入的基于注意力机制（Attention Mechanism）的神经网络模型，同时 Transformer 模型可以进行全局的并行计算，也在一定程度上提高了整个模型的训练和计算效率。

2.3 模型特色

Transformer 大语言模型相比之前的几代产品有了较大的改进和提升，下面简单列举出其几条模型特色：

位置编码（Positional Encoding）：由于 Transformer 不是基于循环结构的，因此需要对输入的位置进行编码，使得模型能够考虑到输入序列中各个位置的信息。位置编码是通过一组正弦函数和余弦函数的和来计算的，得到的编码向量与输入向量相加作为模型的输入。

自注意力机制（Self-Attention Mechanism）：自注意力机制是 Transformer 的核心，能够有效地处理输入序列中各个位置之间的依赖关系。自注意力机制的输入是一个序列的向量集合，其中每个向量表示序列中的一个位置的特征，输出也是一个向量集合，其中每个向量表示序列中每个位置的加权特征。具体来说，自注意力机制将输入的向量集合分别映射到三个不同的向量空间中（即查询向量空间、键向量空间和值向量空间），然后通过计算查询向量与键向量的相似度得到注意力权重，再将权重与值向量进行加权求和得到输出向量。Transformer 采用多头自注意力机制，即将输入序列映射到多个不同的向量空间中进行计算，最终将多个输出向量拼接起来作为最终的自注意力输出向量。

前馈神经网络（Feed-Forward Neural Network）：前馈神经网络用于对自注意力输出向量进行非线性变换。具体来说，前馈神经网络由两层全连接层组成，其中第一层使用 ReLU 激活函数，第二层不使用激活函数。前馈神经网络的输出向量被加权求和后再与自注意力输出向量相加作为最终的输出向量。

残差连接（Residual Connection）和层归一化（Layer Normalization）：Transformer 使用残差连接和层归一化来连接不同的 Transformer Block。残差连接是指将输入向量直接与输出向量相加，从而保证梯度的流动和信息的传递。层归一化则是对每个 Transformer Block 的输出进行归一化，使得模型更易于训练。

总之，Transformer 是一种基于注意力机制的 Seq2Seq 模型，其核心是自注意力机制。Transformer 在自然语言处理领域中取得了很好的效果，例如在机器翻译、文本摘要等任务中表现优秀。

2.4 模型训练与参数微调

Transformer 模型的训练可以分为两个阶段：预训练和微调。

预训练阶段：在预训练阶段，Transformer 模型首先被训练成为一个大规模的语言模型，即输入一个词序列，预测下一个词。在这个阶段，使用了一个称为“掩码语言模型”（Masked Language Model，MLM）的任务，即在输入序列中随机掩盖一些词，让模型预测这些被掩盖的词。同时，还使用了另一个任务称为“下一句预测”（Next Sentence Prediction，NSP），即输入两个句子，让模型预测这两个句子是否相邻。这两个任务都是无监督学习，可以在大规模的语料库上进行训练。在预训练阶段，Transformer 模型通过这两个任务的训练，可以自动地学习到语言的结构和规律。

Transformer 模型预训练是一个需要大量数据支撑的环节，根据神经网络的基本原理，只有训练数据足够多，整个模型才能更趋向于完美，在论文《Attention is all you need》中，作者使用了 WMT 2014 英德翻译任务的数据集，包括约 450 万个句子对。在进行训练时，作者使用了 Adam 优化器和学习率为 0.0001 的学习率调度器，同时还使用了标准的 dropout 技术以减少过拟合。

微调阶段：在微调阶段，将预训练得到的 Transformer 模型应用于特定的自然语言处理任务中，如文本分类、命名实体识别等。微调阶段的训练通常是有监督学习，需要提供标注数据。在微调阶段，可以对预训练得到的 Transformer 模型的参数进行微调，使其更适应于特定的任务。微调通常使用反向传播算法进行优化，通过最小化损失函数来更新模型的参数。在微调阶段，还可以使用一些技巧来进一步优化模型的性能，如加入额外的特征、调整学习率、使用正则化等。

3 实验结果与分析

这里的实验结果我参考了论文《Attention is all you need》中作者的实验结果，这里仅仅分析其实验结果并做出自己的总结。

首先，在英德翻译任务中，作者使用了 WMT 2014 数据集进行训练和评估，使用 BLEU 指标来评估翻译质量。结果显示，使用 Transformer 模型进行翻译的 BLEU 得分超过了之前的 SOTA 结果，达到了 28.4 的分数，证明了 Transformer 模型在机器翻译任务上的有效性和优越性。

其次，在其他几个 NLP 任务上，例如语言建模（language modeling）和问答任务（question answering），Transformer 模型也取得了非常好的结果。在语言建模任务中，使用 Transformer 模型的 perplexity 得分也超过了之前的 SOTA 结果；在问答任务中，使用 Transformer 模型进行答案提取的 F1 得分也显著优于其他模型。

值得注意的是，Transformer 模型相比于其他的 RNN 和 CNN 模型，具有更好的并行性，能够更快地进行训练和预测。同时，Transformer 模型还具有更好的长距离依赖性建模能力，能够更好地处理长文本输入和输出。

下面列出该作者的实验结果：

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Label Smoothing During training, we employed label smoothing of value $\epsilon_{ls} = 0.1$ [30]. This hurts perplexity, as the model learns to be more unsure, but improves accuracy and BLEU score.

4 应用与展望

目前来看，以 ChatGpt 为代表的 Transformer 类大语言模型在自然语言处理领域已经得到了广泛的应用，ChatGpt 在语言生成、对话系统、文本分类以及推荐系统等方面都已经展现出了巨大的优势；同时 ChatGPT 及其他 Transformer 类大语言模型在未来的应用也有很大的潜力，相信随着科学技术的不断发展，ChatGpt 可以逐渐应用到多语言领域，实现多模态应用（结合视觉听觉等信息进行任务处理），也一定会在推理方面实现更大的进步，更好的帮助人们促进科学技术的发展。

参考文献

参考文献

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[2]]Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (pp. 1877-1901).

[4]]Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding with unsupervised learning. Technical report, OpenAI.

[5] Zhang, Y., Sun, Y., Zhang, J., Qi, P., Manning, C. D. (2020). Dialogpt: Large-scale generative pre-training for conversational response generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 5290-5304).

[6] Liu, B., Gao, Y., Zhang, H., Zhu, F., Gao, Y. (2021). A comprehensive survey on transformers. arXiv preprint arXiv:2106.04554

[7] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[8] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2019 EMNLP and the 9th International Workshop on Semantic Evaluation (pp. 353-355).

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998 – 6008, 2017.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[11] Dario Amodei, Samy Bengio, Dustin Tran, Greg Diamos, and Hanjun Dai. Scaling compute for deep learning. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pages 1 – 8. JMLR.org, 2017.

[12] Thang Luong, Eugene Brevdo, and Rui Zhao. Neural machine translation (seq2seq) tutorial. 2017

装 订 线