通过 5 篇论文简单理解大语言模型

胡磊 2021214596

2023 年 6 月 19 日

摘要

最近,以 ChatGPT 为代表的,基于 Transformer 架构的大语言模型几乎彻底改变了自然语言处理领域的研究。本文通过解读对其发展极为重要的 5 篇论文来理解大语言模型，其中包括注意力机制、Transformer、BERT、GPT 和 BART。

关键词: LLMs attention Transformer GPT BERT BART

1 介绍

通常来说，大语言模型指的是那些在大规模文本语料上训练、包含百亿级别（或更多）参数的语言模型，例如 GPT-3，PaLM，LLaMA 等。目前的大语言模型采用与小模型类似的 Transformer 架构和预训练目标。

大模型相较于小模型，主要区别在于模型大小、训练数据和计算资源的不同。同时，当语言模型的规模达到一定程度后，显现出了一些特殊的能力，这些能力被称为 “涌现能力”，比较有代表性的涌现能力包括: 上下文能力、指令遵循、逐步推理等。

本文的目标是了解大语言模型的基础，下面将主要从这五篇论文来了解 LLMs 背后的主要思想:

- - Neural Machine Translation by Jointly Learning to Align and Translate (2014)
- Attention Is All You Need (2017)
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)
- Improving Language Understanding by Generative Pre-Training (2018)
- BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (2019)

2 结合对齐和翻译的神经网络机器翻译模型

机器翻译的发展历经了多次变革,从最初的基于规则的方法，到后面的基于统计的方法，再到后来基于神经网络的方法。而在文章《Neural Machine Translation by Jointly Learning to Align and Translate》之前，机器翻译领域的 sota 模型 Seq2Seq，使用这种基于神经网络的方法需要将整个输入句子编码成一个固定长度的向量，这就使得神经网络难以处理长句子，特别是那些比语料库中的句子更长的句子。

而文章《Neural Machine Translation by Jointly Learning to Align and Translate》提出的 RNNsearch 模型，不试图将整个输入句子编码成单个固定长度的向量。相反，它将输入的句子编码成一个向量序列，并在解码翻译时自适应地选择这些向量的子集。这使神经翻译模型不必将源句子的所有信息压缩到固定长度的向量中。该模型能够学习对齐和翻译，在翻译过程中自动搜索和选择源语言句子中与预测的目标语言单词相关性最高的部分，然后再根据这些源位置信息和之前生成的所有目标单词相关的上下文向量来预测目标单词。

从图一中可以看到，decoder 中每一时刻的输出是由多个变量共同决定的，其中包含了 encoder 中每一时刻的隐藏状态量 $(h_1, h_2, ..., h_n)$，和上一时刻的输出 y_{t-1}，以及当前时刻 decoder 中的隐藏状态向量 S_t，因此公式可表示为:

$$p(Y_i|y_1, ..., y_{i-1}, X) = g(y_{i-1}, s_i, c_i)$$

文章中提出的 attention 机制，在机器翻译领域取得了巨大的成功，是一项里程碑式的工作。这篇文章对文本生成，文本摘要、机器翻译等生成式任务起到了重要影响。Attention 机制除了解决了论文中所提到的信息瓶颈问题外，还使得 RNN（包括 LSTM）在长距离依赖中梯度消失的问题得到了进一步缓解。Attention 被提出以来，很快被应用到了各个领域，从 Seq2Seq

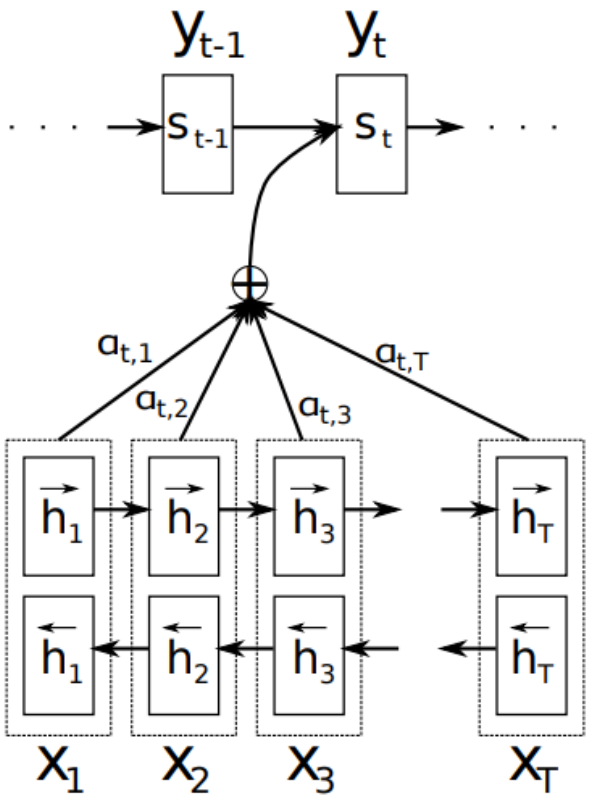

图 1

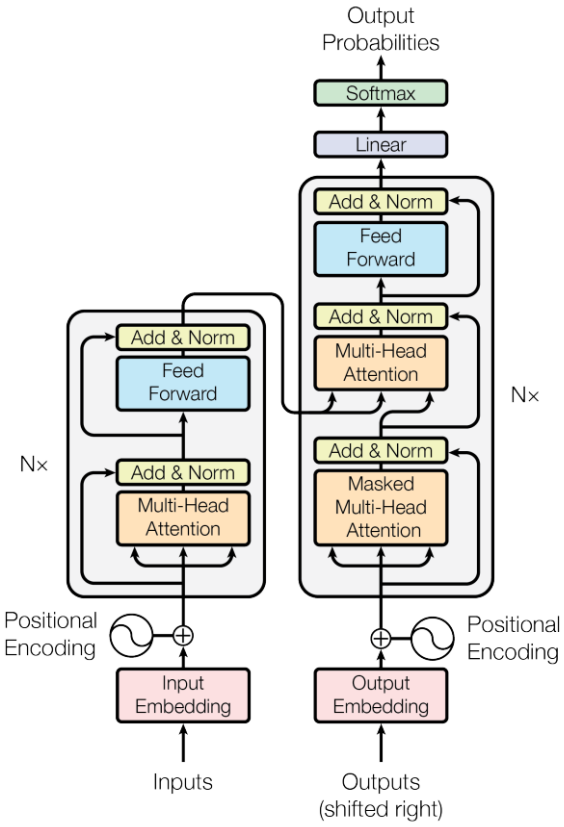

图 2

到整个 NLP，再到 CV，已经成为了一个广泛应用的的技术。

3 注意力机制

在上一篇文章引入的注意力机制的基础上，《attention is all you need》一文在注意力机制的使用方面取得了很大的进步，并且提出了 transformer 模型。transformer 架构是大模型的基础，比如后面的 BERT 就是由几十个 transformer 组成的。

transformer 完全依赖于注意力机制来捕捉输入与输出之间的全局依赖关系，它不依赖于 CNN、RNN 等模型，但仍然具备它们的优点：可以做并行计算、相比于 LSTM 更好地解决了长距离依赖问题。

transformer 有如下创新点：

- 只使用注意力机制：这种做法相较于 RNN、LSTM，在处理序列任务时，能更好的解决长时依赖问题以及并行处理序列。
- 自注意力机制：transformer 中使用缩放点积注意力（Scaled Dot-Product Attention）来实现，使得自注意力机制可以捕获输入序列中的全局依赖关系，而不仅仅是局部的信息。
- 并行计算：由于 Transformer 模型基于多头注意力机制不依赖于序列，来自不同位置的信息可以同时被处理，这显著提高了计算效率。

transfomer 的模型结构见图二。

transformer 模型的整体架构可以分为编码器（Encoder）和解码器（Decoder）两部分。一个编码器由多层相同的层组成，每层有两个子层：一个多头自注意力层（Multi-Head Self-Attention）和一个前馈神经网络层（Feed-Forward Neural Network）。每个子层都有一个残差连接（Residual Connection）和层归一化（Layer Normalization）。解码器也由多层相同的层组成，每层有三个子层：一个掩码多头自注意力层（Masked Multi-Head Attention），一个多头注意力层（Multi-Head Attention）用于处理编码器的输出，和一个前馈神经网络层。每个子层也都有一个残差连接和层归一化。

此外，自注意力机制本身并不能处理顺序信息，因为它是对输入序列的全局处理。而在自然语言处理任务中，单词的顺序是非常重要的，为了解决这个问题，

transformer 模型引入了位置编码(Position Encoding),将顺序信息添加到输入序列的表示中。

Transformer 模型是自然语言处理领域的一项重要创新,它通过自注意力机制,使得模型可以更好地处理序列数据,尤其是长距离的依赖关系。这在很大程度上解决了以前的 RNN 和 LSTM 模型在处理长序列时出现的梯度消失和计算效率低的问题。Transformer 模型的这种优势使得它在许多自然语言处理任务中取得了显著的成绩,包括但不限于机器翻译、文本摘要、情感分析等。例如,Google 的神经机器翻译系统就采用了 Transformer 模型,大大提高了翻译的质量和效率。

此外,Transformer 模型也是许多最新的自然语言处理模型的基础。在原始的 Transformer 模型之后,大语言模型研究开始向两个方向分化:基于编码器结构的 Transformer 模型用于预测建模任务,例如文本分类;而基于解码器结构的 Transformer 模型用于生成建模任务,例如翻译、摘要和其他形式的文本内容生成。如 BERT(Bidirectional Encoder Representations from Transformers)、GPT(Generative Pretrained Transformer)等。这些模型在 Transformer 模型的基础上,加入了预训练,多任务等技术,进一步提高了模型的性能,开创了自然语言处理的新篇章。

论文最后作者也提出 Transformer 的应用不仅仅局限于自然语言处理,它的自注意力机制还被广泛应用于其他领域,例如计算机视觉。许多最新的计算机视觉模型,如 ViT(Vision Transformer),也采用了 Transformer 作为其核心结构,这进一步证明了 Transformer 的通用性和有效性。

4 BERT: 语言理解的深度双向 Transformer 预训练

BERT 采用了基于微调(Fine-turing)的预训练(Pre-tranining)方式。在预训练过程中,模型在不同的预训练任务上对未标记数据进行训练。对于微调,首先使用预训练的参数初始化 BERT 模型,然后使用来自下游任务的标记数据对所有参数进行微调。BERT 的一个显著特征是其跨不同任务的统一架构,预训练的体系结构和最终的下游体系结构之间的差别很小。

使用预训练模型做特征表示的时候一般有两类策略:基于特征的和基于微调的。而这两类策略的代表分别有 ELMo 和 GPT。BERT 和 GPT、ELMo 的区别主要有(图三):

- GPT 因为掩码注意力的限制,考虑的是单向的信息,即用左边的信息去预测未来;而 BERT 使用了左侧和右侧,即双向的信息。
- ELMo 是一个基于 LSTM 的架构,采用的是基于特征(Feature-based)的预训练方式,而 BERT 使用的是 Transformer,所以 ELMo 在用于一些下游任务时,需要对架构做出调整,而 BERT 只需要在语言模型之后再接其他的模型,可以说它学习的是预训练模型提取的高级语义。

BERT 并没有什么本质上的创新,但它是一个集大成者:

- 参考了 ELMo 模型的双向编码思想。
- 借鉴了 GPT 使用 Transformer 作为特征提取器的思路。
- 采用了 Word2vec 所使用的 CBOW 方法。

之前的模型是从左向右输入一个文本序列,或者将 left-to-right 和 right-to-left 的训练结合起来。而 BERT 将双向 Transformer 用于语言模型,实验的结果表明,双向训练的语言模型对语境的理解会比单向的语言模型更深刻。BERT 只使用了 Transformer 的 encoder 部分,一次性读取整个文本序列,而不是从左到右或从右到左地按顺序读取,这个特征使得模型能够基于单词的两侧学习,相当于是一个双向的功能。

BERT 模型使用了两个无监督任务来预训练:Masked LM 和 Next Sentence Prediction。由于 BERT 采用的是双向语言模型,这意味着模型在训练时要预测的前文和后文都已知,在答案已知的情况下训练的模型是不准确的,因此 BERT 采用学习“完形填空”的方式。即掩码语言模型(Marked LM),将文章中的一些词进行掩盖或替换,然后再训练。由于很多下游任务比如问答(QA)、自然语言推断(NLI)等,都是基于理解两个句子之间的关系的,因此 BERT 增加了第二个预训练任务——Nest Sentence Prediction,用来预测下一句是否是上一句的下一句。BERT 采用 50% 是上下句的句对,50% 不是上下句的句对进行训练,并在训练时两个句子之间加入“[SEP]”进行分隔,从而使两个句子合并成一个句子。

这篇论文提出的 BERT 的意义重大。在 BERT 之前,NLP 领域一直没有一个很深的神经网络,能够使得训练好的模型适用于一大片的 NLP 任务。研究者都

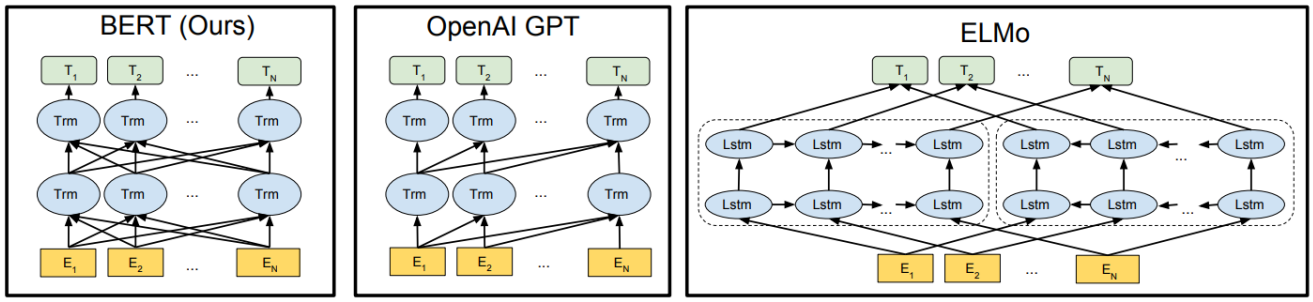

图 3

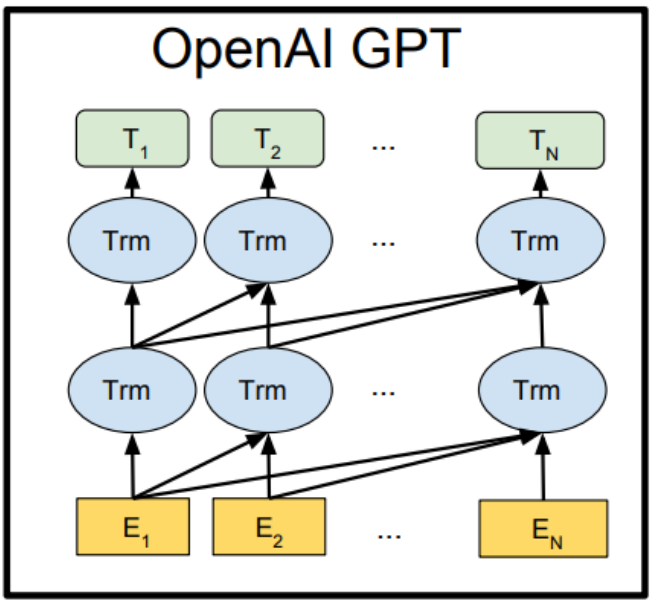

图 4

是对每个任务构建自己的神经网络，在自己的任务上做训练。而 BERT 的出现使得我们可以在一个比较大的数据集上预训练好一个比较深的神经网络，然后应用在很多 NLP 任务上。这样一来既简化了 NLP 任务的训练，又提升了它们的性能。所以 BERT 和它之后的一系列工作使得 NLP 在过去几年里有了一个质的飞跃。

5 GPT: 通过生成式预训练改进语言理解

这篇文章提出的模型主要由两个阶段构成：第一阶段是在大量的语料上用无监督的方式训练一个语言模型，第二阶段是根据下游的任务进行有监督式的微调。模型架构如图四。

第一阶段的优化目标为：

$$L_1(u) = \sum_i logP(u_i|u_{i-k}, ..., u_{i-1}; \Theta)$$

这是一个经典的自回归模型，它利用生成式 (Generative) 的无监督方式预训练 (Pre-Train) 模型来提高语言理解能力。GPT 的模型架构是一个多层堆叠而成的 Transformer 里的解码器部分，并在四种类型的语言理解任务上进行了评估：自然语言推理（Natural Language Inference），问题回答（Question Answering），语义相似度（Semantic Similarity）和文本分类（Text Classification）。在所有任务上，生成式预训练模型都超越了使用特定任务架构或监督学习方法的基线模型，并且只需要很少或没有微调就可以达到最佳效果。

生成式预训练有两个优点：

- 一是可以利用无标注文本中隐含的语法、语义和常识知识来学习语言表示，从而提高模型在各种任务上的泛化能力；
- 二是可以通过生成式目标函数来避免使用特定任务相关的架构或损失函数，从而简化了模型设计和训练过程。

第二阶段的微调是以有监督的形式进行的，将文本输入放进第一阶段训练所得的模型，取最后一层的隐层输出，加一个简单的线性层做映射，最后以 softmax 转化为概率形式，就得到需要的最大化的似然函数表达形式：

$$P(y|x^1, ..., x^m) = softmax(h_t W_y)$$

$$L_2(C) = \sum_{(x,y)} logP(y|x^1, ..., x^m)$$

在微调阶段加入语言模型建模作为辅助目标可以提高模型的泛化能力和收敛速度。微调时只需要增加

输出层的参数 W_y 和分隔符 token 的 embedding。因此，最终的训练损失函数是 L_2 与缩放后的 L_1 的和：

$$L_3(C) = L_2(C) + \lambda * L_1(C)$$

对于文本分类，可以直接对模型进行微调。对于其他有结构化输入的任务，如问答或文本蕴含，需要将输入转换为有序的文本序列。这样可以避免对模型架构做大量的改动。如图五。

6 BART：用于自然语言生成、翻译和理解的降噪序列对序列预训练

BART 模型作为一种 Seq-to-Seq 结构的预训练模型，是由 Facebook 在 ACL 2020 上首次提出。BART 模型结合了 BERT 的双向 transformer 和 GPT 的自回归 transformer，可用于文本生成和理解（图六）。

BART（Bidirectional and Auto-Regressive Transformers，双向自回归变压器）是一种采用序列到序列模型构建的降噪自编码器，适用于各种最终任务。它使用基于标准 transformer 的神经机器翻译架构。BART 的预训练包括：

- 使用噪声函数破坏文本；
- 学习序列到序列模型以重建原始文本。

这些预训练步骤的主要优势在于：该模型可以灵活处理原始输入文本，并学会有效地重建文本。BART 模型架构与 transformer 相同，但它参考了 GPT 模型，将原有 ReLU 激活函数变成了 GeLUs 函数。在预训练时，首先使用多种噪声对原始文本进行破坏，然后通过 seq2seq 模型重建原始文本。BART 共介绍了 5 种破坏原始文本的噪声方法，噪声方式如图七。

在微调阶段，为了将 BART 用于自然语言理解和自然语言生成等下游任务中，采用了图八的方式。

- 左边是分类任务的微调方式，输入将会同时送入 Encoder 和 Decoder，最终使用最后一个输出为文本表示。
- 右边是翻译任务的微调方式，由于翻译任务的词表可能和模型词表不同，所以这里使用一个新的小型 Encoder 替换 BART 中的 Embedding。

BART 吸收了 BERT 的 bidirectional encoder 和 GPT 的 left-to-right decoder 各自的特点，建立在标准的 seq2seq Transformer model 的基础之上，这使得它比 BERT 更适合文本生成的场景；相比 GPT，也多了双向上下文语境信息。同时，相较于 BERT 中单一的噪声类型 (只有简单地用 [MASK] token 进行替换这一种噪声)，BART 在 encoder 端尝试了多种噪声。

BERT 的这种简单替换导致的是 encoder 端的输入携带了有关序列结构的一些信息（比如序列的长度等信息），而这些信息在文本生成任务中一般是不会提供给模型的。BART 采用更加多样的噪声，意图是破坏掉这些有关序列结构的信息，防止模型“依赖”这样的信息。

7 总结与展望

注意力机制是 Transformer 的重点，而 Transforemr 又是后面 BERT、GPT、BART 的基础架构，这是大语言模型发展的重要节点。其中：

- GPT：是一种 Auto-Regressive(自回归) 的语言模型。它也可以看作是 Transformer model 的 Decoder 部分，它的优化目标就是标准的语言模型目标：序列中所有 token 的联合概率。GPT 采用的是自然序列中的从左到右（或者从右到左）的因式分解。
- BERT：是一种 Auto-Encoding(自编码) 的语言模型。它也可以看作是 Transformer model 的 Encoder 部分，在输入端随机使用一种特殊的 [MASK] token 来替换序列中的 token，这也可以看作是一种噪声，所以 BERT 也叫 Masked Language Model。
- BART：吸收了 BERT 的 bidirectional encoder 和 GPT 的 left-to-right decoder 各自的特点；建立在标准的 seq2seq Transformer model 的基础之上，这使得它比 BERT 更适合文本生成的场景；此外，相比 GPT 也多了双向上下文信息。在生成任务上获得进步的同时，它也可以在一些文本理解类任务上取得 SOTA。

本文着重介绍了对语言模型的发展极为重要的五篇论文，它们作为大语言模型里程碑式的论文，从中我们可以基本了解大语言模型背后的主要思想：预训练和微调。从模型架构方面看，由堆叠的多头自注意力层组成的 transformer 已经成为构建 LLMs 的普遍架构。

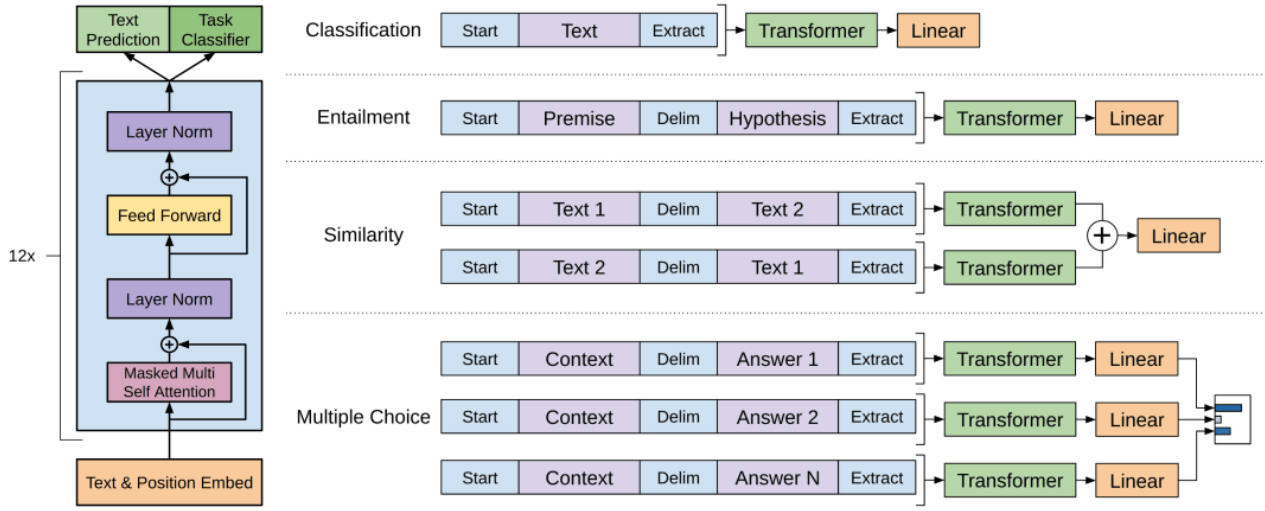

图 5

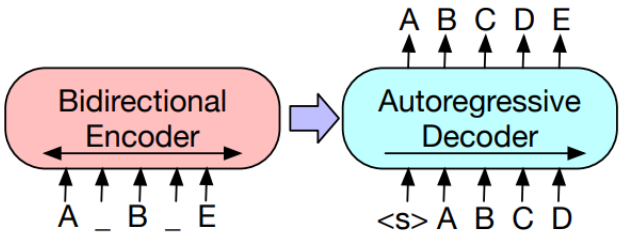

图 6

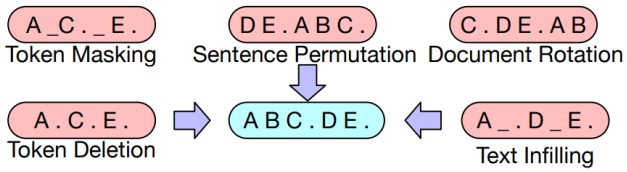

图 7

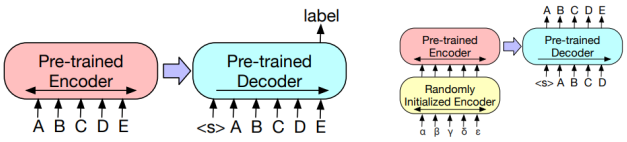

图 8

当前研究表明，当语言模型的参数规模增加到一个临界点时，一些新兴能力会以一种意想不到的方式出现（涌现现象），典型的包括上下文学习、指令跟随和分步推理等。目前 LLMs 的这些涌现能力还难以解释，而这些基本问题对于开发下一代的 LLMs 有着重要作用，值得进一步探索。同时，由于训练数据可能存在的问题以及使用者恶意的激发指令，LLMs 可能产生有害的、有偏见的文本。因此，大语言模型的安全问题也需要得到进一步的完善。

此外，随着 LLMs 的技术升级，它也将被使用到越来越多的领域。作为一个显著的进步，ChatGPT 已经潜在地改变了人类获取信息的方式，这也带来了新必应的发布。在不久的将来，可以预见，LLMs 将对信息搜索技术产生重大影响，包括搜索引擎和识别系统。

8 参考文献

[1] Bahdanau, D. , Cho, K. , Bengio, Y. . (2014). Neural machine translation by jointly learning to align and translate. Computer Science.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[3] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] Radford, A., Narasimhan, K., Salimans, T.,

Sutskever, I. (2018). Improving language understanding by generative pre-training.

[5] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[6] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.