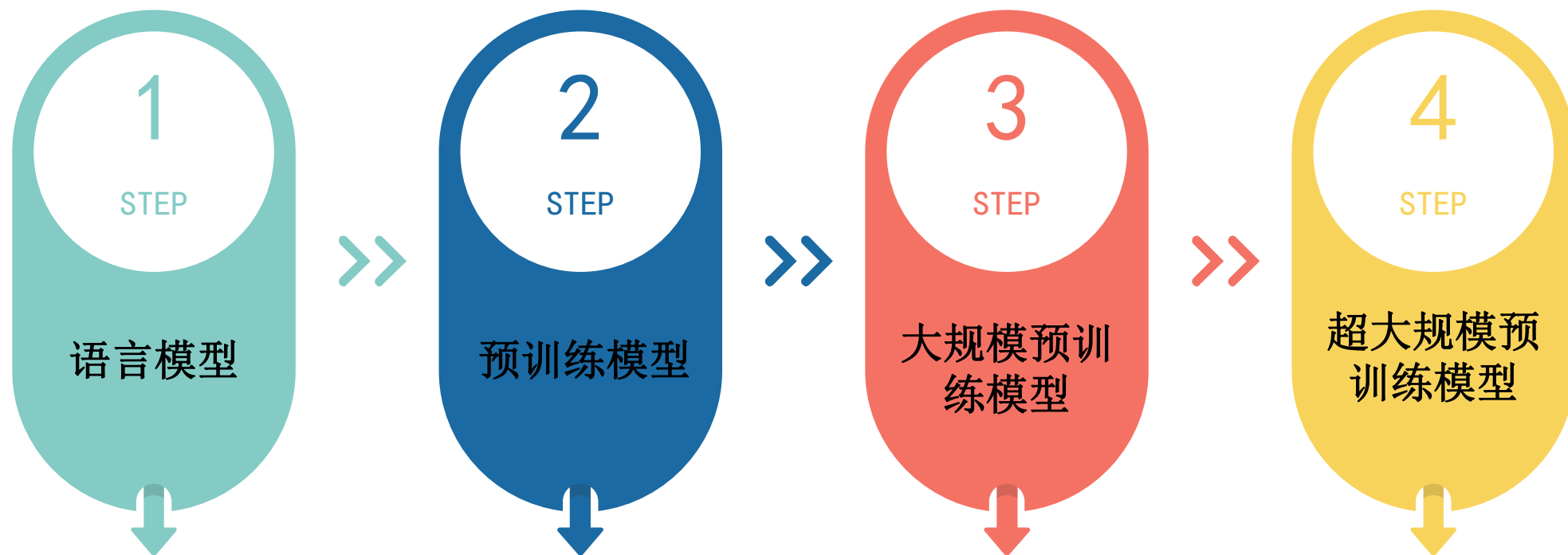


AI大模型发展概况



在自然语言处理中（NLP）占很大地位。主要是为一个长度为 m 的词列确定一个概率分布 P 。

对于很长的词序列，为了更好地捕捉长距离信息，找到一种更强的语言模型方法，由此提出了以transformer结构为基础的预训练语言模型。本质：根据上下文去预测下一个词是什么。

参数量突破。

参数量再突破。

语言模型发展

将神经网络(NN)应用于LM中，
为每个单词学习一个分布式
表征来实现在连续空间上的
建模, 有效避免了数据稀疏
的问题。缺点：只能处理定
长的序列。

除了记忆单元和NN的部分，
LSTM-RNNLM 的架构几乎与
RNNLM 一样。至此，NNLM逐
渐成为主流的语言模型，并
得到了迅速发展。



能实现语言模型的基本作用，
即为一个长度为m的词序列
确定一个概率分布P。缺点：
当n较大时，会出现数据稀
疏问题，导致估算结果不准
确。

可以处理变长的序列。缺点：
RNN训练过程中，可能会发
生梯度消失或者梯度爆炸，
导致训练速度变慢或使得参
数值无穷大，无法解决长时
依赖问题。

预训练语言模型



嵌入语言模型 (ELMO)：解决在实际情况中同一个单词在不同的语境里有不同的含义。



GPT系列模型：

GPT-1：主要包括无监督预训练和有监督微调两个阶段。相比ELMo，GPT-1真正意义上实现了预训练-微调的框架。虽然GPT-1往往只需要简单的微调便能取得非常好的效果，但在未经微调的任务上其泛化能力远远低于经过微调的有监督任务，说明了GPT-1只是一个简单的领域专家，而非通用的语言学家。

GPT-2：提升模型的容量和数据多样性，让语言模型能够达到解决任何任务的程度。

GPT-3：应用时不进行梯度更新或者微调，仅使用任务说明和个别示例与模型进行文本交互，同时，GPT-3也不像GPT-2追求零样本学习。GPT-3参数量达到1750亿。









BERT系列模型：

2018年，Google提出了该模型，它主要有两大创新：一是借助Transformer学习双向表示，将多个Transformer编码器堆叠在一起。二是在预训练方法的基础上，使用掩码语言模型(MLM)和下一句预测(NSP)分别捕捉词语和句子级别的语义表征。因此，经过预先训练的BERT模型只需一个额外的输出层就可以进行微调，从而为各种自然语言处理任务生成最新模型。

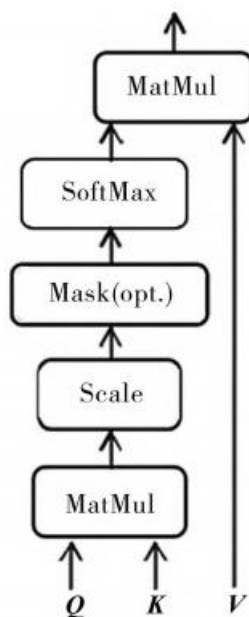
工业大模型发展

自2018年以来，国内外超大规模预训练模型参数指标不断创出新高，相继推出各自的巨量模型。

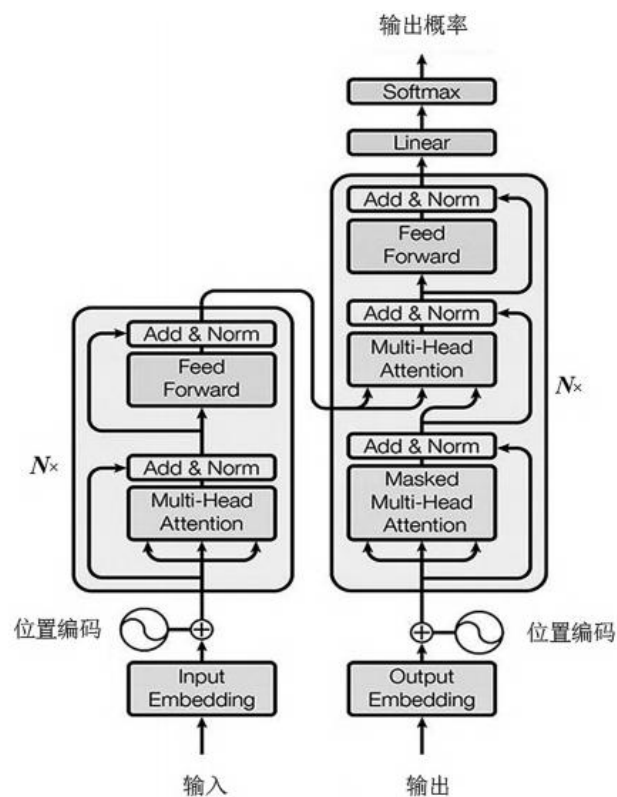
| | | | |
|--|--------------|--------------------|----------------------|
|  M6 | 1000 亿个参数 | 2021-03-01 发布时间 | 查看详情 |
|  Stable Diffusion-1.0 | 9.83 亿个参数 | 2022-08-00 发布时间 | 查看详情 |
|  Whisper | 15.5 亿个参数 | 2022-09-21 发布时间 | 查看详情 |
|  ERNIE-ViLG 2.0 | 240 亿个参数 | 2022-10-27 发布时间 | 查看详情 |
|  Stable Diffusion-2.0 | 9.83 亿个参数 | 2022-11-24 发布时间 | 查看详情 |
|  Whisper V2 | 15.5 亿个参数 | 2022-12-09 发布时间 | 查看详情 |
|  MOSS | 160 亿个参数 | 2023-02-20 发布时间 | 查看详情 |
|  PaLM-E | 5620 亿个参数 | 2023-03-06 发布时间 | 查看详情 |
|  Visual ChatGPT | 1750 亿个参数 | 2023-03-08 发布时间 | 查看详情 |
|  GPT-4 | 1750 亿个参数 | 2023-03-14 发布时间 | 查看详情 |
|  HuggingGPT | 0 亿个参数 | 2023-03-31 发布时间 | 查看详情 |
|  LVDM | 0 亿个参数 | 2023-04-06 发布时间 | 查看详情 |
|  Whisper JAX | 15.5 亿个参数 | 2023-04-14 发布时间 | 查看详情 |

AI大模型的两大核心技术

自注意力机制(Self-Attention), 又称内部注意力机制: 是一种将单个序列的不同位置关联起来以计算同一序列表示的注意力机制。通过引入自注意力机制, 可以捕获同一个句子中单词之间的一些句法特征或者语义特征, 同时也更容易捕获句子中长距离的相互依赖的特征, 对于增加计算的并行性也有直接帮助作用。



Transformer模型结构: 由多个编码器和解码器叠加组成, 其中编码器和解码器均是由基于自注意力的模块叠加而成的, 其输入分别是源序列和目标序列的嵌入表示加上位置编码。



全模态大模型“紫东太初” 2.0

6月16日，在人工智能框架生态峰会2023上，“紫东太初”全模态大模型重磅发布。

该大模型是在千亿参数多模态大模型“紫东太初”1.0基础上升级打造的2.0版本，在语音、图像和文本三模态的基础上，加入了视频、信号、3D点云等模态数据，研究突破了认知增强的多模态关联等关键技术，具备全模态理解能力、生成能力和关联能力。

不同于以ChatGPT背后的GPT为代表的大型语言模型(LLM)，“紫东太初”2.0实现了更全的模态覆盖。

「紫东太初2.0」现已支持多轮问答、文本创作、图像生成、3D理解、信号分析等全面问答任务，拥有更强的认知、理解、创作能力，带来全新互动体验。实现了真正意义上的任意输入，任意输出。可以说大模型的发展已经从单模态、多模态，进化到全模态赛道上了。



未来大模型的应用总结

由于成本问题，训练通用大模型不是未来行业具体应用所围绕的核心。相比而言，垂直大模型作为一种全新的生产力，随着其底层能力的不断突破，必然能实现企业的降本增效，带来上层应用的迭代和变革。未来会是‘底座模型+垂直类’的模型互相融通。”

国内的大模型基本上都聚焦语言模型，而世界70%的数据是靠视觉，10%靠触觉，其他的靠听觉及其他的信 息，要实现低功耗更类人的智能，必须融合多模态的信息。所以全模态大模型“紫东·太初2.0”定位相当于拥有世界知识的底座模型。

在算力、算法和数据中，数据是垂直大模型的难点。公开数据库中有大量无标注数据，既要利用好大量无标注数据，又要利用好少量高精度数据，所以对模型建构提出了较高要求。