浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

华为AI盘古大模型研究框架

华为产业链深度系列研究

行业评级：看好

2023年3月25日

分析师	陈杭	分析师	刘雯蜀	研究助理	安子超
邮箱	chenhang@stocke.com.cn	邮箱	liuwenshu03@stock.com.cn	邮箱	anzichao@stocke.com.cn
证书编号	S1230522110004	证书编号	S1230523020002	电话	18611396466

盘古大模型金字塔架构，算力为底层支撑

资料来源：浙商证券研究所

摘要

目前我们将迎来科技的重大转折点：ChatGPT时刻。而在ChatGPT背后，不断迭代的GPT系列使得大模型成为当下科技企业核心竞争力的重要体现，未来，大模型将成为AIGC时代的核心支撑。华为作为国内科技龙头，2021年发布的盘古大模型有望在AIGC时代中引领潮流。我们将从：昇腾/鲲鹏→MindSpore AI框架→ModelArts→盘古大模型四层架构进行分析：

1、AI算力资源：“鲲鹏+昇腾”，打造盘古算力底座

- 鲲鹏：华为自主芯片→鲲鹏芯片→鲲鹏服务器→欧拉操作系统→高斯数据库→行业应用向外扩张，构建鲲鹏生态，提供算力支撑。
- 昇腾：昇腾AI处理器→CANN异构计算架构→MindSpore AI框架→应用使能→行业应用，助力打造华为昇腾全栈AI软硬件平台，构筑智能世界基石。

2、人工智能框架： MindSpore高效易开发，可实现全场景覆盖

- CANN：作为华为昇腾AI基础软硬件平台的核心，CANN向上支持多种AI框架，向下服务AI处理器与编程，助力芯片使能。
- MindSpore：是国内首个支持千亿参数大模型训练AI计算框架，最佳匹配昇腾处理器算力，支持终端、边缘、云全场景灵活部署，开创全新的AI编程范式，降低AI开发门槛。

3、AI开发平台：ModelArts强势赋能开发者，精度效率双提升

- 为机器学习与深度学习提供海量数据预处理及交互式智能标注、大规模分布式训练、自动化模型生成，及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期AI工作流。

4、盘古大模型：AI落地的重要途径

- 由NLP大模型、CV大模型、多模态大模型、科学计算大模型等多个大模型构成，目前已实现等AI场景落地。

建议关注标的：神州数码、拓维信息、麒麟信安、软通动力、常山北明、海量数据、润和软件

数据来源：华为云官网，浙商证券研究所

风险提示

1、宏观经济下行风险
2、上游晶圆紧缺加剧
3、市场发展不及预期
4、行业竞争风险

目录

CONTENTS

01 AI算力资源

鲲鹏服务器助力满足澎湃算力需求

昇腾全栈AI软硬件平台构筑智能世界基石

02 人工智能框架

CANN-AI异构计算架构芯片使能

MindSpore智能适配盘古大模型

03 AI开发平台

ModelArts强势赋能开发者

落地场景可覆盖完整产业链

04 盘古大模型

NLP大模型　　多模态大模型

CV大模型　　科学计算大模型

01 算力资源

鲲鹏 昇腾

华为鲲鹏生态：华为自主芯片→鲲鹏芯片→鲲鹏服务器→欧拉操作系统→高斯数据库→行业应用

1、鲲鹏芯片：鲲鹏920作为低功耗、高性能的Arm处理器，为鲲鹏服务器主板及整机产品提供芯片支撑，是鲲鹏生态发展壮大的核心所在，在此基础上，华为进一步开启自主研发芯片，为鲲鹏生态发展奠定坚实基础。

2、鲲鹏服务器：华为凭借多年积累的硬件工程能力，打造TaiShan服务器，使能整个产业链，进一步构建完整鲲鹏生态。

3、欧拉操作系统：作为面向B端的电脑服务器操作系统，华为自主研发的EulerOS，以Linux稳定系统内核为基础，南向支持多样性设备，北向覆盖全场景应用，横向对接鸿蒙，通过能力共享实现生态互通。

4、高斯数据库：华为GaussDB是主打政企核心业务负载的金融级分布式数据库，目前已实现助力部分保险及车企数字化转型。

5、行业应用：华为以行业聚合应用，通过平台和生态双轮驱动，形成行业应用矩阵，为众多行业客户提供解决方案。并陆续成立五大军团，不断开发全新应用场景。

华为昇腾AI产业：昇腾AI处理器→CANN异构计算架构→MindSpore AI框架→应用使能→行业应用

1、Ascend：昇腾AI处理器作为基础，通过模块、标卡、小站、服务器等丰富的产品形态，打造面向“端、边、云“的全栈解决方案，为整个昇腾AI产业的底层核心支撑。

2、CANN：作为华为昇腾AI基础软硬件平台的核心，CANN向上支持多种AI框架，向下服务AI处理器与编程，以极致性能、极简开发、开放生态为目标，助力昇腾构建全场景人工智能平台。

3、MindSpore：是国内首个支持千亿参数大模型训练AI计算框架，覆盖包含生物医学在内的多个领域。

4、应用使能：以昇腾AI处理器→CANN异构计算架构→MindSpore AI框架的传导机制，为深度学习、智能边缘以及行业应用解决方案等强势赋能。

数据来源： 昇思官网，浙商证券研究所

算力布局 = 鲲鹏+昇腾

鲲鹏：最强算力异构计算服务器

行业应用 金融 运营商 ……

交通 政府 电力 医疗

中间件 DDM

数据库 Gauss DB

操作系统 To B：欧拉

To C：鸿蒙

服务器 Taishan 200

Taishan 100

芯片
自主化
鲲鹏920

算力布局

盘古大模型的底层算力支撑：昇腾

行业应用 金融 能源 ……

制造 电信 公共 交通

第三方平台 Model Arts

全场景AI框架

MindSpore

DL Edge SDK MindX

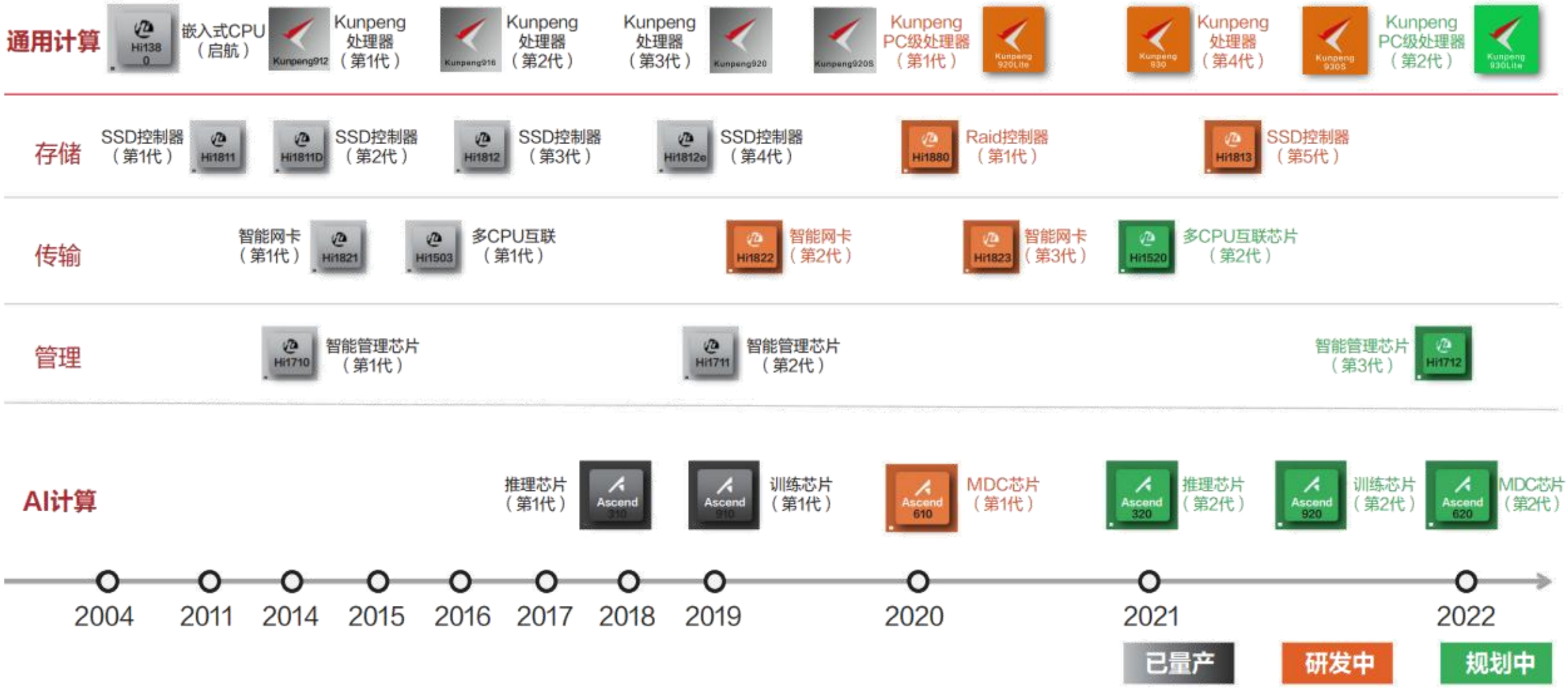
芯片使能 CANN

驱动层

Atlas系列
AI训练卡
AI推理卡
训练服务器

数据来源： 昇思官网，浙商证券研究所

Kunpeng处理器支持鲲鹏计算产业长期演进



数据来源： 鲲鹏计算产业发展白皮书，浙商证券研究所

硬件：鲲鹏服务器实现数据中心高效能计算

鲲鹏服务器

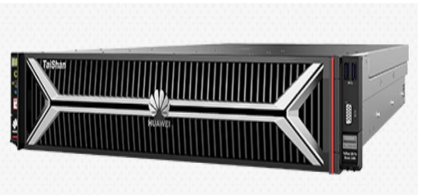
TaiShan100

- 包含2280均衡型和5280存储型等产品型号。
- 基于鲲鹏916处理器的数据中心服务器，具有多核高并发、低功耗等计算优势，适合为大数据、分布式存储等应用高效加速。

TaiShan200

- 包含2280E边缘型、1280高密型、2280均衡型、2480高性能型、5280存储型和X6000高密型等产品型号。
- 基于华为鲲鹏920处理器，旨在满足数据中心多样性计算需求。

TaiShan200 Pro



- 包含2480、2280和1280等三款高端产品型号。
- 基于鲲鹏920 3.0GHz高主频处理器，同时集成三大创新RAS特性，获得权威安全可信认证。

高效能计算

- 搭载具有超强算力的鲲鹏处理器
- 多核计算架构
- 高效加速应用

安全可靠

- 处理器及服务器芯片全自研
- 17年计算工程能力铸就稳如泰山品质

开放生态

- 开放计算平台
- 支持业界主流软件
- 携手合作伙伴，共赢计算新生态

数据来源： 鲲鹏社区官网，浙商证券研究所

澎湃算力携手合作伙伴，共同拓展鲲鹏计算产业

数据来源： 鲲鹏社区官网，浙商证券研究所

昇腾AI底层支撑系列产品

AI模块

开发者套件

芯片：昇腾310

最高算力：22 TOPS

AI加速模块

芯片：昇腾310

最高算力：22 TOPS

加速卡

推理卡

芯片：昇腾310

最高算力：88 TOPS

训练卡

芯片：昇腾910

最高算力：280 TFLOPS

智能边缘

智能小站

芯片：昇腾310

最高算力：22 TOPS

边缘服务器

芯片：鲲鹏920

最高算力：352 TOPS

AI服务器

推理服务器

2*鲲鹏920

最高算力：704 TOPS

训练服务器

8*昇腾910+4*鲲鹏920

最高算力：2.24 PFLOPS

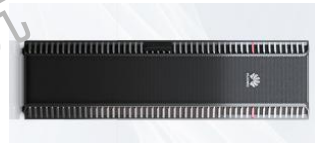
AI集群

AI集群

数千颗昇腾910

算力：256P～1024P FLOPS

AI集群基础单元



64*昇腾910+32*鲲鹏920

形态：47U机柜

数据来源：华为昇腾官网，浙商证券研究所

01 昇腾生态伙伴网络遍及主流厂商

IHV硬件伙伴

ADLINK
ADVANTECH 研華科技
BFA
全爱科技
Dongsheng 东声智能科技

一体机解决方案伙伴

YISA 以萨

整机硬件伙伴

北联国芯 BEILIAN COMPUTING
Wuhan Yangtze Computing Technology 长江计算
华鲲振宇
清华同方 TSINGHUA TONGFANG
H3C
iQOO百信 TRUST
KunLun
宝德 PowerLeader
五舟 WUZHOU
湘江鲲鹏
Huanghe
神州数码 Digital China
安擎 NGINETECH

应用软件伙伴

WHAYER华雁智科
YISA 以萨
DEEPGLINT 格灵深瞳
云从科技 定义智慧生活
intellifusion 云天励飞
ZHY 智洋创新

辅助运营伙伴

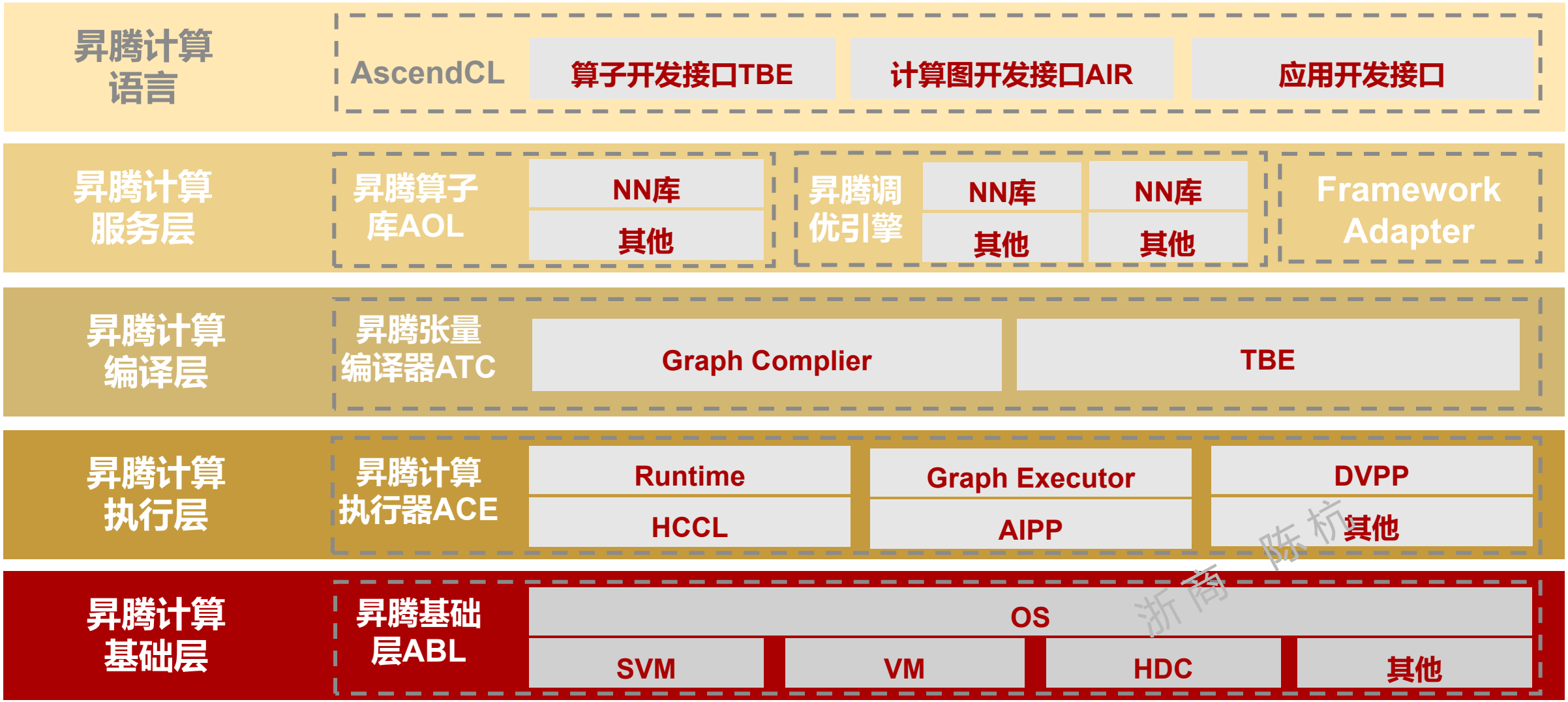
极视 开创AI视觉算法
ICS&S 中软国际

数据来源： 昇腾社区官网，浙商证券研究所

02 人工智能框架

CANN Mindspore

CANN：软硬协同释放澎湃算力



数据来源：华为昇腾官网，浙商证券研究所

昇思 MindSpore 总体架构

昇腾计算语言

- Model Zoo模型库：Vision、NLP、Audio、Rec、Model Hub
- Extend：强化学习/GNN/深度概率编程…
- Sience科学计算：电磁仿真/分子模拟/量子计算
- Insight：网络调试、性能调优、精度调优

开发态友好

- Expression 全场景统一API：函数/OOP编程范式融合；AI+数值计算表达融合；动静表达统一；单机分布式表达统一；NN；Ops；Dataset；Train/Infer；Numpy
- 第三方前端：仓颉、Julia

运行态高效

- Data：数据加载、数据格式、异构加速、数据增强
- Complier AI编译器：类型推导、自动微分、自动并行、二阶优化、量化/剪枝/蒸馏、内存优化、图算融合、流水线执行、IR中间表达层、硬件相关优化、量子自动生成
- Armour AI安全：密态AI、可信AI

全场景部署

- Runtime：端-边-云全场景运行时系统

多样性硬件

- CNN(NPU)、CUDA(GPU)、Eigen(CPU)、oneDNN(CPU)

数据来源： 华为昇腾官网，浙商证券研究所

MindSpore 执行流程

ModelZoo 模型库

MindSpore DevKit开发套件

MindSpore Extend高阶扩展库

MindSpore Hub 预训练模型

下载模型

MindSpore 训练和验证

模型导出

MindIR 端云统一IR

模型导入

MindSpore Servinig 云测推理

MindSpore Lite 端侧推理

端边增量学习 Fine-tuning训练等

调试调优

数据处理

安全机制

MindSpore Insight 可视化

MindSpore Data数据

MindSpore Armour AI安全

端云协同（联邦学习、端云AutoML等）

数据来源： 昇腾计算产业发展白皮书，浙商证券研究所

03

AI开发平台

ModelArts

03 AI开发平台

华为云AI开发生产线ModelArts在AI云服务方面的竞争优势越发明显。历经多年的技术创新，ModelArts已成功在十多个领域进行商业化落地，持续领跑机器学习公有云市场，为AI开发带来变革。

1、简化开发，让AI落地更简单

- **华为云AI开发生产线ModelArts支持全流程MLOps开发，实现行业数据参与AI持续迭代，大幅提升AI应用的二次开发效率。**
- **发起AI生态伙伴计划D-Plan，提供“人”“货”“场”服务，和行业ISV一起，加速AI解决方案在行业的落地。**
- **基于D-Plan的AI项目实践，华为云已在AI Gallery上沉淀了9大行业场景Usecase，覆盖生产、销售、服务和运营等企业运作全场景。**
- **AI Gallery还汇聚了2000多个覆盖零售、医疗、游戏等数十个商业领域的优质模型，助力千行百业智能升级。**

2、深耕技术，让AI应用更高效

- **ModelArts沉淀了知识计算、盘古大模型和天筹AI求解器三项AI根技术，持续构建大模型训练及推理加速能力、分布式训练能力等，从算力资源调度、AI业务编排、AI资产管理以及AI应用部署，提供数据处理、算法开发、模型训练、模型管理、模型部署等AI应用开发全流程技术能力。**

3、应用实例

- **互联网领域：**华为云ModelArts基于算法优化、语音质检等途径，有效提升了T3出行司乘安全检测模型的准确率和召回率，使危险驾驶事件率下降38.6%，同时大幅降低模型开发和交付周期。
- **自动驾驶领域：**针对AI算法训练，华为云ModelArts支撑端到端训练效率提升；分布式多级缓存技术可以将训练时长缩短50%；针对大规模集群训练，拓扑感知调度和动态软路由技术可以提升训练性能30%。

数据来源： IDC，华为云公众号，浙商证券研究所

一站式AI开发平台ModelArets易用且高效

数据来源： 华为官网，浙商证券研究所

1、高效率

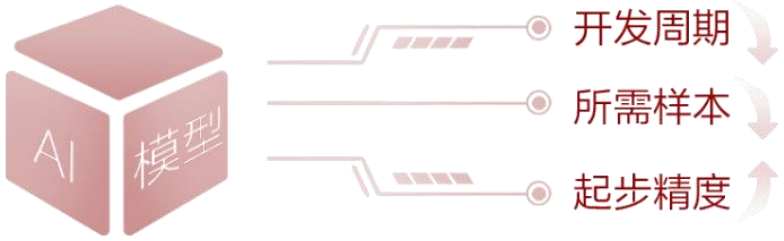
大模型加速AI业务上线

训练推理效率指数级增长

开发周期一降再降（月级>>天级）

所需样本显著降低（万个>>几十）

小样本下，起步精度提升90%



2、更低TCO

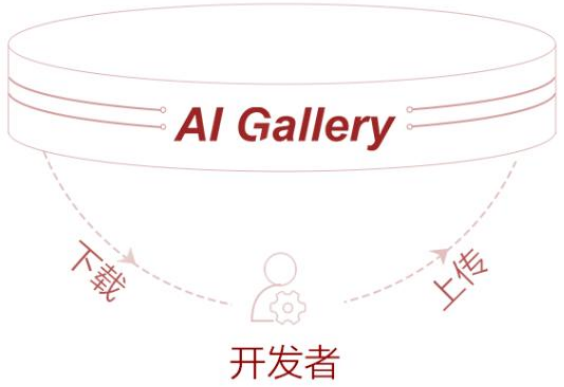
AI Gallery, AI资源无限扩展

常见算法、工具，AI Gallery直达

轻松下载、开发更便利

算法上传，资源易共享

AI资产，高效沉淀和流通



3、易部署

MLOps，AI全流程生命管理

提供各角色无缝协作平台

提升业务价值产出

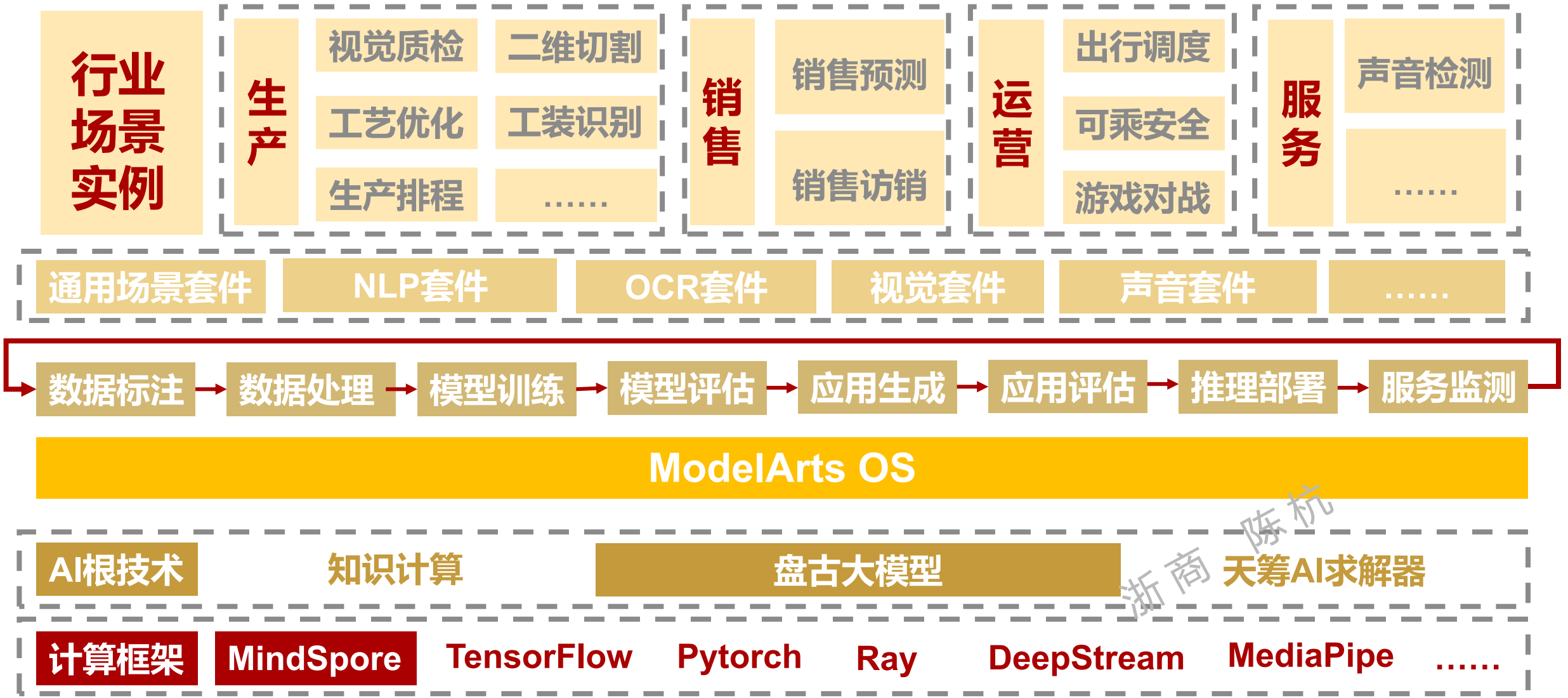
向导式完成AI服务运维和更新

降低运维、更新门槛

实现全流程一键化运行

数据来源： 华为云公众号，浙商证券研究所

03 ModelArts应用场景实例



数据来源： 华为云公众号，浙商证券研究所

04

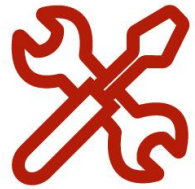
盘古大模型

NLP　　多模态

CV　　科学计算

04 “盘古”开天记，AI落地时

天堑：“小作坊式”的AI开发困境



过去开发模式“三高问题”：开发人员专业性要求高、综合成本高、不可控程度高。
需要拥有作为底座的“重型机械”——预训练大模型：提前将知识、数据、训练成果沉淀到一个模型中，然后将这个基础释放到产业。

登山：盘古大模型背后的人与事

2020年3月：田奇加入华为云-**8月**：核心专家加入-**9月**：推动立项
两大门槛：技术门槛+资源门槛
盘古大模型核心设计原则：模型大；网络结构强；优秀的泛化能力
选择赛道：NLP+机器视觉
未来规划：多模态+科学计算大模型
一系列挑战：算力资源不足、行业数据磨合、内部团队“超人发挥”

翻越：“盘古”究竟强在何处？

2021年4月，盘古大模型正式对外发布。
盘古NLP大模型：业界首个千亿参数的中文预训练大模型，在CLUE实现了业界领先。
盘古CV大模型：业界首次实现模型按需抽取，在ImageNet上小样本学习能力业界第一。
优势和能力点：兼顾架构+小样本学习能力、微调能力、集成行业知识的能力更强+以商业价值驱动研发创新的“实干模式”大模型+生态化、协同创新

灯火：大模型的落地进行时

物流场景：协助浦发银行构建了“物的银行”——**浦慧云仓**。人员行为、货物检测性能提升5-10%，开发效率提升，成本降低。
落地行业：已在100多个行业场景完成验证，包括能源、零售、金融、工业、医疗、环境、物流等等。

思索：AI工业化的虹吸与变革

目前我们处在AI工业化开发模式起步后的快速发展阶段，而大模型是最有希望将AI进行落地的方向。
盘古大模型的价值：推动AI的低成本、可复制。
多重力量的携手与跨界：产学研的纵向融合，不同行业领域的横向协同，诞生AI工业化的虹吸效应。

数据来源： 华为云公众号，浙商证券研究所

NLP模型助力各行提升效率，气象模型全面胜于传统方法

NLP模型

首次使用Encoder-Decoder架构

- 小样本学习超越GPT系
- 通用知识 x 行业经验
- 生成与理解性能领先

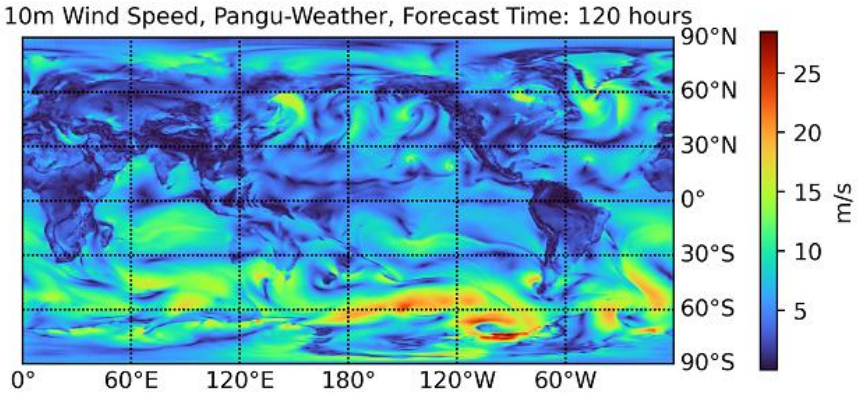
智能舆情

行业效果
精准舆情分析
企业运营软件分析

智能营销

行业效果
取代上升的人力成本
取代低端客服与营销

盘古气象大模型



关键技术

- 3DEST网络结构
- 分层时间聚合算法

效果

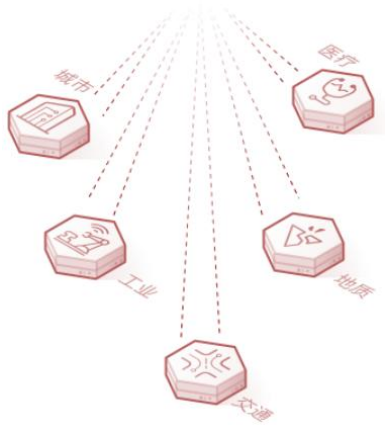
- 精度超传统方法
- 速度超传统方法千倍
- 预测台风路径降低20%位置误差

数据来源： 华为云官网，浙商证券研究所

盘古大模型：CV+多模态+语言

CV大模型：分类 | 分割 | 检测

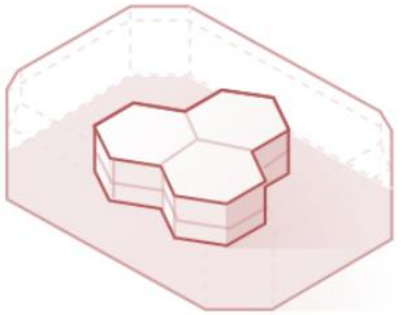
亿级图片数据



预训练

数据处理 → 模型生成 → 模型优化

CV大模型



业界最大预训练CV模型：30亿参数，10亿级图像

判别与生成联合预训练：底层/高层视觉预训练任务通用

100+场景验证：研发成本降低90%

小样本学习性能领先：ImageNet10%标签分类精度业界第一

智能巡检

智慧物流

多模态大模型

跨模态检索

跨模态生成

看图说话

语言大模型

语音识别

语音分类任务

语音回归任务

持续推出

数据来源： 华为云官网，浙商证券研究所

应用实例：盘古时尚多模态大模型

技术层面

亿级自然图像

预训练

十万级时尚产业数据

微调

优化策略：模型并行、数据并行、混合精度运算、稀疏训练等

节点并行调度算法

盘古时尚多模态大模型

天级单位完成训练

应用实例

HUAWEI

+

时谛智能 4D·SHOETECH

=

为四川大凉山孩子制作爱心冬衣；依靠AI、云计算等技术支撑，将设计时间从三周缩短为5-7天。

未来

1、帮助设计师洞察流行趋势

对十万级时尚产业数据的颜色、版型、图案等元素进行分析后，批量生成与流行趋势接近的服饰，将当下流行元素视觉化并提炼给设计师，洞察用户消费意向，预判行业流行趋势。

2、支持生成多样化服饰

盘古多模态大模型基于大规模时尚产业数据，向设计师推荐服饰图片。设计师搜集素材的时间占整个制作周期的70%以上。应用盘古大模型，设计师能够在创作前期根据不同的推荐结果快速积累灵感，快速进入二次创作。

3、帮助批量生成符合要求的服饰图片

盘古时尚多模态大模型当前可支撑一站式批量呈现128张推理生成的服饰图片。

4、具备分钟级设计推理能力

当前盘古多模态时尚大模型利用Attention Cache等技术加快推理速度，在华为云提供的单卡V100支持上可支撑分钟级推理速度，快速反馈推理结果。

数据来源： 华为云公众号，浙商证券研究所

应用实例：药物分子大模型

华为云AI辅助药物设计服务

基于盘古药物分子大模型训练

过去：研发周期 数年；成本 高

将来：研发周期 一个月；成本 ↓70%

AI+制药降本增效 Drug X迎来重大突破

独有“图-序列不对称条件自编码器”架构

把药物分子结构转换成可量化的数值，可以更好地在数值空间定量地对药物分子结构与性质进行预测与推荐。

海量数据训练

学习了17亿个药物分子的化学结构，能够对药物分子的80多种化合物理化性质进行预测，包括水溶性、吸收、代谢活性、排泄速率、毒性等。

计出化合物的新颖性可以达到99.68%

成药性预测的准确性提升20%

科技平台与科研团队联手 AI与生物医药双向赋能

创新药行业同质化靶点扎堆问题严峻，能给新药研发带来颠覆性变化的AI+制药被寄予厚望。

AI+制药，尤其药物设计环节，是复杂软件工程，涉及到非常大规模的计算。

华为云的优势：大数据与AI相结合

华为云未来计划：结合硬件，持续打造软硬件能力兼备的药物辅助设计平台。

全流程辅助药物设计 AI未来大有可为

华为云盘古药物分子大模型”包含多方面的创新，该大模型在药物研发领域推出的预训练大模型，对实现全流程的 AI 辅助药物设计会大有帮助。

经典药物研发过程

靶点验证、化合物筛选、药物优化

各期临床试验等阶段

开创新的药物研究范式

数据来源： 华为云公众号，浙商证券研究所

行业评级与免责声明

行业的投资评级

以报告日后的6个月内，行业指数相对于沪深300指数的涨跌幅为标准，定义如下：

1、看好 ：行业指数相对于沪深300指数表现＋10%以上；

2、中性 ：行业指数相对于沪深300指数表现－10%～＋10%以上；

3、看淡 ：行业指数相对于沪深300指数表现－10%以下。

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。投资者不应仅仅依靠投资评级来推断结论

行业评级与免责声明

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。

本报告仅供本公司的客户作参考之用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本报告仅反映报告作者的出具日的观点和判断，在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、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的意见或建议不一致的投资决策。

联系方式

浙商证券研究所

上海总部地址：杨高南路729号陆家嘴世纪金融广场1号楼25层

北京地址：北京市东城区朝阳门北大街8号富华大厦E座4层

深圳地址：广东省深圳市福田区广电金融中心33层

邮政编码：200127

电话：(8621)80108518

传真：(8621)80106010

浙商证券研究所：http://research.stocke.com.cn