InfoQ 极客传媒

端侧智能
模型架构设计与算法改进

演讲人：刘凡平
ROCK AI，CEO

AiCon
全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

01 端侧智能的兴起与挑战

引入：从云端智能到端侧智能的演进路径

端侧大模型定义

端侧大模型是一种在终端设备上进行本地私有化部署的人工智能模型，其核心能力在于基于多模态感知实现**自主学习**与**记忆**，以提供个性化服务并保障数据隐私与运行安全。

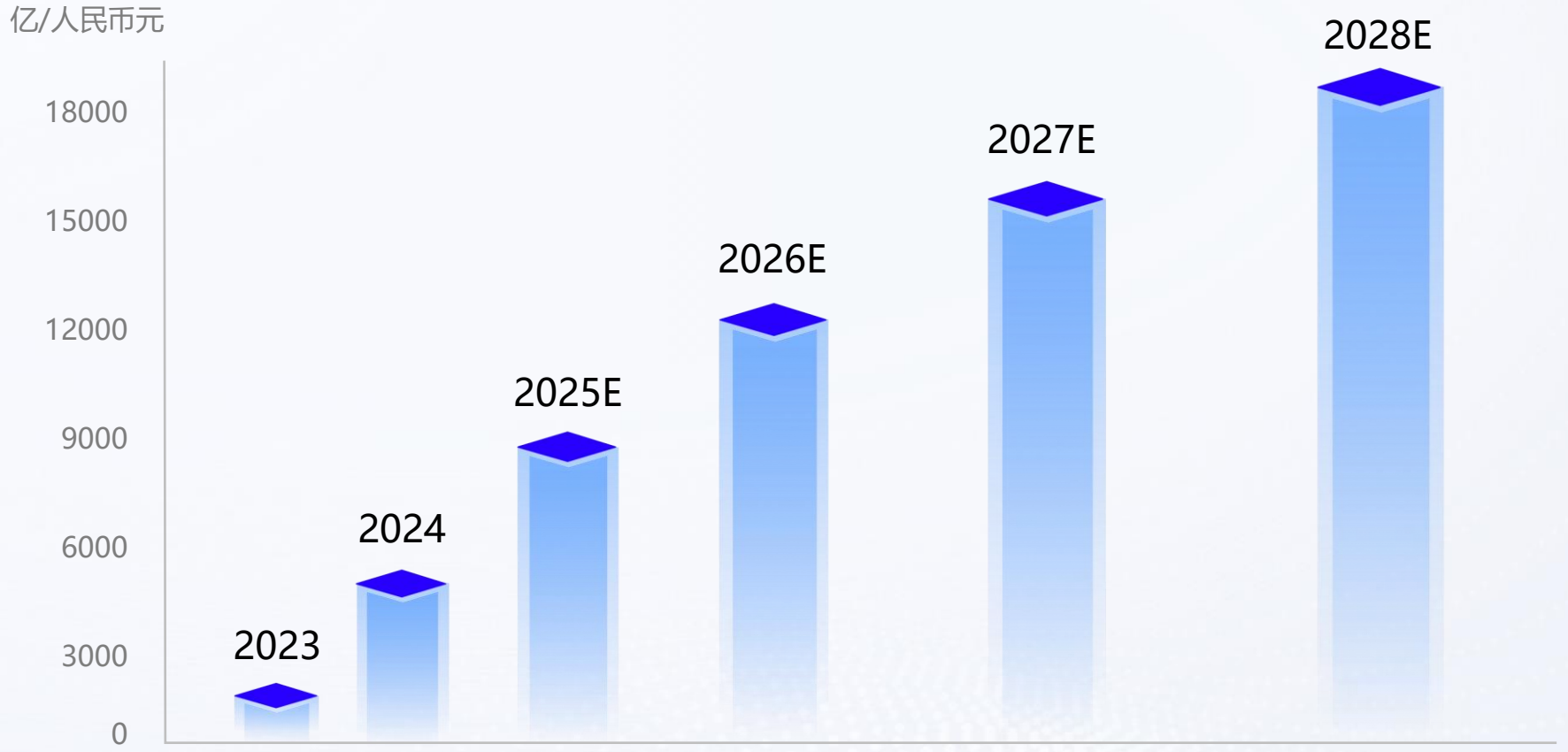
端侧大模型不等于云端大模型的小参数版本

自主学习和记忆能力才是核心！

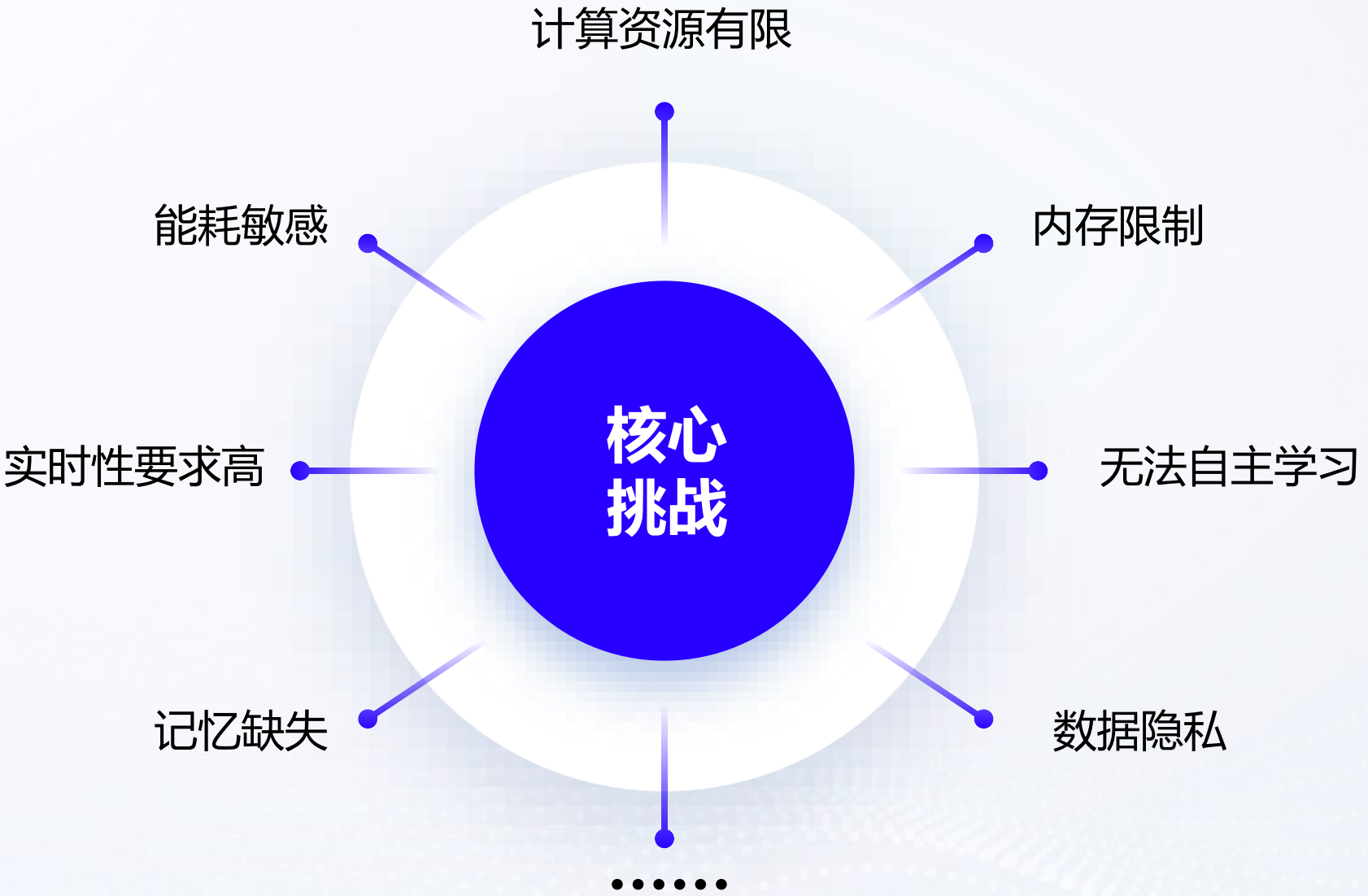
低延迟	实时交互体验（语音助手、智能输入法、AR应用）
数据隐私	用户数据无需上传云端，本地处理更安全
离线可用	无网络或弱网络环境下也能工作
降低成本	减少对云端计算资源的依赖
个性化与定制	基于本地交互数据进行自主学习

端侧AI持续持续扩大

端侧AI行业规模

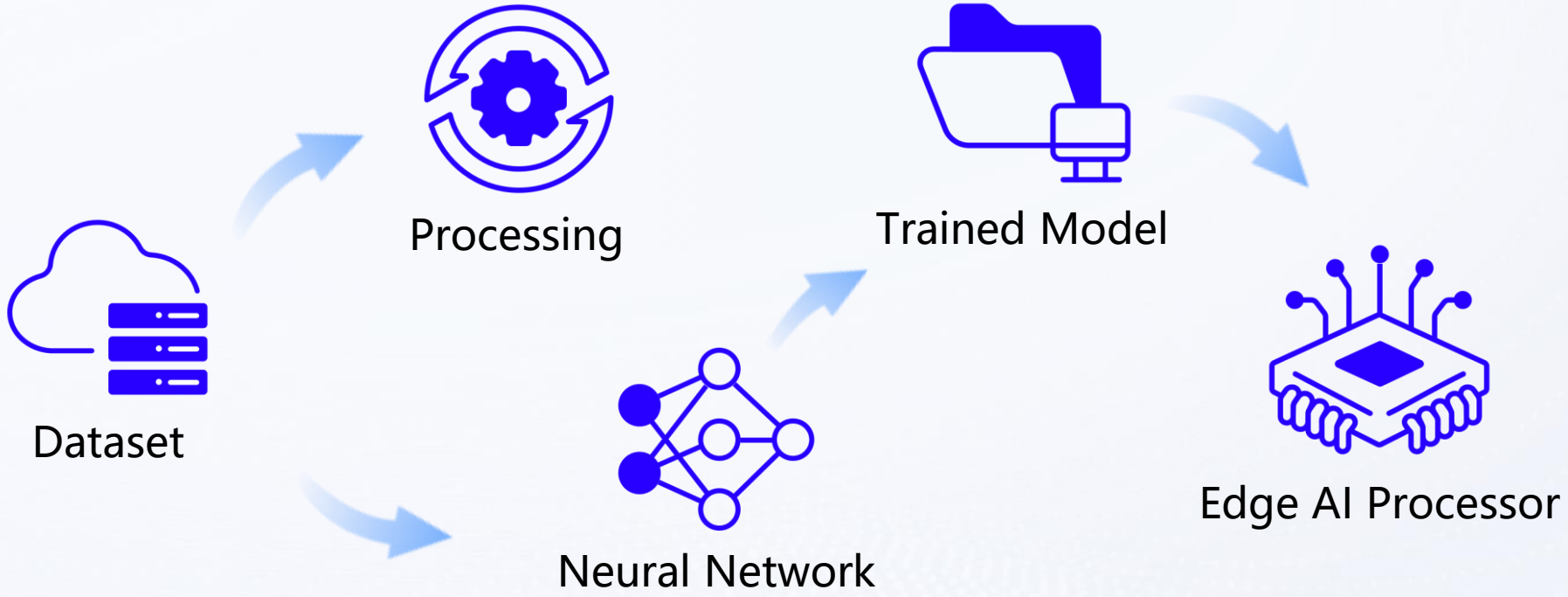


数据来源：深圳新闻网，中安网，中国知网，数字信息安防产业技术创新联盟，长沙优来电子科技有限公司，OPPO，联想，三星，嘉德智能


计算资源有限
能耗敏感
内存限制
核心挑战
实时性要求高
无法自主学习
记忆缺失
数据隐私
.....

云端大模型难以部署到端侧

算力鸿沟、内存溢出、运行时延、功耗爆炸



让世界上每一台设备拥有自己的智能！

欢迎友商跟随，让端侧应用成为行业共识。

02

端侧大模型架构设计方向

传统模型优化的基本原则：少算、快算、省能

1、激活函数与优化器的轻量化改进（如ReLU6，AdamW替代）；

2、数据稀疏性与动态计算路径；

3、模型微调方式：LoRA、Adapter、Prompt Tuning 的端侧适配；

4、模型鲁棒性增强：对抗样本防御与小样本学习策略；

5、注意力机制的高效替代：线性Attention、Performer、Linformer等；

6、高效卷积替代：Depthwise Conv、Group Conv等；

7、核心模型+轻量子模型。

… …

如何让“大模型”跑在“小设备”上?

糟糕的三部曲：

剪枝

量化

蒸馏

1 网络架构优化：轻量模型设计
（MobileNet, ShuffleNet, EfficientNet-Lite）

Transformer在端侧的轻量化尝试
（TinyBERT, MobileBERT, DistilBERT） 2

网络架构轻量化设计?

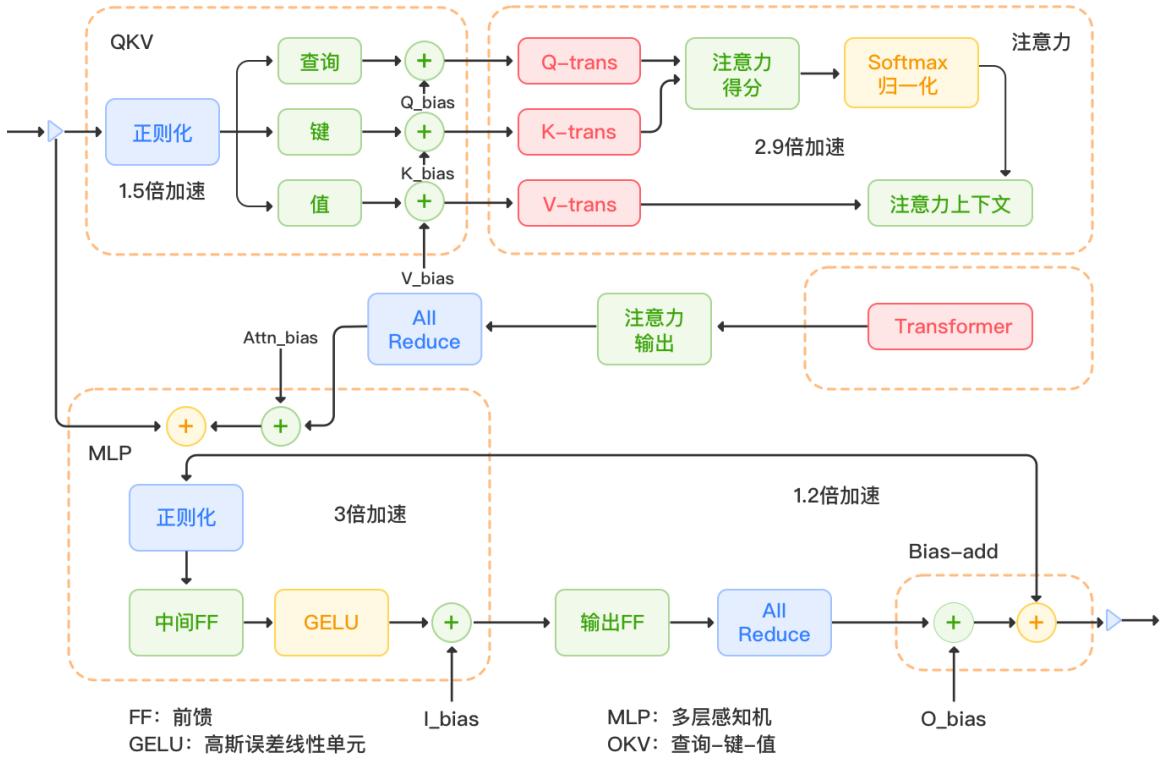
3 神经架构搜索（NAS）在端侧的实践
（ProxylessNAS, Once-for-All）

CNN与Transformer混合架构 4

……

算子融合示例

针对Transformer的结构特点，算子融合主要分为4类：归一化层和QKV横向融合，自注意力计算融合，残差连接、归一化层、全连接层和激活层融合，偏置加法和残差连接融合。



Transformer层中的算子融合示意图

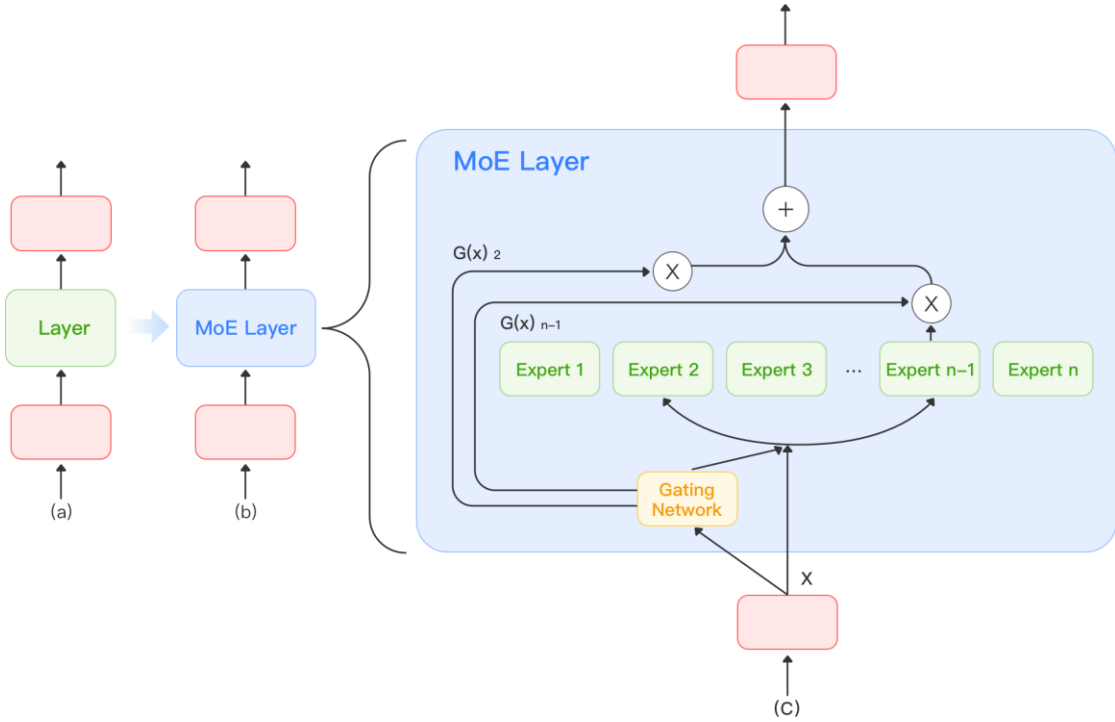
混合专家机制?

动态路由选择，减少计算负载。

1991年的论文《Adaptive Mixtures of Local Experts》：

“This idea was first presented by Jacobs and Hinton at the Connectionist Summer School in Pittsburg in 1988.”

Google在2017年1月发布了《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》，把MoE带进了LSTM，训出了最大137B参数，专家数达到128k的LSTM模型。



结构优化思路

- 轻量化网络设计：MobileNetV3、EfficientFormer、TinyBERT；
- 混合专家机制（MoE）：动态路由选择，减少计算负载；
- 层级裁剪（Structured Pruning）：通道、头部、Block级别剪枝；
- 蒸馏策略（Task-aware Distillation）：强化特定下
- 游任务表现。

参数压缩与量化策略

- Post-training Quantization（PTQ）：离线量化，
- 部署灵活；
- Quantization-aware Training（QAT）：训练阶段
- 模拟量化误差；
- Bit-width探索：INT8、INT4、甚至Binary；
- 混合精度策略：关键路径高精度，非关键路径低精度。

稀疏性与a结构感知优化

- 激活稀疏性（Activation Sparsity）：ReLU后的0值跳过；
- 权重稀疏性（Weight Pruning）：Static vs. Dynamic稀疏；
- 结构感知剪枝：保持模型结构对称性，利于硬件并行；
- 软硬结合优化：软件模型剪枝配合硬件编译优化（如NPU）。

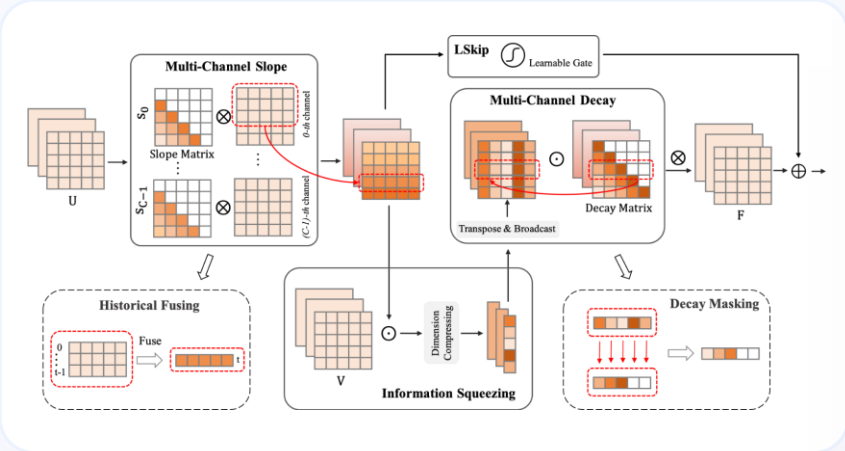
推理算法与执行策略

- Transformer加速：Linformer、Performer（低秩注意力）；
- 图优化：ONNX Graph Optimization、TensorRT Fusion；
- 异构并行：CPU+NPU/GPU协同执行调度；
- KV缓存优化；
- 精简缓存长度、分块KV存储。

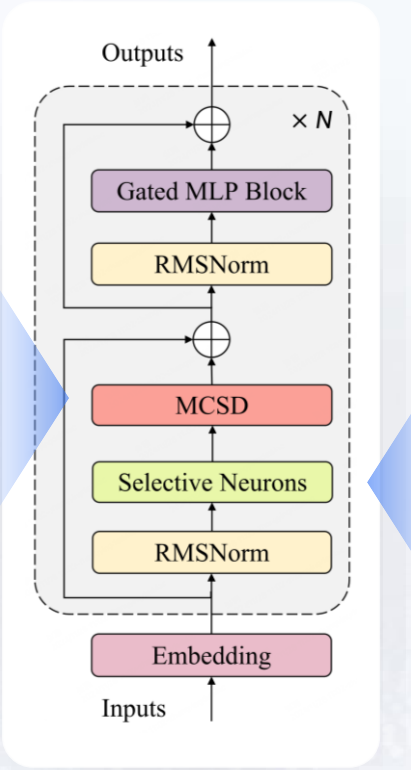
MCSD与类脑激活机制

- **MCSD：** 实现更高效的特征提取

MCSD整体架构，具有快速训练、推理能力，以解决Transformer全局注意力导致的模型计算复杂度高、推理速度慢等问题。

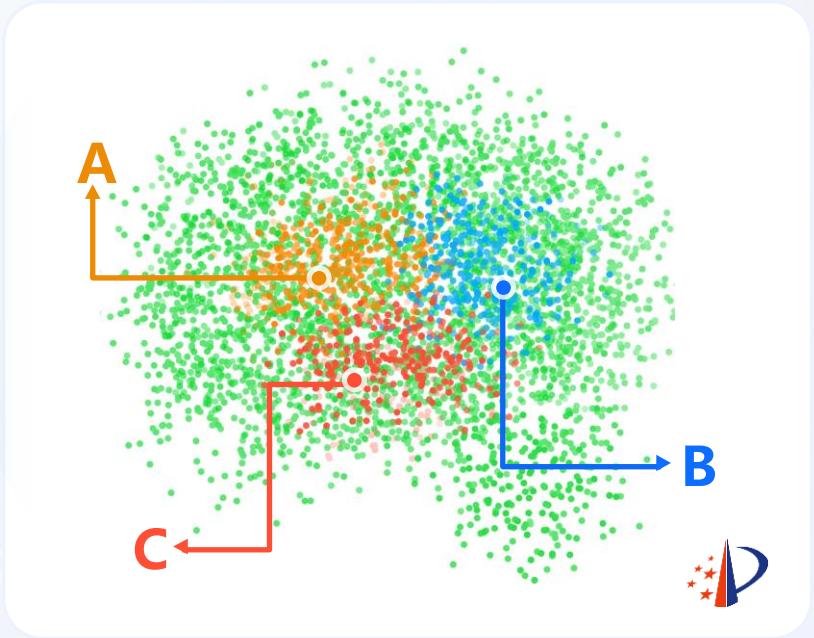


通过斜率变换和指数衰减两个部分提取输入数据中的当前和历史信息，并进行位置感知的多通道特征融合。



- **类脑激活机制：** 大幅减少计算冗余

模拟大脑中的神经元激活模式，更有效地处理复杂数据和任务，显著提升计算效率和精度，为解决现实复杂问题提供了新的工具。



燃油汽车 VS 新能源汽车

欢迎友商跟随，让非Transformer成为行业共识。

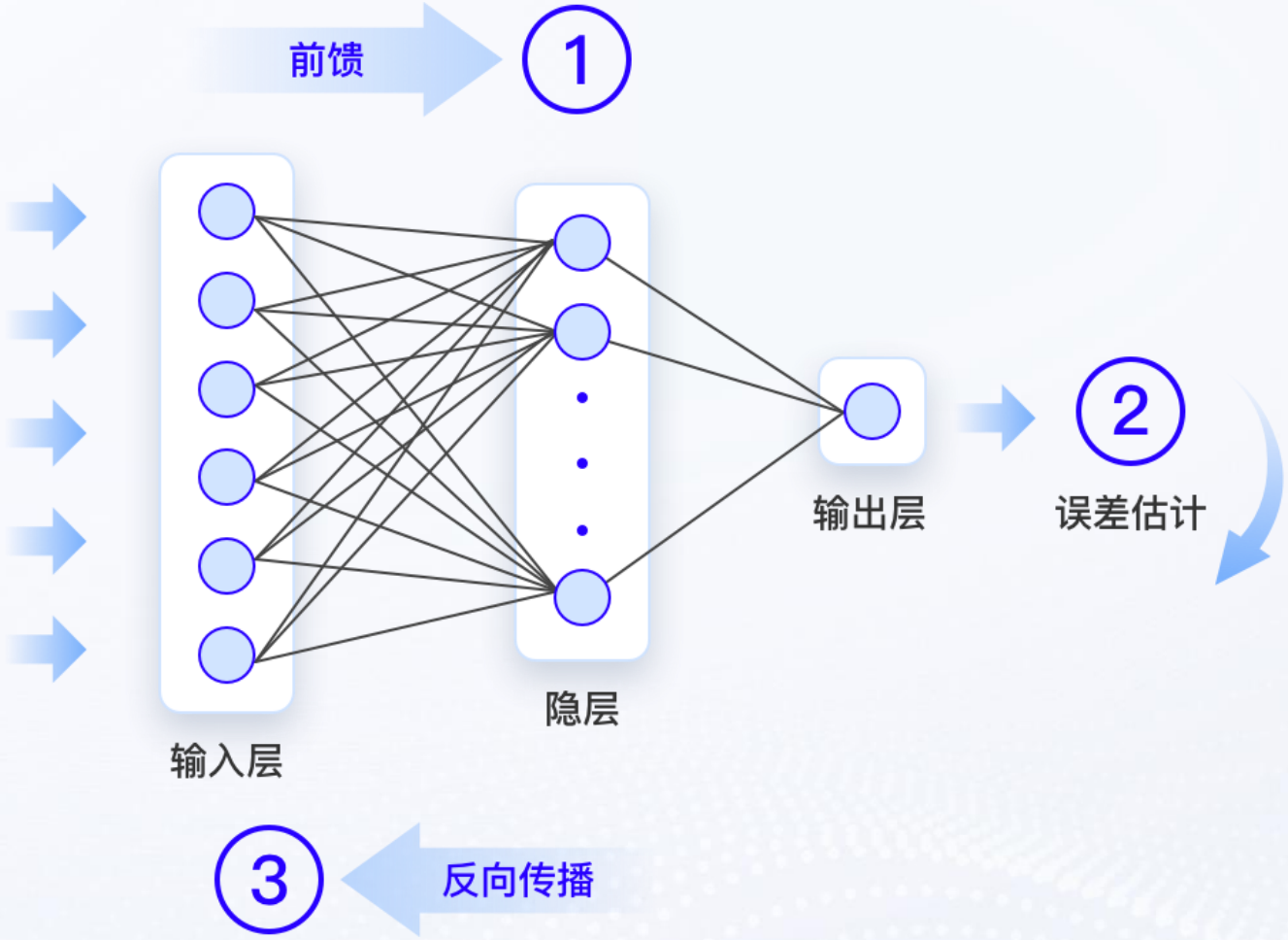
03

端侧场景的基础算法改进思路

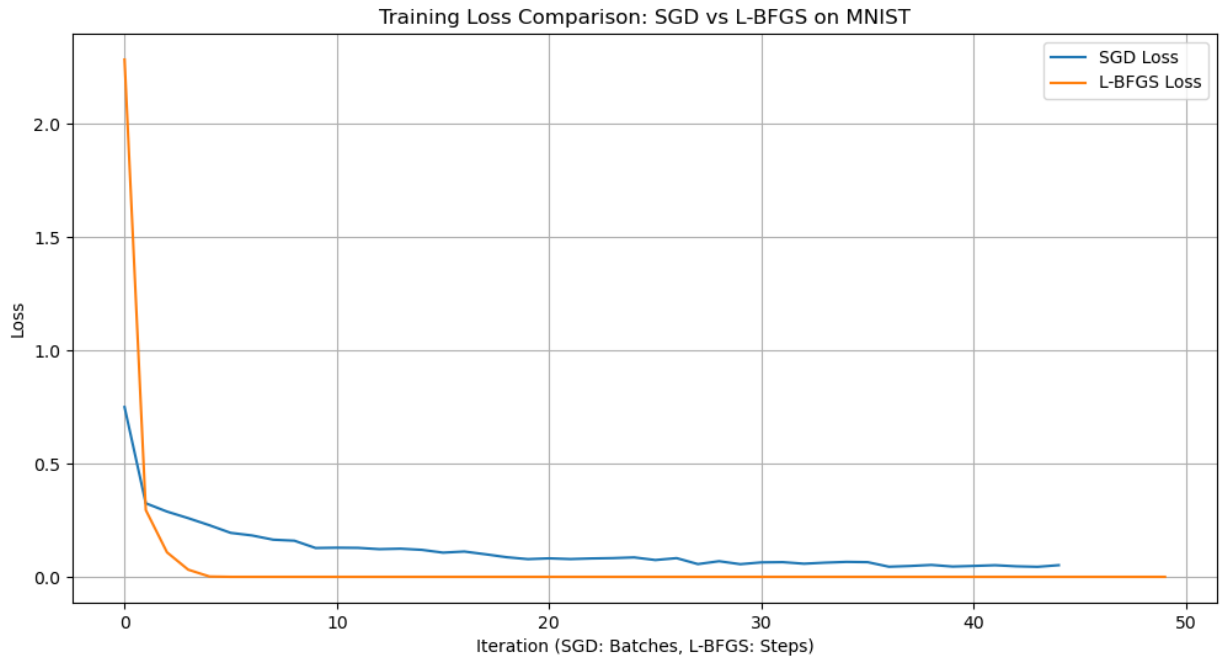
到底是谁在“拖后腿”？

不要只想着端侧推理。

爱恨交织的反向传播算法！



牛顿法、拟牛顿法、梯度下降法



梯度下降（如SGD）每次迭代只使用一个（或一小批）样本的梯度来更新模型参数，计算成本低，尤其适用于大规模数据集。

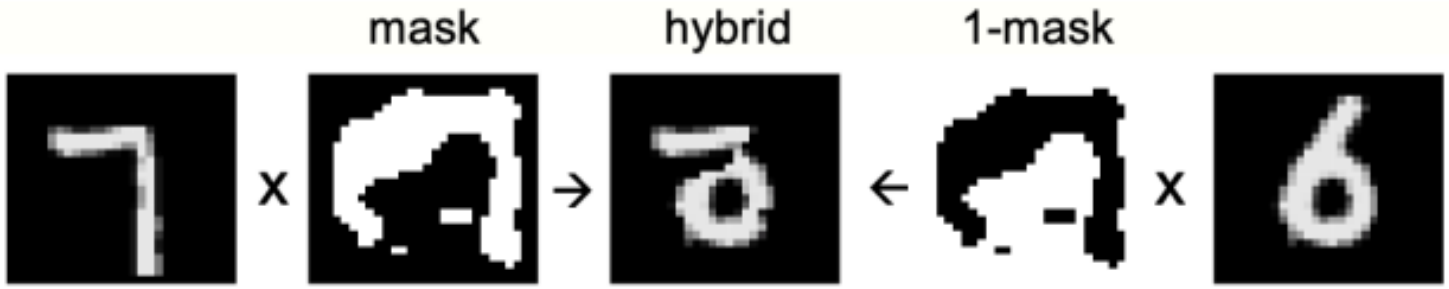
拟牛顿法（如L-BFGS）通过估计或逼近目标函数的Hessian矩阵（二阶导数信息）来决定搜索方向，相较于只使用梯度信息的方法，通常收敛更快。

Hinton：Forward-Forward

传统的反向传播算法通过计算损失函数对网络参数的梯度来更新权重，但这种方法存在一些问题，例如梯度消失、梯度爆炸以及训练过程较为复杂等。

Forward-Forward 算法的核心思想是将神经网络的训练过程分解为多个前向传播阶段，并在每个阶段中逐步调整网络的权重，而不是通过反向传播计算梯度。

它利用了“对比学习”（Contrastive Learning）的思想，通过比较输入数据在不同阶段的输出来调整权重。

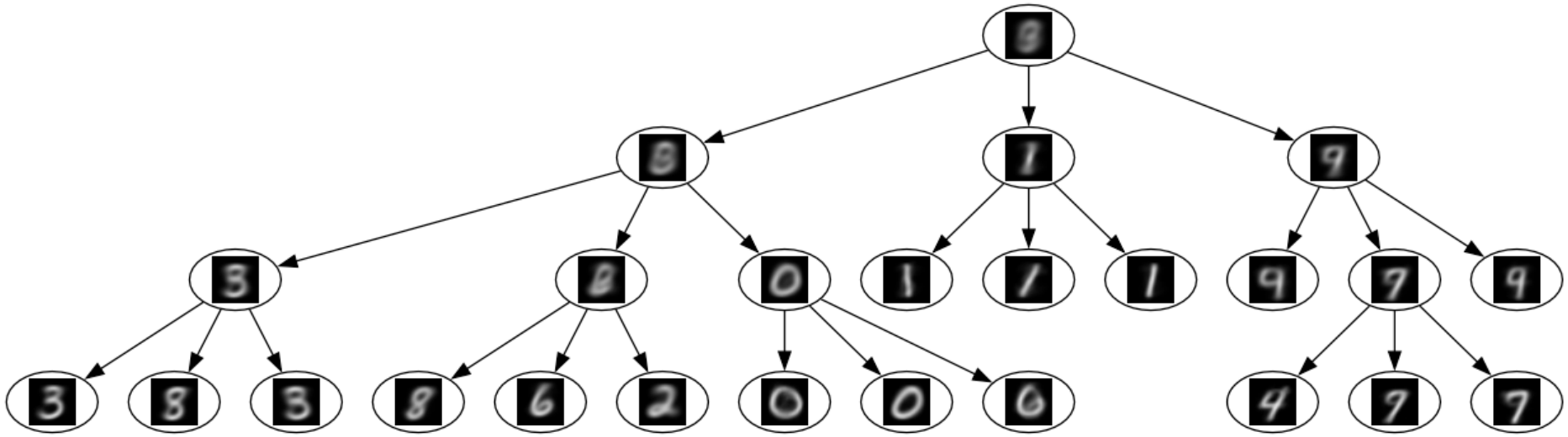


负样本的构造

*测试错误率：1.37%

可视化的理解机器智能：以MNIST为例

ROCK AI在非Transformer之外，也在实验室探索非反向传播之外的可理解的机器学习方式，针对MNIST数据集，在保持测试准确率95%以上的情况下，实现可以通过可视化的方式理解预测逻辑。



MNIST的可视化理解推理示例（局部）

04

端侧智能的未来发展趋势展望

群体智能

定义：具备自主学习的若干智能单元，通过环境感知、自我组织、互动协作 共同解决复杂问题，并在不断变化的环境中实现整体智能提升。

期待群体智能成为通用人工智能之路的新共识

从质疑到认可，从认可到行动！

早期

RNN/LSTM

为大语言模型的开发奠定了基础。

2017 Google

Transformer架构诞生

引入自注意力机制，为预训练模型提供了技术基础和新的路径。

幻觉严重　缺乏记忆　无法持续学习

2018

BERT

Google发布首个基于Transformer的双向预训练模型。

2022

ChatGPT

OpenAI发布对话型AI，引发了公众对大语言模型的广泛关注。

2024.01

已通过中央网信办备案

国产非Tranformer架构的大模型诞生

Yan架构大模型与Transformer架构大模型对比：

1、7倍训练效率、5倍推理吞吐、3倍记忆能力；

2、低幻觉，CPU无损部署；

3、Yan-3B模型超过Llama3-8B水平。提交52项发明专利，涵盖大模型的模型结构设计、基础算法改进等核心环节。

2024.06

全球首个树莓派上运行的多模态大模型

2024年06月世界人工智能大会上，ROCK AI对外发布了能够在树莓派上运行的多模态大模型，证明超低算力设备也可以拥有智能。

2025

DeepSeek-R1

融合了深度推理，参数规模超万亿，开源设计，应用成本相对降低。

亟需突破智能边界
改变机器学习模式

2025-2026

基于训推同步的自主学习

赋予模型在推理时进行训练的能力，以实现**持续进化**和**环境适应**。

2027

协同化群体智能

ROCK AI相比于OpenAI、Meta等的“造神”路径，ROCK AI更专注于**群体学习**能力、**协同进化**的群体智能路径。

OpenAI　Deep Seek　Meta

AGI

搭载Yan架构大模型的手机

搭载Yan架构大模型的无人机

搭载Yan架构大模型的机器人

让世界上每一台设备拥有自己的智能

智能重新定义硬件！

硬件-算法协同设计

不再是你传统认知的端云结合！

端侧不够，云端来凑？我们必须放弃大模型和小模型的概念。

InfoQ 极客传媒

北京

QCon 全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

北京

AiCon 全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

上海

QCon 全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月 — 5月 — 6月 — 8月 — 10月 — 12月

上海

AiCon 全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

深圳

AiCon 全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

北京

AiCon 全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

InfoQ 极客传媒

THANKS

ROCK AI：让世界上每一台设备拥有自己的智能。

AiCon
全球人工智能开发与应用大会