电子AI+系列专题报告（一）
AI大语言模型的原理、演进及算力测算

行业研究 · 行业专题

电子

投资评级：超配（维持评级）

证券分析师：胡剑
021-60893306
hujian1@guosen.com.cn
S0980521080001

证券分析师：胡慧
021-60871321
huhui2@guosen.com.cn
S0980521080002

证券分析师：周靖翔
021-60375402
zhoujingxiang@guosen.com.cn
S0980522100001

证券分析师：李梓澎
0755-81981181
lizipeng@guosen.com.cn
S0980522090001

联系人：詹浏洋
010-88005307
zhanliuyang@guosen.com.cn

AI大语言模型的原理、演进及算力测算

国信证券 GUOSEN SECURITIES

- **机器学习中模型及数据规模增加有利于提高深度神经网络性能。**

人工智能致力于研究能够模拟、延伸和扩展人类智能的理论方法及技术，并开发相关应用系统；其最终目标是使计算机能够模拟人的思维方式和行为。机器学习是一门专门研究计算机如何模拟或实现人类的学习行为、以获取新的知识或技能、重新组织已有的知识结构使之不断改善自身性能的学科，广泛应用于数据挖掘、计算机视觉、自然语言处理等领域。深度学习是机器学习的子集，主要由人工神经网络组成。与传统算法及中小型神经网络相比，大规模的神经网络及海量的数据支撑将有效提高深度神经网络的表现性能。

- **Transformer模型架构是现代大语言模型所采用的基础架构。**

Transformer模型是一种非串行的神经网络架构，最初被用于执行基于上下文的机器翻译任务。Transformer模型以Encoder-Decoder架构为基础，能够并行处理整个文本序列，同时引入“注意机制”（Attention），使其能够在文本序列中正向和反向地跟踪单词之间的关系，适合在大规模分布式集群中进行训练，因此具有能够并行运算、关注上下文信息、表达能力强等优势。Transformer模型以词嵌入向量叠加位置编码作为输入，使得输入序列具有位置上的关联信息。编码器（Encoder）由Self-Attention（自注意力层）和 Feed Forward Network（前馈网络）两个子层组成，Attention使得模型不仅关注当前位置的词语，同时能够关注上下文的词语。解码器（Decoder）通过Encoder-Decoder Attention层，用于解码时对于输入端编码信息的关注；利用掩码（Mask）机制，对序列中每一位置根据之前位置的输出结果循环解码得到当前位置的输出结果。

AI大语言模型的原理、演进及算力测算

国信证券 GUOSEN SECURITIES

- **GPT是基于Transformer架构的大语言模型，近年迭代演进迅速。**

构建语言模型是自然语言处理中最基本和最重要的任务之一。GPT是基于Transformer架构衍生出的生成式预训练的单向语言模型，通过对大量语料数据进行无监督学习，从而实现文本生成的目的；在结构上仅采用Transformer架构的Decoder部分。自2018年6月OpenAI发布GPT-1模型以来，GPT模型迭代演进迅速。GPT-1核心思想是采用“预训练+微调”的半监督学习方法，服务于单序列文本的生成式任务；GPT-2在预训练阶段引入多任务学习机制，将多样化的自然语言处理任务全部转化为语言模型问题；GPT-3大幅增加了模型参数，更能有效利用上下文信息，性能得到跨越式提高；GPT-3.5引入人类反馈强化学习机制，通过使用人类反馈的数据集进行监督学习，能够使得模型输出与人类意图一致。

- **大语言模型的训练及推理应用对算力需求带来急剧提升。**

以GPT-3为例，GPT-3参数量达1750亿个，训练样本token数达3000亿个。考虑采用精度为32位的单精度浮点数数据来训练模型及进行谷歌级访问量推理，假设GPT-3模型每次训练时间要求在30天完成，对应GPT-3所需运算次数为3.15*10^23FLOPs，所需算力为121.528PFLOPS，以A100 PCIe芯片为例，训练阶段需要新增A100 GPU芯片1558颗，价值量约2337万美元；对应DGX A100服务器195台，价值量约3880.5万美元。假设推理阶段按谷歌每日搜索量35亿次进行估计，则每日GPT-3需推理token数达7.9万亿个，所需运算次数为4.76*10^24FLOPs，所需算力为55EFLOPs，则推理阶段需要新增A100 GPU芯片70.6万颗，价值量约105.95亿美元；对应DGX A100服务器8.8万台，价值量约175.12亿美元。

- **产业链相关公司：**工业富联、沪电股份、寒武纪、海光信息、国芯科技、全志科技。

- **风险提示：**宏观AI推广不及预期，AI投资规模低于预期，AI服务器渗透率提升低于预期 ，AI监管政策收紧等。

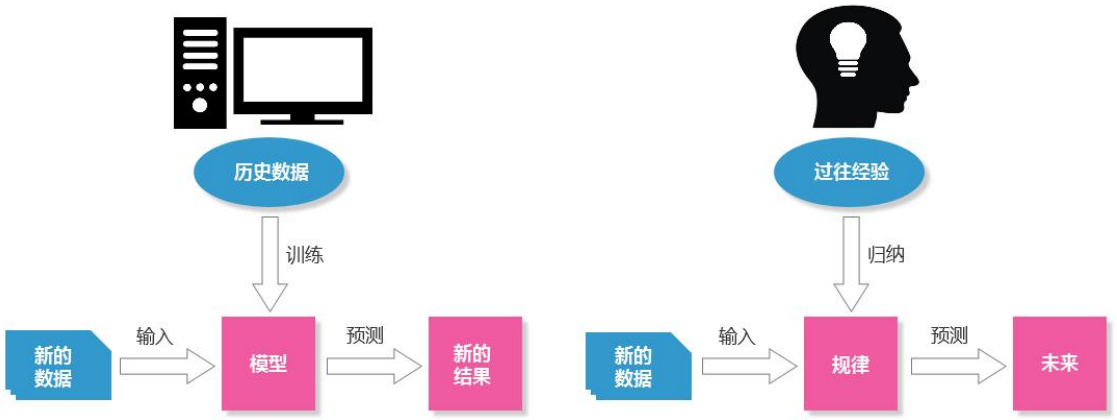
目录

一、人工智能、机器学习与神经网络简介

机器学习是实现人工智能的途径之一

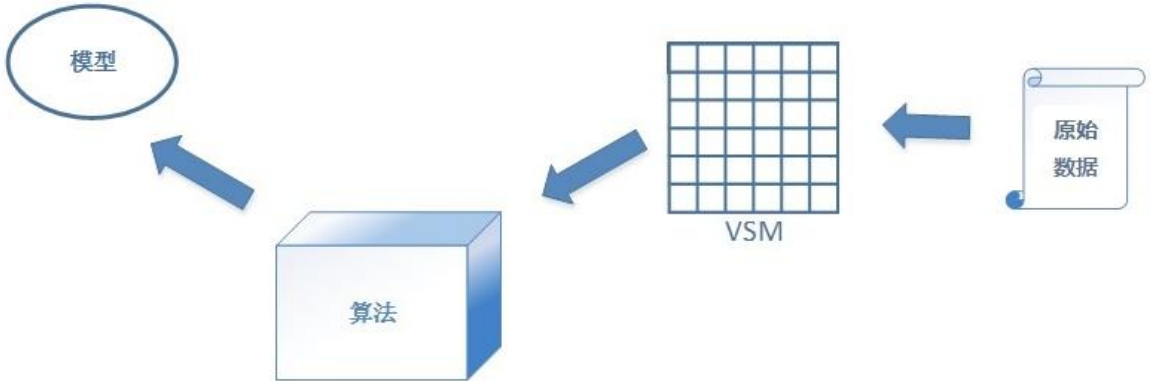
- **人工智能（Artificial Intelligence，AI）**是研究、开发用于模拟、延伸和扩展人类智能的理论、方法、技术及应用系统的一门新的技术科学。人工智能的最终目标是使计算机能够模拟人的思维方式和行为。
- **机器学习（Machine Learning，ML）**是实现人工智能的一种途径，是一门专门研究计算机如何模拟或实现人类的学习行为、以获取新的知识或技能、重新组织已有的知识结构使之不断改善自身性能的学科。
- **机器学习包括数据、模型、算法三要素。**从实践上来看，机器学习是在大数据的支撑下，通过各种算法让机器对数据进行深层次的统计分析以进行“自学”（训练模型），使人工智能系统获得了归纳推理和决策能力。机器学习作为一套数据驱动方法，已广泛应用于数据挖掘、自然语言处理、机器视觉、搜索引擎、医学诊断、生物特征识别、DNA序列测序、证券市场分析等领域。

图：机器学习训练与推理示意图



资料来源：woshipm，国信证券经济研究所整理

图：机器学习三要素



资料来源：gitbook，国信证券经济研究所整理

模型及数据规模增加有利于提高深度神经网络性能

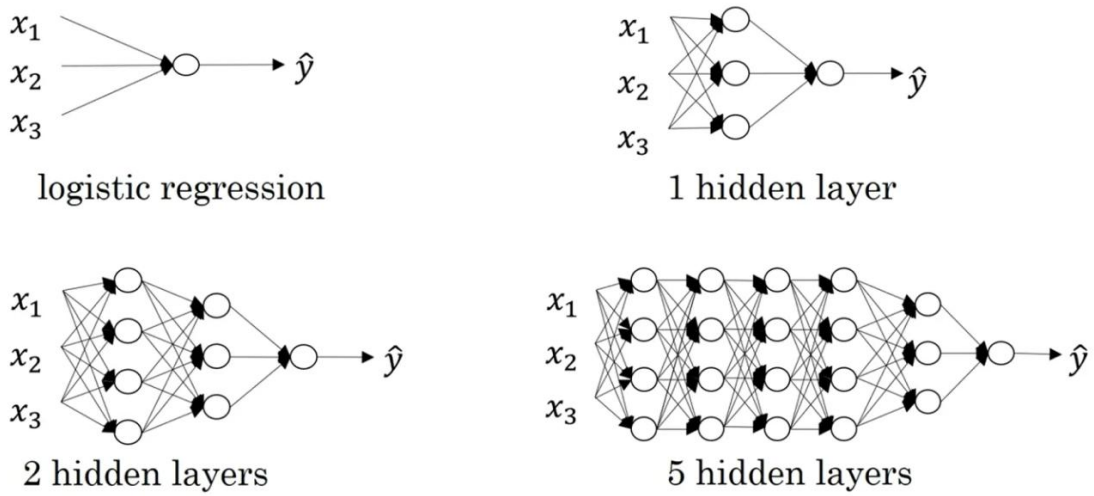
国信证券 GUOSEN SECURITIES

- **深度学习（Deep Learning，DL）是机器学习的子集，**由人工神经网络（ANN）组成。深度学习模仿人脑中存在的相似结构，其学习是通过相互关联的“神经元”的深层的、多层的“网络”来进行的。

- **典型的神经网络从结构上可以分为三层：输入层、隐藏层、输出层。**其中，输入层（input layer）是指输入特征向量；隐藏层（hidden layer）是指抽象的非线性中间层；输出层（output layer）是指输出预测值。深层神经网络即包含更多隐藏层的神经网络。

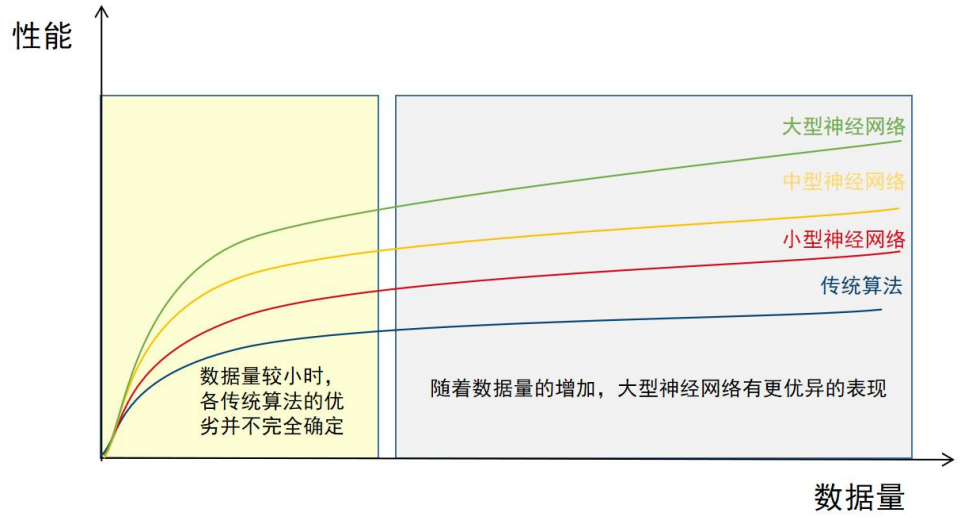
- **相比于传统机器学习模型，深度学习神经网络更能在海量数据上发挥作用。**若希望获得更好的性能，不仅需要训练一个规模足够大的神经网络（即带有许多隐藏层的神经网络，及许多参数及相关性），同时也需要海量的数据支撑。数据的规模及神经网络的计算性能，需要有强大的算力作为支撑。

图：不同深度的神经网络模型结构示意图



资料来源：Coursera，吴恩达《深度学习》，国信证券经济研究所整理

图：不同神经网络模型在不同数据量下性能曲线



资料来源： ShowMeAI，国信证券经济研究所整理

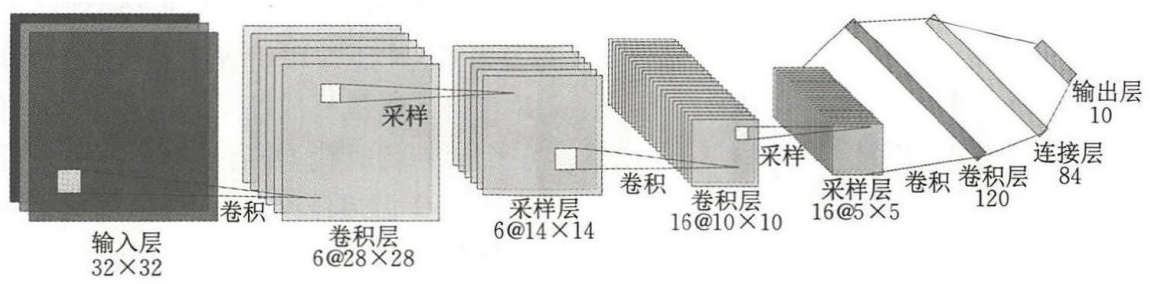
CNN和RNN是常见的神经网络模型

- **传统常见的神经网络模型包括卷积神经网络（CNN）和循环神经网络（RNN）等。**其中，卷积神经网络（Convolutional Neural Network，CNN）多用于计算机视觉、自动驾驶、人脸识别、虚拟现实、医学领域、人机交互、智能安防等图像应用；相比于标准神经网络，CNN能够更好地适应高纬度的输入数据，卷积设计有效减少了模型的参数数量。

- **循环神经网络（Recurrent Neural Network，RNN）常用于处理序列数据**（例如含有时间成分的音频和文本），获取数据中的时间依赖性。由于语言（无论是英语字母还是汉语汉字）都是逐个出现的，同时语言是时序前后相互关联的数据，因此语言作为最自然表达出来的序列数据，适合应用RNN进行语音识别、情感分类、机器翻译、语言生成、命名实体识别等应用。

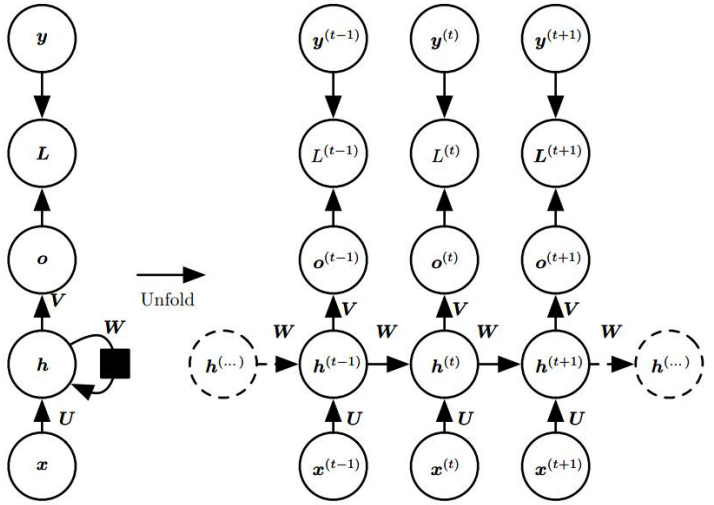
- **循环神经网络（RNN）曾是自然语言处理的首选解决方案。**RNN能够在处理单词序列时，将处理第一个词的结果反馈到处理下一个词的层，使得模型能够跟踪整个句子而非单个单词。但RNN存在缺点：由于这种串行结构，RNN无法对于长序列文本进行有效处理，甚至可能当初始单词过远时“遗忘”相关信息。

图：卷积神经网络示意图



资料来源：《深度卷积神经网络的发展及其在计算机视觉领域的应用》，国信证券经济研究所整理

图：循环神经网络示意图



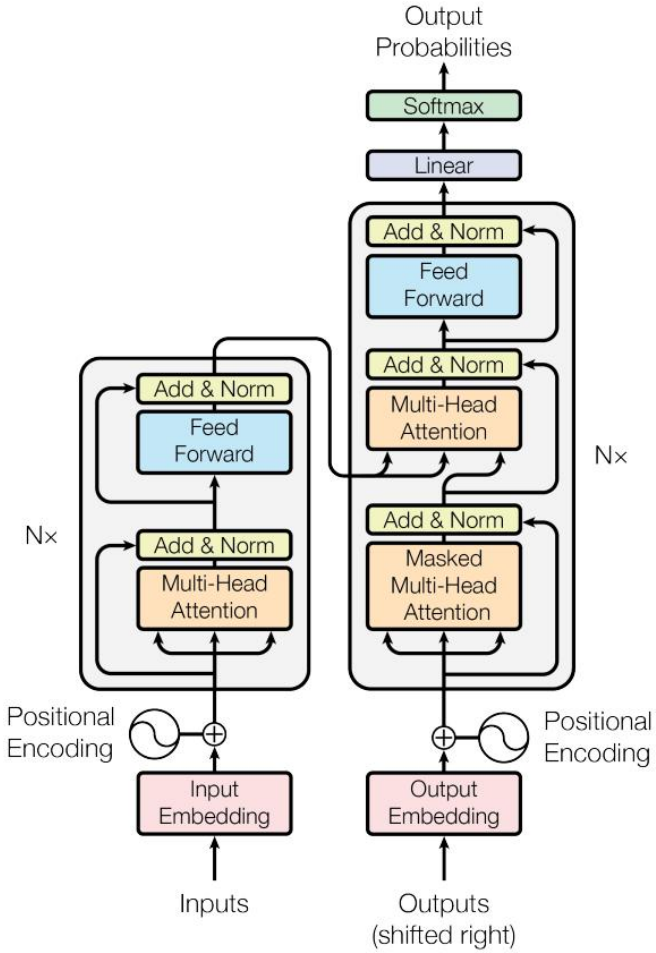
资料来源：《深度学习（中文版）》，国信证券经济研究所整理

二、Transformer模型结构分析

Transformer模型以Encoder-Decoder架构为基础

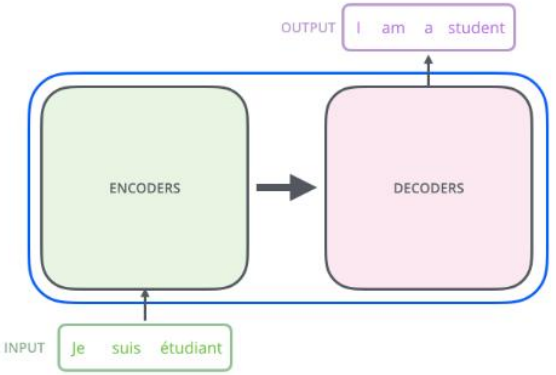
- **《Attention is all your need》 by OpenAI**
- 作为与传统的CNN、RNN不同的深度学习模型架构，**Transformer模型最初是被用于基于上下文的机器翻译模型。**由于Transformer模型非串行结构，能够并行处理整个序列；同时引入“注意机制”（attention），能够在文本序列中正向和反向地跟踪单词之间的关系，适合在大规模分布式集群中进行训练。
- **Transformer以Encoder-Decoder架构为基础。**其中，编码组件由多层编码器（Encoder）组成。解码组件也是由相同层数的解码器（Decoder）组成。Encoder用于提取源端语言的语义特征，而用Decoder提取目标端语言的语义特征，并生成相对应的译文。
- **Transformer模型具有能够并行运算、关注上下文信息、表达能力强等优势。**

图：Transformer模型介绍



图：Transformer最初用于机器翻译

图： Transformer以Encoder-Decoder架构为基础



资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

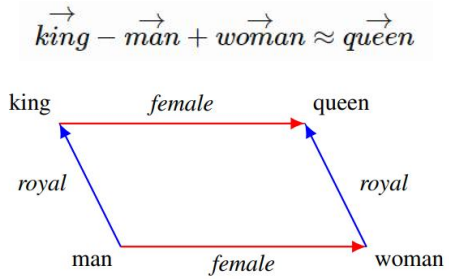
资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

资料来源：《Attention Is All You Need》，国信证券经济研究所整理

Transformer模型结构分析——词嵌入（Embedding）

- **词嵌入是NLP最基础的概念之一，表示来自词汇表的单词或者短语被映射成实数向量。**最早的词嵌入模型是word2vec等神经网络模型，属于静态词嵌入（不关注上下文）。例如大模型诞生前常用的RNN模型所用的输入便是预训练好的词嵌入。词向量能够将语义信息与空间向量关联起来（例如经典的词类比例子：king、queen、man、woman对应词向量的关系）。

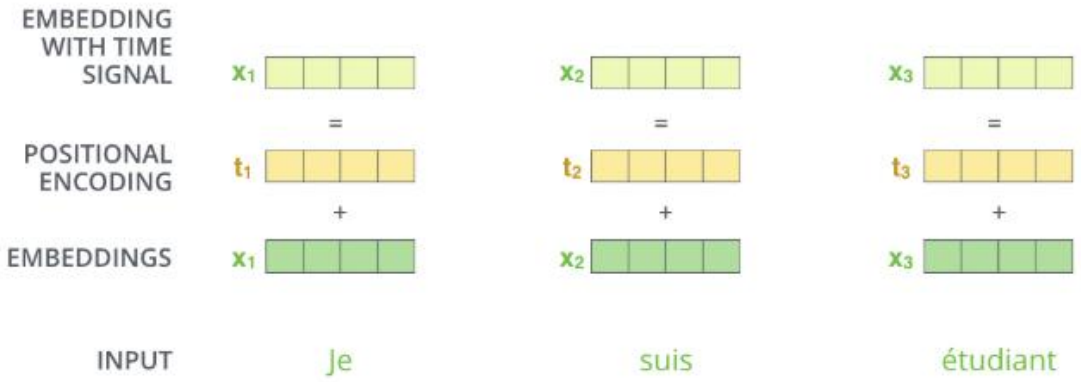
图：经典的词类比例子



资料来源：《Towards Understanding Linear Word Analogies》，国信证券经济研究所整理

- **词嵌入产生要素及步骤：**
  - Vocabulary：所有的token组成集合。
  - 词向量表：token与词向量的一一对应关系。词向量可以由预训练产生，也可以是模型参数。
  - 查表：输入的token都对应一个固定维度的浮点数向量（词嵌入向量）。
- **位置编码：**表示序列中词的顺序，具体方法为为每个输入的词添加一个位置向量。
  - 根据位置编码对应计算公式，pos表示位置，i表示维度。位置编码能够让模型学习到token之间的相对位置关系。

图：带有位置编码的词嵌入向量生成方法



资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：位置编码对应计算公式

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

资料来源：《Attention Is All You Need》，国信证券经济研究所整理

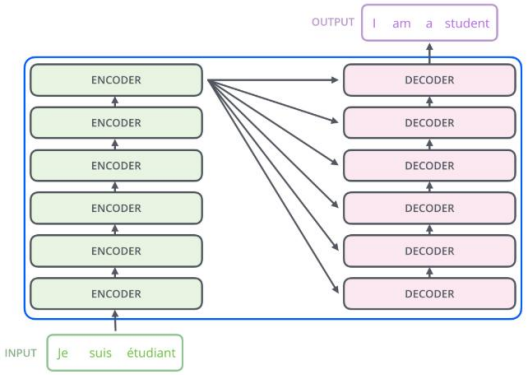
Transformer模型结构分析——Encoder

- **编码组件可由多层编码器（Encoder）组成**，同样解码组件也由相同层数的解码器（Decoder）组成。
  - 一般来讲，对于中间层的输出向量，底层Encoder输出的表示浅层含义，顶层Encoder输出的表示深层含义。
- **每个Encoder由两个子层组成：**Self-Attention层（自注意力层）和 Feed Forward Network（FFN，前馈网络）组成。
- 对于每一层Encoder，词嵌入向量输入会首先进入**Self-Attention层**，Encoder对词向量进行编码时，会对整句输入的上下文均进行Attention操作，从而关注并使用到输入句子的上下文的信息。
  - Decoder端存在Cross-Attention层（Encoder-Decoder Attention层），用于解码时对输入部分的信息进行Attention关注。
- 经过Self-Attention层的输入进入**前馈网络**，前馈网络一般是全连接层网络（并经过非线性的激活函数，如ReLU函数）。
  - 全连接层是最基本的神经网络，每一个结点都与上一层的所有结点相连。
  - ReLU函数：即修正线性单元（Rectified linear unit），又称线性整流函数，通常指以斜坡函数及其变种为代表的非线性函数。
  - 激活函数：为使神经网络具有拟合函数的能力而引入非线性；如不引入非线性，则无论多少层神经网络都相当于一个线性映射。
- 下一个Encoder的输入是上一个Encoder的输出，以此类推。

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

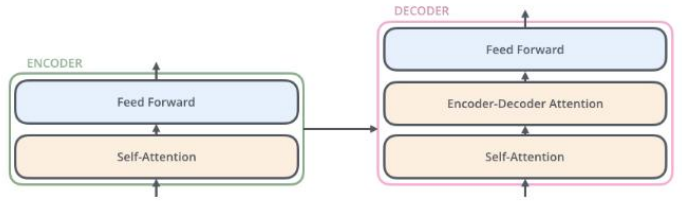
图：数据在Encoder中流动示意图

图：编码组件和解码组件均可由多层Encode/Decoder组成

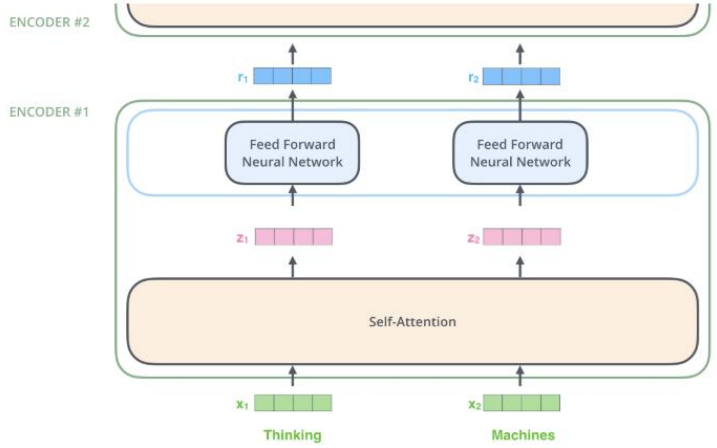


资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：Encoder由Self-Attention和FFN两个子层组成



资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

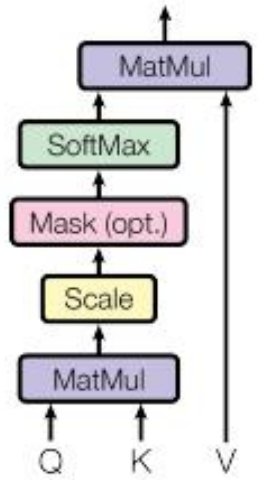


资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

Transformer模型结构分析——Attention

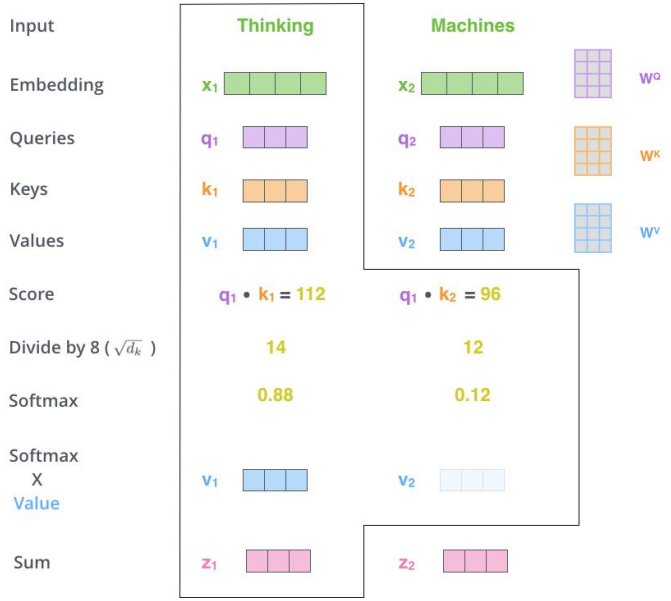
- **Attention机制：**编码时，模型不仅能够关注当前位置的词语，同时能够关注上下文的词语。
- Attention由**Q（Query）、K（Key）、V（Value）**三个矩阵实现（分别对应q、k、v三组向量；其中产生Q、K、V的三个权重矩阵W^Q、W^K、W^V为模型参数，通过训练获得）。
- 对于计算某一词向量x1与其他词向量（包括自身）的注意力分数时，用该词向量的q1分别与其他词向量（包括自身）的k向量点积，得到注意力分数；以该注意力分数经过Softmax函数进行归一化处理后，得到对应权重，表示为该词向量x1与所有位置词向量的注意力权重。以该权重对对应词向量的v向量进行加权求和，得到Self-Attention层在该位置的输出。

图：以Q、K、V矩阵计算缩放点积Attention流程图



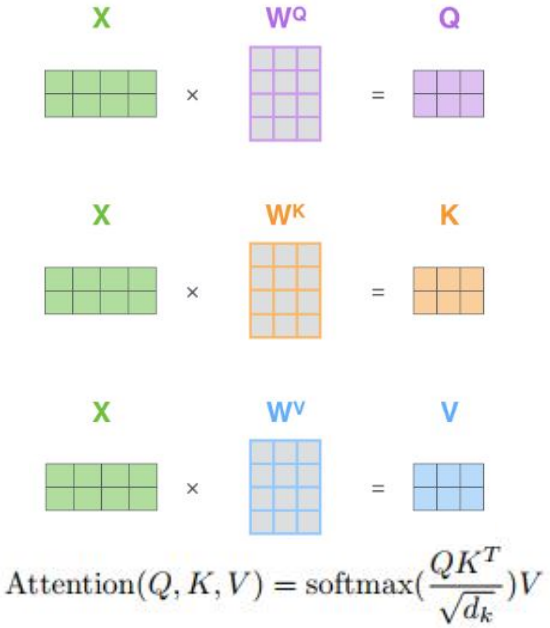
资料来源：《Attention Is All You Need》，国信证券经济研究所整理

图：词向量进入Self-Attention层后q、k、v向量计算步骤图



资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：以矩阵表示的Attention计算示意图

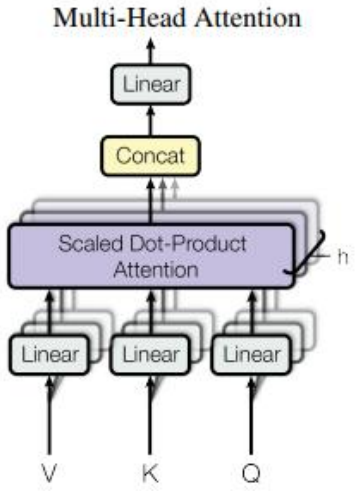


资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

Transformer模型结构分析——Multi-head Attention

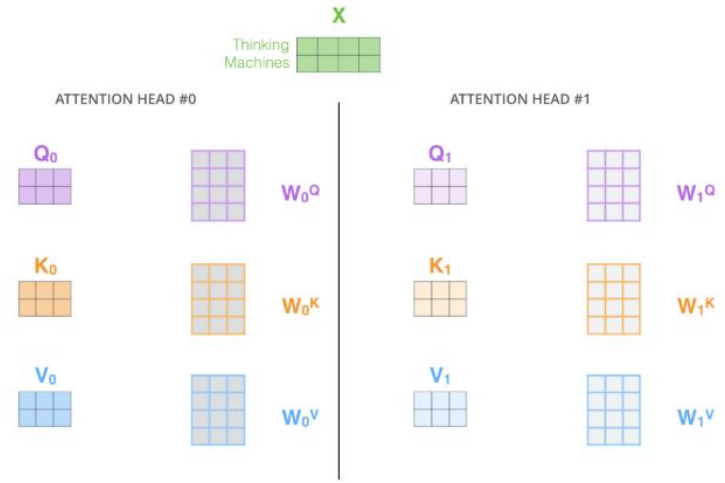
- **Multi-head Attention即多头注意力机制，**采用多组不同的线性变换对Q、K、V矩阵进行映射并分别计算Attention，再将不同的Attention结果拼接起来进行线性变换。
- **Multi-head Attention本质**是在参数总量保持不变的情况下，将Q、K、V映射到高维空间的不同子空间进行Attention计算，防止过拟合。

图：Multi-head Attention原理示意图



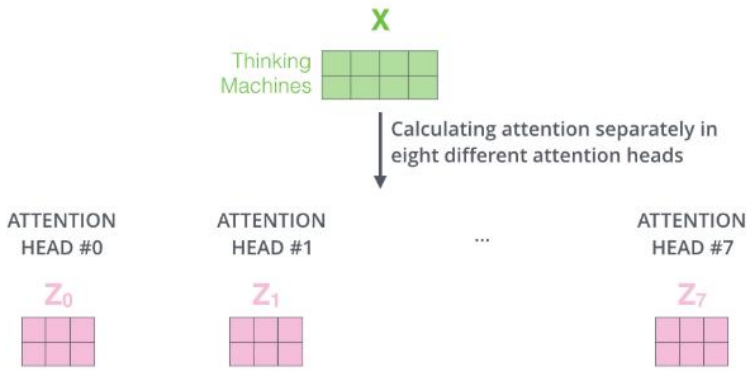
资料来源：《Attention Is All You Need》，国信证券经济研究所整理

图：词嵌入矩阵与不同权重W^Q、W^K、W^V矩阵运算得到不同Q、K、V矩阵



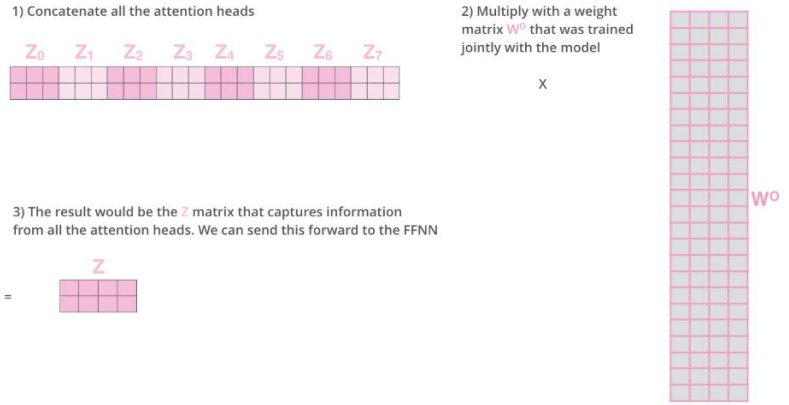
资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：不同子空间Attention运算后得到对应输出结果



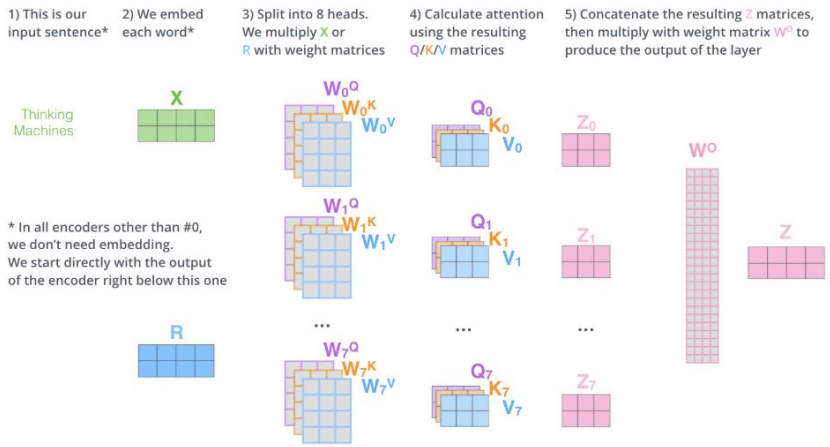
资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：不同子空间输出进行整合运算的方法



资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：Multi-head Attention计算方法与流程总结

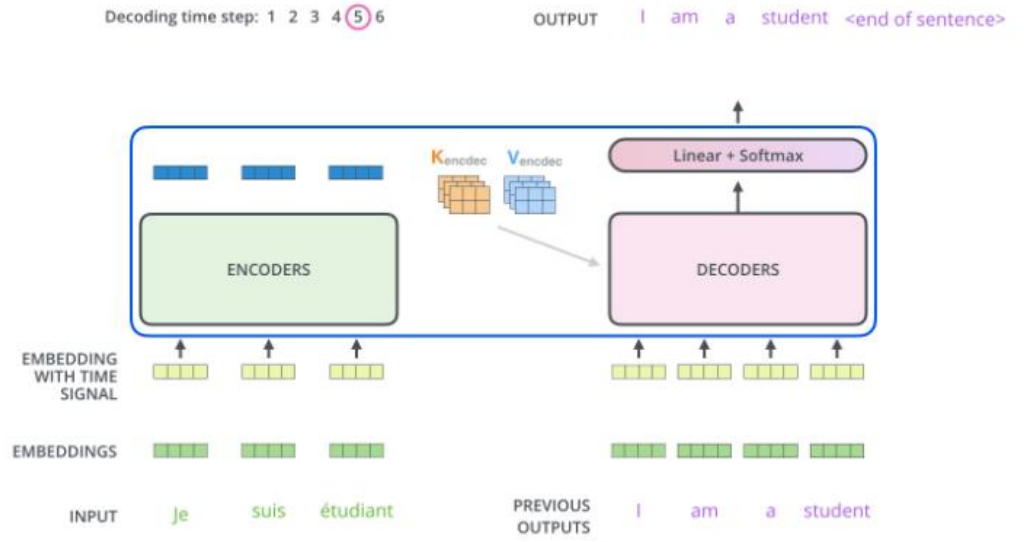


资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

Transformer模型结构分析——Decoder

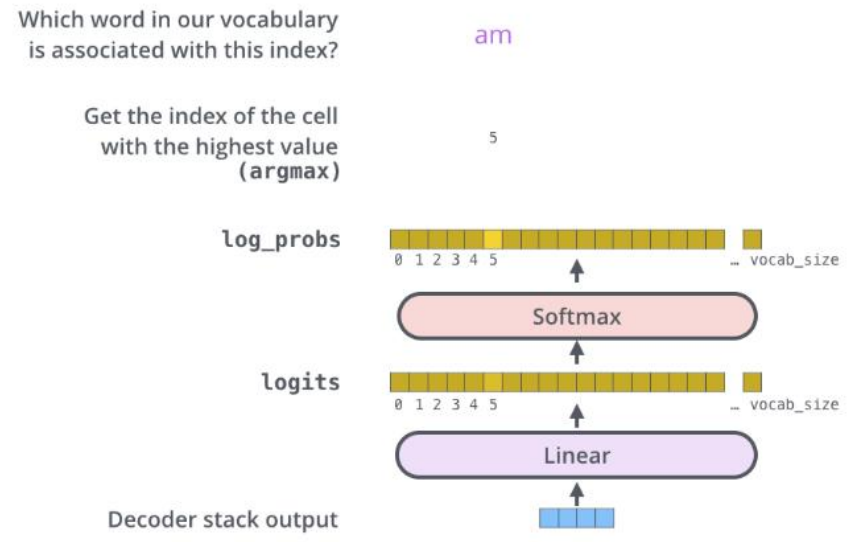
- **训练时：**输入样本句子（即翻译后的正确结果）；**推理时：**输入待定词（即mask）组成的句子。
- **Mask（掩码）**分为Padding Mask和Sequence Mask。其中，Padding Mask用于填充输入序列长度，保持输入序列对齐；Sequence Mask用于Masked Multi-Head Attention层中，使得Decoder不能获取未来的信息，从而在每个位置上仅能根据之前位置的输出结果及Encoder-Decoder Attention得到当前位置的输出。
- 当Decoder某一序列位置产生输出结果后，首先通过线性层将该输出向量映射成为维度数与vocabulary内的词数一致的向量，并通过Softmax层将每一维数字归一化为概率（即归一化后每一维数字代表对应token的概率）。
  - 训练时，构造损失函数，训练模型参数使得衡量输出概率与样本概率分布之差的损失函数值最小。
  - 推理时，根据概率采样（例如最大概率所对应）的输出词即为该位置的输出结果。

图：Decoder数据输入与输出流程示意图



资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

图：Decoder通过计算概率预测下一个token



资料来源：《The Illustrated Transformer》，GitHub，国信证券经济研究所整理

三、大规模语言模型算力需求测算（以GPT-3为例）

BERT和GPT是基于Transformer架构的两种大规模语言模型

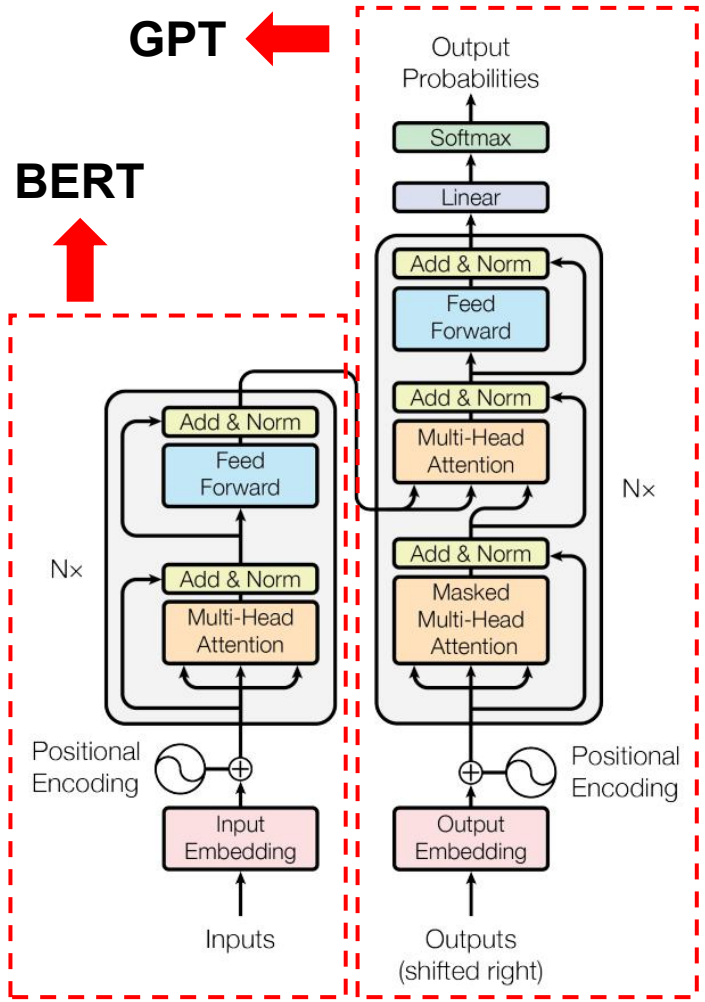
- **构建语言模型（Language Model，LM）是自然语言处理（Natural Language Processing，NLP）中最基本和最重要的任务之一，**自然语言处理基于Transformer架构衍生出了两种主流大语言模型（Large Language Model，LLM）——BERT和GPT。二者都是无监督预训练的大语言模型。

- **BERT（Bidirectional Encoder Representations from Transformer）**能够生成深度双向语言表征，是采用带有掩码（mask）的大语言模型，类似于完形填空，根据上下文预测空缺处的词语。结构上，BERT仅采用Transformer架构的Encoder部分。

- **GPT（Generative Pre-training Transformer）**是生成式预训练的单向语言模型。通过对大量语料数据进行无监督学习，从而实现文本生成的目的。结构上，GPT仅采用Transformer架构的Decoder部分。

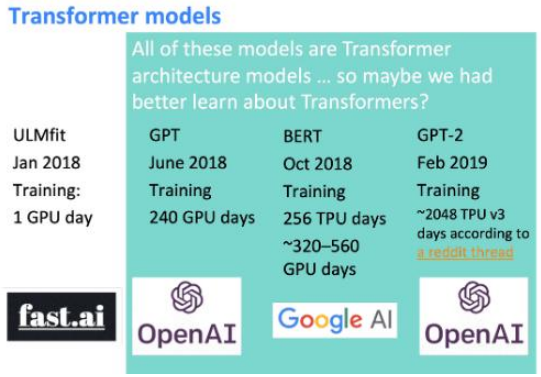
- **自2018年6月起OpenAI发布GPT-1模型以来，GPT更新换代持续提升模型及参数规模。**随着OpenAI于2022年11月30日发布ChatGPT引爆AI领域，海内外科技公司纷纷宣布发布大语言模型。用户爆发式增长对大语言模型的算力需求带来挑战。

图：基于Transformer架构的BERT和GPT模型示意图



资料来源：《Attention Is All You Need》，国信证券经济研究所整理

图：2018年6月以来发布的模型BERT和GPT以Transformer架构为主



资料来源：ShowMeAI，国信证券经济研究所整理

图：国内外科技企业发语言模型发布情况

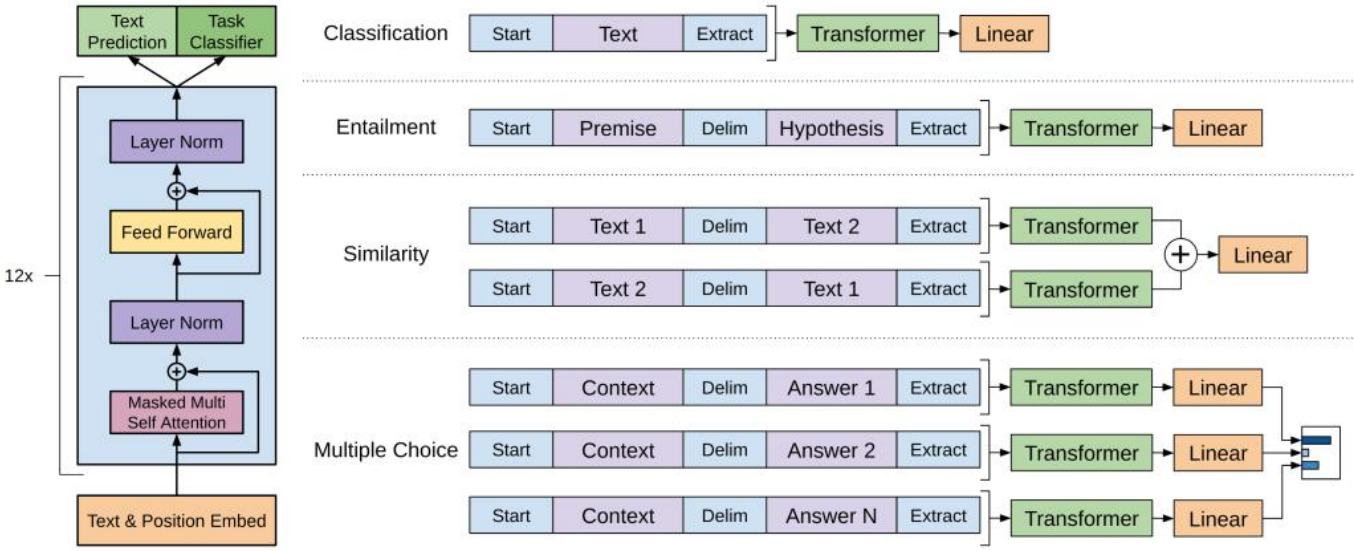
公司	产品	(拟)发布日期	阶段	链接
OpenAI	ChatGPT	2022年11月30日	开放注册	https://chat.openai.com/
Google	Bard	2月8日	公开测试	http://bard.google.com/
复旦大学	Moss	2月21日	公开测试(目前升级中)	https://moss.fastnlp.top/
澜舟科技	孟子	3月14日	已发布	https://www.langboat.com/portal/mengzi-model
百度	文心一言	3月16日	企业用户内测	https://yiyan.baidu.com/
达观数据	曹植	3月21日	可申请使用	http://www.datagrand.com/products/aigc/
清华大学	ChatGLB-6B	3月28日	已开源	https://github.com/THUDM/ChatGLM-6B
阿里巴巴	通义千问	4月7日	企业用户内测	https://tongyi.aliyun.com/
360	360智脑	4月10日	企业用户内测	http://www.360dmodel.com/
商汤科技	日日新	4月10日	即将邀请内测	https://www.sensecore.cn/
昆仑万维	天工3.5	4月17日	即将邀请内测	http://tiangong.kunlun.com/
科大讯飞	1+N认知智能大模型	5月6日	即将发布	-
网易有道	子曰	近期	即将发布	-
华为	盘古NLP模型	近期	即将发布	-
腾讯	混元助手	近期	未开放	-
京东	言犀	今年	未开放	-

资料来源：金十数据，国信证券经济研究所整理

GPT-1：预训练+微调的半监督学习模型

- 《Improving Language Understanding by Generative Pre-Training》 by OpenAI
- **GPT-1是生成式预训练模型，核心思想是“预训练+微调”的半监督学习方法，**目标是服务于单序列文本的生成式任务。
  - 生成式：表示模型建模的是一段句子出现的概率，可以分解为基于语言序列前序已出现单词条件下后一单词出现的条件概率之乘积。
    - 例如：P(一颗苹果)=P(一)P(颗|一)P(苹|一颗)P(果|一颗苹)；P(一苹颗果)=P(一)P(苹|一)P(颗|一苹)P(果|一苹颗)。
  - 预训练（无监督学习）：在无标注语料上进行无监督的预训练，通过最大化似然函数从而得到标准的GPT模型。$L_1(\mathcal{U}) = \sum_i \log P(u_i|u_{i-k}, \ldots, u_{i-1}; \Theta)$
  - 微调（有监督学习）：针对特定下游任务，采用特定的有标签数据进行微调，得到专用于情感分析、机器翻译等特定功能的模型。
- **四大常见应用：分类、蕴含、相似、选择**
  - 分类：每段文本具有对应标号，将文本按标号进行分类
  - 蕴含：给出一段文本和假设，判断该段文本中是否蕴含该假设
  - 相似：判断两段文本是否相似（用于搜索、查询、去重等）
  - 选择：对有多个选项的问题进行回答

图：GPT-1“预训练+微调”模型示意图



资料来源：《Improving Language Understanding by Generative Pre-Training》，国信证券经济研究所整理

GPT-2：强调多任务的预训练模型

国信证券 GUOSEN SECURITIES

- 《Language Models are Unsupervised Multitask Learners》 by OpenAI

- 预训练+微调的范式只能对于特定自然语言处理任务（例如问答、机器翻译、阅读理解、提取摘要等）使用特定的数据集进行有监督学习，单一领域数据集缺乏对多种任务训练的普适性。

- **GPT-2在预训练阶段便引入多任务学习机制，**通过加入各种NLP任务所需要的数据集，在尽可能多的领域和上下文中收集属于对应任务的自然语言。由此得到的GPT-2模型可以以zero-shot的方式被直接应用于下游任务，而无需进行有监督的精调。

- **GPT-2将多样化的的NLP任务全部转化为语言模型问题。**语言提供了一种灵活的方式来将任务，输入和输出全部指定为一段文本。对文本的生成式建模就是对特定任务进行有监督学习。

  - 即，所有NLP任务中的样本都能归结为一句自然语言文本。
  - 例如，翻译训练样本可以写成序列“翻译为法语，英语文本，法语文本”。同样，阅读理解训练的例子可以写成序列“回答问题，文档，问题，答案”。

图：GPT-2在部分自然问题集上生成的答案及对应正误、概率情况

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

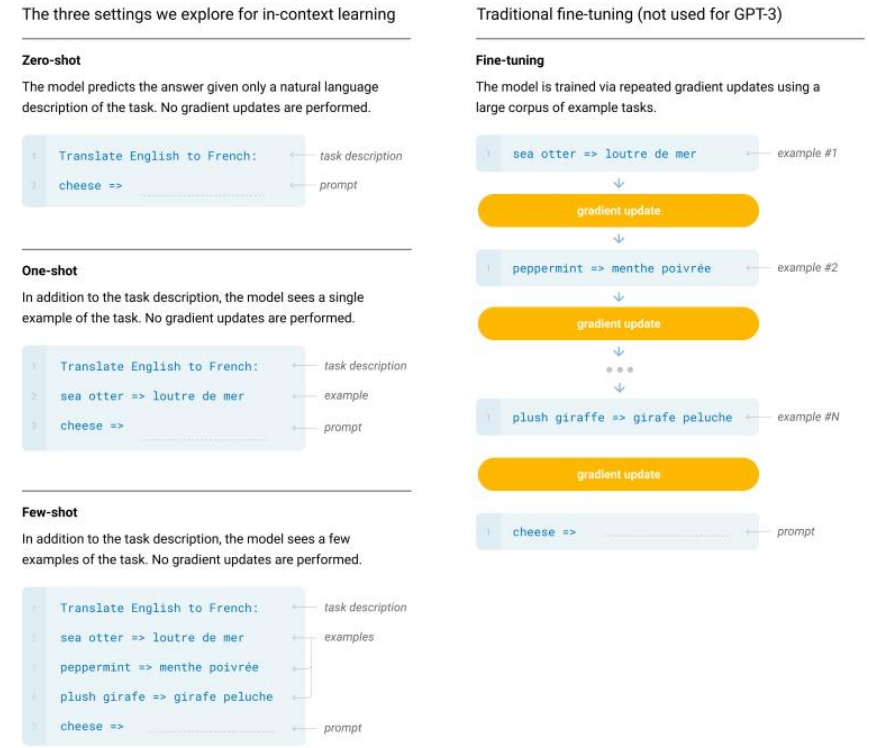
资料来源：Bloomberg，国信证券经济研究所整理

GPT-3：能够举一反三的大语言模型

国信证券 GUOSEN SECURITIES

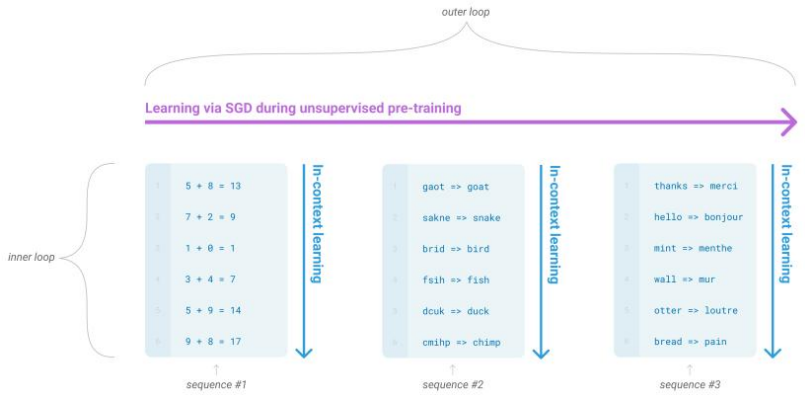
- 《Language Models are Few-Shot Learners》 by OpenAI
- 相比GPT-2，GPT-3大幅增加了模型参数。**GPT-3是具有1750亿个参数的自回归语言模型，**更能有效利用上下文信息。对于特定的下游任务，GPT-3无需进行任何梯度更新或微调，仅需通过与模型交互并提供少量范例即可。
- 特点：1、模型规模急剧增加（使得模型性能提升迅猛）；2、实现few-shot learning。
- **in-context learning：**对模型进行引导，使其明白应输出什么内容。
  - Q：你喜欢吃苹果吗？A1：我喜欢吃。A2：苹果是什么？A3：今天天气真好。A4：Do you like eating apples?
  - 采用prompt提示语：　　汉译英：你喜欢吃苹果吗？　　请回答：你喜欢吃苹果吗？
- **in-context learning三种方式：**不需要进行参数更新，仅需把少量标注样本作为输入文本的上下文
  - 仅提示zero-shot(0S)：仅需给出任务描述
  - 一个范例one-shot(1S)：仅需给出任务描述和一个示例
  - 多个范例few-shot(FS)：仅需给出任务描述和少量示例

图：zero-shot、one-shot和few-shot与传统微调形成对比



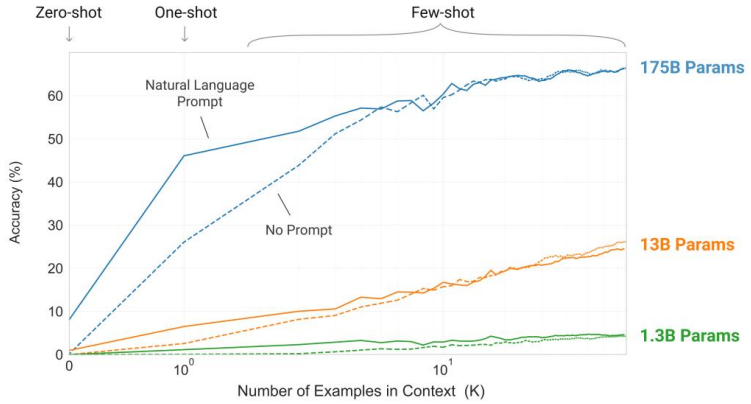
资料来源：《Language Models are Few-Shot Learners》，国信证券经济研究所整理

图：语言模型的元学习



资料来源：《Language Models are Few-Shot Learners》，国信证券经济研究所整理

图：一个简单任务中不同参数模型上下文学习性能



资料来源：《Language Models are Few-Shot Learners》，国信证券经济研究所整理

图：GPT-1至GPT-3模型参数

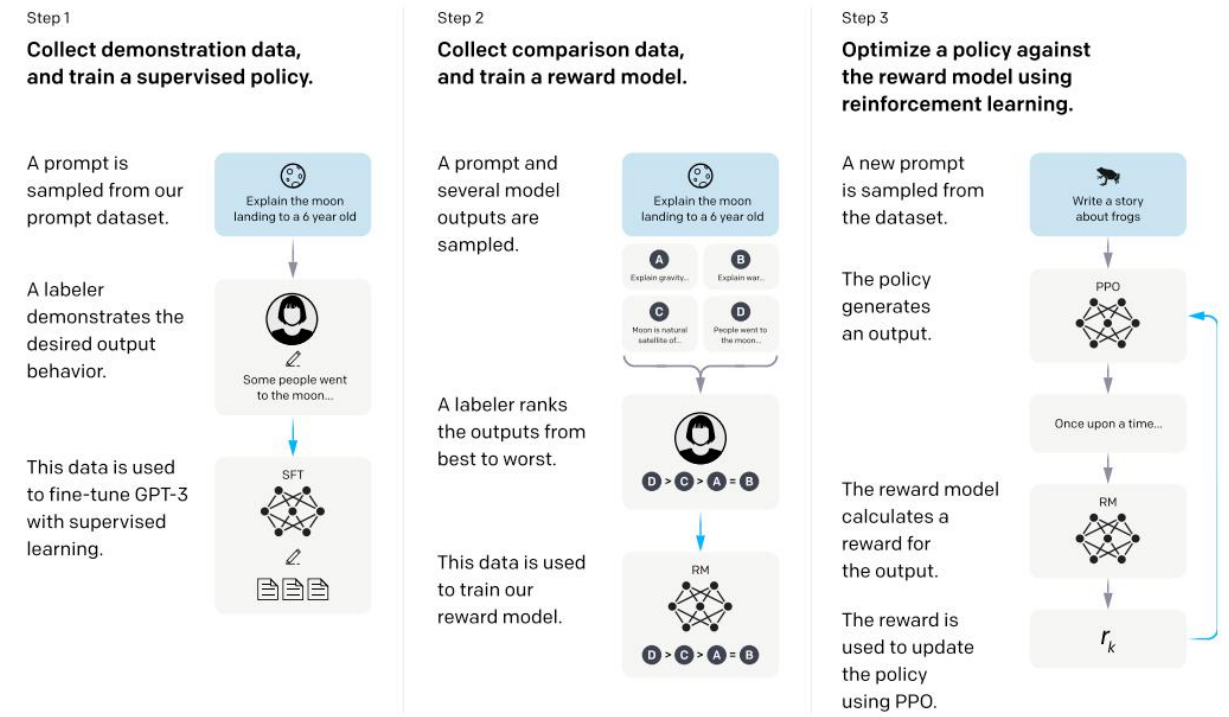
版本	GPT-1	GPT-2	GPT-3
时间	2018年6月	2019年2月	2020年5月
参数量	1.17亿	15.4亿	1750亿
预训练数据量	5GB	40GB	45TB
训练方式	Pre-training+Fine-tuning	Pre-training	Pre-training
序列长度	512	1024	2048
# of Decoder Layers	12	48	96
Size of Hidden Layers	768	1600	12288

资料来源：腾讯云开发者，国信证券经济研究所整理

GPT-3.5（ChatGPT）：引入人类反馈强化学习机制

- 《Training language models to follow instructions with human feedback》 by OpenAI
- 过往GPT模型存在的问题：语料库偏差，继续使用无监督学习扩大模型无法达到使用目的（编造事实、有偏见文本等）。
- **GPT-3.5通过使用人类反馈的数据集进行监督学习**（RLHF，即reinforcement learning from human feedback），对GPT模型进行微调。主要分为以下三步：
  - 1、根据人工标注的数据集构造示范样本，进行有监督的微调，训练出有监督的微调模型。
  - 2、构造Reward模型，通过人工对输出结果标注并进行比较排序打分，训练Reward模型，学习对输出进行排序打分。
  - 3、采用PPO（Proximal policy optimization，近端策略优化，一种强化学习算法），通过不断与环境交互（如ChatGPT不断从训练集中抽取问题并生成解答）以训练GPT模型，使Reward模型打分最大化。
- **结果显示：**通过构建人类反馈的数据集，使用有监督学习微调模型，能够使得模型输出与人类意图一致。

图：GPT-3.5模型架构示意图



资料来源：《Training language models to follow instructions with human feedback》，国信证券经济研究所整理

大语言模型带来的算力需求测算——以GPT-3为例（假设部分）

国信证券 GUOSEN SECURITIES

- **算力有效性：**在4月GTIC 2023中国AIGC创新峰会上，NVIDIA消费互联网行业解决方案架构师负责人徐添豪提出，NVIDIA的NeMo Framework在训练GPT-3过程中能使得硬件算力有效性能达到50%以上。
- **GPT-3模型参数量与样token数：**GPT-3参数量达1750亿个，训练样本token数达3000亿个。
- **训练时间要求：**假设GPT-3模型每次训练时间要求在30天完成。
- **推理访问次数：**按谷歌每日搜索量35亿次进行估计；假设每次访问提问4次，每次提问+回答需处理字数425字，平均每个字转换为token比例为4/3，则每日GPT-3需推理token数为79330亿个。

图：算力测算参数假设

阶段	参数	值
训练阶段	算力有效性	50%
	GPT-3训练样本token数	300B
	GPT-3模型参数量	175B
	每个token训练所需运算次数	6*参数量FLOPs
	训练时间要求	30天
推理阶段	每位用户提问次数	4
	每次提问字数	25
	每次回答字数	400
	平均每个字数对应token数	4/3
	谷歌每日搜索量	35亿次
	每个token推理所需运算次数	2*参数量FLOPs

资料来源：OpenAI官网，国信证券经济研究所整理

图：所需CPU数量计算公式

$$\text{训练所需CPU数} = \frac{\text{训练样本token数} * \text{单token训练所需运算次数} / \text{训练时间要求}}{\text{单颗芯片最大算力} * \text{算力有效性}}$$

$$\text{推理所需CPU数} = \frac{\text{推理访问次数} * \text{单次访问处理字数} * \text{字数与token转换倍数} * \text{单token推理所需运算次数} / \text{规定推理时间}}{\text{单颗芯片最大算力} * \text{算力有效性}}$$

资料来源：国信证券经济研究所整理

大语言模型带来的算力需求测算——以GPT-3为例（结论部分）

国信证券 GUOSEN SECURITIES

- **训练阶段：**考虑采用精度为32位的单精度浮点数数据进行训练和推理。以A100 PCIe芯片为例（H100 PCIe芯片同理），根据前述公式，GPT-3训练所需运算次数为：样本token数3000亿个*6*参数量1750亿个=315*10^21FLOPs；考虑训练时间要求在30天完成（训练时间为2592000秒），则对应GPT-3训练所需算力为121528TFLOPS；结合A100有效算力78TFLOPS，**得到所需GPU数量为1558个，对应AI服务器为195台。**

- **推理阶段：**根据前述公式，GPT-3每日需推理token数为79330亿个，则推理所需运算次数为4760*10^21FLOPs；考虑推理时间以每日为单位（推理时间为86400秒），则对应GPT-3推理所需算力为55*10^6TFLOPS；结合A100有效算力78TFLOPS，**得到所需GPU数量为706315个，对应AI服务器为8.8万台。**

图：训练阶段算力需求测算过程及结论

		A100 PCIe	H100 PCIe
GPU相关	Tensor Float 32(TF32)	156TFLOPS	756TFLOPS
	有效算力	78TFLOPS	378TFLOPS
	GPT-3训练所需运算次数	315*10^21FLOPs	315*10^21FLOPs
	GPT-3训练所需算力	121528TFLOPS	121528TFLOPS
	所需GPU数量	1558	322
	GPU单价	1.5万美元	3.65万美元
	对应GPU价值	2337万美元	1175.3万美元
		DGX A100	**DGX H100**
AI服务器相关	单个服务器对应GPU数量	8	8
	所需服务器数量	195台	40台

资料来源：英伟达官网，国信证券经济研究所整理及预测

图：推理阶段算力需求测算过程及结论

		A100 PCIe	H100 PCIe
GPU相关	Tensor Float 32(TF32)	156TFLOPS	756TFLOPS
	有效算力	78TFLOPS	378TFLOPS
	GPT-3推理所需运算次数	4760*10^21FLOPs	4760*10^21FLOPs
	GPT-3推理所需算力	55*10^6TFLOPS	55*10^6TFLOPS
	所需GPU数量	706315	145748
	GPU单价	1.5万美元	3.65万美元
	对应GPU价值	105.95亿美元	53.2亿美元
		DGX A100	**H100**
AI服务器相关	单个服务器对应GPU数量	8	8
	所需服务器数量	8.8万台	1.8万台

资料来源：英伟达官网，国信证券经济研究所整理及预测

注：部分名词解释

- **训练：**指利用大数据训练神经网络，通过大量数据确定网络中的权重和偏置的值，使其能够适应特定功能。
- **推理：**指利用训练好的模型，使用新的数据推理和判断出各种结论。
- **token：**语言模型的最基本单位，将长文本分解为基本数据结构，再根据映射规则进行计算。
- **浮点数：**一种计算机系统数字表示标准，指一个数的小数点的位置不是固定的，可以浮动，利用科学计数法来表示实数。常见浮点数根据精度不同分为双精度浮点数FP64、单精度浮点数FP32、半精度浮点数FP16等。
- **FLOPS**（loating-point operations per second）：每秒浮点运算次数，用于大量浮点运算的科学计算领域中。
- **TFLOPS**（teraFLOPS）：每秒1万亿（=10^12）次浮点运算。

图：英伟达GPU芯片售价

	芯片型号	售价（美元）	售价（元）	中国供应情况
中低端	A10	3200	22080	
	A16	3500	24150	
	A30	4700	32430	
	A40	5300	36570	
	L40	7600	52440	
高性能	V100	10000	69000	
	A800	12000	82800	缺货
	A100	15000	103500	美国政府禁止供应中国
	H100	36500	251850	

资料来源：财经十一人，国信证券经济研究所整理

图：英伟达A100 GPU参数

	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS| 1248 TOPS*	
GPU 显存	80GB HBM2e	80GB HBM2e
GPU 显存带宽	1935GB/s	2039GB/s

资料来源：英伟达官网，国信证券经济研究所整理

图：英伟达H100 GPU参数

	H100 SXM	H100 PCIe
FP64	34 TFLOPS	26 TFLOPS
FP64 Tensor Core	67 TFLOPS	51 TFLOPS
FP32	67 TFLOPS	51 TFLOPS
TF32 Tensor Core	989 TFLOPS*	756 TFLOPS*
BFLOAT16 Tensor Core	1,979 TFLOPS*	1,513 TFLOPS*
FP16 Tensor Core	1,979 TFLOPS*	1,513 TFLOPS*
FP8 Tensor Core	3,958 TFLOPS*	3,026 TFLOPS*
INT8 Tensor Core	3,958 TOPS*	3,026 TOPS*
GPU memory	80GB	80GB
GPU memory bandwidth	3.35TB/s	2TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max thermal design power (TDP)	Up to 700W (configurable)	300-350W (configurable)

资料来源：英伟达官网，国信证券经济研究所整理

GPT-3模型对GPU与AI服务器需求展望

国信证券 GUOSEN SECURITIES

- 根据结论，1个参数量为1750亿个的GPT-3模型在训练阶段需要新增1558颗A100 GPU芯片，对应价值为2337万美元，**需要195台DGX A100服务器**；在推理阶段需要新增70.6万颗A100 GPU芯片，对应价值为105.95亿美元，**需要8.8万台DGX A100服务器**。考虑一台DGX A100服务器售价19.9万美元，则**在训练阶段DGX A100服务器价值量为3880.5万美元，推理阶段DGX A100服务器价值量为175.12亿美元。**

- 英伟达（NVIDIA）是一家人工智能计算公司，其GPU产品和架构为科学计算、人工智能（AI）、数据科学、自动驾驶汽车（AV）、机器人、元宇宙和3D互联网应用创建平台。FY23英伟达收入为269.74亿美元。若按上述结论，**GPT-3新增GPU价值达到英伟达公司FY23收入的39.4%。**

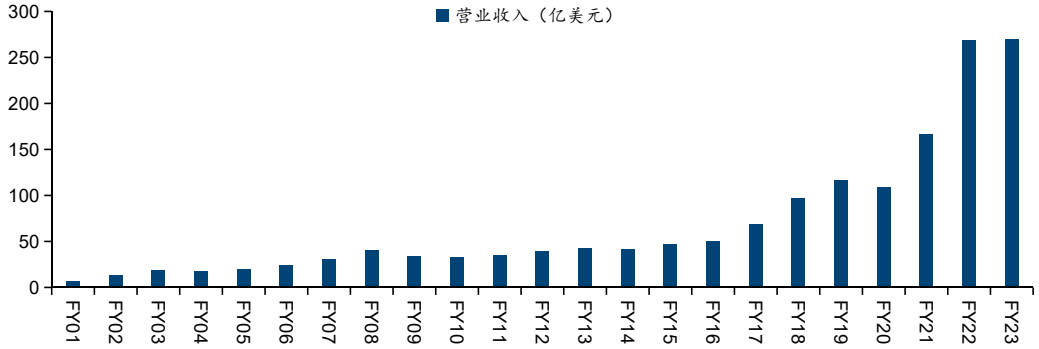
- 据IDC数据，受益于全球经济的快速复苏，2021年用户对数据中心基础设施的投资持续上涨，全球服务器市场出货量为1353.9万台。据TrendForce数据，截至2022年底预计搭载GPGPU（General Purpose GPU）的AI服务器年出货量占整体服务器比例近1%。若采用上述数据大致估算，**GPT-3新增AI服务器数量达到2021年全球AI服务器数量的65.35%。**

图：英伟达历年GPU产品对比

产品型号	发布时间	制程	双精度浮点运算性能（TFLOPS）	单精度浮点运算性能（TFLOPS）	半精度浮点运算性能（TFLOPS）	整型定点运算性能（TOPS）	显存	显存带宽	最大功耗
H100 SXM	2022	4nm	26	51	1979(Tensor Core)	3958(Tensor Core)	80GB	3.35TB/s	700W
A100 80GB SXM	2020	7nm	9.7	19.5	624(Tensor Core)	1248(Tensor Core)	80GB	2039GB/s	400W
V100S PCIe	2019	12nm	8.2	16.4	-	-	32GB	1134GB/s	250W

资料来源：英伟达官网，国信证券经济研究所整理

图：英伟达历年收入情况



资料来源：Bloomberg，国信证券经济研究所整理

四、产业链相关公司

工业富联：电子设备制造（EMS）行业龙头

国信证券 GUOSEN SECURITIES

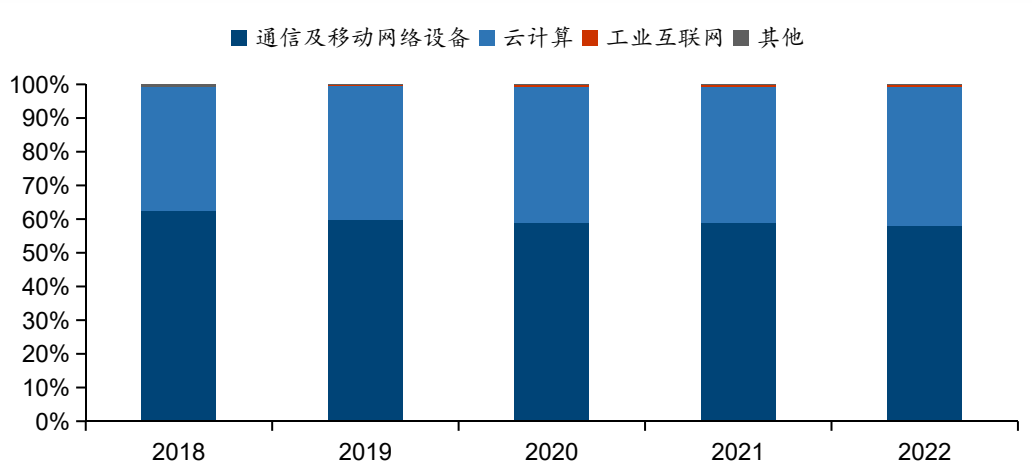
- **公司是依托于工业互联网为全球知名客户提供智能制造和科技服务解决方案的制造业龙头企业。**公司主要由母公司鸿海精密集团旗下通信网络设备、云服务设备、工业互联网三大业务整合而成，业务范围覆盖数字经济产业全品类。

- **公司云计算业务收入连续五年保持成长趋势：**2022年公司云计算业务收入2124.44亿元，同比增长19.6%，占总营收比例为41.5%。公司云计算产品包括服务器、存储设备及云服务设备高精密机构件，主要为苹果、亚马逊、谷歌、戴尔、HPE、思科等国内外领先的云服务商和品牌商提供云计算设备代工制造服务。服务器ODM厂商直接供货具备产能充足、交付速度快、定制性强、价格低廉等优势，近年来云服务厂商逐渐倾向于直接向服务器ODM厂商采购定制化服务器产品。

图：公司产品布局

资料来源：工业富联公司公告，国信证券经济研究所整理

图：公司历年营收占比



资料来源：Wind，国信证券经济研究所整理

沪电股份：服务器、交换机高阶硬板核心供应商

国信证券 GUOSEN SECURITIES

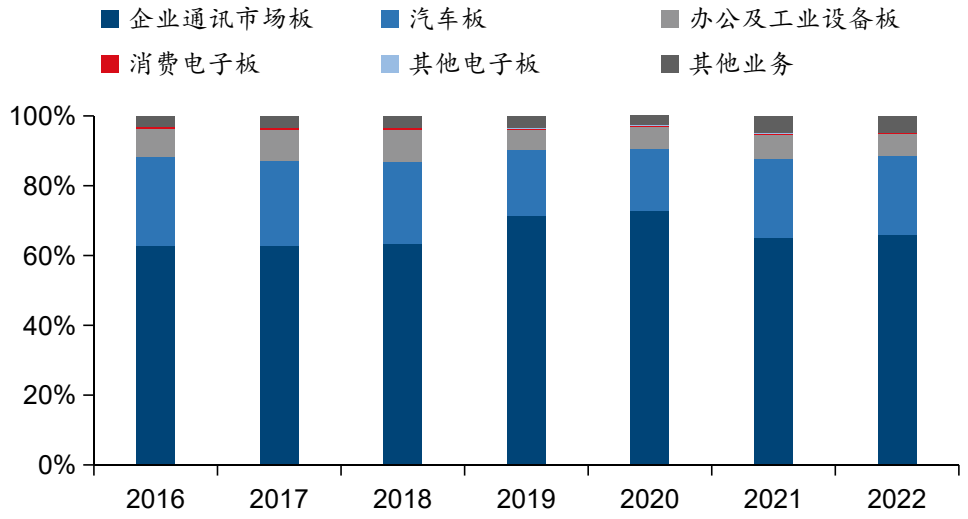
- **公司是全球领先的PCB厂商，**在中高阶板领域具有显著的领先优势，持续深耕通信通讯设备、数据中心基础设施以及汽车电子应用领域的核心产品市场，产品在服务器、交换机中市占率较高。

- **公司通讯市场板业务占公司PCB营收70%：**公司EGS级服务器领域产品已实现量产；HPC领域，应用于AI加速、Graphics的产品，应用于GPU、OAM、FPGA等加速模块类的产品以及应用于UBB、BaseBoard的产品已批量出货，正在预研UBB2.0、OAM2.0的产品；交换机领域，应用于Pre800G的产品已批量生产，应用于800G的产品已实现小批量的交付；基于数据中心加速模块的多阶HDI Interposer产品，已实现4阶HDI的产品化，目前在预研6阶HDI产品，同时基于交换、路由的NPO/CPO架构的Interposer产品也同步开始预研；在半导体芯片测试线路板部分重点开发0.35mm以上Pitch的高阶产品。

图：公司产品布局

资料来源：沪电股份官网，国信证券经济研究所整理

图：公司历年营收占比



资料来源：Wind，国信证券经济研究所整理

寒武纪：国内人工智能芯片公司

国信证券 GUOSEN SECURITIES

- **公司是国产人工智能芯片厂商，**专注于人工智能芯片产品的研发与技术创新，致力于打造人工智能领域的核心处理器芯片，提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。公司主要产品包括：终端智能处理器IP、云端智能芯片及加速卡、边缘智能芯片及加速卡、基础系统软件平台等。

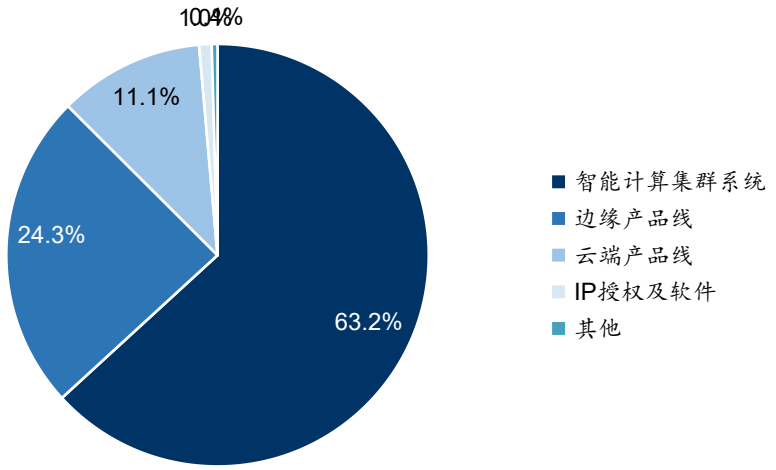
- **公司具备广泛的产品体系：**目前已推出的产品体系覆盖了云端、边缘端的智能芯片及其加速卡、训练整机、处理器IP及软件，可满足云、边、端不同规模的人工智能计算需求。2022年3月，公司正式发布新款训练加速卡MLU370-X8。MLU370-X8搭载双芯片四芯粒思元370，集成寒武纪MLU-Link多芯互联技术，主要面向训练任务，在业界应用广泛的YOLOv3、Transformer等训练任务中，8卡计算系统的并行性能平均达到350W RTX GPU的155%。

图：公司产品布局

产品线	产品类型	寒武纪主要产品	推出时间
云端产品线	云端智能芯片及加速卡	思元100（MLU100）芯片及云端智能加速卡	2018年
		思元270（MLU270）芯片及云端智能加速卡	2019年
		思元290（MLU290）芯片及云端智能加速卡	2020年
		思元370（MLU370）芯片及云端智能加速卡	2021年
	训练整机	玄思1000智能加速器	2020年
边缘产品线	边缘智能芯片及加速卡	思元220（MLU220）芯片及边缘智能加速卡	2019年
IP授权及软件	终端智能处理器IP	寒武纪1A处理器	2016年
		寒武纪1H处理器	2017年
		寒武纪1M处理器	2018年
	基础系统软件平台	寒武纪基础软件开发平台（适用于公司所有芯片与处理器产品）	持续研发和升级，以适配新的芯片

资料来源：寒武纪公司公告，国信证券经济研究所整理

图：公司2021年营收占比



资料来源：Wind，国信证券经济研究所整理

海光信息：国产服务器CPU芯片龙头

国信证券 GUOSEN SECURITIES

- **公司是国内少数几家同时具备高端通用处理器（CPU）和协处理器（DCU）研发能力的集成电路设计企业。**公司掌握了高端处理器核心微结构设计、高端处理器SoC架构设计、处理器安全、处理器验证、高主频与低功耗处理器实现、高端芯片IP设计、先进工艺物理设计、先进封装设计、基础软件等关键技术，专注于研发、设计和销售应用于服务器、工作站等计算、存储设备中的高端处理器，建立了完善的高端处理器的研发环境和流程，产品性能逐代提升，功能不断丰富。

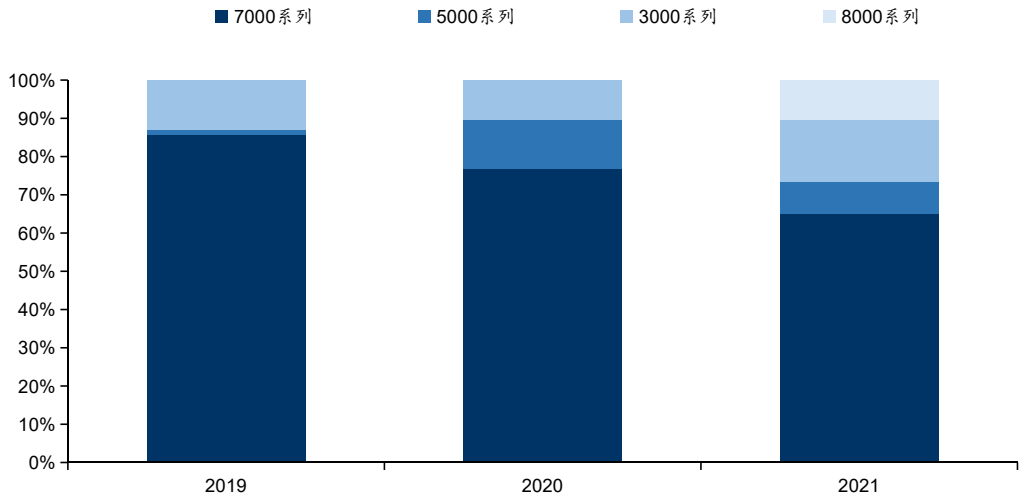
- 海光信息CPU系列产品兼容x86指令集以及国际上主流操作系统和应用软件，性能优异，软硬件生态丰富，安全可靠，得到了国内用户的高度认可；**2022年，公司成功推出 CPU 产品海光三号。**DCU系列产品以GPGPU架构为基础，兼容通用的“类 CUDA”环境以及国际主流商业计算软件和人工智能软件，软硬件生态丰富，**DCU 产品深算一号在2022年度实现了在大数据处理、人工智能、商业计算等领域的商业化应用。**

图：公司产品布局

产品类型	主要产品	指令集	产品特征	典型应用场景
高端处理器	通用处理器-海光CPU	兼容x86指令集	内置多个处理器核心，集成通用的高性能外设接口，拥有完善的软硬件生态环境和完备的系统安全机制。针对不同应用场景对高端处理器计算性能、功能、功耗等技术指标的要求，分别提供海光7000系列产品、5000系列产品、3000系列产品	云计算、物联网、信息服务等
	协处理器-海光DCU	兼容“类CUDA”环境	内置大量运算核心，具有较强的并行计算能力和较高的能效比，适用于向量计算和矩阵计算等计算密集型应用	大数据处理、人工智能、商业计算等

资料来源：海光信息公司公告，国信证券经济研究所整理

图：公司历年营收占比



资料来源：Wind，国信证券经济研究所整理

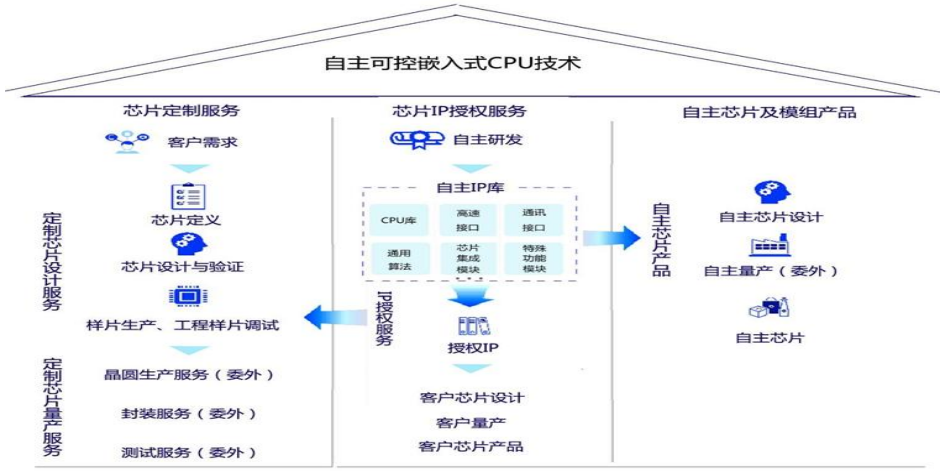
国芯科技：国产自主可控嵌入式CPU芯片设计公司

国信证券 GUOSEN SECURITIES

- **公司嵌入式CPU IP积累深厚，**自成立以来一直采用Fabless的经营模式，提供IP授权、芯片定制服务和自主芯片及模组产品。产品主要应用于信息安全、汽车电子以及工业控制、边缘计算和网络通信三大领域。

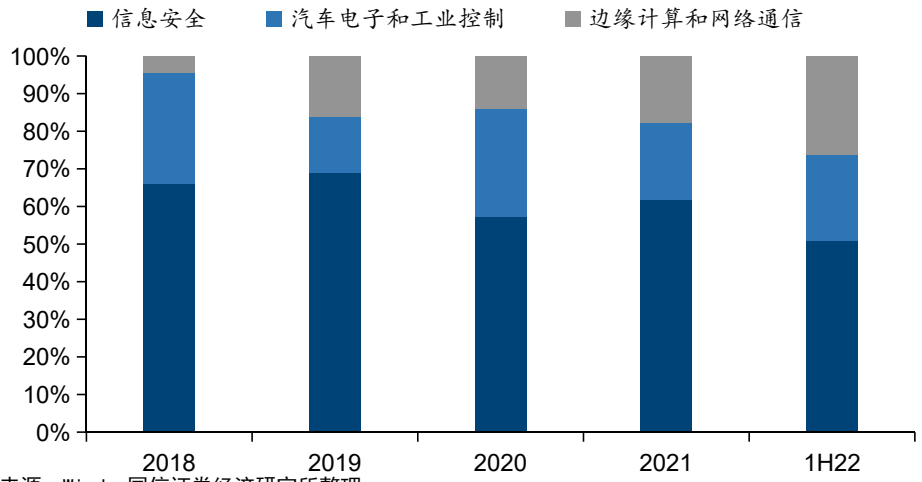
- **公司云安全、边缘计算、云存储相关产品受益于算力需求提升：**在云安全领域，公司的高性能CCP907T、CCP908T芯片及密码卡已完成研发并进入市场推广阶段，产品满足国密算法需求，性能达到国际先进水平，适用于安全网关、VPN设备、密码服务器、可信服务器和云存储服务器等应用。在边缘计算和网络通信领域，公司正在研发的S1020芯片具备多核计算、网络路径和协议加速引擎、路由转发以及多种高速通信接口，适用于边缘计算与网络通信的计算、安全及通信需求。在云存储领域，RAID芯片成功完成研发，具有高性能、大缓存、低功耗等特点，可广泛应用于图形工作站、服务器数据库存储、金融数据库存储等领域，可望实现该领域Raid芯片产品的国产化替代。第一代量产版Raid芯片已正式投片。在此基础上，公司正在瞄准国际一流公司产品水平，积极开展第二代Raid芯片的设计工作，将采用12nm先进工艺技术和高性能高速接口IP技术实现高性能Raid芯片。

图：公司产品布局



资料来源：国芯科技招股书，国信证券经济研究所整理

图：公司历年营收占比



资料来源：Wind，国信证券经济研究所整理

全志科技：国内SoC龙头厂商

国信证券 GUOSEN SECURITIES

- **公司主营业务为智能应用处理器SoC、高性能模拟器件和无线互联芯片的研发与设计，**主要产品为智能应用处理器SoC、高性能模拟器件和无线互联芯片，广泛适用于工业控制、智能家电、智能硬件、平板电脑、汽车电子、机器人、虚拟现实、网络机顶盒以及电源模拟器件、无线通信模组、智能物联网等多个产品领域。

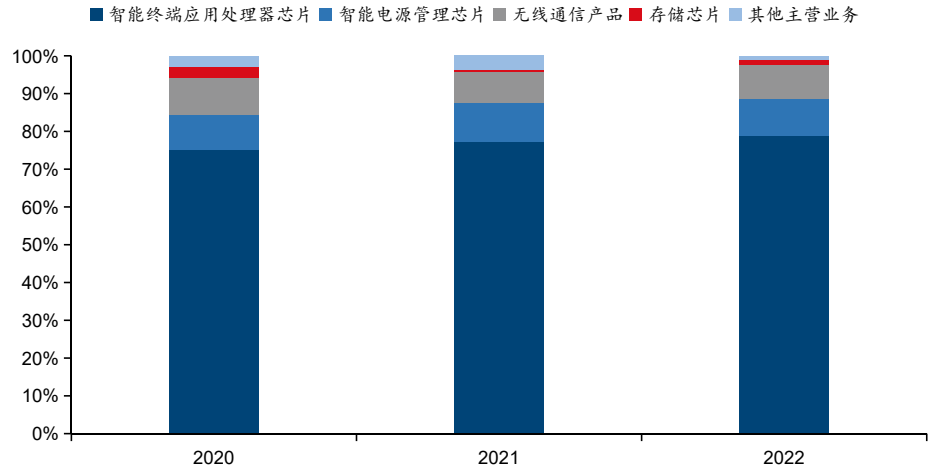
- **公司聚焦AI语音、AI视觉应用的完整链条，实现细分AI产品量产落地：**在AIOT领域，公司产品覆盖智能音箱、智能清洁机器人、智能家电、智能视觉等市场，R系列芯片产品已实现带屏、无屏音箱全面量产。在智能汽车电子领域，公司发布T113芯片产品及解决方案，已在车载人机交互和仪表类应用落地。同时，搭载公司产品的AR-HUD，APA类智能化产品已与前装市场客户合作实现量产上市。在智能工业领域，公司推出了T系列AI处理器新品，与标杆客户打造的easy系列工业PLC控制器获得良好市场表现。在智能解码显示领域，公司推出了智慧屏芯片TV303，后续将逐步在智能电视、智能投影、智能商显领域投入量产。在通用智能终端领域，公司积极拓展通用智能终端相关衍生市场，布局包括电子相册、教育设备等产品，均取得良好的市场反馈。

图：公司产品布局

产品大类	产品系列	主要型号产品	主要应用领域
智能终端应用处理器芯片	R系列	R16、R328、R329、R818、MR813	智能音箱、智能白电、扫地机器人、3D打印机、词典笔等
	V系列	V3、V526、V533、V536、V831、V851、V853	智能安防摄像机、低功耗电池摄像机、多目枪球摄像头、行车记录仪、运动相机、智能扫描笔及泛视觉AI产品等
	H系列	H3、H6、H313、H133、H616、H700、H618	智能机顶盒、智能投影、商业显示、云解码、多屏互动等
	A系列	A33、A64、A100、A133、A133P	平板电脑、电子相册、教育设备、电子书等
	F系列	F1C100S、F1C200S、F133	车载仪表/播放器、人机交互智能控制HMI、视频机等
	T系列	T3、T7、T5、T113	智能座舱、辅助驾驶、智慧工业、行业智能、智能电网等
	其他	B300、D1、B810	电子书、视频一体机、开发板等
智能电源管理芯片	AXP系列	AXP221S、AXP223、AXP707、AXP305、AXP858、AXP717、AXP313	提供智能的供电、电池管理等功能，与主控芯片配套使用
无线通信产品	XR系列	XR8052、XR819、XR829、XR872、XR806	智能家电、智能早教机、儿童机器人、智能机器人、低功耗IPC、无线图传、智能门铃等；
语音信号芯片	AC系列	AC107、AC108、AC101、AC102	提供高集成度的语音信号编解码、信号转换等功能，与主控芯片配套使用

资料来源：全志科技公司公告，国信证券经济研究所整理

图：公司历年营收占比



资料来源：Wind，国信证券经济研究所整理

五、风险提示

风险提示

国信证券 GUOSEN SECURITIES

1、宏观AI应用推广不及预期。AI技术在应用推广的过程可能面临各种挑战，比如：（1）AI技术需要更多的时间来研发和调试，而且在应用过程中可能会受到数据质量、资源限制和技术能力等因素的制约；（2）AI技术的实施需要更多的资源和资金支持；（3）市场竞争可能也会影响企业在AI应用推广方面的表现。因此，投资者应审慎评估相关企业的技术实力、资金实力以及管理能力，相关企业的AI应用存在推广进度不及预期的风险。

2、AI投资规模低于预期。尽管AI技术在过去几年中受到广泛关注，但AI相关领域的企业投资回报并不总是符合预期。部分企业在AI领域可能缺乏足够的经验和资源，难以把握市场机会。此外，市场竞争也可能会影响企业的投资力度。因此，存在AI领域投资规模低于预期，导致企业相关业务销售收入不及预期的风险。

3、AI服务器渗透率提升低于预期。虽然AI服务器的应用已经较为广泛，但AI服务器渗透率提升的速度存在低于预期的风险，这与企业对AI技术的投资意愿有关，也可能与市场需求和技术进展的速度有关。

4、AI监管政策收紧。由于AI技术的快速发展和广泛应用，监管机构可能会加强对AI技术的监管力度。监管机构可能会制定严格的AI技术使用规定，以保障人们的隐私和数据安全，这些监管政策可能会对企业的业务模式和发展战略造成影响。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级		
类别	级别	定义
股票投资评级	买入	预计6个月内，股价表现优于市场指数20%以上
	增持	预计6个月内，股价表现优于市场指数10%-20%之间
	中性	预计6个月内，股价表现介于市场指数±10%之间
	卖出	预计6个月内，股价表现弱于市场指数10%以上
行业投资评级	超配	预计6个月内，行业指数表现优于市场指数10%以上
	中性	预计6个月内，行业指数表现介于市场指数±10%之间
	低配	预计6个月内，行业指数表现弱于市场指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032