领军路径分歧：更大的模型还是更低的成本？

——AI全行业赋能系列深度研究之五

证券分析师：刘洋A0230513050006、洪依真A0230519060003、施鑫展A0230519080002、周海晨A0230511040036

2021.12.13

结论

以商汤、旷视为代表的AI明星在平台上有什么特别的布局？

- **自研了深度学习训练、推理平台，大量算法积累，AIDC算力支持；**
- **商汤科技：**SenseCore大模型+小模型，降低AI应用落地成本；
- **旷视科技：**Brain++平台支撑了跨行业AIoT解决方案。

深度学习开源训练框架格局如何？国内AI公司自研深度学习框架有何种意义？

- **TensorFlow、PyTorch、MXNet、CNTK已可以满足工业界、学界的绝大部分要求；**
- **技术遗留问题：**静态图、动态图技术方案都还有缺陷，有同时解决的可能性；
- **国产平台由于技术遗留问题、国产化等适配性等原因可能仍有空间。**

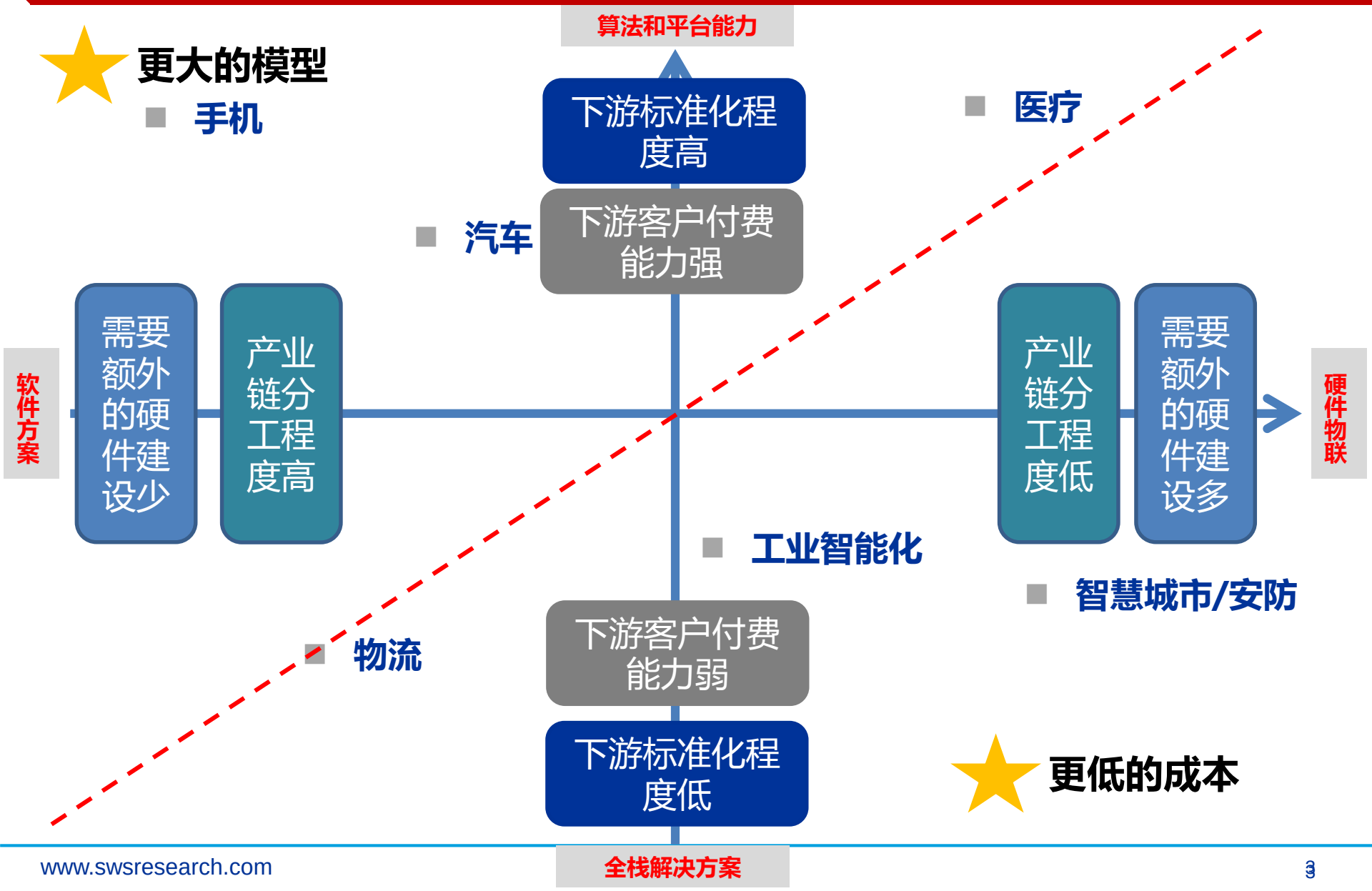
为何AI大模型成为趋势？复现大模型的难度在哪里？

- 以GPT为代表的大模型能大幅降低对数据量的需求，**预训练大模型+细分场景微调，更适合长尾落地。**
- 但对存储、算力要求极高，普通机构难以复现。**商汤、旷视在复现大模型上有算力、训练推理平台优势。**

对于必然的碎片化AI落地，不同公司路径差异在哪里？哪种路径可能胜出？

- **更大的模型路径：**较高软件占比，硬件外采；大规模参数的通用模型，极高的首次开发成本；模型长尾投入理想状态接近0；适合额外硬件建设较少，下游需求标准化程度强，产业链已有分工度高的行业；
- **更低的成本路径：**自有生产线压缩硬件成本；小模型、小算力，较低的首次开发成本；中台复用等方式控制成本。适合已有硬件基础差，需求标准化程度低，产业链已有分工度低的行业。

两种路径更适合的场景


更大的模型
手机
医疗
汽车
物流
工业智能化
智慧城市/安防
更低的成本
算法和平台能力
下游标准化程度高
下游客户付费能力强
下游客户付费能力弱
下游标准化程度低
软件方案
硬件物联
需要额外的硬件建设少
产业链分工程度高
产业链分工程度低
需要额外的硬件建设多
全栈解决方案

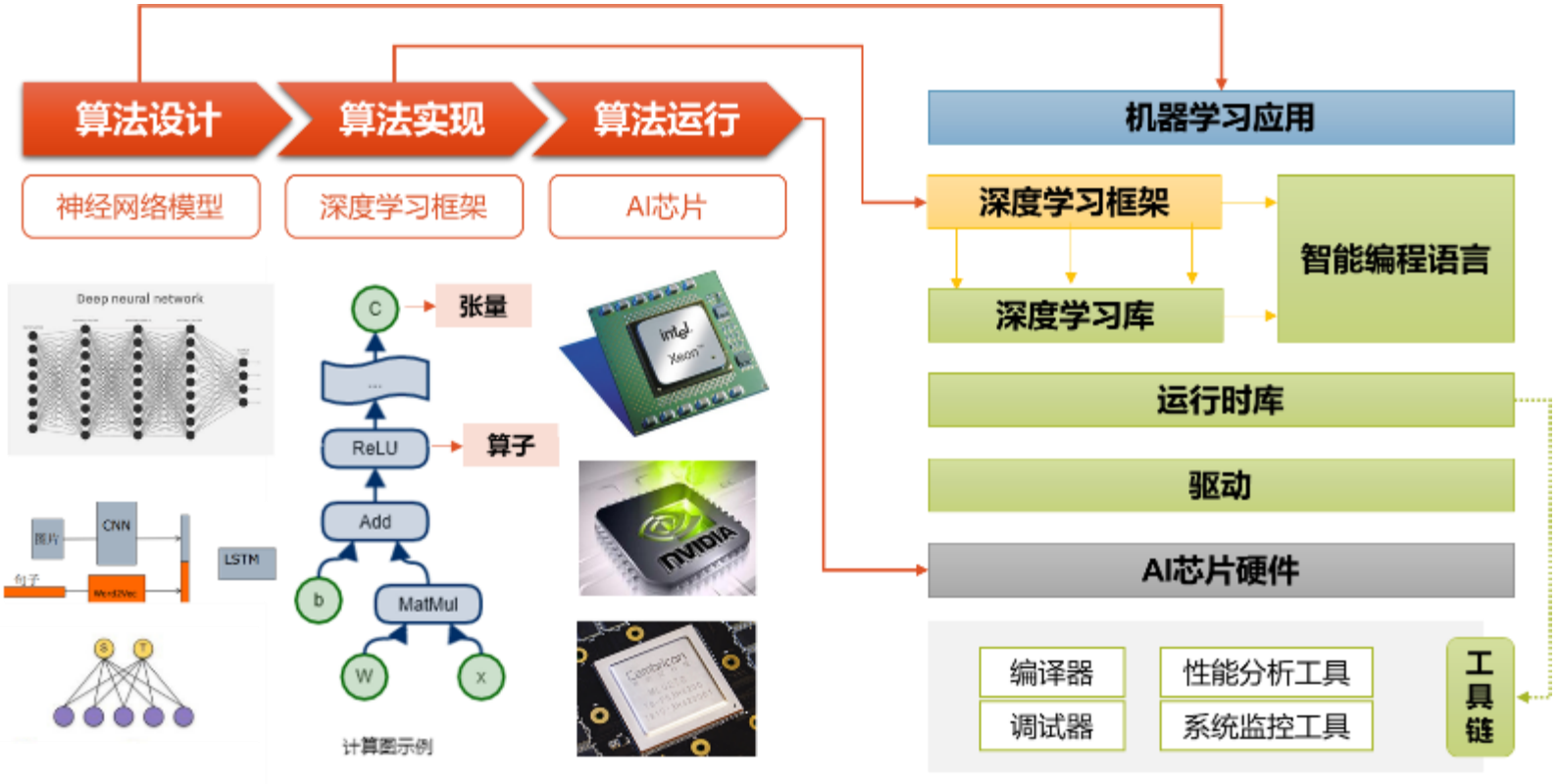
目录

1.1 AI行业产业链——工作流程视角

设计、实现、运行：

- **算法设计环节**：机器视觉、语音识别、自然语音处理、知识图谱；
- **算法实现环节**：深度学习框架，训练、推理部署，对模型的调参优化；
- **算法运行环节**：AI芯片和AIDC超算中心，提供硬件基础。

图：算法实现工作流程



1.2 AI行业产业链——代表厂商

图：AI行业更领域代表性公司

1.2 AI芯片：突破Nv壁垒的三种可能性

技术及应用标准与规范

科研学术机构与各层次人才

通信与信息网络

数据存储设施

物联网与微型传感器

AI芯片的整体性能=AI芯片硬件性能×（AI芯片厂商算子库和工具链+深度学习框架支持）

生态壁垒产生的根源在于软硬件的高度耦合

AI芯片公司需要做的软件

最大的工作量是对海量算子和特定芯片的支持

突破AI生态壁垒的三种可能机会

从苹果MacOS X案例可以推测，当任何一家AI芯片公司能够从英伟达手中逐渐拿走**10%份额**的时候，深度学习框架厂商也大概率会去逐渐投入为该厂商的AI芯片去做特定优化。

1. AI编译器的成熟

解决算子自动开发优化的问题，以及将训练好的模型部署到各种AI芯片上的问题，从而实现AI软硬件的解耦

2. **芯片性能和英伟达相比具备明显优势**

AI芯片整体的性能是由基础软件和芯片硬件本身共同决定的，虽然软件生态存在短板，但是如果硬件本身性能非常突出是有可能获得整体上的优势的

3. 等待英伟达失误

类比PC时代AMD在英特尔两次犯错时分别获得了明显市占率提升的机会

AI芯片公司所采取不同突围路线

AMD	兼容CUDA的路线，借力英伟达生态
华为	深度学习框架+AI芯片”自研路线
谷歌	深度学习框架+AI芯片”自研路线
寒武纪	在走英伟达的路线

AI芯片公司遇到了不同的瓶颈

难点在于其更新迭代速度永远跟不上CUDA并且很难做到完全兼容

华为：没有深度学习框架厂商主动支持，其自研的MindSpore框架尚未解决TF/PyTorch面临的共同痛点

谷歌：TPU本身性能还有进一步提升空间以及过于专用的问题

芯片本身性能以及算子库丰富程度还有追赶空间，没有深度学习框架厂商基于芯片做特定优化

1.2 AI平台层：巨头必争之地

AI平台层：

- 支撑AI大规模训练生产、部署的技术体系；
- 包括**训练框架、模型生产平台、推理部署框架、数据平台。**

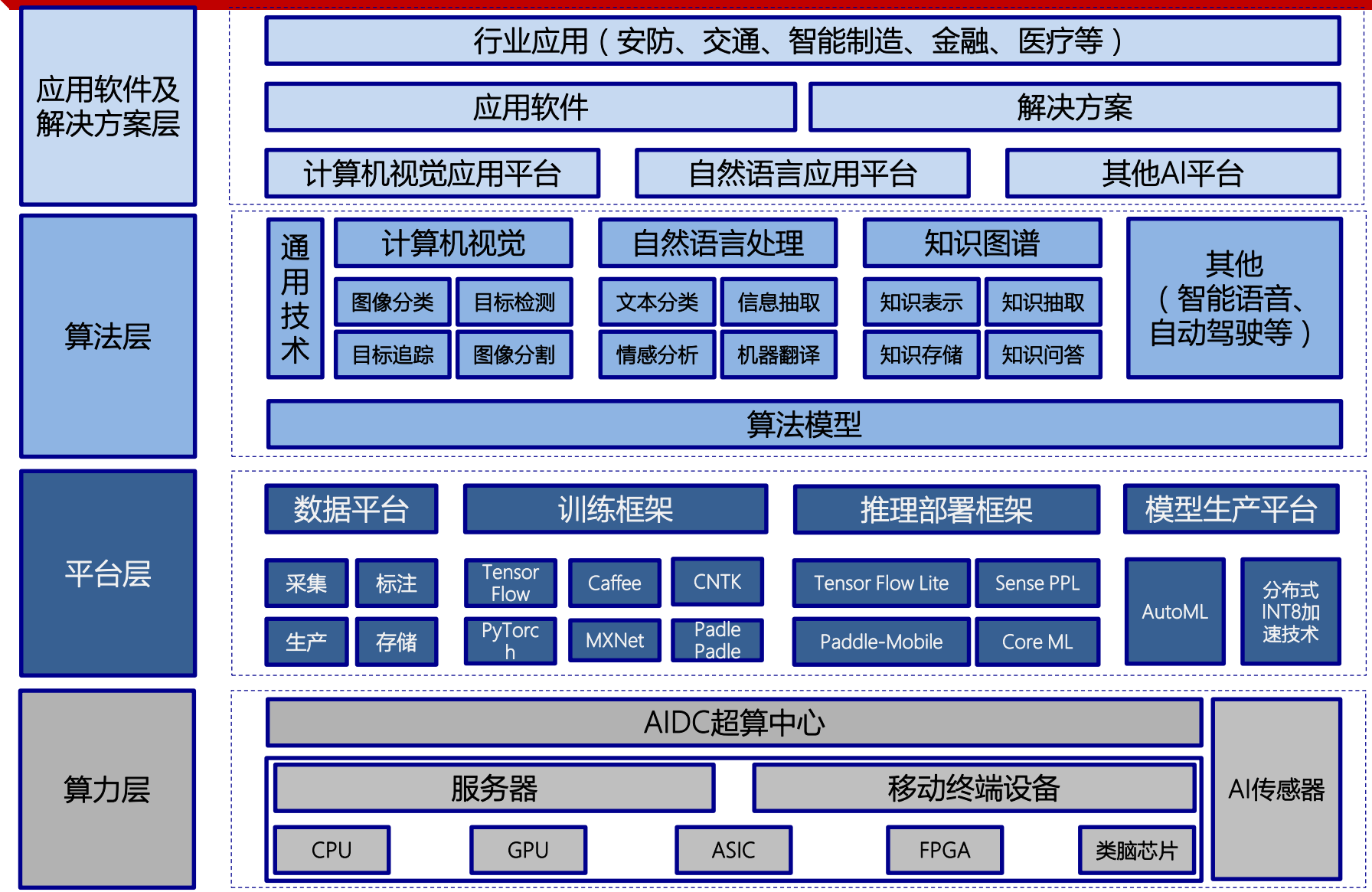
训练、推理部署框架是核心：

- **机器学习框架或深度学习框架：**AI开发依赖的环境安装、部署、测试以及不断迭代改进准确性和性能调优，框架目的是为了简化、加速和优化这个过程。
- **避免重复发明轮子**，而专注于技术研究和产品创新。
- **巨头竞争的核心点**，各大厂建设算法模型数据库，将其封装为软件框架，为应用开发提供集成软件工具包，为上层应用开发提供了算法调用接口。

1.2 AI应用层：百花齐放，工程和变现能力为核心

- **AI+安防、AI+金融是标配：**
  - 智慧城市和安防仍然是AI机器视觉最成熟的落地场景；
  - 安防+金融合计收入在四小龙中占比都在**50%**以上。
  - **云从科技**：继续探索AI在社区、政务、金融**更深层次全栈应用。**
- **AI+手机仍然是最理想的收费场景：**
  - 虹软、商汤、旷视该业务毛利率可能都在80%以上，纯SDK收费，理想的场景；
  - **但规模后续增长有限**
- **AI+汽车、 AI+教育、AI+零售可能为新的增长空间：**
  - **商汤科技**：AI+**汽车**，探索机器视觉在L2+自动驾驶应用。
  - **旷视科技**：探索**AIoT**在物流、智造等多行业的广泛应用

1.3 AI行业产业链——整体图谱



目录

2 本节结论：开源框架规模效应，国产平台仍有空间

四大顶级深度学习框架阵营可以满足绝大部分开发者要求

- **TensorFlow、PyTorch、MXNet、CNTK已可以满足工业界、学界的绝大部分要求。**
- **开源框架规模效应：**维护力量、贡献人员决定了算法库扩展及时性、API水平，软件框架规模效应较强。

目前深度学习框架发展趋势和遗留的技术问题

- **发展趋势：**增加对Python的支持、动态图应用；支持分布式和移动端运行平台；前端的编程接口更加灵活，训练速度不断提高：对网络优化，减少训练耗时，提升底层计算硬件单元的计算能力；
- **技术遗留问题：**静态图、动态图技术方案都还有缺陷，有同时解决的可能性；在网络结构、设备兼容、性能与功耗均衡和各种自动化设计等有提升空间
- 动态图：其核心特点是计算图的构建和计算同时发生（Define by run）。优点是调试方便，缺点是难以对整个计算图进行优化。PT
- 静态图：将计算图的构建和实际计算分开（Define and run）。优点是对全局的信息掌握更丰富，可以做的优化更多，缺点是无法实时观察中间结果。TF

国产平台由于技术遗留问题、国产化等适配性等原因可能仍有空间

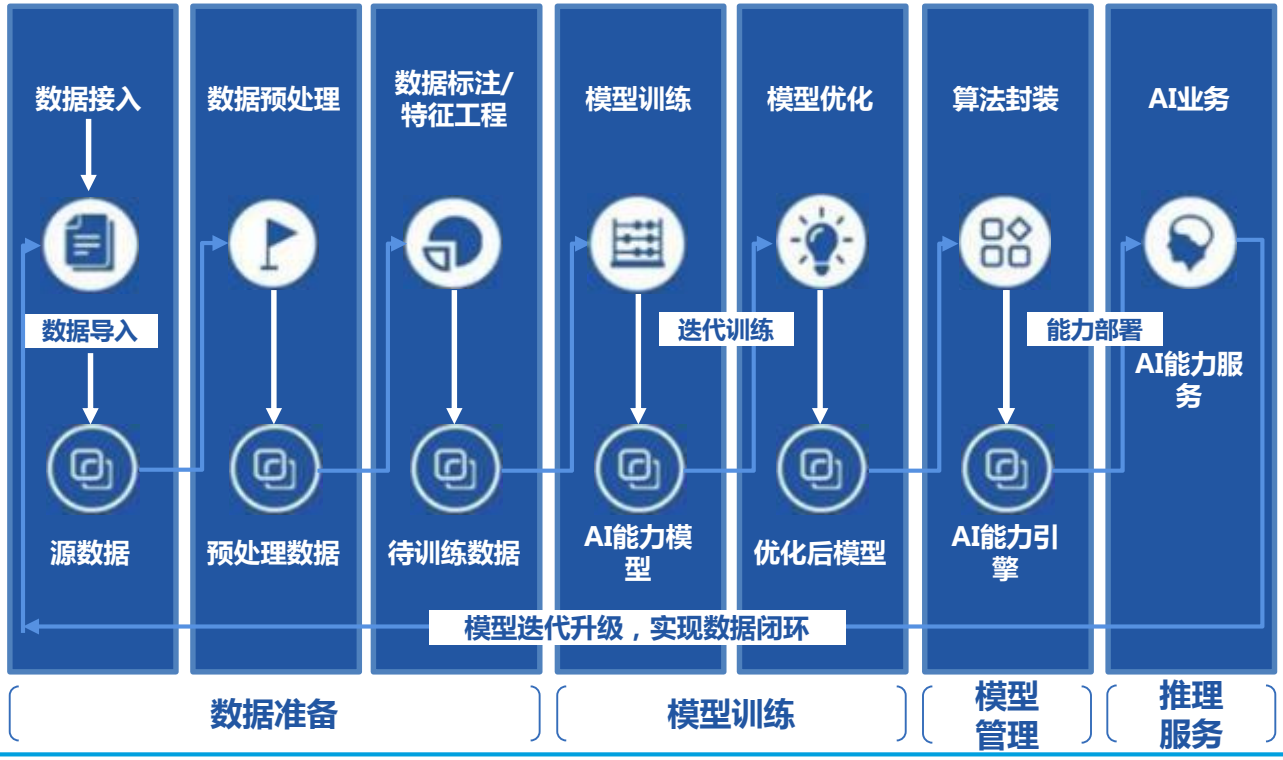
- 特定场景框架可能更优；开源平台可能工业包不共享的问题；国产芯片和适配，中文环境的API
- 国内百度、华为、商汤、旷视在自研框架初期就考虑到训练速度要求提高带来的各种问题，同时**适应国产服务器芯片等环境**

2.1 AI平台：少量企业参与的AI高地之争

平台层：

- **训练软件框架**：实现深度学习训练算法的模块化封装。
- **模型生产平台**：实现模型的工业级生产。
- **推理部署框架**：实现模型生产完成后的工业级高效、自动的部署。
- **数据平台**：包括数据采集、数据标注、数据生产、数据存储等功能

图：AI模型训练部署全流程示意图



2.2 训练框架：调节参数，生成参数

训练框架是AI的重要基石，也是AI发展战略的制高点

- 当算法变成改造甚至颠覆软件行业的力量时，最后核心就是看这些AI的公司有没有平台化的能力，即“能够批量、高效、比竞争对手更及时地供应优质算法”

训练框架的功能

- 1、基于图（Graph）的张量计算引擎（基础的概率统计、线性代数的计算模块）
- 2、大量的外围库（训练样本库、应用数据库、模型参数库、模型代码库）
- 3、大量的领域模型（以文字处理、语音识别、图像处理、目标识别等为主）

表：深度学子超参数对模型的影响

超参数	如何影响模型容量	原因	注意事项
学习率	调至最优，提升有效容量	过高或者过低的学习率，都会由于优化失败而导致降低模型有效容限	学习率最优点，在训练的不同时间点都可能变化，所以需要一套有效的学习率衰减策略
损失函数	调至最优，提升有效容量	损失函数超参数大部分情况都会可能影响优化，不合适的超参数会使即便是对目标优化非常合适的损失函数同样难以优化模型，降低模型有效容限。	对于部分损失函数超参数其变化会对结果十分敏感，而有些则并不会太影响。在调整时，建议参考论文的推荐值，并在该推荐值数量级上进行最大最小值调试该参数对结果的影响。
批样本数量	过大过小，容易降低有效容量	大部分情况下，选择适合自身硬件容量的批样本数量，并不会对模型容限造成。	在一些特殊的目标函数的设计中，如何选择样本是很可能影响到模型的有效容限的，例如度量学习（metric learning）中的N-pair loss。这类损失因为需要样本的多样性，可能会依赖于批样本数量。
丢弃法	比率降低会提升模型容量	较少的丢弃参数意味着模型参数量的提升，参数间适应性提升，模型容量提升，但不一定能提升模型有效容限	
权重衰减系数	调至最优，提升有效容量	权重衰减可以有效的起到限制参数变化的幅度，起到一定的正则作用	
优化器动量	调至最优，可能提升有效容量	动量参数通常用来加快训练，同时更容易跳出极值点，避免陷入局部最优解。	
模型深度	同条件下，深度增加，模型容量提升	同条件，下增加深度意味着模型具有更多的参数，更强的拟合能力。	同条件下，深度越深意味着参数越多，需要的时间和硬件资源也越高。

资料来源：CSDN、申万宏源研究

2.2 主流训练框架对比

软件框架是整个AI技术体系的核心，巨头以开源软件框架为核心打造生态：

- 通过使用者和贡献者之间的良好互动和规模化效应，形成实质标准体系和生态;
- 除苹果等少数公司外，开源框架是主流。

主流训练软件框架：

- TensorFlow（谷歌）、pyTorch（脸书），Caffe/2（脸书，图像处理领域生态积累深厚）、MXNet（亚马逊）、CNTK（微软）
- PaddlePaddle（百度）、 计图（清华）、SenseParrots（商汤）、天元（旷视）

表：主流开源训练框架编程语言和能力评价

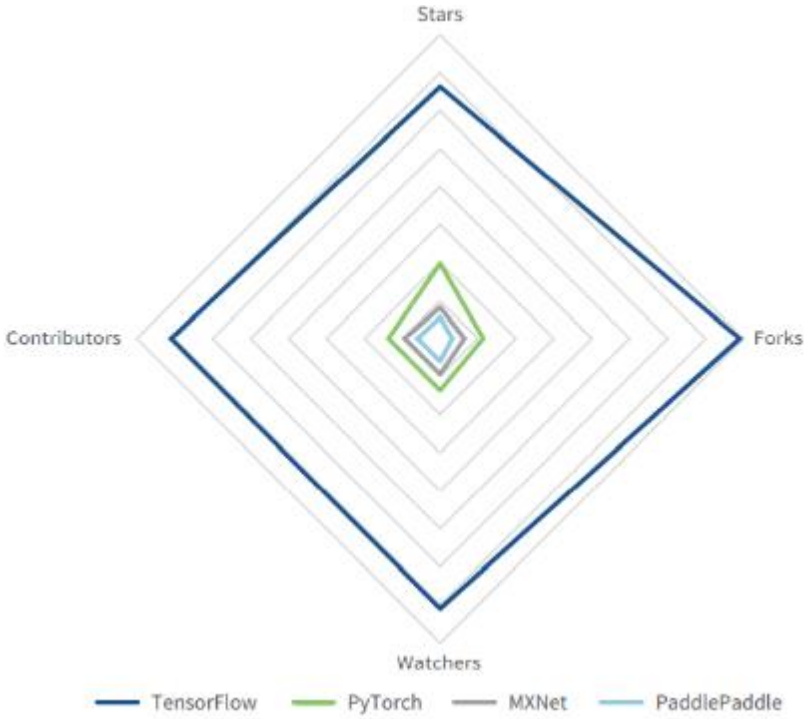
	编程语言	教程和培训材料	CNN模型能力	RNN模型能力	架构：易用性和模块化前端	速度	多GPU支持	Keras兼容性
Theano	Python,C++	++	++	++	+	++	+	+
Tensor-Flow	Python	+++	+++	++	+++	++	++	+
Torch	Python,Lua	+	+++	++	++	+++	++	
Caffe	C++	+	++		+	+	+	
MXNet	R,python,Julia,Scala	++	++	+	++	++	+++	
Neon	Python	+	++	+	+	++	+	
CNTK	C++	+	+	+++	+	++	+	

资料来源：CSDN、申万宏源研究

2.2 海外巨头背书开源训练框架对比

- **现有格局，海外开源框架四巨头**
- **（1）TensorFlow**
  - 前端框架Keras，背后巨头Google；
- **（2）PyTorch**
  - 前端框架FastAI，背后巨头Facebook；
- **（3）MXNet**
  - 前端框架Gluon，背后巨头Amazon；
- **（4）Cognitive Toolkit (CNTK)**
  - 前端框架Keras或Gluon，背后巨头Microsoft。

图：GitHub各框架流行度指标比例雷达图



图：深度学习框架发展时间表

2.3.1 从Theano到tensorflow

Theano：较早的Python深度学习框架，奠定计算图为框架核心、GPU加速理念

- 始于2007，最老牌和最稳定的库之一，第一个有较大影响力的Python深度学习框架；
- **优点：作为早期深度学习框架，结合了CAS和优化编译器，优势明显**，用于定义、优化和求值数学表达式，效率高，非常适用于多维数组。会对用符号式语言定义的程序进行编译，来高效运行于 GPU 或 CPU上。
- **缺点：工程设计薄弱。**Theano不支持分布式计算，在工程设计上有较大的缺陷，有难调试，构建图慢的缺点。2017年后不再维护。

Tensorflow：全工具支持的AI开源框架

- 2015年11月Google推出机器学习开源工具TensorFlow。
- **TensorFlow和Theano设计理念相近**：有很大一批共同的开发者，都是**基于计算图实现自动微分系统**。TensorFlow 使用数据流图进行数值计算。
- **基于计算图实现自动微分系统**，使用数据流图进行数值计算，图中的节点代表数学运算，图中的线条则代表在这些节点之间传递的张量（多维数组）。
- **主流编程工具基本全支持：**支持Python、C++、Java、Go、R等。库可在ARM架构上编译和优化，用户可以在各种服务器和移动设备上部署自己的训练模型。
- **背后Google巨大影响力：**很多企业都在基于TensorFlow 开发自己的产品或将 TensorFlow整合到自己的产品中去，如Airbnb、Uber、Twitter、英特尔、高通、小米、京东等。

2.3.1 Tensorflow出现的问题

Tensorflow：过于复杂和全面的设计导致实际使用生产力低下

- **过于复杂的系统设计**：TensorFlow在GitHub代码仓库的总代码量超过100万行，维护和学习难度极大；
- **频繁变动的接口**：TensorFlow的接口一直处于快速迭代之中，并且没有很好地考虑向后兼容性；
- **接口设计过于晦涩难懂**：创造了图、会话、命名空间、PlaceHolder等诸多抽象概念；
- **文档混乱脱节**：TensorFlow作为一个复杂的系统，文档和教程众多，但缺乏明显的条理和层次

Keras：TensorFlow的默认高级API层

- **在Tensorflow上层封装的高级API层**：纯Python编写而成，以TensorFlow、Theano或CNTK为底层引擎。2017年成为第一个被Google添加到TensorFlow核心中的高级别框架，这让Keras变成TensorFlow的默认API，使Keras + TensorFlow的组合成为Google官方认可并大力支持的平台。
- **优点，提升易用性**：Keras的目标是只需几行代码就能构建一个神经网络，提升易用性。学习使用Keras很容易。
- **缺点：难以学到真正深度学习内容。**开发者大多数时间都在学习如何调用接口，难以真正学习到深度学习的内容，Keras层层封装让用户在新增操作或获取底层的数据信息时过于困难，存在过度封装导致缺乏灵活性的问题，性能也存在瓶颈。
- Keras有助于快速入门，但想了解深度学习需要进一步学习使用TensorFlow。

2.3.2 从Caffe到PyTorch

Caffe：早期有较高完备性和易用性的框架

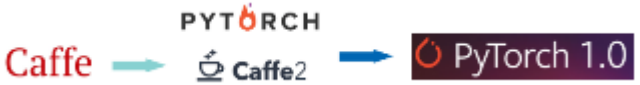
- Convolutional Architecture for Fast Feature Embedding，用于特征提取的卷积架构；最初发起于2013年9月，核心语言C++。作者贾扬清，曾参与过TensorFlow开发。
- **优点：在于较为完备和易用性。**代码和框架都比较简单，代码易于扩展，运行速度快，也适合深入学习分析。在Caffe之前，深度学习领域缺少一个完全公开所有的代码、算法和各种细节的框架。
- **缺点：Caffe不支持分布式，不够灵活。**套用原有模型很方便，但个性化就要读源代码，常常需要用C++和 CUDA编程，Caffe网络结构都是以配置文件形式定义，缺乏以计算图为代表的相对自由灵活、可视化的算法表达。
- **随时间发展，对大型神经网络使用繁琐缺点显现。**截止 2015 年，以 152 层的 ResNet 为代表的一些大型神经网络已经出现，而恰恰针对这种对于大型神经网络，Caffe 使用起来会变得十分繁琐。

Caffe2：针对工业界的轻量化、模块化深度学习算法框架

- 贾扬清在2016年2月加入Facebook，推出Caffe2go。2017年4月Facebook开源Caffe2。
- **优点：定位于工业级、可跨平台部署，将AI生产工具标准化。**Caffe2开发重点是性能和跨平台部署，更注重模块化，支持大规模的分布式计算，支持跨平台。

2.3.2 从Caffe到PyTorch

Torch：适用于卷积神经网络的深度学习框架



- 2002年诞生于纽约大学Torch，后续加入了深度学习的内容，Torch7是Facebook和DeepMind一开始使用的深度学习工具。
- **更高的灵活度，适用于卷积神经网络。**Torch是命令式的，因此与TensorFlow和Theano相比，Torch的灵活度更高，而前两者是陈述式的（declarative），必须declare一个计算图。Torch非常适用于卷积神经网络，第三方的扩展工具包提供了丰富的递归神经网络RNN模型。
- **缺点：基于Lua语言**，但Python很明显已经抢先统治了机器学习领域

PyTorch 1.0：前端PyTorch+后端Caffe2

- PyTorch重新设计了model和intermediate中间变量的关系，使用Python，相比lua提升debug功能。
- **在Facebook的AI双平台定位中专注于快速原型设计和研究的灵活性。**Caffe2的开发重点是性能和跨平台部署，PyTorch 则专注于快速原型设计和研究的灵活性。此前独立发展，但是组件已经被大量共享；
- **PyTorch 1.0 = Caffe2 + PyTorch。**合并后可以将 PyTorch 前端的灵活用户体验与 Caffe2 后端的扩展、部署和嵌入式功能相结合。
- 2018年12月Facebook 正式发布 PyTorch 1.0稳定版。

FastAI：提升PyTorch易用性的高级API层

- 目标是只需几行代码就能让你构建一个神经网络。实测中用5行代码就可以完成Keras用31行才能解决的事情。

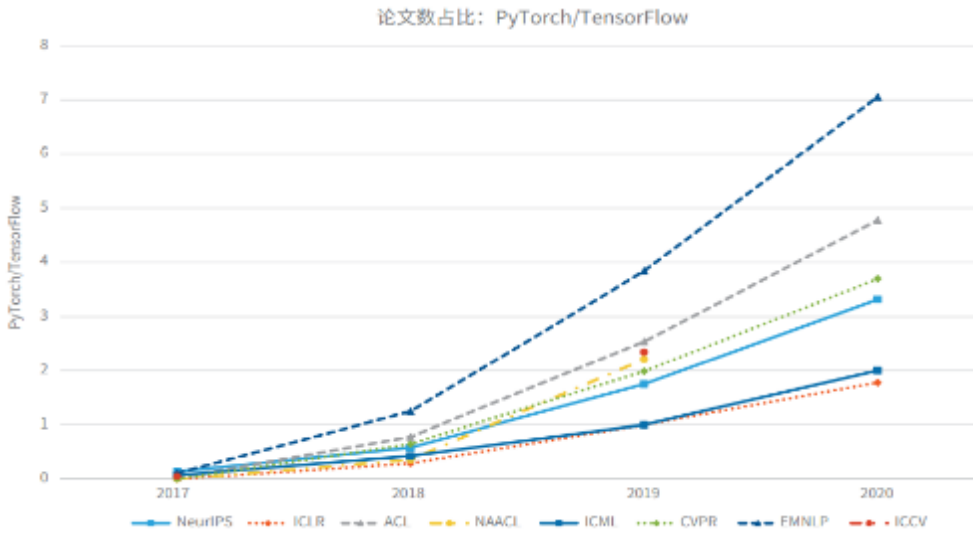
2.3.2* 为何PyTorch可能反超TF?

易用性 和 适配度 的互相取舍

- 截止 2020 年底PyTorch 项目的贡献者大约 1626 人、下游项目 45k + 个，论坛用户34k；
- **学术界PyTorch超过TensorFlow已成定局：** PyTorch以易用性、快速上手取胜，可以快速验证自己的idea；
- **工业界Tensorflow暂时无法替代：**和推理部署框架有更好的兼容性，如Tensorflow和英伟达支持的TensorRT（章节2.5中详细描述）良好兼容，而工业界TensorRT对PyTorch兼容还需要时间；
- **未来工业界谁占优尚无定论：**1）推理部署框架可以在一段时间后得以更好的兼容支持PT，2）即使对于工业界，TF的上层API仍然过于复杂，tf. Keras，tf.layer，tf.contrib等API接口。

图：2018-2020 年中国市场各框架市场认知与份额调研

图：每年各AI顶级研究会议接收的PyTorch论文数和TensorFlow论文数比例



2.3.3 MXNet和CNTK

MXNet：轻量级、可移植、灵活的分布式框架

- Amazon官方主推，支持CNN、RNN和LTSM。诞生于2015年9月，作者是当时在卡耐基梅隆大学CMU读博士的李沐， 2016年11月被亚马逊选为官方开源平台；
- **优点：尝试结合命令式编程（PyTorch）和声明式编程（TensorFlow）。**命令式编程上提供张量运算，声明式编程中支持符号表达式。 同样模型MXN往往占用更小的内存和显存；
- 多语言支持：Python、C++、R、Scala、Julia、Matlab 和 JavaScript。
- **缺点：**文档更新速度较慢，导致新用户难以上手。
- **Gluon：**模仿了PyTorch的接口设计，成为主推的MXNet使用的上层API。

CNTK*：数据包来自微软自己大规模生产

- Computational Network Toolkit，2016年1月在GitHub上开源
- **优点：微软自产数据包。**最初面向语音识别，发展后处理图像、手写字体和语音识别都支持。微软的人工智能工具包跟其他工具包最大的不同在于数据，数据都来自于微软自己的大规模生产数据。包括Cortana、Bing以及Cognitive Services中的Emotion API。
- 基于C++架构，Python或C++编程接口，支持跨平台的CPU/GPU 部署。
- 缺点：CNTK现在还不支持ARM 架构，使其在移动设备上的功能受到了限制。

2.4.1 国内开源架构：百度Paddle、清华Jittor

PaddlePaddle：国内第一个开源神经网络框架

- **2016年8月，百度**在Github上100%开源内部使用多年的深度学习平台PaddlePaddle；
- **中文环境下较多的优势：1）**能够应用于自然语言处理、图像识别、推荐引擎等多个领域，其优势在于开放的多个领先的预训练中文模型，适应中文环境。2）模型库丰富，来自百度各个业务部门贡献；3）较多企业级的包，可以直接在产业界落地使用；4）兼容大量国产AI芯片；
- **整体来看反馈使用感受类似PT，我们对国产开源深度学习框架有极大期待！**
- **劣势：**使用习惯、社区人群数、普及度和海外框架相比有差距，部分模型实现过程有优化空间

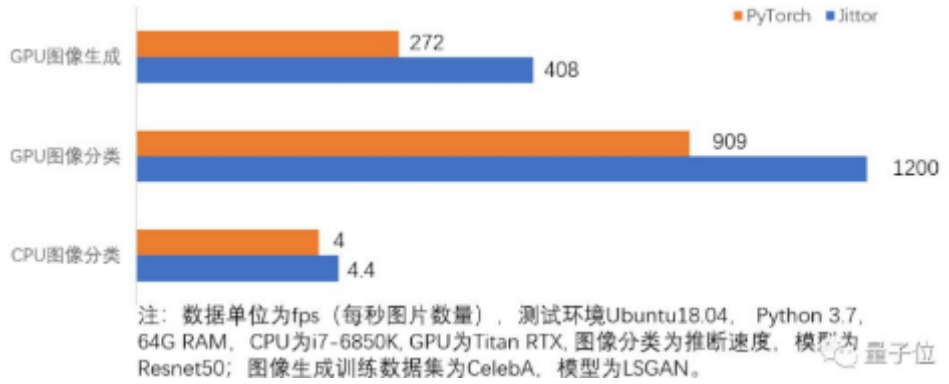
Jittor计图：目的为兼顾易使用、可定制、高性能

- **2020年3月，清华**大学计算机系图形实验室自研深度学习框架Jittor计图对外开源；
- **通过元算子和统计计算图提升易用性：**易用且可定制用户只需要数行代码，就可定义新的算子和模型，在易用的同时，不丧失任何可定制性。支持统一内存、异步接口。

图：Jittor和其它深度学习框架比较

	Tensorflow 1.0	Tensorflow 2.0	PyTorch	Jittor
自动微分	√	√	√	√
动态图	×	√	√	√
同步接口	×	√	√	√
异步接口	√	√	×	√
统一内存	√	√	×	√
跨迭代融合	×	×	×	√

图：Jittor与PyTorch推理与训练速度对比



注：数据单位为fps（每秒图片数量），测试环境Ubuntu18.04，Python 3.7，64G RAM，CPU为i7-6850K，GPU为Titan RTX，图像分类为推断速度，模型为Resnet50；图像生成训练数据集为CelebA，模型为LSGAN。

2.4.2 国内开源架构：华为Mindspore，旷视天元

Mindspore：云边端同步适配

- 2018年10月10日，**华为**首次展示CANN算子库、MindSpore深度学习框架、AI开发平台ModelArts；
- **2020年3月**华为在码云开源MindSpore，企业级AI应用开发者套件ModelArts Pro在华为云上线；
- **设计思路：**着重提升易用性并降低AI开发者的开发门槛，端、边缘和云都适应，并能够在按需协同的基础上，通过实现AI算法即代码。适配华为昇腾AI处理器，也支持GPU、CPU等其它；
- **社区反馈问题：**算子和PyTorch接近但不完全一致，对于静态图理解需要引导，无中文版文档，文档相比tf和pt不够详细。

深度学习天元MegEngine：特色是训练推理一体，静态图动态图都有优化

- 2015年开始搭建，针对当年Caffe架构不足，旷视Brain++在一开始就确立了要以计算图的方式来进行框架搭建的思路，大思路正确；**2020年3月开源MegEngine；**2020年9月推出Brain++商业版。
- **优势：**1）训练推理一体化，训练结果可直接进行用于产品推理、封装。部署时自动删除冗余代码；2）静态图性能高、占用资源少且易于部署、动态图简单灵活、方便调试且易于上手；3）具备Pythonic的API，支持PyTorch Module，直接导入方便；在特定领域如机器视觉模型ResNet 18、ResNet50、MobileNet v2和 ShuffleNet V2上优于其它主流框架。
- **社区反馈问题：**部分支持还不够完善，模型、数据集不够丰富

2.5 推理框架：与硬件和设备端紧密相关

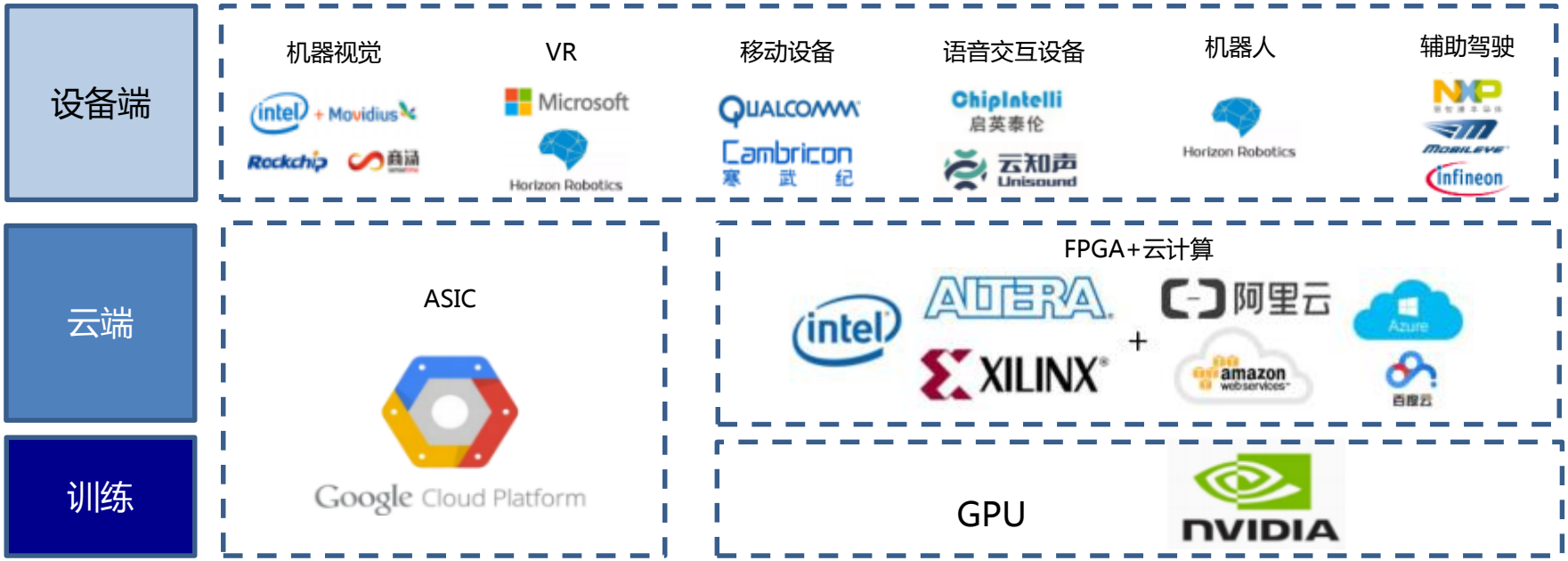
推理框架：主要部署在设备端和云端，因此与硬件厂商紧密相关

表：海外主流推理框架特性比较

模型推理部署框架	应用平台	支持深度学习模型			
		TensorFlow	Caffe	Mxnet	Pytorch
OpenVINO	CPU,GPU嵌入式平台都可以使用，CPU上首选OpenVINO。DephAI嵌入式空间AI平台	√	√	√	
TensorRT	只能用在NIVDIA的GPU上的推理框架。NIVDIA自家的Jeston平台	√	√	√	√
Mediapipe	服务端，移动端，嵌入式平台。TPU	√			

资料来源：CSDN、申万宏源研究

图：推理框架在云端和设备端部署



2.6 开源框架的选择：规模效应与生态

四大顶级深度学习框架阵营可以满足绝大部分开发者要求

- **社区规模效应：**维护力量、贡献人员决定了算法库扩展及时性、API水平，软件框架规模效应较强。
- 科研和工程落地，前者需要有足够的灵活度和易用性，而后者需要的是部署和性能，PT和TF分别对应两种特性，可以满足绝大部分使用者要求。

目前深度学习框架发展趋势

- 1、增加对Python的支持，动态图应用；
- 2、支持分布式和移动端运行平台；
- 3、前端的编程接口更加灵活，设计需要兼容简单高效的命令式和逻辑清晰的声明式；
- 4、训练速度不断提高：支持单机多卡/多机多卡等训练方式；对网络优化减枝以减小训练耗时的同时；提升底层计算硬件单元的计算能力

表：主流开源训练AI框架核心指标对比

	是否支持分布式计算，是不是分布式框架？	是否支持移动端部署？	命令式编程(imperative programming)还是声明式语言(declarative programing)？	基于动态计算图还是静态计算图	是否有强大的社区和生态支持	社区评价
TensorFlow	√	√	声明式	静态计算图	Google	广泛适配，适合工业界
PyTorch	√	√	命令式	动态计算图	Facebook	轻量易上手，适合学术界
MXNet	√	√	命令式	动态计算图	Amazon	优化云端分布式部署
CNTK	√			静态计算图	Microsoft	简单配置易上手
Theano			声明式	静态计算图		
Caffe			声明式	静态计算图		
Caffe2	√	√		静态计算图		

资料来源：GitHub，CNDN，申万宏源研究

2.6 开源框架的选择：国产自研深度学习框架原因

1、技术遗留问题

- **静态图、动态图技术方案都还有缺陷，有同时解决的可能性**
- *动态图：其核心特点是计算图的构建和计算同时发生（Define by run）。优点是调试方便，缺点是难以对整个计算图进行优化。PT*
- *静态图：将计算图的构建和实际计算分开（Define and run）。优点是对全局的信息掌握更丰富，可以做的优化更多，缺点是无法实时观察中间结果。TF*
- 在网络结构、设备兼容、性能与功耗均衡和各种自动化设计等有提升空间

2、国内特色问题

- 特定场景框架可能更优
- 国产芯片和适配
- 开源平台可能工业包不共享的问题
- 中文环境的API

- 国内百度、华为、商汤、旷视在自研框架初期就考虑到训练速度要求提高带来的各种问题，同时**适应国产服务器芯片等环境**

目录

3 本节结论：深度学习热点“大模型”优缺点同时存在

以GPT为代表的“大模型”是什么

- 大规模预训练：GPT(Generative Pre-Training)是OpenAI在2018年提出的模型，基于Transformer模型。采用Pre-training + Fine-tuning训练模式，使大量无标记数据得以利用。

优势显著：大幅提升对数据要求，长尾场景落地新思路

- **自监督学习功能，大幅降低对数据量的需求：**GPT舍弃Fine-tuning ，先使用海量数据预训练大模型，得到一套模型参数，然后用这套参数对模型进行初始化，再进行训练。大幅降低后续对数据量的需求。
- **预训练大模型+细分场景微调，更适合长尾落地：**大规模预训练可以有效地从大量标记和未标记的数据中捕获知识，通过将知识存储到大量的参数中并对特定任务进行微调，极大扩展模型的泛化能力。
- **有望进一步突破现有模型结构的精度局限：**可能继续突破精度上限。

但对自然语言逻辑理解仍有缺陷

- “尽管GPT-3观察到它读到的单词和短语之间的统计关系，但不理解其含义。”

对存储、算力要求极高，普通机构难以复现

- 据 NVIDIA 估算如果要训练 GPT-3 ，用 8 张 V100 的显卡，训练时长预计要 36 年；以微软与OpenAI合作建造 的Azure A工智能算力平台为例，该算力平台投资约10亿美元，使用该超算中心训练一次超大模型GPT-3大约花费1200万美元。
- **解决分布式训练问题：**上百台服务器之间的通信、拓扑、模型并行、流水并行等问题，**模型训练是显存峰值问题。**GPT-3发布一年后，只有 **NVIDIA 、微软等**大企业可以复现 。

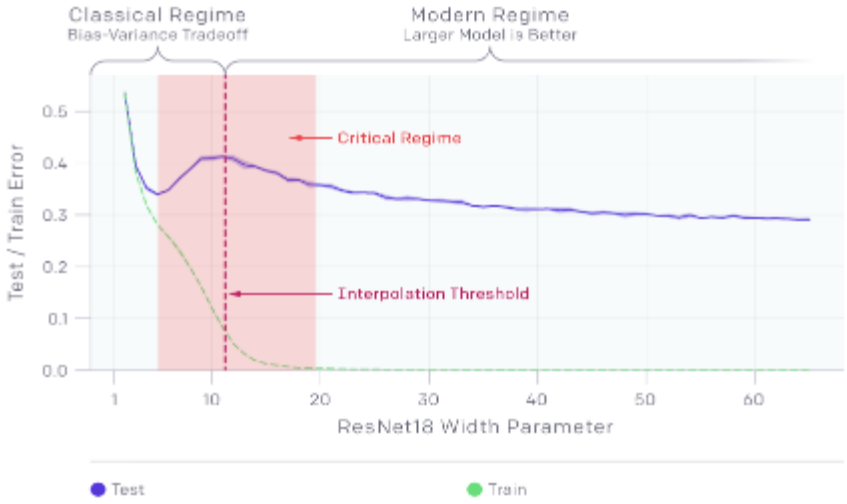
3.1 行业更新热点在“大模型”

- 此前AI算法基于深度学习创新，从业者使用DNN、CNN、RNN等模型以及变种，加上attention、GRU等机制，产生巨大的收益。但是近年很多算法与策略都已经使用过，前沿创新、业务演进减缓。
- 最后一次底层算法创新被认为是2017年Google的**Transformer**（3.1*）。

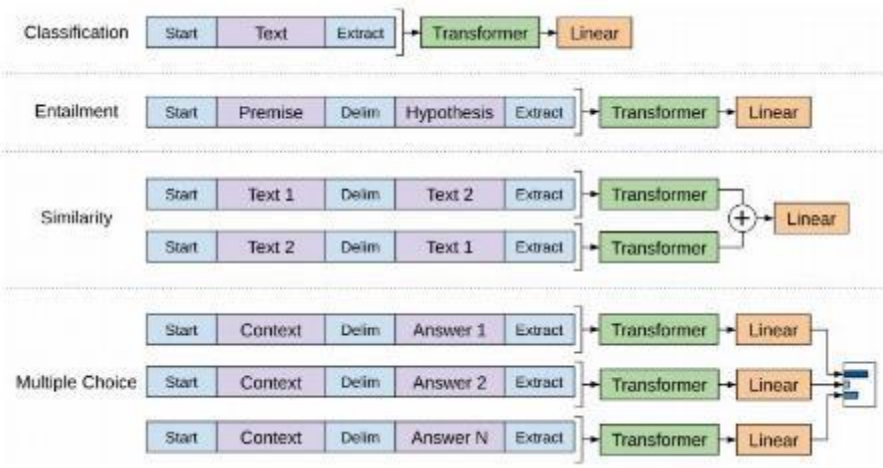
大规模预训练模型（大模型）成为AI算法领域的热点

- **Double Descent现象。**传统机器学习里，模型过小则欠拟合，模型过大则过拟合。深度学习里Double Descent现象在2018年揭示，随着模型参数变多，Test Error是先下降，再上升，**然后第二次下降；**原则上，在成本可接受的情况下，模型越大，准确率越好。
- **大规模预训练：**GPT(Generative Pre-Training)，是OpenAI在2018年提出的模型，利用Transformer模型来解决各种自然语言问题，例如分类、推理、问答、相似度等应用的模型。GPT采用了**Pre-training + Fine-tuning**的训练模式，使得大量无标记的数据得以利用，大大提高了这些问题的效果。

图：深度学习中的Double Descent现象



图：对于不同问题进行不同的预训练

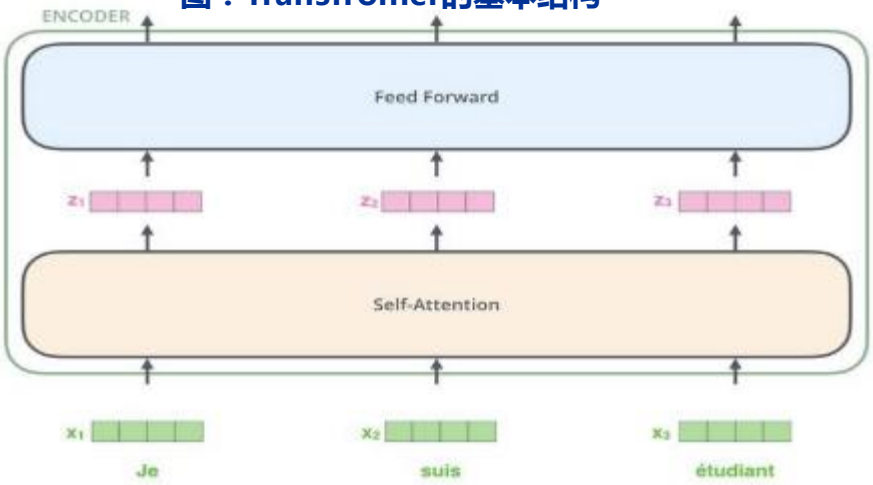


3.1* Transfomer对RNN的改进

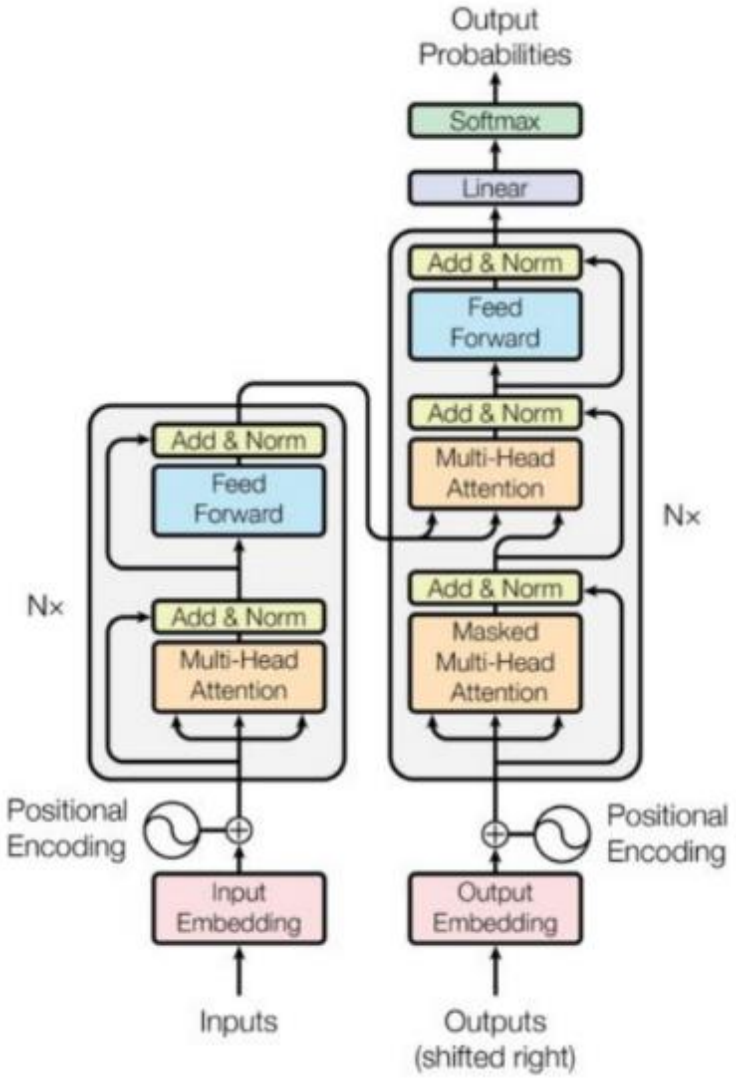
Transformer取代RNN

- **RNN存在问题：**1）效率问题：需要逐个词进行处理，后一个词要等到前一个词的隐状态输出以后才能开始处理，2）如果传递距离过长还会有梯度消失、梯度爆炸和遗忘问题
- **Transformer。**Google Brain 2017的提出，针对RNN的弱点进行重新设计，解决了RNN效率问题和传递中的缺陷等，在很多问题上都超过了RNN的表现。
- **N进N出的结构，**Transformer解决了效率问题和距离问题。**Self-Attention和Feed Forward Networks**
- **在机器翻译任务上，Transformer表现超过了RNN和CNN只需要编/解码器就能达到很好的效果。在CV领域也有应用**

图：Transfromer的基本结构



图：Transformer的详细结构



3.1 行业更新热点在“大模型”

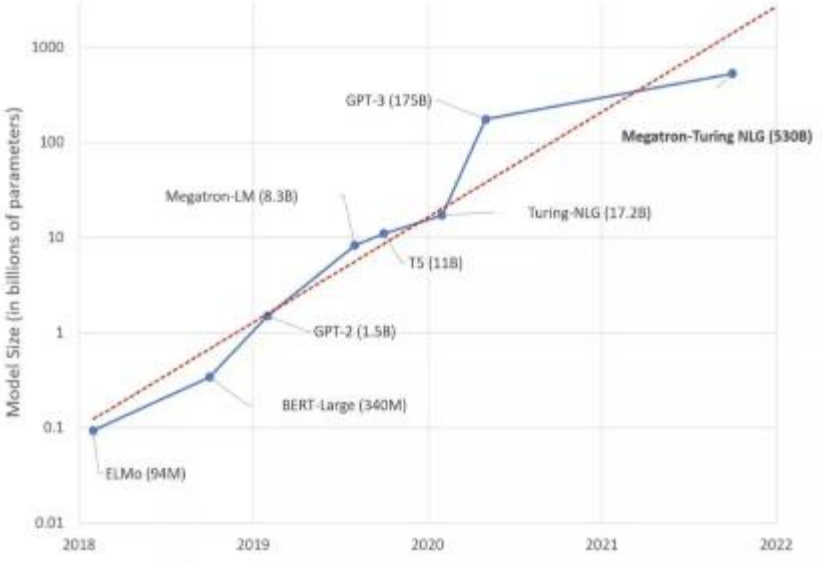
大模型2018至今快速迭代

- 1）2018年，OpenAI基于Transformer提出了GPT；2）2019年，Google推出了GPT的升级版BERT；3）2019年，OpenAI推出了GPT的升级版GPT2.0；4）2020年，GPT-3；5）2021，Switch Transformer、MT-NLG。

分为单体模型、混合模型两类

- **单体/稠密模型：**OpenAI GPT-3，华为盘古/鹏程盘古α（MindSpore支撑）；模型规模的扩展是全结构的扩容；
- **混合/稀疏模型：**Google Switch Transformer，智源悟道2.0，阿里M6。一般来说是选择一个基础的稠密模型，通过MoE稀疏结构扩展FFN部分，以此来达成模型的扩容。

图：2018年后大模型迭代与参数量快速提升



资料来源：Nvidia Developer《Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World’s Largest and Most Powerful Generative Language Model》，申万宏源研究

表：主流大模型和参数对比

单体模型				
公司	模型名	参数量	数据量	领域
OpenAI	GPT-3	1750亿	570GB高质量数据集	专注自然语言理解（NLP）
浪潮	源1.0	2457亿	5000GB高质量中文数据集	专注自然语言理解（NLP）
微软-英伟达	MT-NLG	5300亿	835GB高质量数据集	专注自然语言理解（NLP）
混合模型				
公司	模型名	参数量	数据量	领域
谷歌	Switch Transformer	1.6万亿	/	专注自然语言理解（NLP）
智源研究院	悟道	1.75万亿	/	中文、多模态、认知、蛋白质预测等系列模型

3.2 OpenAI： GPT-3

GPT-3：更少的领域数据、且不经过精调步骤去解决问题

- **GPT-2舍弃了模型Fine-Tuning过程**，不再规定任务，转向容量更大、无监督训练、更加通用；
- **GPT-3继续增加参数**：具有**1,750亿**个参数的自然语言深度学习模型（GPT-2 100倍）
- 该模型经过了将近0.5万亿个单词的预训练，并且在不进行微调的情况下，可以在多个NLP基准上达到最先进的性能。
- GPT-3 在许多 NLP 数据集上均具有出色的性能，包括翻译、问答和文本填空任务，这还包括一些需要即时推理或领域适应的任务，例如给一句话中的单词替换成同义词，或执行3位数的数学运算。

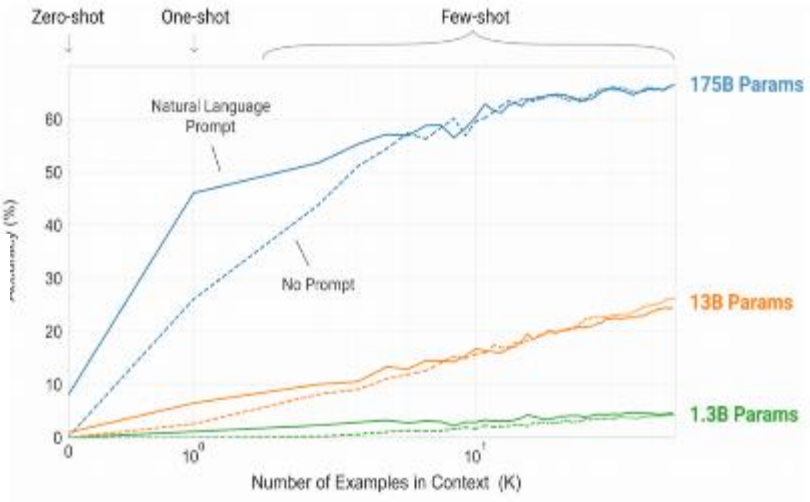
图：GPT-3的训练数据集庞大

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

图：GPT-3不同尺寸模型效果对比

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

图：Few-shot下GPT-3有很好的表现

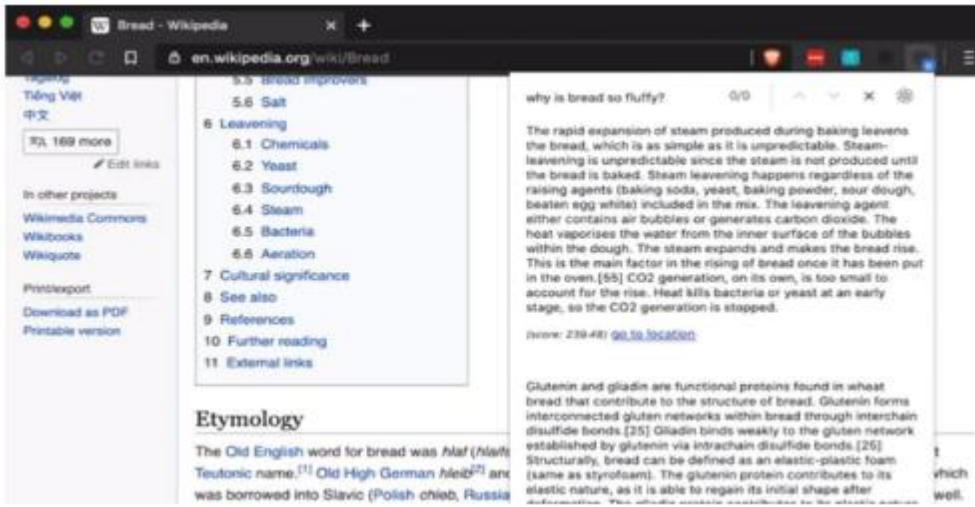


3.2 GPT-3在自然语言领域能做什么

人们使用GPT-3创建的一小部分示例

- **1、基于问题的搜索引擎：**键入问题，GPT-3会将定向到相关的维基百科URL作为答案。
- **2、与历史人物交谈的聊天机器人：**启动GPT-3，使其像哲学家罗素一样讲话。
- **3、仅需几个样本，即可解决语言和语法难题。**
- **4、基于文本描述的代码生成：**用简单的文字描述你选择的设计元素或页面布局，GPT-3会弹出相关代码。
- **5、回答医疗问题：**医疗保健问题不仅给出了正确答案，还正确解释了潜在的生物学机制。
- **6、基于文本的探险游戏。**
- **7、文本的风格迁移：**以某种格式编写的输入文本，GPT-3可以将其更改为另一种格式。
- **8、自行生成音乐：**编写吉他曲谱。
- **9、写创意小说。**
- **其它**

图：使用GPT-3的问题搜索引擎



图：GPT-3自动生成的新闻文章

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

3.2 GPT-3工具DALL·E和“逻辑”优化

DALL•E：可以按照文字描述、生成对应图片

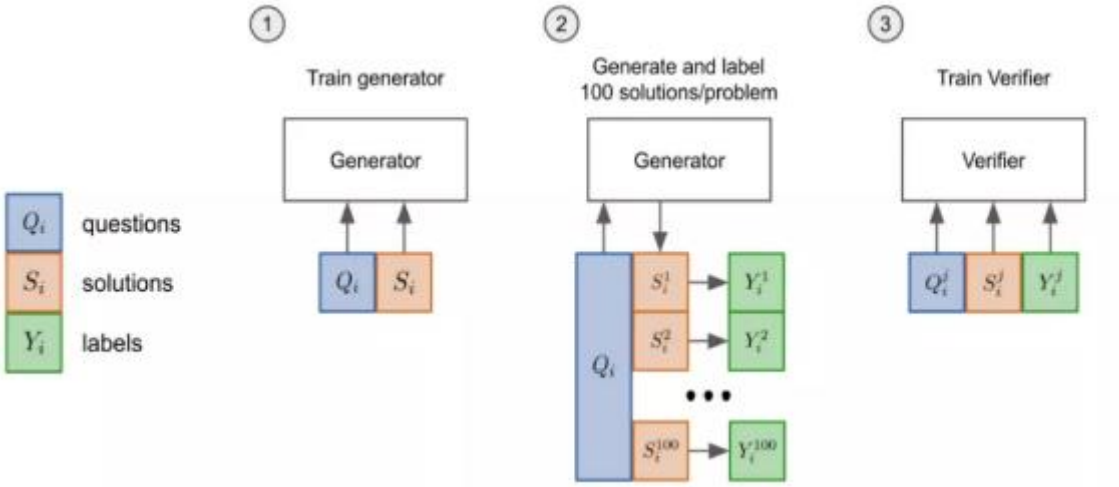
- 基于GPT-3构建，仅使用了120亿个参数样本，相当于GPT-3参数量的十四分之一；
- 意义：降低了深度学习需要的数据标注量，文本和图像理解结合起来。

通过验证器优化复杂逻辑问题

- 为解决GPT-3产生逻辑错误，OpenAI提出了一个训练验证器（verifier）来判断模型完成的正确性；
- 该研究训练了一个解决小学数学问题的系统，其准确率约是经过微调的 GPT-3 模型的两倍。它能像真正的学生一样可以解决 90% 的数学应用问题。

图：DALL•E按要求设计“一颗白菜穿着芭蕾舞裙在遛狗”

图：Open AI训练验证器判断模型生成的解决方案的正确性



3.3 微软和英伟达：MT-NLG

2021年10月微软和英伟达推出迄今为止训练最大最强的语言模型MT-NLG

- MT-NLG是最大最强的**生成语言模型**（Generative Language Model）；
- **基础设施**：英伟达 A100 Tensor Core GPU 和 HDR InfiniBand 网络支撑的 SOTA 超级计算集群；
- **软件设计**：使用来自 Megatron-LM 的 tensor-slicing 来扩展节点内的模型，并使用来自 DeepSpeed 的 pipeline 并行来跨节点扩展模型。
- **擅长应用**：完成预测（Completion prediction）；阅读理解（Reading comprehension）；常识推理（Commonsense reasoning）；自然语言推论（Natural language inferences）；词义消歧（Wordsense disambiguation）

图：用于训练 MT-NLG 模型的数据集

Dataset	Tokens (billions)	Weights (%)	Epochs
Books3	25.7	14.3	1.5
OpenWebText2	14.0	19.3	3.6
Stack Exchange	11.6	5.7	1.4
PubMed Abstracts	4.4	2.9	1.8
Wikipedia	4.2	4.8	3.2
Gutenberg (PG-19)	2.7	0.9	0.9
BookCorpus2	1.5	1.0	1.0
NIH ExPorter	0.3	0.2	1.8
Pile-CC	49.8	9.4	0.5
ArXiv	20.8	1.4	0.2
GitHub	24.3	1.4	0.2
CC-2020-50	68.7	13.0	0.5
CC-2021-04	82.6	15.7	0.5
RealNews	21.9	9.0	1.1
CC-Stories	5.3	0.9	0.5

图：MT-NLG 在不同句法结构下的数学运算推理和自然语言推理的样本

Tasks	Sample 1	Sample 2
Inferring mathematical operations	**Prompt:** Q: 11 # 2 = ? A: 5.5 Q: 18 # 9 = ? A: 2 Q: 35 # 4 = ? A: 8.75 Q: 27 # 9 = ? A: 3 Q: 75 # 6 = ? A: **MT-NLG:** 12.5 Q: 45 # 5 = ? A: 9 Q: 18 # 6 = ? A: 3	**Prompt:** Q: 11 $^% 2 = ? A: 22 Q: 3 $^% 9 = ? A: 27 Q: 12 $^% 4 = ? A: 48 Q: 7 $^% 7 = ? A: 49 Q: 8 $^% 5 = ? A: **MT-NLG:** 40 Q: 9 $^% 6 = ? A: 54 Q: 10 $^% 5 = ? A: 50
Natural language inference on different syntactic structures	**Prompt:** The banker forgot that the judge advised the professors. Question: The judge advised the professors. True or False? Answer: **MT-NLG:** True. The banker forgot that the judge advised the professors.	**Prompt:** Famous professors supported the secretary. Question: Professors supported the secretary. True or False? Answer: **MT-NLG:** True. The secretary was supported by famous professors.

- 即使符号被严重混淆（sample 2），该模型也可以从上下文中推断出基本的数学运算（sample 1）。
- 虽然称不上拥有了算术能力，但该模型似乎超越了仅记忆算术的水平

资料来源：Nvidia Developer《Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model》，申万宏源研究

3.4 Google：Switch Transformers

1.6万亿参数的Switch Transformers

- **稀疏激活模型**：此模型可以保证计算成本基本保持不变的同时允许网络拥有巨量的参数。谷歌改进了专家混合范式(MoE，Mixture-of-Experts)层；
- 通过简化MoE，得到了一个易于理解、易于训练的体系结构，该结构还比同等大小的密集模型具有更大的采样效率。

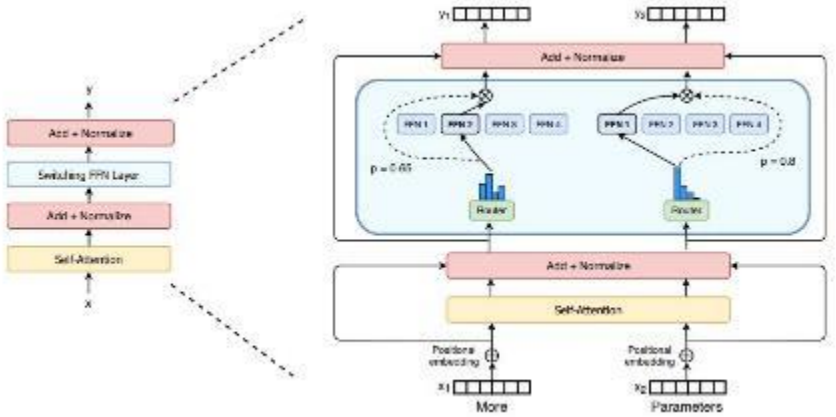
可扩展、高效的自然语言学习模型

- **预训练、微调和多任务训练**表现出色。

但是参数量和任务效果并非完全等比例扩大

- Switch-Base是基于T5-Base的MoE稀疏扩展，参数规模是T5-Large的10倍，也就是说内存开销是T5的10倍，算力开销是T5-Large的29%；
- 右表格的下游任务对比来看，在同样的算力开销下，Switch-Base的效果比T5-Base整体上要好，这个优势是通过33倍的内存开销换取的；
- 但是同时，**Switch-Base在参数量比T5-Large大了10倍的情况下，效果比T5-Large要差一些。**

图：Switch Transformer编码块



图：Switch Transformer和T5下游任务对比结果

Model	GLUE	SQuAD	SuperGLUE	Winogrande (XL)
T5-Base	84.3	85.5	75.1	66.6
Switch-Base	**86.7**	**87.2**	**79.5**	**73.3**
T5-Large	87.8	88.1	82.7	79.1
Switch-Large	**88.5**	**88.6**	**84.7**	**83.0**

Model	XSum	ANLI (R3)	ARC Easy	ARC Chal.
T5-Base	18.7	51.8	56.7	**35.5**
Switch-Base	**20.3**	**54.0**	**61.3**	32.8
T5-Large	20.9	56.6	**68.8**	**35.5**
Switch-Large	**22.3**	**58.6**	66.0	**35.5**

Model	CB Web QA	CB Natural QA	CB Trivia QA
T5-Base	26.6	25.8	24.5
Switch-Base	**27.4**	**26.8**	**30.7**
T5-Large	27.7	27.6	29.5
Switch-Large	**31.3**	**29.5**	**36.9**

3.5 华为云：盘古大模型

盘古：最大中文语言预训练模型

- 2021年4月发布，千亿参数40TB训练数据的全球最大中文语言（NLP）预训练模型，30亿参数的全球最大视觉（CV）预训练模型。

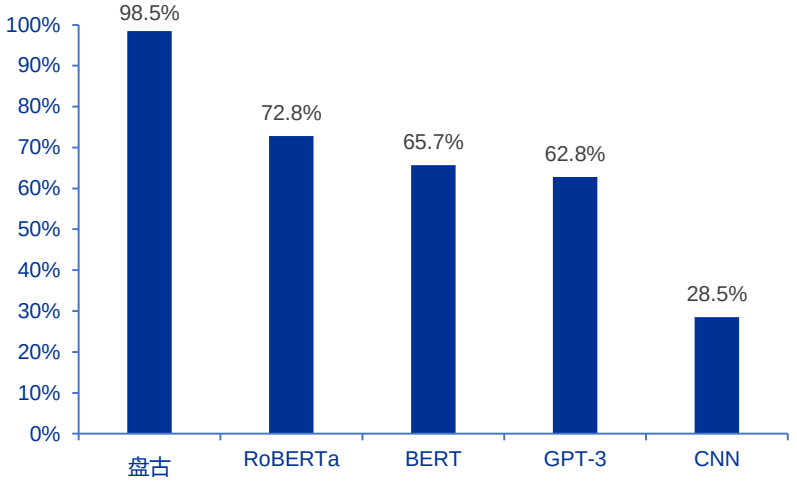
基于GPT思路提升商业落地可用性

- 将 P-tuning、priming 等最新技术融入到盘古的微调框架中，**提升微调效果**；
- 在**样本极少**的情况下，盘古的少样本学习能力远超GPT系列和BERT系列；
- 英文The Pile 825GB数据，中文数据集最大开源项目CLUECorpus2020只包含100GB高质量数据集；
- 要得到相同的F1结果，盘古所需的数据量仅为中文GPT-3的1/9，实现了近10倍的生产效率提升。

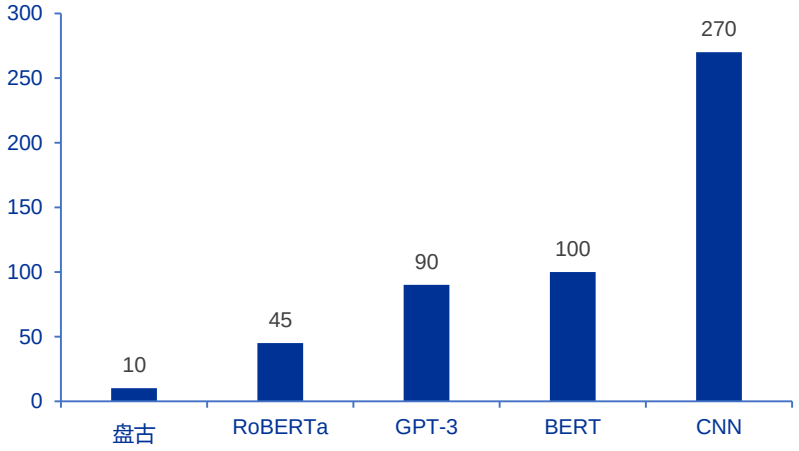
为产业落地做出的更多努力

- 深入具体场景
- 打造通用 API

图：复杂商用场景实测不同模型少样本学习达到的 F1 结果（100%表示跟 full label 结果相同）



图：各模型复杂商用场景实测得到目标 F1 结果所需的平均样本量



资料来源：华为云官网，申万宏源研究

3.6 以GPT为例，大模型的优势和局限

自监督学习功能，大幅降低对数据量的需求：

- 传统的模型训练方式是反向传播算法，先对网络中的参数进行随机初始化（预训练大模型中不是随机初始化的），再利用随机梯度下降等优化算法不断优化模型参数，这种方式下对数据需求量较大。
- **GPT-3先使用海量数据预训练大模型，得到一套模型参数，然后用这套参数对模型进行初始化，再进行训练。大幅降低后续对数据量的需求。**

预训练大模型+细分场景微调，更适合长尾落地

- 大模型提供了一种“预训练大模型+下游任务微调”的方式。大规模预训练可以有效地从**大量标记和未标记的数据中捕获知识**，通过将知识存储到大量的参数中并**对特定任务进行微调**，极大扩展模型的泛化能力。
- 例：在NLP领域，预训练大模型共享了预训任务和部分下游任务的参数，在一定程度上解决了通用性的难题，可以被应用于翻译，问答，文本生成等自然语言任务。

有望进一步突破现有模型结构的精度局限

- 此前深度学习模型精度提升，主要依赖网络在结构变革。例如，从AlexNet到ResNet50，再到NAS搜索出来的EfficientNet，ImageNet Top-1 精度从58提升到了84。但近年来提升有限。
- **大模型可能继续突破精度上限。**以谷歌2021年发布的视觉迁移模型Big Transfer为例，扩大数据规模也能带来精度提升，使用JFT-300M训练ResNet152x4，精度可以上升到87.5%，相比ILSVRC-2012+ResNet50结构提升了10.5%

有可能新的更佳的商业模式：未来可能部分API收费，不排除按调用量收费

3.6 以GPT为例，大模型的优势和局限

对自然语言逻辑理解仍有缺陷

- 基于2020年7月OpenAI首席执行Sam Altman论点：“GPT-3仍然存在严重的弱点，有时还会犯一些非常愚蠢的错误。尽管GPT-3能观察到它读到的单词和短语之间的统计关系，但不理解其含义。”
- 复杂商用场景的少样本学习能力较弱；
- 对于微调并不友好，在落地场景中难以进一步优化；
- GPT-3 只能进行直接的、端到端的生成（把知识库做成很长的一段文字，直接放进 prompt 中），难以融入领域知识。

对存储、算力要求极高，普通机构难以复现

- 据 NVIDIA 估算，如果要训练 GPT-3 ，即使单个机器的显存 / 内存能装得下，用 8 张 V100 的显卡，训练时长预计要 36 年；即使用 512 张 V100 ，训练也需要将近 7 个月；如果拥有 1024 张 80GB A100，那么完整训练 GPT-3 的时长可以缩减到 1 个月。
- 以微软与OpenAI合作建造 的Azure A工智能算力平台为例，该算力平台投资约10亿美元，使用该超算中心训练一次超大模型GPT-3大约花费1200万美元。
- **解决分布式训练问题：**上百台服务器之间的通信、拓扑、模型并行、流水并行等问题，**模型训练是显存峰值问题。**
- GPT-3发布一年后，只有 **NVIDIA 、微软等**大企业可以复现 。

目录

4 本节结论：商汤和旷视的自研平台各有特色

商汤科技：SenseCore大模型+小模型，降低AI应用落地成本

- **算力层上海临港大型AI计算中心AIDC**：预计能够产生每秒3.74百亿亿次浮点运算的总算力。
- **自研视觉算法训练框架SenseParrots**：动态实图时编译。所有的代码，都是在运行过程当中即时编译，并且放到引擎上大规模地并行执行。同时具备传统的静态深度网络的伸缩性，也具备了当代的动态编程模型的灵活性。
- **开源推理平台SensePPL**：商汤 HPC 团队从2015年开始研发的推理部署引擎；对模型进行加载转换，生成有向图和执行计划，进行图级别优化，运行时调用深度调教过的算子库进行推理计算。
- **开源算法库OpenMMLab**：超过22000个算法模型。

旷视科技：Brain++平台支撑了跨行业AIoT解决方案

- **开源框架天元MegEngine**：训练推理一体化，训练结果可直接进行用于产品推理、封装。部署时自动删除冗余代码；静态图性能高、占用资源少且易于部署、动态图简单灵活、方便调试且易于上手。
- **云计算平台MegCompute**：E级算力资源调度、EB级海量数据存储管理、400G RDMA高速骨干网络。
- 三类算法：深度学习算法、计算机视觉算法、AIoT算法
- **AI平台使得极度碎片化的AIoT具备大规模生产可能性**

4.1 商汤：Sense Core算力、平台、算法全覆盖

整体定位：高效率、低成本、规模化的新型人工智能基础设施

算力层：大型AI计算中心AIDC

- 在上海临港建设大型人工智能计算中心（AIDC），预计能够产生**每秒3.74百亿亿次浮点运算**的总算力。

平台层：SenseParrots

- 视觉算法训练框架。高效利用GPU集群算力，训练单个大模型时可以在一千块GPU上取得超过**90%的加速效率**，在业内处于领先水平;

算法层：超过22000个算法模型

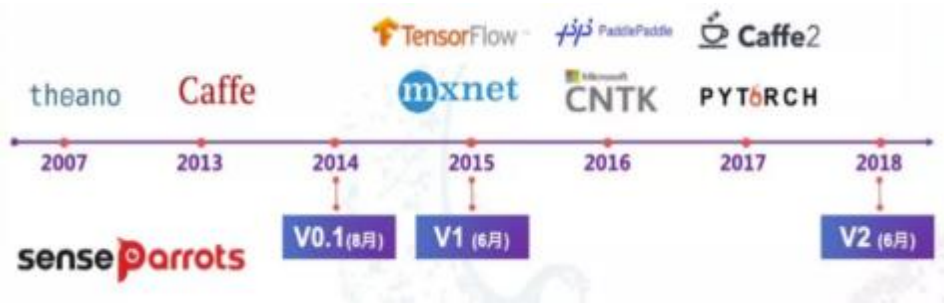
- 与香港中文大学商汤联合实验室共同打造了算法开源计划OpenMMLab，开源算法训练及推理模型，与外部社区共同构建创新生态;
- GitHub上超40000颗星，**亚洲星数最高**，与国内其他开源框架总星数相当。

4.1 商汤： Sense Core平台层从parrots到PPL

SenseParrots：商汤训练框架，也是商汤基础设施SenseCore的核心

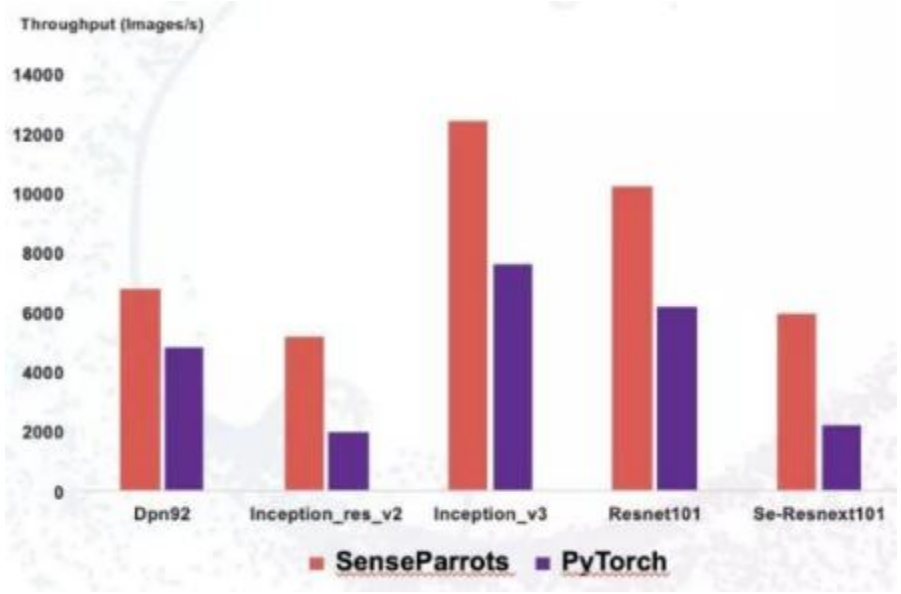
- 第一代专注于静态网络，规模达到1207层卷积神经网络，这是已公开的最深的卷积神经网络；
- **第二代升级为当前AI模型体系**
- **动态实图时编译。**所有的代码，都是在运行过程当中即时编译，并且放到引擎上大规模地并行执行。
- 同时具备传统的静态深度网络的**伸缩性**，也具备了当代的动态编程模型的**灵活性。**

图：SenseParrotsV0.1在TF之前出现



图：SenseParrots训练框架是商汤人工智能基础设施SenseCore的核心

图：64卡V100训练吞吐量与PT对比



资料来源：商汤集团招股书、商汤科技官网，申万宏源研究

4.1 商汤： Sense Core平台层从parrots到PPL

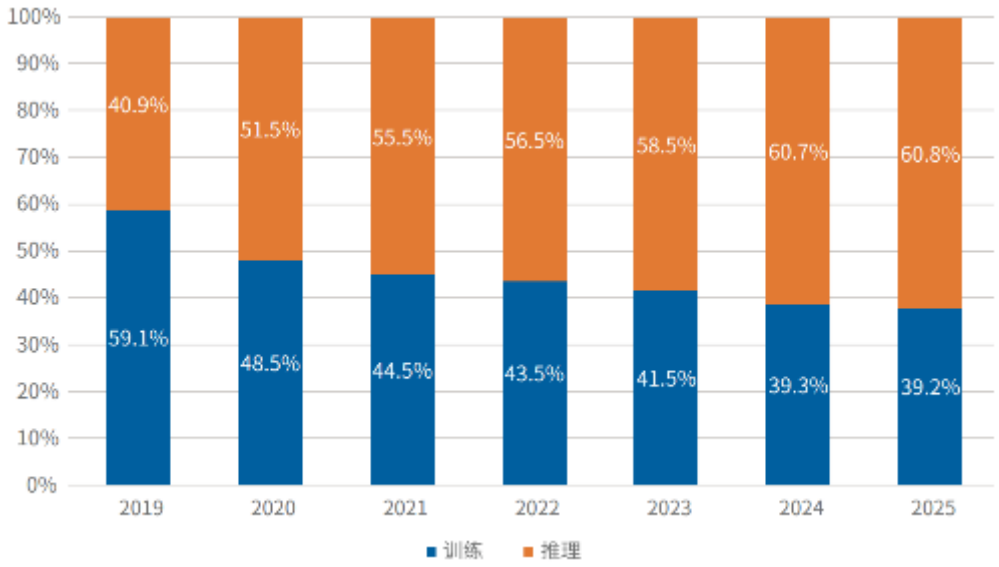
SensePPL：商汤开源推理平台

- SensePPL 是商汤 HPC 团队从2015年开始研发的推理部署引擎，早期定位于**算子优化库。2021**年开源；
- 训练平台训练好的模型，可以通过转换成 onnx 等标准格式，支持CPU、AMD Zen架构和**国产多款 x86 处理器**，支持NVIDIA GPU，支持**ARM server**；
- 对模型进行加载转换，生成有向图和执行计划，进行图级别优化，运行时调用深度调教过的算子库进行推理计算。所有核心框架和算子库的代码，**完全自研。**

随着未来算法成熟、工业化成为行业重点，IDC预测推理负载占比提升，推理框架重要性可能不断增加

图：PPL核心计算库和特性

图：IDC预测未来推理框架负载占比将增加



4.1 商汤： Sense Core算法层OpenMMLab

开源算法库OpenMMLab：

- 算法层的模块化组装，解决长尾问题，提升客户付费意愿。对于**商业化AI企业而言，成熟算法库是非常重要的工具。**
- **年生产商用模型库：**由2019年的1,152个模型，提升至2020年的9,673个，截至2021年上半年提高到了8,377个；**研发人员年人均生产商用模型数：**从0.44提高到3.45又继续提高到了5.24个。
- **积累商业模型：**积累超过22,000模型，涉及多个垂直行业。
- 目前商汤科技尚未开源SenseParrots，但是2019年开始陆续开源OpenMMLabs，**GitHub上超40000颗星，亚洲星数最高**，与国内其他开源框架总星数相当。
- 包括**计算机视觉基础库、物体检测工具箱、行为理解工具箱和服饰分析工具箱**等

图：基于PyTorch开源的目标检测工具箱MMDetection

表：商汤OpenMMLab发展历程

发布时间	发布内容	特点	Github Fork	Github Star
2018年10月	MMCV（计算机视觉基础库）	OpenMMLab的核心，它拥有通用的IO接口，支持图像处理、视频处理，提供良好的图像和标注可视化支持	390+	1600+
	MMDetection物体检测工具箱（目标检测库）	相比于其它开源检测库，MMDetection有多项优点，包括高度模块化设计，多种算法框架支持，显著提高训练效率和密切同步最新算法支持	3600+	10500+
	MMAction行为理解工具箱（动作识别和检测库）	已实现多种动作分析算法框架，持有多种流行的数据集，并全面支持视频动作分析的各种任务	273	1300+
2019年6月	MMDetection物体检测工具箱（目标检测库）升级至1.0	提供了一大批新的算法实现。越来越多的MMLab以外的研究团队开始把MMDetection作为实现新的目标检测算法的基础，并不断把他们开发的新算法回馈到MMDetection，有效促进目标检测领域的应用和新方法研究发展	3600+	10500+
	MMSkeleton基于骨架的视频分析	拥有领先的在视频中进行骨骼识别的能力，支持框架预训练模型并提供多数据集	601	1800+
2019年9月	MMSR超分辨率工具箱	一个针对图像和视频超分辨率（就Super—Resolution，简称超分）的工具箱。拥有统一的框架，支持图像超分和视频超分的处理，也支持去噪、去模糊等其他底层视觉任务。MMSR涵盖了一系列比赛冠军的算法，同时具有良好扩展性能	233	1200+
2019年11月	MMFashion服饰分析工具箱	一款针对视觉时尚分析的工具包。实现了服饰领域的三大主流任务；服饰属性预测；给定服饰找同款；服饰关键点定位。目前开源的内容包括项目代码和预训练模型。Mmfashion旨在给服饰分析领域提供一个通用性广，易于拓展功能性强的工具包。	96	492
2020年7月	战略升级	OpenMMLab战略升级成为人工智能算法开放体系，并发布更多库和工具箱		

资料来源：Github、申万宏源研究

资料来源：商汤科技公众号、申万宏源研究

4.1 商汤：大模型+小模型降低AI落地成本

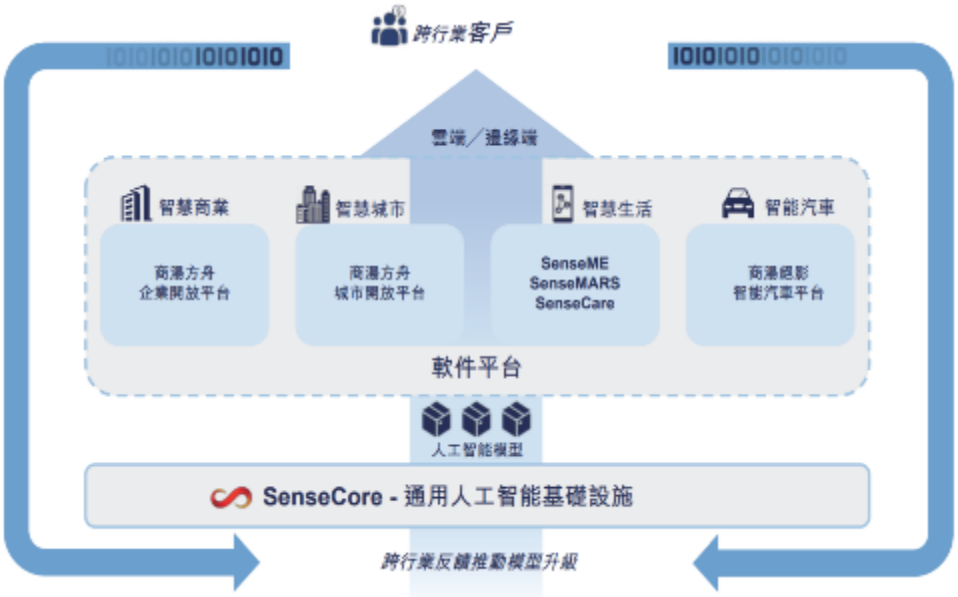
底层自研框架使得GTP等大模型的复现更加可行

- **大模型是业界发展趋势，但是在复现上有众多问题。**如3.6，GTP-3在复现时对存储、算力硬件要求极高，需要对分布式并行计算十分了解，同时再模型层改进以解决峰值显存等问题。
- **底层自研框架，相比使用开源框架，在大模型复现上有更多优势。**目前海外开源框架在网络结构、设备兼容、性能与功耗均衡和各种自动化设计等均出现局限性。商汤通过底层技术定制优化，可更好支撑超大模型训练。重算搜索显存压缩、 Distributed Model States、多粒度稀疏通信等解决存显、通信、并行问题。
- **SenseParrots大规模并行训练。**SenseParrots可将训练任 务扩展至数千个GPU。它在1,024个GPU上实现了**91.5%**的高并行效率，而主流训练框架的效率约为**25%**。根据沙利文报告，商汤已训练的一个包含超过300亿个参数用于视觉识别的AI模型。

大模型+小模型，降低AI应用落地成本

- SenseCore算力、算法、数据三重保证，先对大模型预训练（较少的Fine-Tuning和微调），产生较通用模型；
- 再通过量化、剪枝、知识蒸馏等模型压缩方法把大模型变小，高效的进行模型生产；
- SenseCore把非结构化的数据结构化；
- **前期阶段投入较大，**无论自研训练和推理平台搭建、生态维护、AIDC建设、大模型训练，都需要巨额投入；
- 但**避免了“手工作坊”式模型生产的长尾投入。**

图：商汤在SenseCore通用基础上搭载行业软件平台



4.2 旷视：Brain++ AI生产力平台

深度学习天元MegEngine：特色是训练推理一体，静态图动态图都有优化

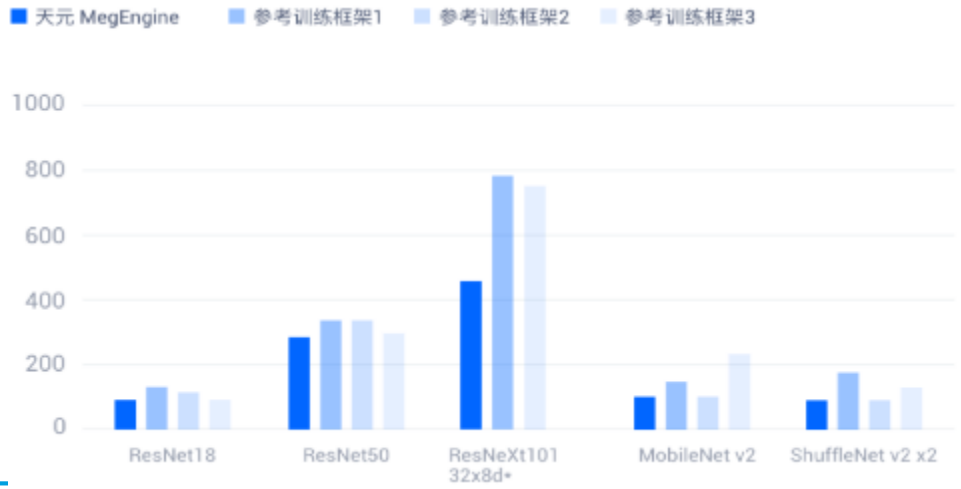
- 2015年开始搭建，针对当年Caffe架构不足，**旷视**Brain++在一开始就确立了要以计算图的方式来进行框架搭建的思路，大思路正确；**2020年3月**开源MegEngine；2020年9月推出Brain++**商业版**。
- **优势**：1）训练推理一体化，训练结果可直接进行用于产品推理、封装。部署时自动删除冗余代码；2）静态图性能高、占用资源少且易于部署、动态图简单灵活、方便调试且易于上手；3）具备Pythonic的API，支持PyTorch Module，直接导入方便；在特定领域如机器视觉模型ResNet 18、ResNet50、MobileNet v2和 ShuffleNet V2上优于其它主流框架。
- **社区反馈问题**：部分支持还不够完善，模型、数据集不够丰富

深度学习云计算平台MegCompute：自主研发的大规模人工智能算力平台

- 提供E级算力资源调度、EB级海量数据存储管理、400G RDMA高速骨干网络。

图：旷视MegEngine 的整体架构

图：旷视MegEngine 和主流开源框架8卡训练耗时（ms/it）对比



资料来源：机器之心，旷视天元官网、申万宏源研究

4.2 旷视：Brain++ AI生产力平台

三类算法：深度学习算法、计算机视觉算法、AIoT算法

- **深度学习算法**：AI基础算法；
- **计算机视觉算法**：大多是使用深度学习算法并结合计算机视觉的具体问题和相关数据所训练出来的（但也有一些传统非深度学习的计算机视觉算法）；
- **AIoT算法**：AIoT和硬件或系统结合更紧密的算法，其可以是基于深度学习的，**也可以是基于其他原理（如最优化）的。**

官网开源14个预训练模型

表：旷视科技开源的预训练模型库

	预训练	领域
ShuffleNet V2	ImageNet 预训练权重	计算机视觉
FReLU	ImageNet 预训练权重	计算机视觉
ResNet	ImageNet 预训练权重	计算机视觉
WeightNet - ShuffleNet V2	ImageNet 预训练权重	计算机视觉
DeepLabV3+	Pascal VOC2012或Cityscapes预训练权重	计算机视觉
BERT for Finetune	BERT	自然语言处理
ATSS	COCO2017预训练权重	计算机视觉
Faster-RCNN	COCO2017预训练权重	计算机视觉
FreeAnchor	COCO2017预训练权重	计算机视觉
RetinaNet	COCO2017预训练权重	计算机视觉
FCOS	COCO2017预训练权重	计算机视觉
SimpleBaseline	COCO 预训练权重	计算机视觉
MSPN	COCO 预训练权重	计算机视觉
Generative Adversarial Networks	生成对抗网络 Cifar10 预训练权重	计算机视觉

表：旷视科技三大类算法

名称	具体分类	简要描述
深度学习算法	云端深度学习算法（ResNet）	ResNet由旷视研究院院长孙剑参与发明，是世界上第一个上百层深度神经网络，开创深度学习领域里程碑。 2015 ：在 ImageNet 大规模图像分类任务上超过人类，ImageNet 以及 COCO 两大学术竞赛中包揽五项冠军。已广泛应用在谷歌 DeepMind 的AlphaGo Zero等学术和工业界 ResNet 根本性解决了多层神经网络训练难题，显著提高精度并降低复杂度。旷视多个云端深度模型都是ResNet改进版本
	移动端深度学习算法（ShuffleNet）	2017 ：高效 ShuffleNet（轻量化卷积神经网络）——大幅降低模型计算复杂度同时保持较高精度。 2018 ：第二代卷积神经网络 ShuffleNet V2——速度与精度大幅提升，目前已获得广泛应用
	边缘端深度学习算法（DorefaNet）	2016 ：第一个梯度量化DorefaNet（低位宽卷积网络）——让CPU、GPU、FPGA 甚至ASIC 上训练神经网络成为可能。 基于 DorefaNet 神经网络广泛应用，完成对CPU、GPU、FPGA 和 ASIC 等全计算平台覆盖。
	自动机器学习技术（AutoML）	自研AutoM技术——充分利用Brain++算力优势，自动帮助算法研究员对深度神经网络的构架进行搜索和参数调优，极大提高算法研究员快速产出最优算法能力。
	其他深度学习算法	应用领域：自监督特征学习、无标签数据自训练、无监督领域自适应、半监督学习、不同粒度下统一的度量学习、长尾数据学习、神经架构搜索、动态卷积、非对称训练、多级模型蒸馏、安全可信的分布式远程训练、高分辨率特征学习等
计算机视觉算法		聚焦图像分类、物体检测、语义分割、视频跟踪、动作/事件识别等核心基础问题。 根据数据和场景的不同，每年持续产出数百个细分算法或算法提升。
AIoT算法		聚焦软硬一体化、多设备协同、大数据分析三方面。 软硬一体化——单体物联网设备有更加智能的感知能力或自主力 多设备协同——多物联网设备高效协同、提升整体效率 大数据分析——大量的物联网设备获取信息并进行分析以支持决策。

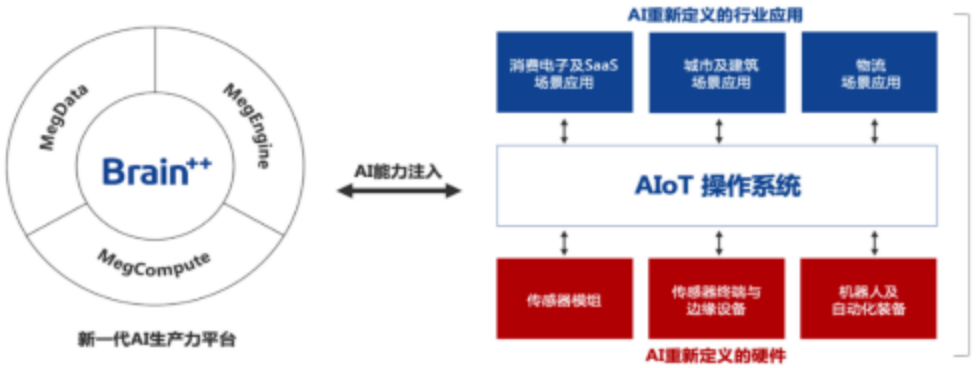
4.2 旷视：平台支撑了跨行业AIoT解决方案

AI平台使得极度碎片化的AIoT具备大规模生产可能性

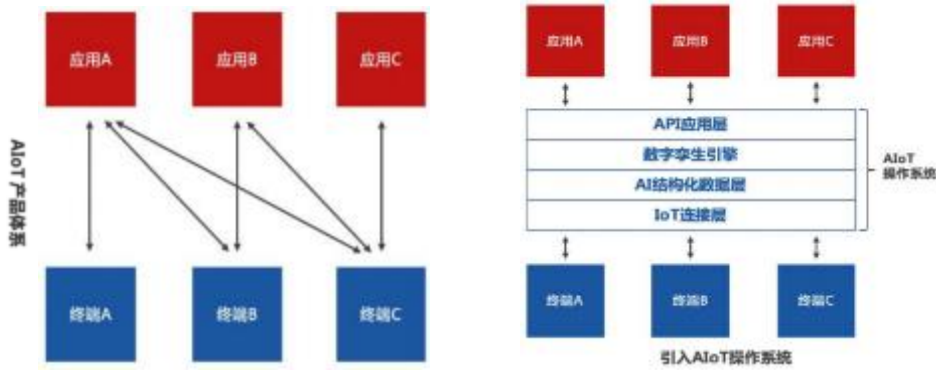
- 当前，物联网更多强调应用、设备之间的直接互联，缺乏智能的感知、分析和协同能力。而 AIoT 操作系统则是在应用和设备之间增加的一个新的操系统层，使其具备连接、分析和协同能力。
- 1）IoT连接层，2）AI结构化数据层，3）数字孪生引擎，在真实空间和数字世界建立双向反馈，4）API应用层，针对不同物联网应用，提供统一开发环境。

在此基础上的城市物联网、供应链物联网AIoT 操作系统

图：旷视Brain++支撑AIoT操作系统



图：从传统物联网到引入AIoT操作系统



表：旷视科技城市物联网平台

名称	简要描述
IoT 设备统一接入、管理和控制层	统一物联网设备抽象模型与接入管理平台：可接入传统网络摄像机、录像存储机、智能摄像机和分析盒等多种旷视科技及第三方物联网设备，并实现统一管理和配置。
视图数据的统一管理、存储和智能结构化层	视图接入数据：统一管理和存储； 云侧硬件算力：统一池化管理调度； 操作系统：内置数据检索、碰撞、聚类等基础大数据应用能力
基于结构化数据的空间数字孪生	可在虚拟的数字化城市、建筑空间中提供离线仿真、在线监控、系统联动和智能运维等功能
API 应用层	支持个性化空间仿真、大数据挖掘和应用，提升内部或第三方团队开发细分领域应用效率

目录

5 本节结论：解决AI落地碎片化必然性的两种路径

碎片化场景可能才是客户付费意愿关键点

- 以占AI上市公司收入重要（达到近50%）的政府场景为例：标准化人脸识别已经无法形成差异化，碎片化场景如：特殊的交通事故、道路塌陷及火灾等，伴随着极具体的要求，客户付费意愿显著提升；

路径分歧：更大的模型或者更低的成本

- **更大的模型路径：**较高软件占比，硬件外采；大规模参数的通用模型，**极高的首次开发成本**；模型长尾投入理想状态接近0；
- **更低的成本路径：**自有生产线压缩**硬件**成本；**小模型、小算力**，较低的首次开发成本；中台复用等方式控制成本。

分析了智慧城市、物流、手机、汽车四个典型行业

两种路径可能适合不同的场景

- **更大的模型适合：**行业需要额外硬件建设较少，下游客户付费能力、需求标准化程度强，产业链已有分工度高；
- 例如：手机、医疗等。
- **更低的成本适合：**行业已有硬件基础差，需要额外硬件建设较多，下游客户付费能力、需求标准化程度低，产业链已有分工度低；
- 例如：智慧城市、智能制造等。
- **目前尚无明确哪种路径更佳：**物流、汽车等

5.1 AI落地和碎片化需求同时出现

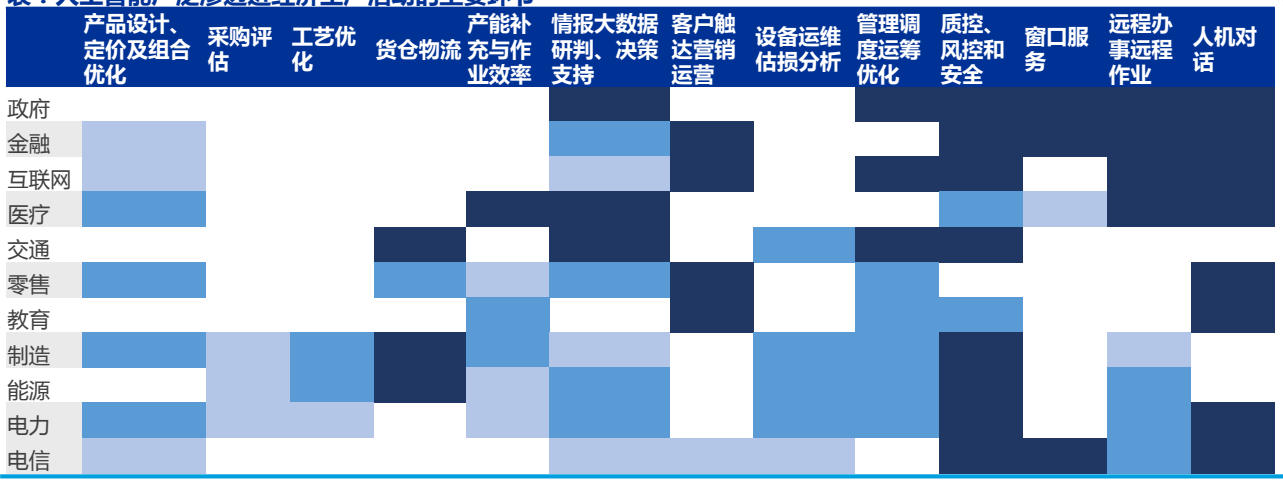
业界在2015-2017年后开始发生变化，AI走出实验室和商业模式转变：

- **工业场景更复杂的逻辑：**编程模式从静态网络结构描述向动态计算过程转变；
- **SDK商业模式可能无法在国内落地：**早期试图复制海外SDK销售路线，但这一模式无法在国内复制；
- 从单纯追求模型精度到平衡：不再不惜代价地追求大模型高精度，而是更多地关注性能和代价的平衡；并开始让AI去解决AI研发过程中的重复劳动

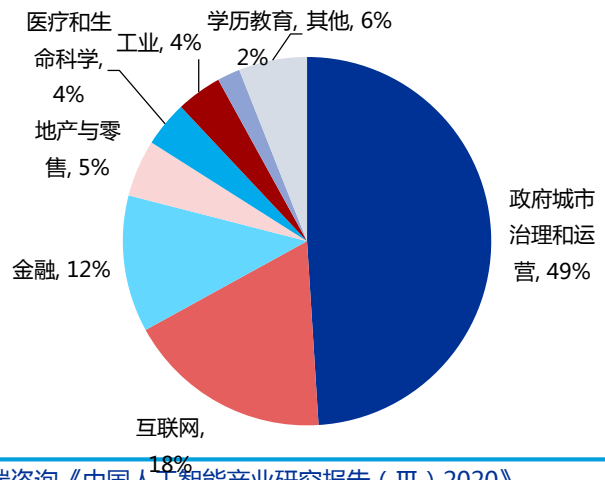
碎片化场景可能才是客户付费意愿关键点

- 以占AI上市公司收入重要（达到近50%）的政府场景为例：标准化人脸识别已经无法形成差异化，碎片化场景如：特殊的交通事故、道路塌陷及火灾等，伴随着极具体的要求，客户付费意愿显著提升；
- 无论是安防还是银行，客户需要的**不是单个模块或开发包，也不具备集成SDK的能力**，而是一套定制化的解决方案。

表：人工智能广泛渗透进经济生产活动的主要环节



图：2020年中国人工智能市场行业份额



AI价值得到验证，推进规模化落地

AI示范项目增加，形成典型应用场景

尝试应用AI

资料来源：艾瑞咨询《中国人工智能产业研究报告（Ⅲ）2020》

5.1 不可避免的，碎片化挑战AI公司盈利能力

工业场景的高性能AI模型生产成本高昂

- 设计并训练工业级的高性能人工智能模型需要大量的成本投入和深厚的技术沉淀，包含(i)多场景汇总的海量数据，(ii)复杂的模型设计和训练算法，以及(iii)包括复杂的软件框架和硬件系统在内的大型计算基础设施来支持大规模计算。

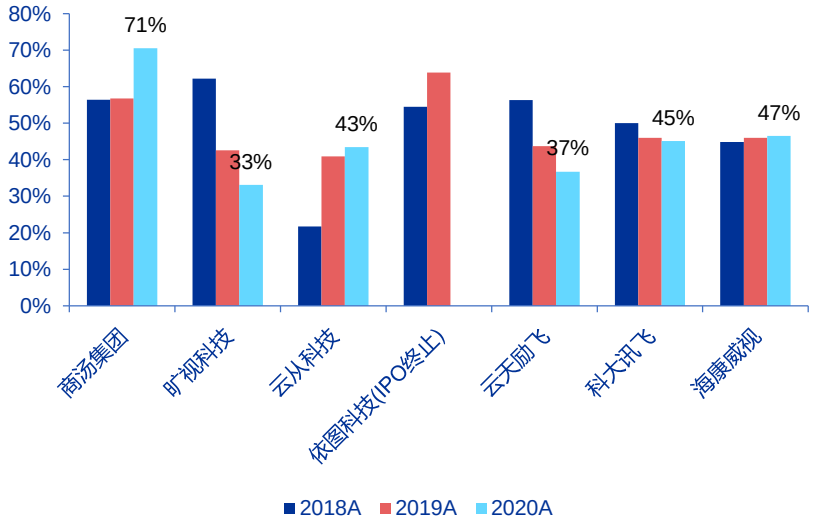
AI模型开发效率低，且开发出现人力密集趋势

- 长尾需求数量庞大，单个场景发生频次低、可用数据少。且由于每种模型的生产都需要大量的算力及人力，因此行业呈现出人力投入增加和资源密集的趋势。

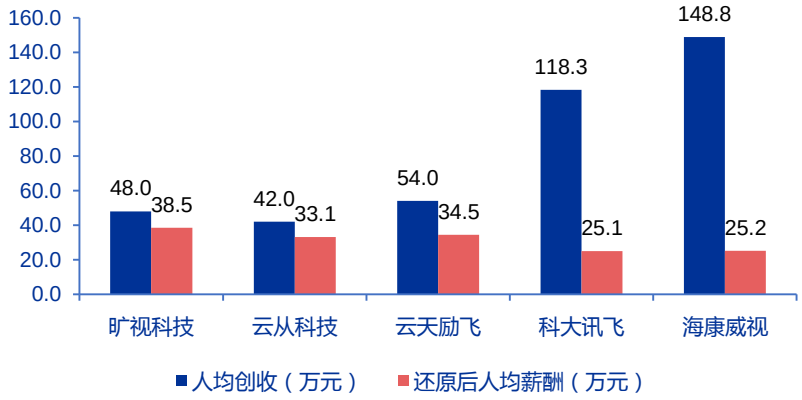
实际AI商业模式毛利率并不很高

而AI行业所需模型和数据人力成本远高于其它传统信息化行业

图：2018-2020AI软件企业毛利率



图：2020年AI软件企业人均创收与还原后人均薪酬情况



注：当期还原后薪酬=当期支付给职工以及为职工支付的现金+当期应付职工薪酬-上一期应付职工薪酬

资料来源：Wind，申万宏源研究

5.1 更大的模型还是更低的碎片成本？两种路径

- 较高软件占比，硬件外采
- 大规模参数的通用模型，极高的首次开发成本
- 模型长尾投入理想状态接近0

- 自有生产线压缩硬件成本
- 小模型、小算力，较低的首次开发成本
- 中台复用等方式控制成本

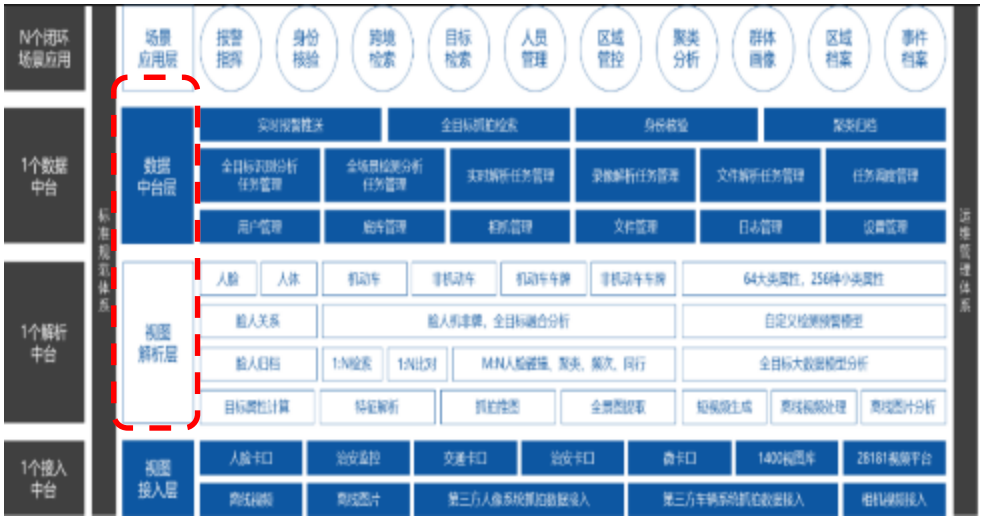
- 案例
  - 智慧城市 商汤、旷视 VS 云从、海康
  - 物流 旷视 VS 大华
  - 手机 商汤、旷视 VS 虹软、汇顶
  - 汽车 商汤 VS 西威、创达

以商汤、旷视为例，自研深度学习平台，是否可以成为AI企业取得更好盈利的路径？

5.2 智慧城市：AI企业标配行业，但实现路径不同

图：商汤“方舟”城市开放平台

图：旷视“昆仑” 城市物联网平台



图：云从智慧社区解决方案

图：海康云-边-端一体化方案

资料来源：商汤集团招股书、旷视科技招股书、云从科技招股书、海康威视年报，申万宏源研究

5.2 智慧城市：AI企业标配行业，但实现路径不同

算法和平台能力

商汤“方舟”城市开放平台

- **模型能力：**内含14,000多个人工智能模型；
- **底层AI平台支持：**SenseCore在线增量训练引擎，提供AI-as-a-Service；
- **结构化数据分析：**将原始的城市数据实时转化成运营洞察、事件警报及管理行动；
- 由人力密集型向人机交互型、由经验导向型向数据驱动型、由被动处置型向主动发现型转变。

旷视“昆仑” 城市物联网平台

- **视觉感知和数据智能**
- **设备统一接入：**将城市中各类视图传感设备统一接入管理、点位规划、智能运维；
- **全目标解析和数据中台：**接入视图全目标要素提取，统一数据中台支持提供全目标数据关联、归档、挖掘、研判、检索、预警等全场景业务原子级数据应用，满足城市管理各场景组合应用

软件方案

硬件物联

云从智慧社区解决方案

- **全栈能力：**人脸识别、车辆识别、OCR、人脸聚类、可视化建模、知识图谱、大数据分析等技术，
- **除视觉外其它信息源：**面向用户提供泛感知数据采集能力和多种社区数字化治理模型；
- **操作系统强调人机协同；**
- **更细化的客户聚类方案：**智慧治理领域，除智慧社区外，文旅、应急、检察院、法院、政法、公安、环保、卫健委、教委等，分别准备不同解决方案。

海康云-边-端一体化方案

- **PBG业务的云边融合、物信融合**
- **完备的硬件产品线：**截至2021年超过万种硬件SKU，并且仍在不断迭代
- **碎片化场景处理：**在解决方案层面，从大行业到细分领域，都有个性化的定制方案；
- **统一软件架构积累模块：**依托统一软件架构平台，提升软件复用率。

全栈解决方案

5.3 物流：传统企业方案包含大量自研硬件

大华智慧物流优势：

- **除了AI视觉外，传统RFID/工业面阵相机/OCR等传统机器视觉能力也有大量积累，**有单独的机器视觉子公司华睿支持；

通过软硬件一体化方案控制成本：

- **方案宣传亮点：强调客户ROI角度投入产出，节约的人力等成本；**
- **以某纺织行业项目为例：**
- 在印染布转运劳动强度大，人工成本及库存成本高的问题场景下，公司通过部署34台防水式地牛AGV，基于AGV智能调度平台与MES对接，完成了三个车间印染布室内外混合智能转运；
- 减少了约**54人工成本和2/3库存**。同时，高端制造应用拉动了客户的订单增长，**每年为客户创收超过500万元。**

表：大华智慧物流一体化方案

环节	方案	核心产品	形式	研发	功能描述	图示
卸车	自动抓取包裹方案	双目红外3D相机	硬件	自研	采集视野范围内包裹的**3D点云数据**，输出各个包裹的位置信息，并联动机械臂**快速抓取包裹并放置到传输线上**	
分拣	单件分离方案	双目红外3D相机	硬件	自研	**红外相机**稳定性高、抗环境光干扰；**高精度包裹轮廓识别算法**，支持软包、纸箱等包裹类型；左右、前后、高低、密集等多种包裹粘连状态的自动分离；最大支持6000件/小时；	
		高精度包裹轮廓识别算法	软件	自研	**自动对包裹进行整位、分离、智能排队**，将批量无序包裹变成单件排列，形成整齐“阵型”，联动自动分拣设备实现包裹全自动传输和分拣	
	DWS六面读码方案	高分辨率工业相机及线扫读码模块	硬件	自研	全方位、无死角的动态读码作业，分拣效率高达**3600件/小时**；DWS系统实时**对包裹信息进行智能追溯**，并对接到ERP系统或服务器	
搬运	搬运机器人	搬运型AGV（地牛）	硬件	自研	采用激光SLAM导航，通过栈板识别、测重、测偏，可在小于2m的通道内自由取放货，定位精度达±10mm，**适用于各种仓库、产线等标准川字形栈板搬运**	
	牵引机器人	200kg牵引型AGV	硬件	自研	根据料架形式实现自动脱挂料架，依托二维码导航实现AGV与料架同步弧度转弯及无缝对接MES系统	

资料来源：大华股份官网整理，申万宏源研究

5.3 物流：AI企业更强调机器视觉和AI软件能力

旷视：智慧物流操作系统河图

- **强调AI视觉作用：**将AI视觉技术应用于货品从入库、出库到收货的各个环节，获取其文字、条码、位置、破损、缺陷等各方面信息；
- **“河图”操作系统：**向下连接硬件（可以是自研或者第三方的），向上连接企业ERP/MES等系统，内置设备调度、路径优化、业务调度、工作站任务动态分配等一系列算法。

AIoT能力亮点：全连接、大规模

- **统一调度能力，最大支持半径**
- **以某全球500强中领先的鞋服类企业客户为例：**协助完成首个物流配送中心的建设，包含700余台移动机器人在内的10类近4,000台智慧物流装备。项目建成后，物流中心拣选出库效率每天可达**40万件，支持200多家门店业务。**
- **尽管AI软件企业强调对不同类型不同外部厂商的硬件连接能力，但猜测实际方案中仍然需要进行大量硬件设备前期外采**

表：旷视智慧供应链解决方案

环节	方案	核心产品	形式	研发	功能描述	图示
仓储	智能仓储系统	AS/RS Miniload Multi-shuttle Pallet-shuttle 货到人	软件	自研	仓储环节提供的核心设备体系，用AI技术替代传统技术； 兼容多种柔性存储设备和复杂密集存储设备，通过算力算法来提高存储效率和出入库效率； 面向仓储环节涵盖多子系统存储能力可达普通仓3-5倍以上	
运输	智能输送系统	托盘输送系统 箱输送系统	软件	自研	解决“长距离搬运”，AI感知提升系统输送能力、调度逻辑以及系统监测效率 用AI技术部分替代物理的或光电的技术，实现无盲区的感知 智能化和柔性化布局能力	
分拣	智能分拣系统	AI+分拣机 智能圆形播种机	AI嵌入硬件	软件自研、硬件可能外包	视觉技术：对货物跟踪、计数，分拣准确率可达99.99%。本体采用交叉带分拣机，稳定可靠 高ROI：相同的占地面积内分拣口多；配合河图流程拣选技术，可实现出库人力减半	
机器人	AGV	T系列潜伏式二维码AGV、F系列托盘堆垛叉车AVG等	硬件	自研	AI算法赋能机器人实现“环境感知、智能决策、智能执行”三位一体的柔性化与智能化	
	AMR	S系列背负式AMR				

资料来源：旷视科技官网整理、申万宏源研究

表：大华和旷视部分业务毛利率对比

	2018	2019	2020
旷视-供应链物联网解决方案	12.85%	41.34%	5.76%
大华-国内2B业务	-	41.74%	41.77%

资料来源：公司公告、申万宏源研究
注：旷视2020年客户公司E的整仓项目收入按进度确认，导致其毛利率为-14.80%。扣除后供应链物联网解决方案业务毛利率为38.55%。

5.4 手机：高度分工标准件SDK收费最理想场景之一

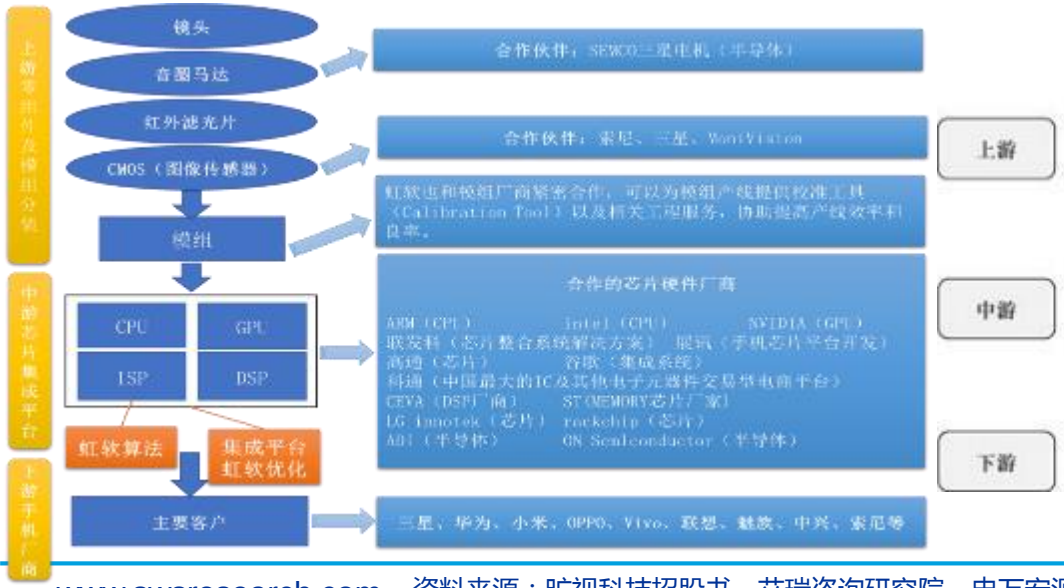
视觉算法在手机中主要应用

- **1、为图像处理器ISP服务，**对于传感器传回的图像进行算法处理，使其能够在光线较差，或者移动中拍摄出画质更好的照片。
- 2、与高通、联发科、展讯等各主流移动芯片公司**针对CPU、GPU、DSP做优化**，或者AI硬件平台（NPU/APU等）做优化。

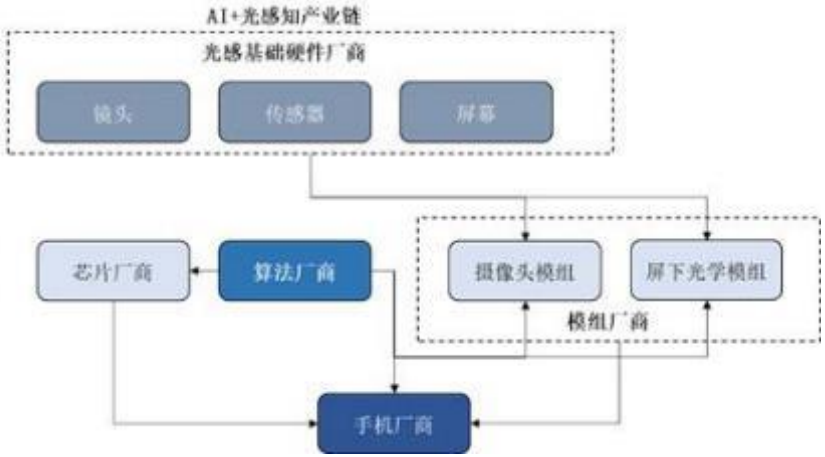
手机行业特点是产业链高度分工，软件硬件也分工明确

- **上游是零组件：**镜音圈马达、红外滤光片、图像传感器。**下游为手机厂商。 AI软件企业位于中游。**
- 即使对于不同手机厂商，**所需的计算摄影方法也有相似性，**不需要大量定制化二开。

图：以虹软科技为例，所在手机产业链位置上下游分工明确



图：AI+光感知产业链中游模组、芯片、算法厂商分工明确



5.4 手机：高度分工标准件SDK收费最理想场景之一

行业较好的收费模式可从财务指标验证

- **收入体量接近**：在3-6亿元规模；
- **毛利率在80%以上**：以SDK或纯软件收入为主

然而方案较为相似，三家企业直接竞争

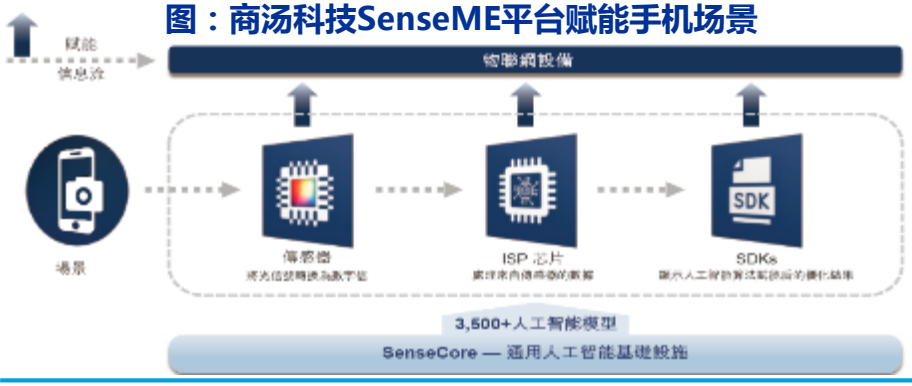
- **旷视**：计算摄影及视频处理解决方案使消费者用移动智能终端拍摄出高质高清的相片及视频。**人脸识别领域，与虹软、商汤形成直接竞争。**
- **虹软**：图像和协助厂商在既有的摄像头硬件能力基础上全面提升摄像头的成像质量。
- **商汤**：SenseME软件平台由超过3,500个AI模型组成，ISP芯片的输出结果经过SDK处理提供感知智能和内容增强。

表：三家企业消费AI收入和毛利率对比

	2018		2019		2020	
	收入（百万元）	毛利率（%）	收入（百万元）	毛利率（%）	收入（百万元）	毛利率（%）
旷视科技 消费物联解决方案	265.11	81.32	358.34	81.68	256.84	81.26
商汤集团 智慧生活	330.30	-	413.50	-	433.90	-
虹软科技 智能手机视觉解决方案	438.95	-	543.32	94.82	599.02	94.93

图：以虹软科智能手机拍摄解决方案

图：商汤科技SenseME平台赋能手机场景



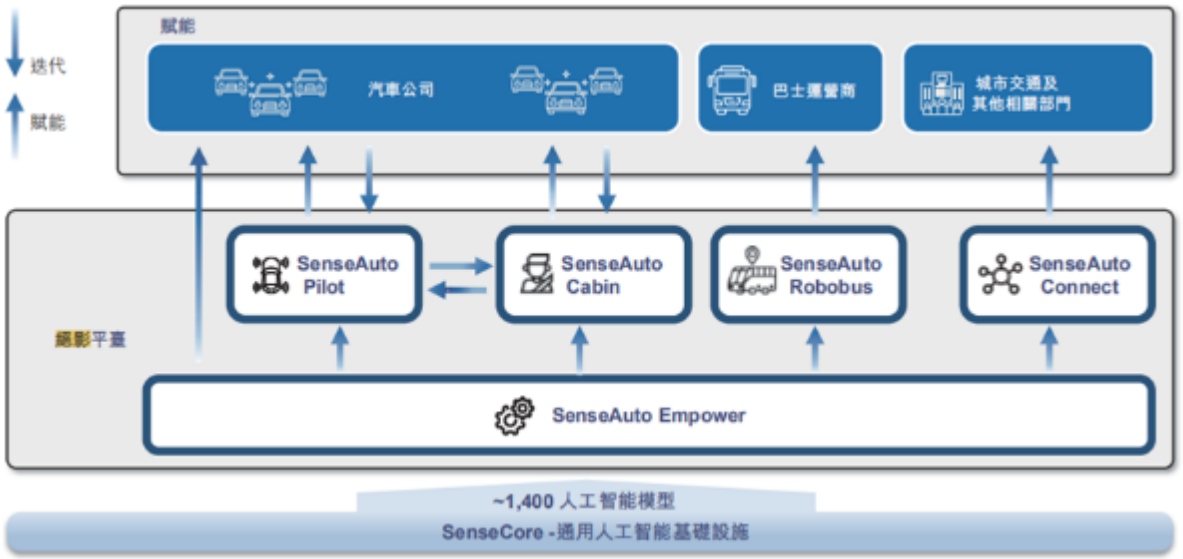
图：以虹软科智能手机拍摄解决方案

5.5 汽车：视觉优势应用于驾驶座舱和自动驾驶软件

四小龙中仅商汤大规模布局汽车IT，且目前布局较为全面

- **1）SenseAuto Pilot：**支持ADAS产品、L4自动驾驶计划及SenseAuto Connect产品，正为传统及新能源车企开发L2+ADAS产品；
- **2）SenseAuto Cabin：**DMS、OMS及车载信息娱乐系统；
- **3）SenseAuto Empower：**约1,400多个人工智能模型，AIDC大算力平台；
- **4）SenseAuto Robobus：**L4 Auto Robobus，巴士自动班车，商业园区、旅游景点、试验区应用。
- **从驾驶座舱到自动驾驶L2-L4，从软件平台、算法，到云端算力支持基本布局完成。**
- **商汤目前和30多家车厂建立合作关系，生态合作伙伴达到了50+，定点车数量已经超过了2000万辆**

图：商汤科技智能汽车绝影平台



资料来源：商汤科技招股书、申万宏源研究

5.5 汽车：驾驶座舱商汤虹软科技等直接竞争

DMS产品基本高度同质化

- 标准方案包括：驾驶员身份识别、睡意检测、专注度检测、缺席检测及异常情况检测等；除商汤、虹软外，未动科技、地平线等**未上市公司大量同质化竞争**；

虹软科技：开始向前装拓展，DMS定点出货预计2022后加快

- 与高通、联发科、瑞萨等合作。前装定点：长城、长安新能源、上汽、理想、一汽、东风、合众多款量产车型。基于高通Qualcomm的长城智能座舱定点项目21Q3量产；

商汤科技：特色是OMS，可能外增加30%单车价值量

- 商汤OMS儿童检测产品已于2020年在客户车型上使用，是中国市场上首个同类产品。
- OMS将在DMS的基础上增加40-80元左右。新增额外30%左右的市场规模；

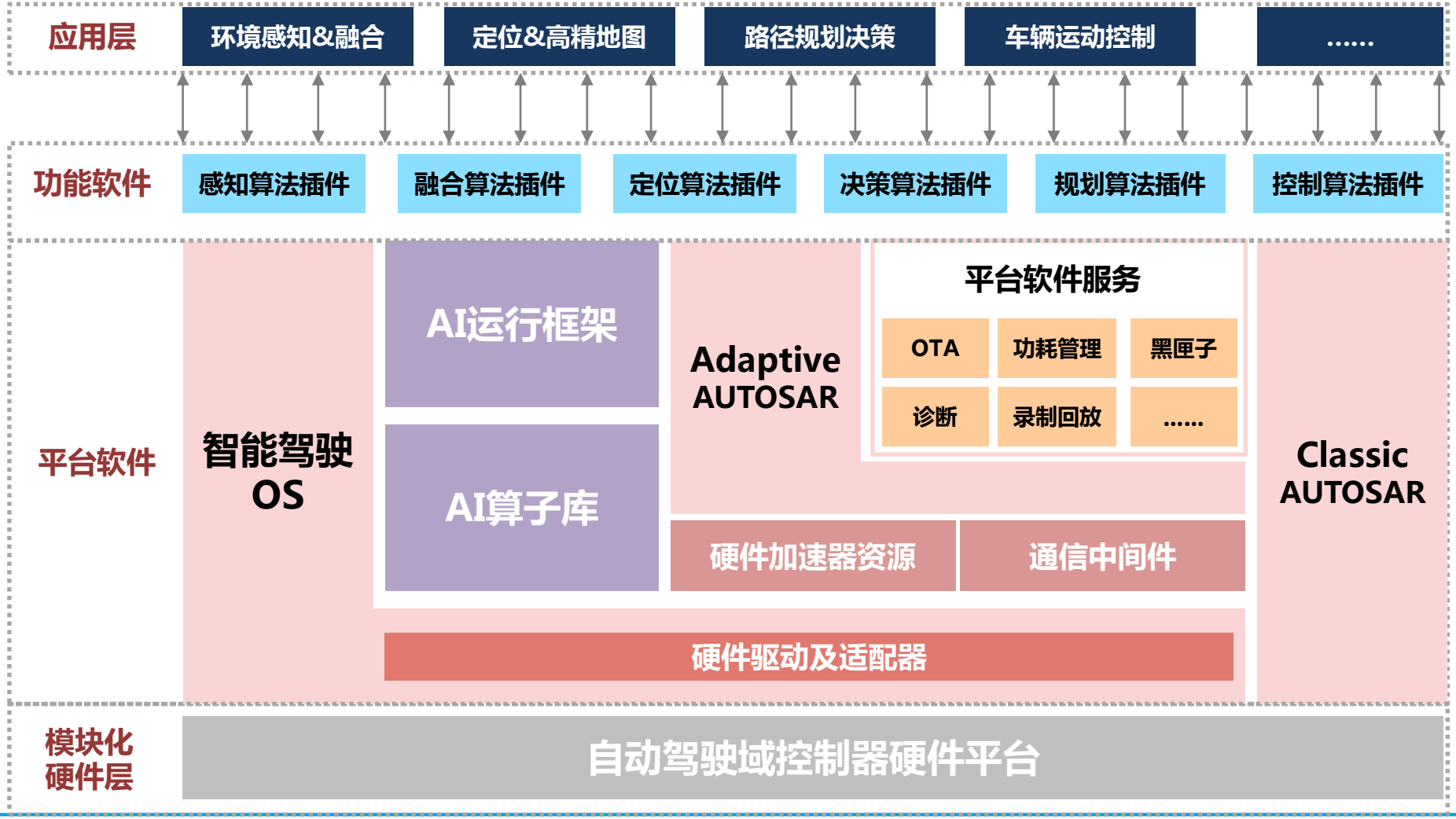
图：高工智能汽车2020年度DMS国产供应商TOP10

图：虹软科技DMS产品线

图：商汤科技智能驾驶座舱产品线

5.5 汽车：自动驾驶探索应用层机会

- **SenseAuto Pilot**：商汤科技主要做**应用层**。定位于**高速场景较多**。



5.5 汽车：自动驾驶探索应用层机会

SenseAuto Pilot

- **基于视觉的高性价比系统**：检测200米以内的车辆、150米以内的行人。基于感知摄像头DVR；
- **多传感器融合系统**： 更广视角及高清特征的多传感器融合系统。融合包括激光雷达在内的传感器。
- **L2+ ADAS全栈产品**：预装商汤L2+ ADAS产品的车型预计将在**2022年**实现量产；
- **功能**，自适应巡航控制(ACC)、车道居中控制(LCC)、交通拥堵辅助(TJA)及自动领航辅助驾驶(NoP)

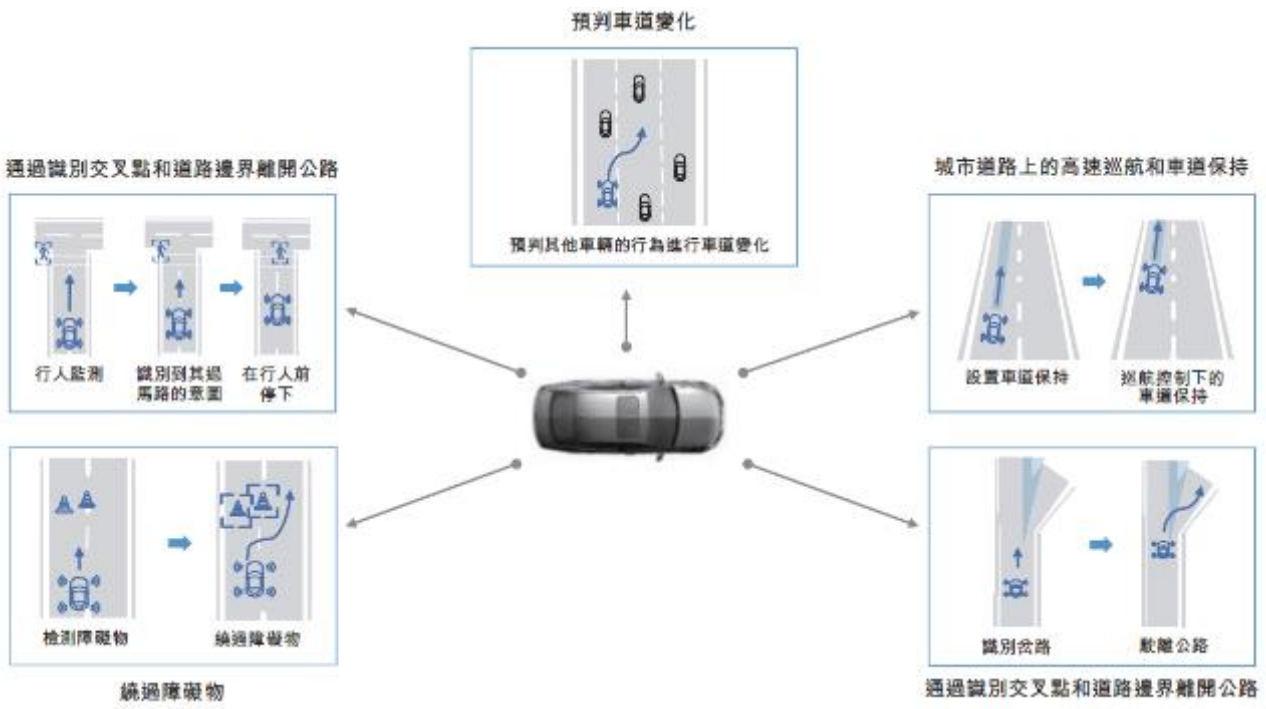
SenseAuto Empower

- AIaaS：为车企提供约**1,400多**个人工智能模型，或AI**联合研发**服务；AIDC基础设施建设和数据服务能力

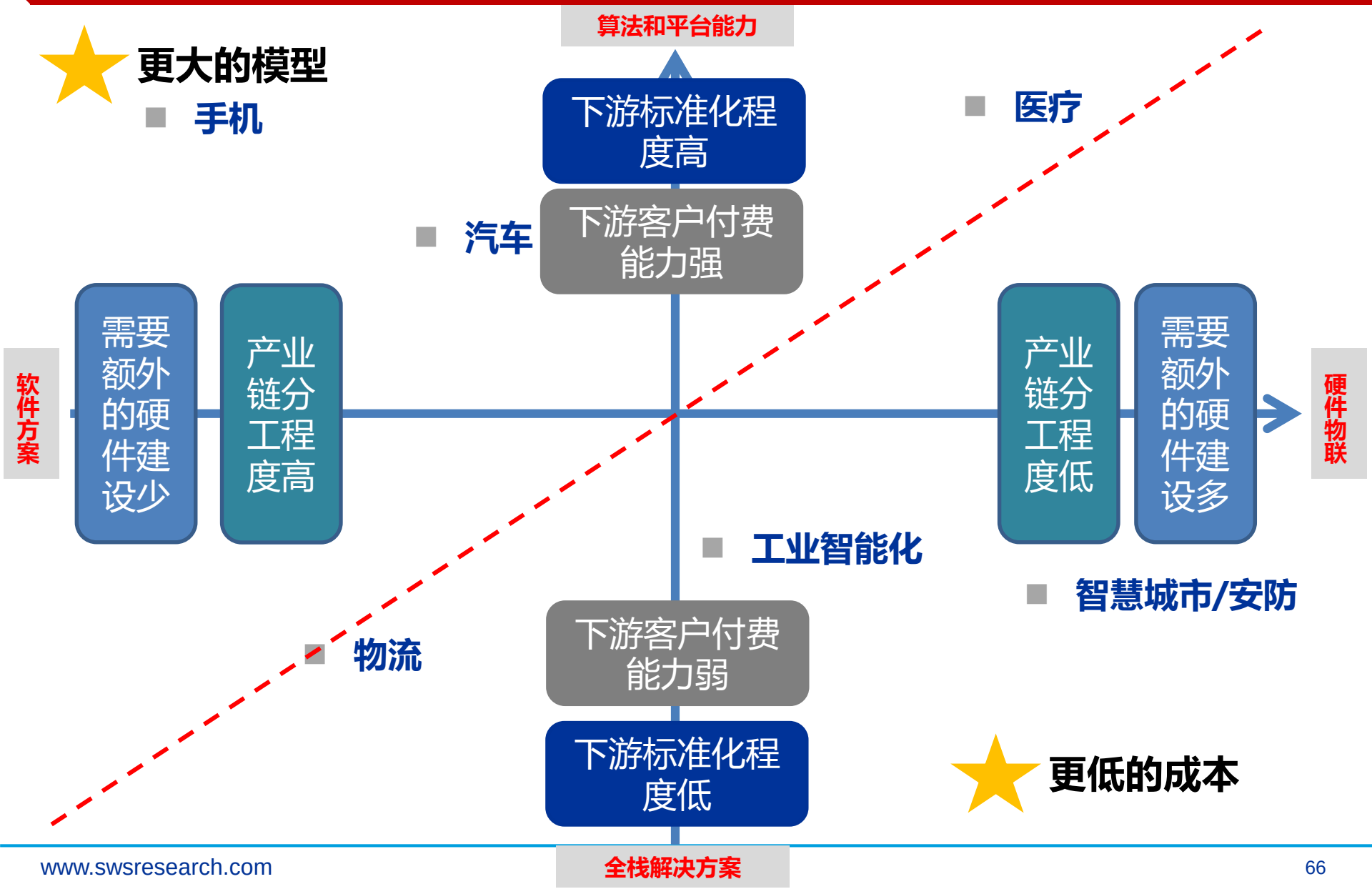
SenseAuto Robobus

- L4自动驾驶用于巴士自动班车。商业园区、旅游景点、自动驾驶试验区。

图：SenseAuto Pilot ADAS系统所面临的常见场景



5.6 两种路径的总结


更大的模型
手机
算法和平台能力
下游标准化程度高
医疗
汽车
下游客户付费能力强
软件方案
需要额外的硬件建设少
产业链分工程度高
产业链分工程度低
需要额外的硬件建设多
硬件物联
工业智能化
智慧城市/安防
物流
下游客户付费能力弱
下游标准化程度低
更低的成本
全栈解决方案

5.6 AI相关标的和风险提示

表：AI行业重点公司估值表

证券代码	证券简称	2021/12/10		PB	申万预测EPS				PE		
		收盘价(元)	总市值（亿元）	2020A	2020A	2021E	2022E	2023E	2021E	2022E	2023E
002230.SZ	科大讯飞	53.8	1238	7.26	0.61	0.80	1.09	1.49	67	49	36
688088.SH	虹软科技	46.53	189	10.64	0.62	0.51	0.63	0.94	91	74	50
002415.SZ	海康威视	53.5	4995	9.48	1.44	1.83	2.27	2.76	29	24	19
002236.SZ	大华股份	25.62	767	3.22	1.33	1.51	1.74	2.13	17	15	12
688256.SH	寒武纪-U	102	408	8.97	-1.15	-1.95	-2.02	-1.52	-	-	-

资料来源：Wind资讯、申万宏源研究
注：寒武纪使用Wind一致预期EPS

核心风险假设

- **1、深度学习技术发展瓶颈风险：**
- 参与深度学习框架建设的主要为海外互联网巨头，国内AI公司投入研发力量可能仍有差距，在开源框架建设上后发劣势；同时机器视觉、NLP等算法仍然存在算法逻辑不明确等问题，可能存在技术发展的瓶颈；
- **2、在业务落地上大量同质化竞争，高薪酬导致成本降低难度大：**
- 在安防、金融等传统AI落地场景上，商汤、旷视等AI企业与海康、大华等传统企业存在同质化竞争，解决方案的差异化有限。同时AI人才仍然薪酬较高，若前期在算法研发、算力储备等投入过大，可能导致成本长期难以下降，企业盈利能力受限；
- **3、海外出口与清单限制风险：**
- 此前美国财政部将商汤加入“中国军工符合体企业”清单，同时海康、大华也存在类似贸易清单限制，可能对以上企业的长期海外化等发展产生影响。

信息披露

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告,并对本报告的内容和观点负责。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

与公司有关的信息披露

本公司隶属于申万宏源证券有限公司。本公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司关联机构在法律许可情况下可能持有或交易本报告提到的投资标的，还可能为或争取为这些标的提供投资银行服务。本公司在知晓范围内依法合规地履行披露义务。客户可通过compliance@swsresearch.com索取有关披露资料或登录www.swsresearch.com信息披露栏目查询从业人员资质情况、静默期安排及其他有关的信息披露。

机构销售团队联系人

华东A组	陈陶	021-33388362	chentao1@swhysc.com
华东B组	谢文霓	021-33388300	xiewenni@swhysc.com
华北组	李丹	010-66500631	lidan4@swhysc.com
华南组	陈左茜	0755-23832751	chenzuoxi@swhysc.com

A股投资评级说明

证券的投资评级：

以报告日后的6个月内，证券相对于市场基准指数的涨跌幅为标准，定义如下：

买入（Buy）	：相对强于市场表现20%以上；
增持（Outperform）	：相对强于市场表现5%～20%；
中性（Neutral）	：相对市场表现在－5%～＋5%之间波动；
减持（Underperform）	：相对弱于市场表现5%以下。

行业的投资评级：

以报告日后的6个月内，行业相对于市场基准指数的涨跌幅为标准，定义如下：

看好（Overweight）	：行业超越整体市场表现；
中性（Neutral）	：行业与整体市场表现基本持平；
看淡（Underweight）	：行业弱于整体市场表现。
本报告采用的基准指数	：沪深300指数

港股投资评级说明

证券的投资评级：

以报告日后的6个月内，证券相对于市场基准指数的涨跌幅为标准，定义如下：

买入（BUY）：	：股价预计将上涨20%以上；
增持（Outperform）	：股价预计将上涨10-20%；
持有（Hold）	：股价变动幅度预计在-10%和+10%之间；
减持（Underperform）	：股价预计将下跌10-20%；
卖出（SELL）	：股价预计将下跌20%以上。

行业的投资评级：

以报告日后的6个月内，行业相对于市场基准指数的涨跌幅为标准，定义如下：

看好（Overweight）	：行业超越整体市场表现；
中性（Neutral）	：行业与整体市场表现基本持平；
看淡（Underweight）	：行业弱于整体市场表现。
本报告采用的基准指数	：恒生中国企业指数（HSCEI）

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。申银万国使用自己的行业分类体系，如果您对我们的行业分类有兴趣，可以向我们的销售员索取。

法律声明

本报告由上海申银万国证券研究所有限公司（隶属于申万宏源证券有限公司，以下简称“本公司”）在中华人民共和国内地（香港、澳门、台湾除外）发布，仅供本公司的客户（包括合格的境外机构投资者等合法合规的客户）使用。本公司不会因接收人收到本报告而视其为客户。有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司http://www.swsresearch.com网站刊载的完整报告为准，本公司并接受客户的后续问询。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为作出投资决策的惟一因素。客户应自主作出投资决策并自行承担投资风险。本公司特别提示,本公司不会与任何客户以任何形式分享证券投资收益或分担证券投资损失，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。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。市场有风险，投资需谨慎。若本报告的接收人非本公司的客户，应在基于本报告作出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。

简单金融•成就梦想

AVirtueofSimpleFinance

申万宏源研究微信订阅号

申万宏源研究微信服务号

上海申银万国证券研究所有限公司
(隶属于申万宏源证券有限公司)

洪依真
hongyz@swsresearch.com