

体系化人工智能(Holistic AI)技术探索

中国移动研究院 张世磊

2023.11.24

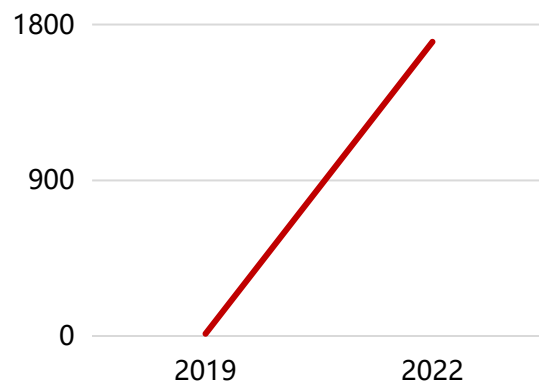
人工智能产业发展的主要矛盾

日趋泛在的智能化需求和智能化技术赋能成本高之间的矛盾

日趋泛在的智能化需求

企业智能化需求持续增长

中国移动商用落地的智能化项目数量三年增长100多倍



智能化技术赋能成本高

核心技术研发成本高

GPT-3大模型训练成本

费用成本

460万美元

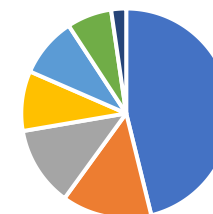
时间成本

1 GPU × 355年

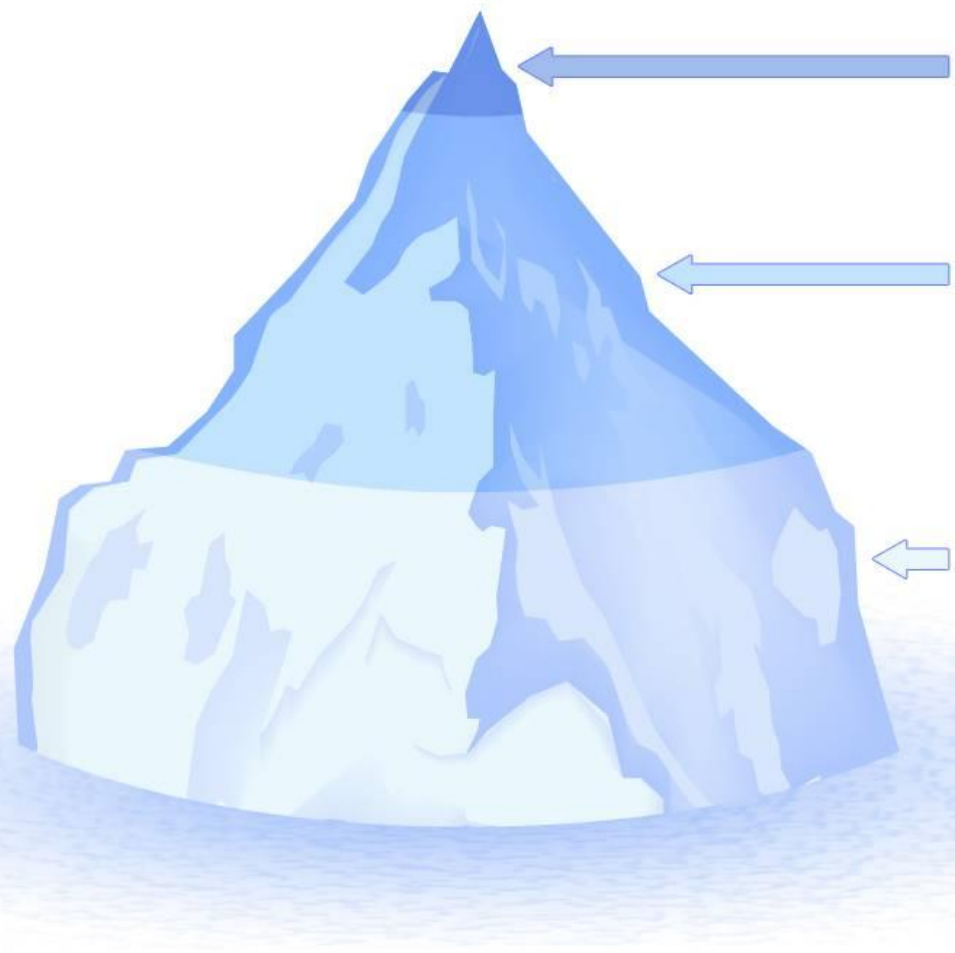
定制化、商务、运维成本高

典型AI商用定制化项目成本构成

- 定制化研发
- 部署交付测试
- 售后运维
- 售前解决方案
- 数据采集
- 合同验收
- 需求沟通



AI应用的挑战



人工智能的应用

需求复杂、迭代优化、运营成本

数据成本、算力成本、算法成本、人才成本

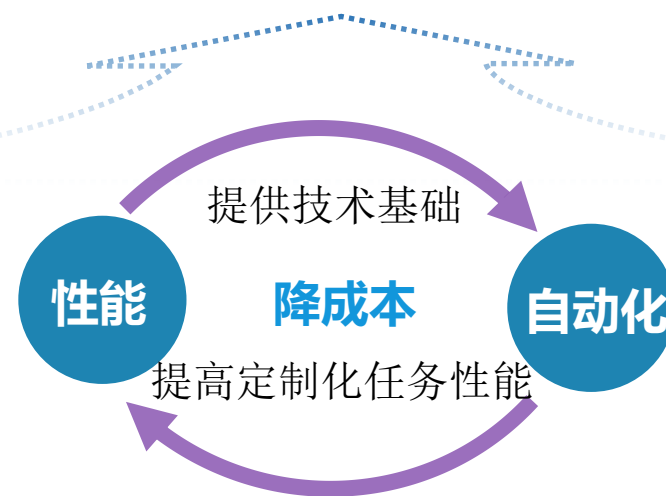
弱人工智能



强人工智能

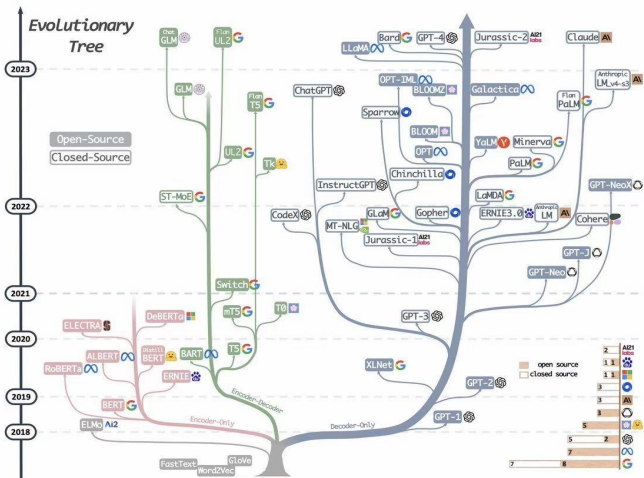
(限定领域、人工参与)

(通用领域、自动化)



通用人工智能发展

LLM



LLM使能的自主智能体

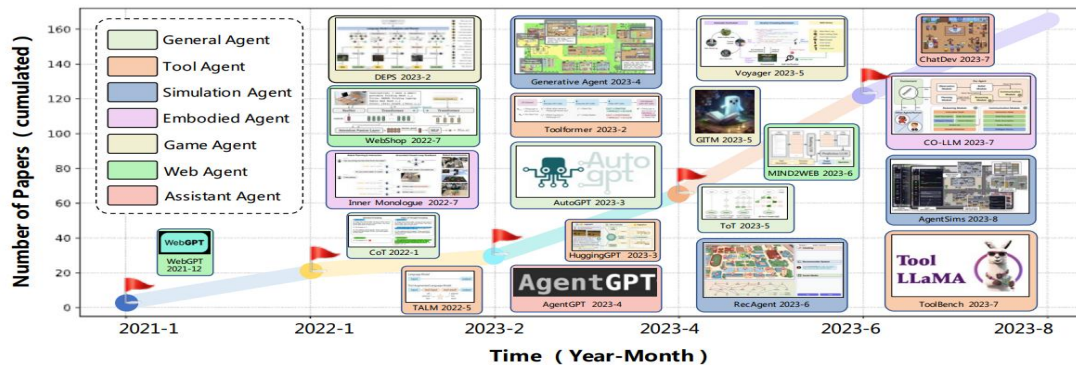
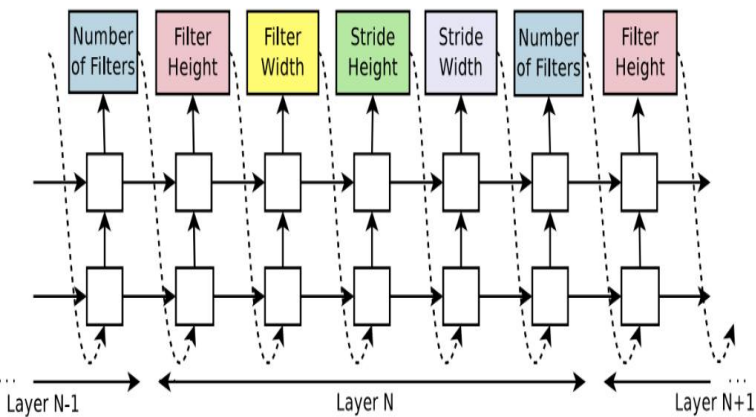


Figure 1: Illustration of the growth trend in the field of LLM-based autonomous agents.

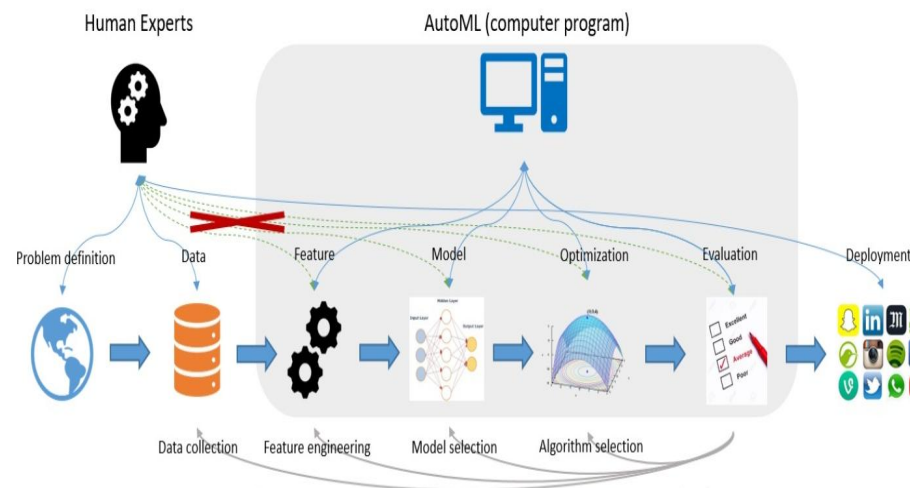
A Survey on Large Language Model based Autonomous Agents

神经网络架构搜索



单模型的通用化

自动机器学习



基于AI任务的自动化

小模型时代：人工智能走向规模化应用的核心点

支撑环境适宜： 选择环境， 培育环境

共性能力： 合理评估AI能力的可达性， 构建可达的共性AI能力

平台化： 实用便捷的工具， 运营运维

业务本身是规模化的： 客户规模， 经济规模

中国移动构建出全面的九天人工智能产品体系，实现规模化赋能价值

九年耕耘，打磨出完善的九天人工智能产品体系；“两给两出”，激发出九天团队人才和研发科技创新活力

规模化应用 41 个

CHBN赋能价值



核心能力 322 个

通用能力

智能语音 智能推荐 机器视觉 智能数据分析 自然语言理解

网络智能化能力簇

感知智能 预测智能 诊断智能 决策智能 控制智能

平台型产品 8 个

⑥ 网络智能化平台

⑦ 九天·毕昇教育平台

⑧ 城市AI平台（合作）

③ 智能交互平台

④ 可视化建模平台

⑤ 智能推荐平台

① 九天深度学习平台

② 九天AI能力平台

中国移动构建“九天”人工智能大模型体系

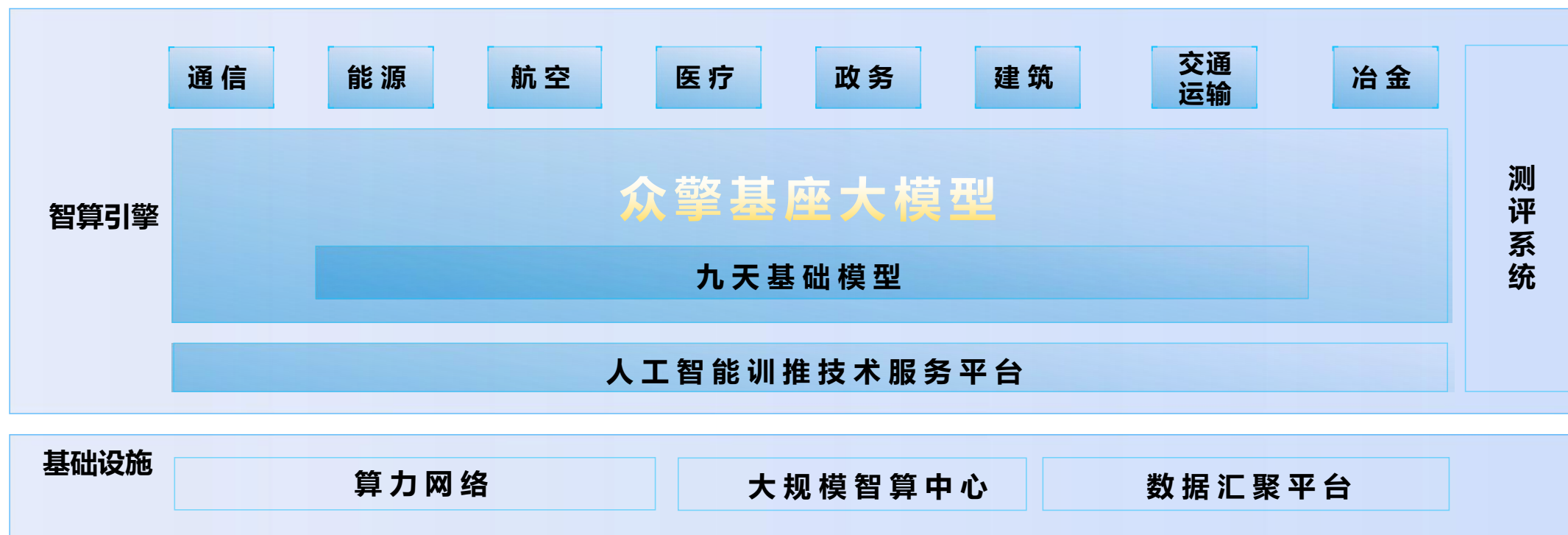
基础大模型：加快构建适用于泛场景的自主可控通用基础大模型，打造通用智能底座

行业大模型：聚焦供给侧，加快构建行业大模型，加速国民经济主体行业的智能化转型升级，促进我国整体生产力跃升



九天·众擎基座大模型

以九天基础模型为基础，联合通信、能源、航空等行业的骨干企业，共建共享九天·众擎基座大模型



CCTV 2
财经

CCTV 央视网
com



正点财经



新 强

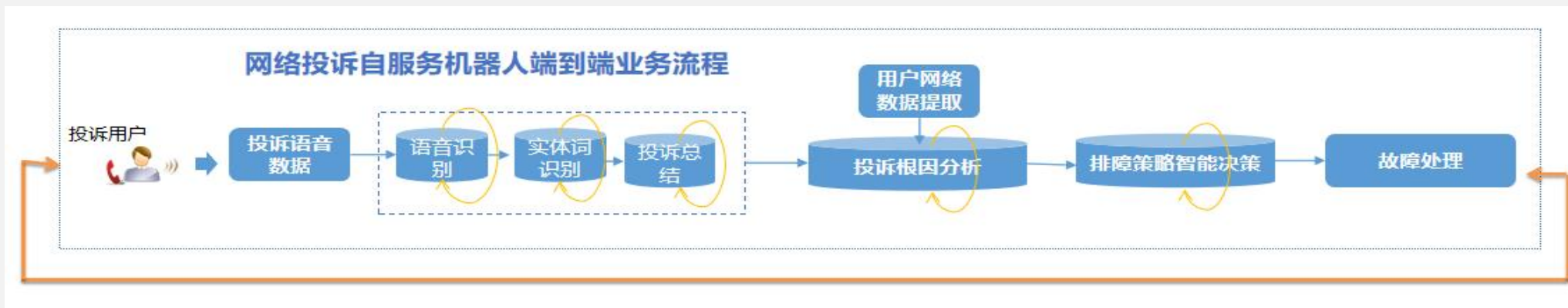
9:00

MSCI中国A股

生产型应用样例



网络问题投诉 级联优化



通常需要在满足计算、传输、安全、可控性等多项约束前提下，组合使用多个模型或能力，包括基础模型、行业模型或面向特定任务的小模型，并能够端到端优化服务于业务目标

体系化人工智能

体系化人工智能（**Holistic AI, HAI**）是中国移动研究院九天团队原创技术的攻关方向，依托泛在的网络和AI算力，在开放环境中实现对AI能力进行灵活且高效的配置、调度、训练和部署，以满足日益丰富的数智化业务需求，同时确保AI业务可信可控安全，其主要特征为**AI服务大闭环、AI能力原子化重构、网络原生AI及安全可信AI**。

根据智能化业务需求，按需对AI能力进行调度、配置和运行监控，使其能在最合理的算网资源上运行和服务



■ 体系化人工智能（Holistic AI）的主要特征

1、“大闭环”（Big Loop AI）

“大闭环”AI以业务端到端的大闭环优化为目标，重点攻关多能力级联与并联优化、开放动态环境中AI能力优化的基础理论和技术，从而达到AI产业闭环。

2、AI技术原子化重构（Atomized AI）

AI能力依据高复用、易调度、自闭环、易适配等原则进行原子化拆解和重构。一个典型的原子化AI能力包含通用智能层、适配层、接口层，通用智能层可多个能力共享。AI原子化重构是体系化人工智能得以实现的基础。

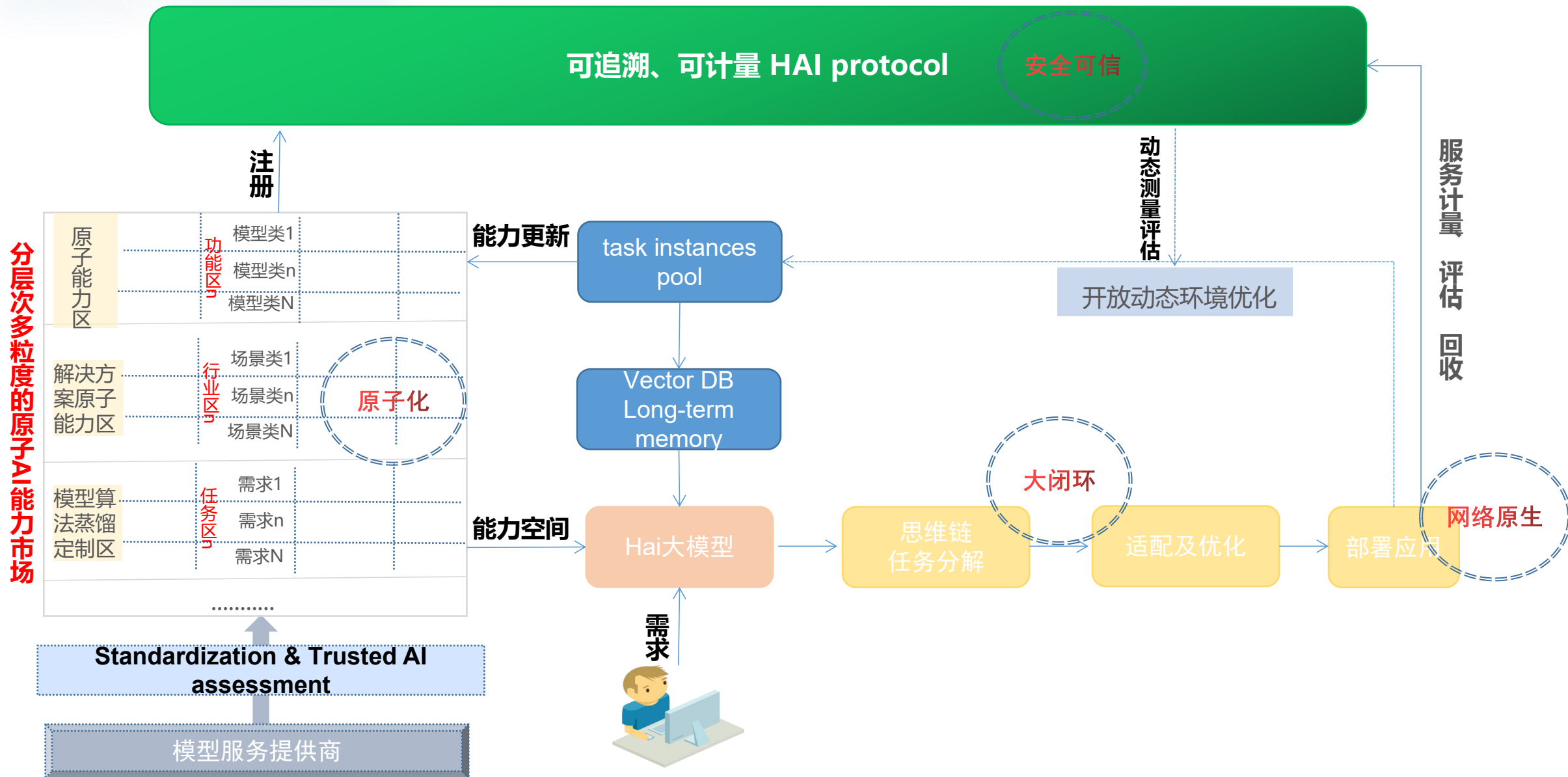
3、网络原生（Network Native AI）

网络原生AI将AI能力与算力通过标准化的方式接入网络、按需调度，重点攻关AI模型自动伸缩的理论和机制，制定AI计算资源、数据、模型、能力、服务的功能、流程、接口和计量的标准，实现AI能力在网云边端弹性部署、计算和迭代。

4、安全可信（Trusted AI）

AI数据、模型、能力、业务的安全可信是体系化人工智能服务的重要基础，重点攻关AI服务可追溯、可互信、可审计、抗攻击的基础理论与方法。

体系化AI系统架构图



体系化AI总体技术方案

$$\begin{aligned} \max & f(D, M, A, P, O, E, S, F, U, T) \\ \text{s. t.} & c(D, M, P, O, E, S, F, U, T) \leq C \end{aligned}$$

其中：

f 是一个复杂的函数，表示体系化人工智能的内部逻辑和流程。

数据集 $D=\{d_1, d_2, \dots, d_n\}$ ，每个数据 d_i 都有一个类型 $t_j \in \{0, 1, 2, \dots, t\}$ ，表示文本、图像和语音等异构数据类型。

模型集 $M=\{m_1, m_2, \dots, m_k\}$ ，每个模型 m_i 都有一个类型 $k_j \in \{0, 1, 2, \dots, s\}$ ，表示分类模型、预测模型和生成模型等不同模型。

原子能力集 $A = \{a_1, a_2, \dots, a_l\}$ ， a_i 是第 i 个能力，表示语音识别，语音增强，图像分割，机器翻译等不同的能力。

真实环境数据分布集 $P=\{p_1, p_2, \dots, p_n\}$ ，每个分布 p_i 都有一个类型 $q_j \in \{0, 1, 2, \dots, q\}$ ，表示高斯分布、均匀分布和其他复杂分布。

原子化评估集 $E=\{e_1, e_2, \dots, e_k\}$ ，每个评估 e_i 都有一个指标 $v_j \in \{0, 1, 2, \dots, v\}$ ，表示不同的评估指标。

标准规范入库集 $S=\{s_1, s_2, \dots, s_k\}$ ，每个入库 s_i 都有一个条件 $w_j \in \{0, 1, 2, \dots, w\}$ ，表示入库准则。

真实场景数据漂移集 $F=\{f_1, f_2, \dots, f_n\}$ ，每个漂移 f_i 都有一个类型 $x_i \in \{0, 1, 2, \dots, x\}$ ，表示协变量漂移、先验漂移和概念漂移等。

数据传输的演化更新集 $U=\{u_1, u_2, \dots, u_k\}$ ，每个更新 u_i 都有一个方法 $y_j \in \{0, 1, 2, \dots, y\}$ ，表示校准模型、和主动学习或迁移学习等方式。

用户需求服务集 $T = \{t_1, t_2, \dots, t_z\}$ ，表示用户提出动态的需求和任务；是一个动态的集合。

C 为算力存储资源、网络资源，以及数据隐私等各种资源约束阈值； c 表示每个流程中对应消耗和占据的资源函数。

体系化AI总体技术方案

考虑体系化人工智能的内部流程和逻辑，进一步可以将 f 分解为以下几个子函数：

$$\max f(g, h, i, j, k, \ell, o)$$

$$s. t. c(g, h, i, j, k, \ell, o) \leq C$$

其中：

端到端跨模态异构数据建模： $g(D, M)$ ；

模型学习机理的优化建模： $h(D, M, P, U)$ ；

模型的原子化表征和建模： $i(M, E)$ ；

模型的标准规范入库： $j(M, S)$ ；

数据漂移的优化建模： $k(D, P, F)$ ；

模型数据传输的演化更新： $\ell(M, F, U)$ ；

运行架构优化建模： $o(A, T, C)$ ，如何在资源约束和安全可信的前提下的完成整体业务流程。



■ 原子化

■ 端到端优化

■ 基于大模型的调度体系

体系化AI原子模型

$$M_i = \{intro_i, funtion_i, input_output_i, interface_i, adapter_i, performance_i, constrain_i\}$$

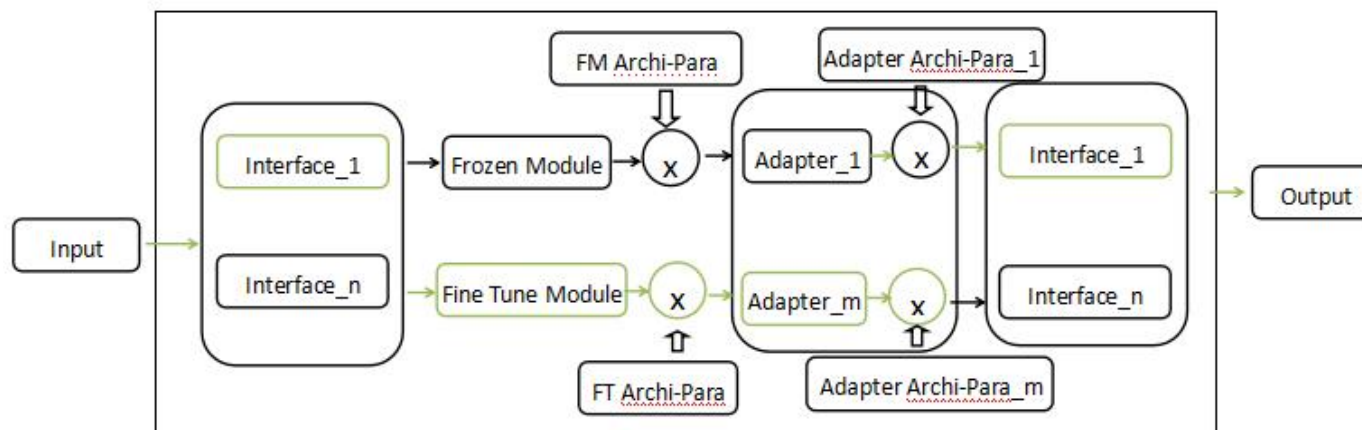


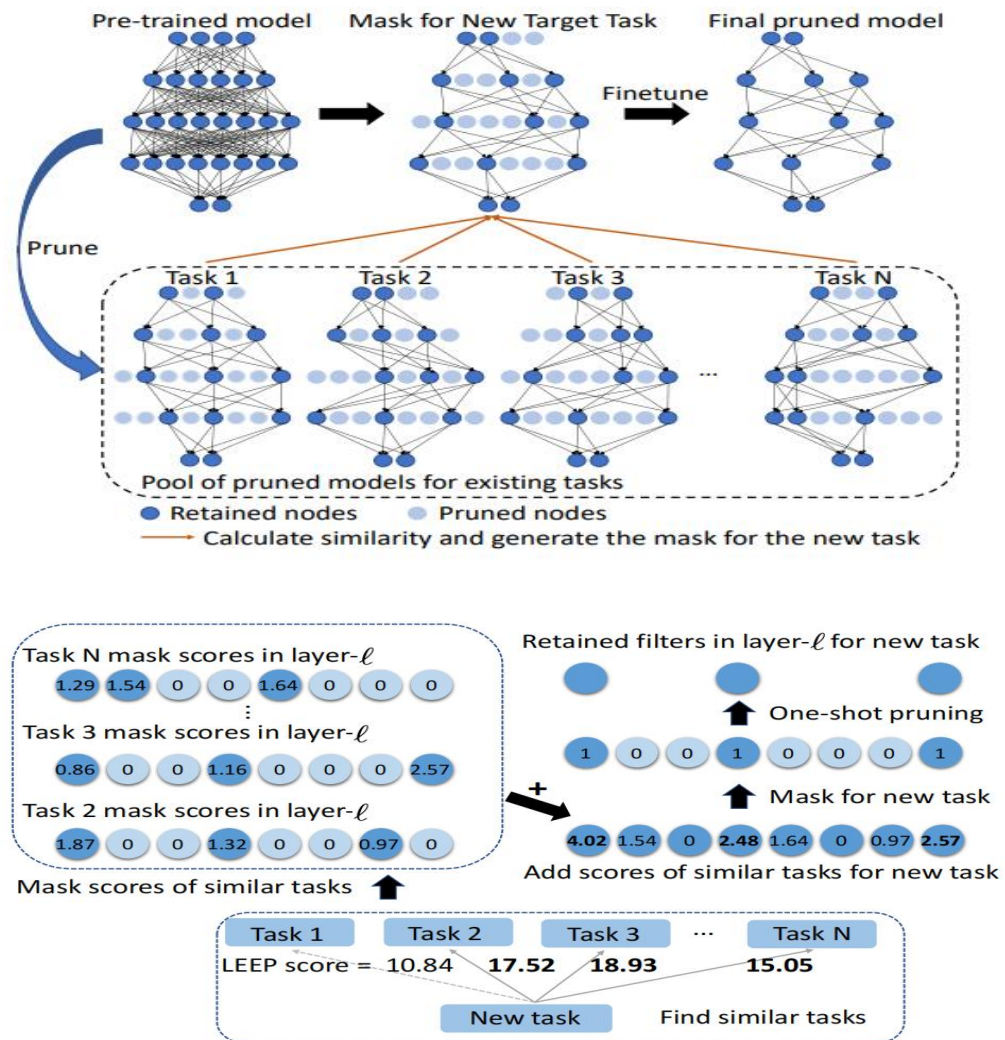
Fig1. 体系化原子模型示意图 (HAI Atomic Model, HAI-AM)其中绿色部分为其中一条可能的路径

原则

- (1) 重用度高
- (2) 输入输出清晰, 功能清晰
- (3) 不过于细小导致模型协同成本高于计算成本
- (4) 适合于独立攻关
- (5) 和基础模型能力互补

模型介绍	模型的类型 (通用型, 特定任务型), 模型结构及参数量, 应用领域, 模态, 构建时长及机构
功能描述	主要完成的功能描述和列表
输入输出	输入输出样例可以是一对多, 一对一, 多对一等组合
接口	模型的前向和后向接口及信息
适配器	适配器选择
性能准确率	性能, 准确率, 及测试方法
约束	应用环境的约束条件。

原子化方式



基础模型的功能解耦

Table Comparison between SMSP with baselines for ViT.

Methods	Accuracy(%)	FLOPs(T)	Training Iterations
AMP	89.83±0.46	81.50	1000
UVC [28]	81.32±0.87	26.73	100
PoWER [5]	78.61±1.32	20.86	100
SMSP(ours)	90.24±0.35	3.25	100

- ✓ **Automatic Mask Pruning (AMP)**: automatically identify task-specific filters/nodes for different tasks in the pre-trained model.
- ✓ apply the **Log Expected Empirical Prediction (LEEP)** which is used to evaluate the transferability of representations learned by the source task to the target task.
- ✓ **Scalable Mask Selection Pruning (SMSP)**: fast-adapt the pre-trained model to downstream tasks.

原子化方式

Decouple one Model into Atomized networks

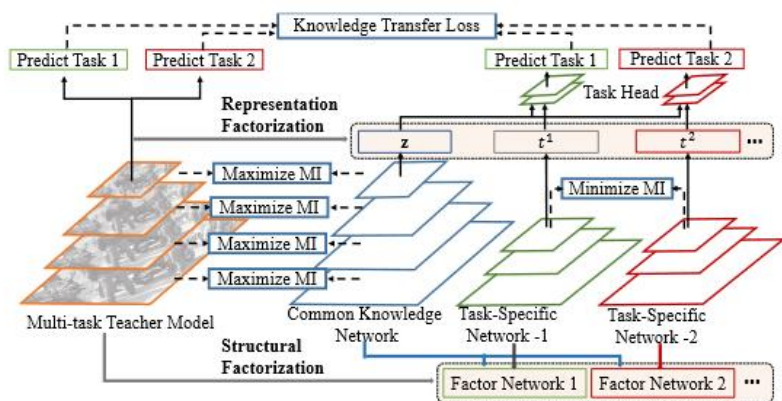


Fig. 2: The overall framework of the proposed knowledge factorization. The factor networks are trained to mimic the prediction of the teacher. The CKN learns to maximize the mutual information between input and its features, whereas the TSNs are dedicated to minimizing the task-wise mutual information.

“Factorizing Knowledge in Neural Networks”, Xingyi Yang, Jingwen Ye, Xinchao Wang, ECCV 2022.

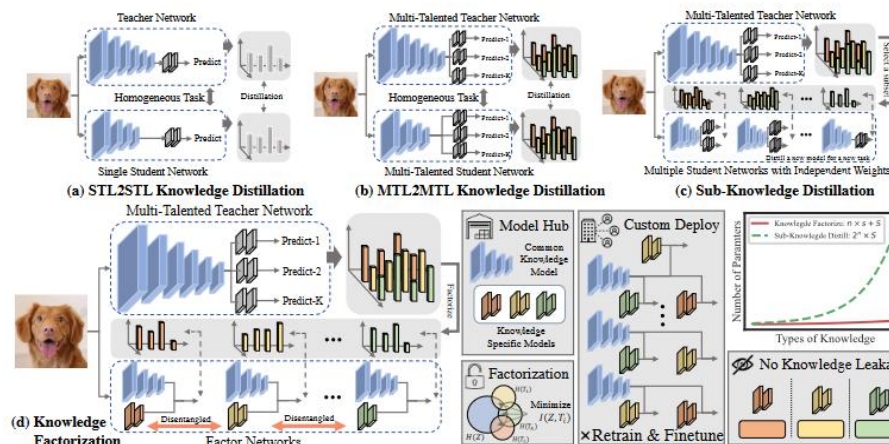


Fig. 1: Illustration of (top) 3 types of Knowledge Distillation and (bottom) our proposed Knowledge Factorization. (a) Single-Task Learning to Single-Task Learning (STL2STL) KD refers to distill a single-task student from a single-task teacher, (b) Multi-Task Learning to Multi-Task Learning (MTL2MTL) KD stands for distilling a multi-task student from a multi-task teacher and (c) Sub-Knowledge Distillation distill a subset of the teacher’s knowledge to its student model.

- 知识分解：包含结构分解和表征分解
- 每个因子网络包含两部分：通用知识网络（CKN）和特定任务网络（TSN）
- 一种新的信息衡量指标-InfoMax Bottleneck (IMB)，使输入和通用特征间互信息最大（最大限度保留大模型的通用知识），使不同特定任务特征间互信息最小（使特定任务网络之间尽可能解耦）。

原子化方式

模型蒸馏

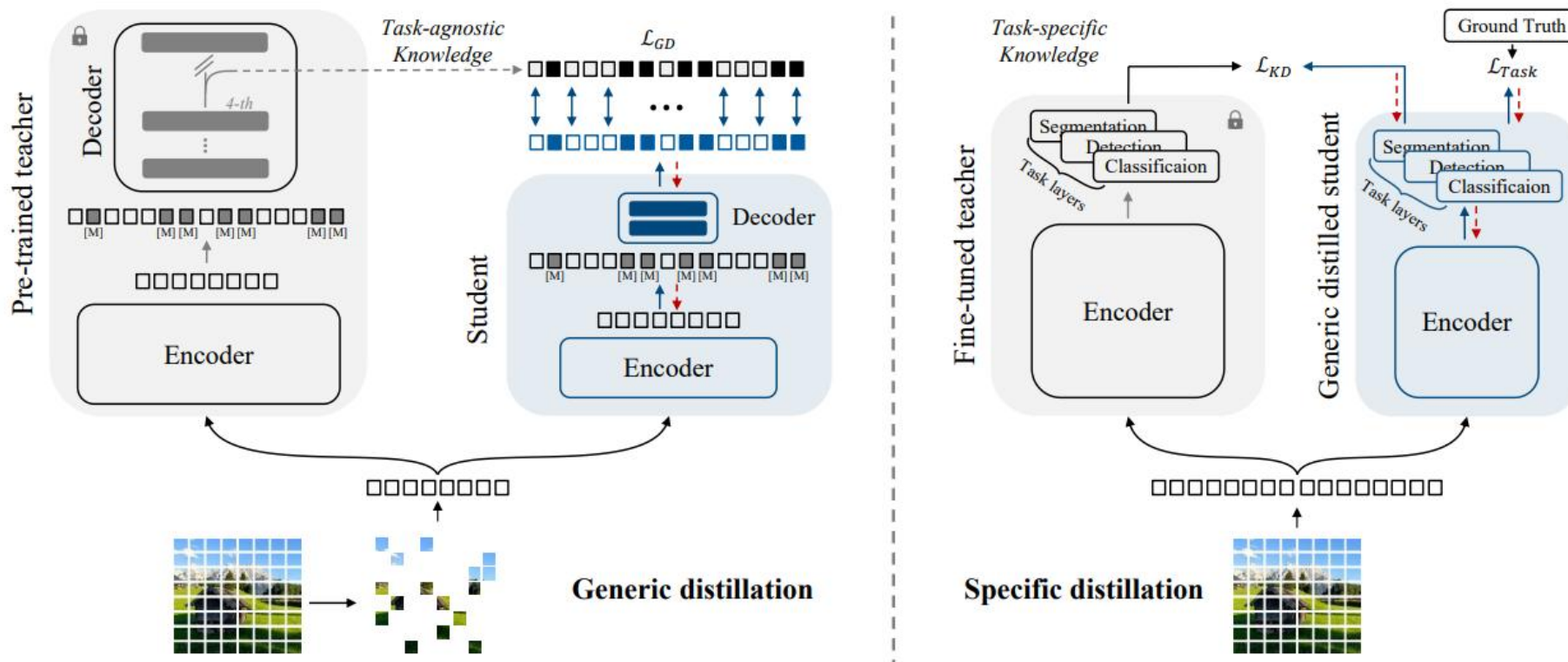
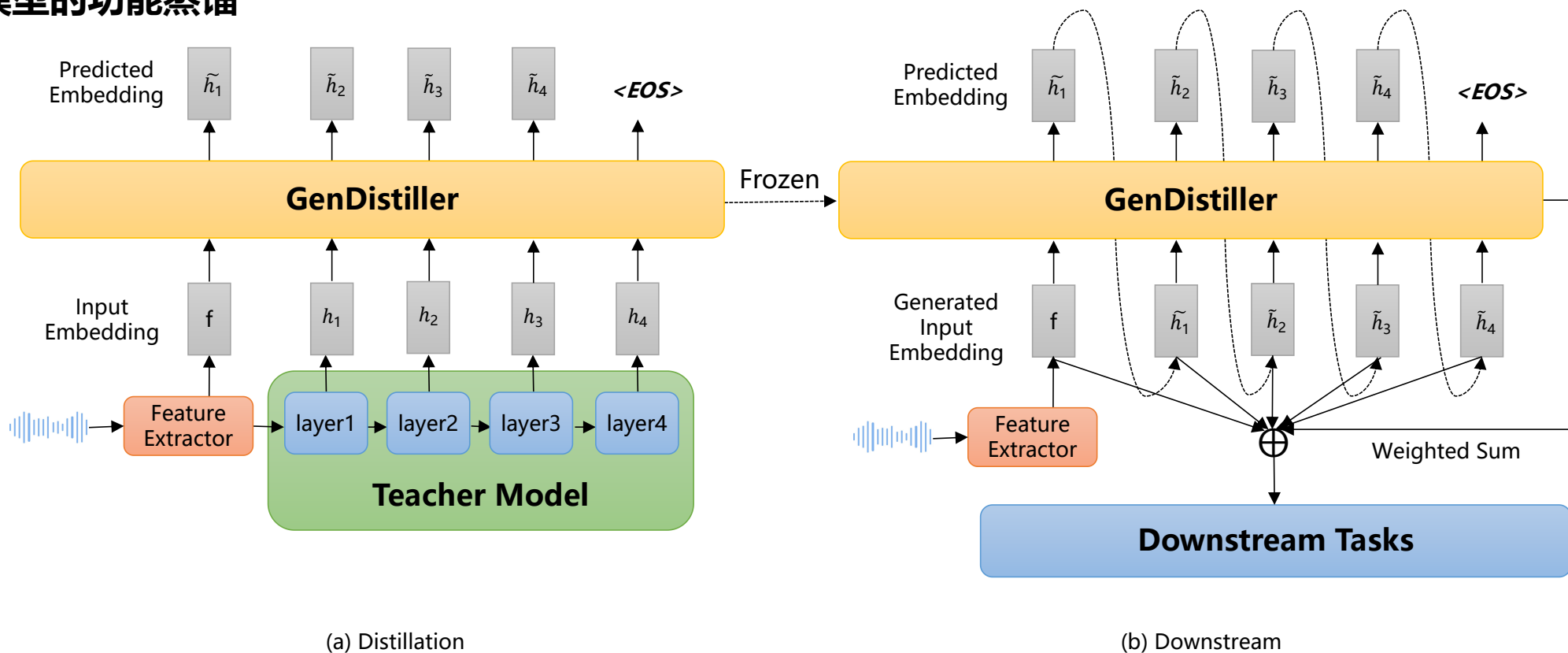


Figure 2. Diagram of the proposed generic-to-specific distillation (G2SD). [M] denotes mask token. In the generic distillation stage (left), masked images are converted to patches and fed to both the teacher and student encoders for feature extraction. Feature predictions of the student decoder are aligned with those of the teacher at both visible and predicted patches. In the specific distillation stage (right), student models are trained to have consistent predictions with teacher models fine-tuned on the specific task.

"Generic-to-Specific Distillation of Masked Autoencoders". Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, & Jianbin Jiao, Qixiang Ye. (2023). 15996-16005. 10.1109/CVPR52729.2023.01535

原子化方式

基础模型的功能蒸馏



GenDistiller: Distilling Pre-trained Language Models based on Generative Models, Y.Gao, Shilei Zhang, Zihao Cui, Chao Deng, Junlan Feng*. Archive-2023

原子化方式

基础模型的功能蒸馏

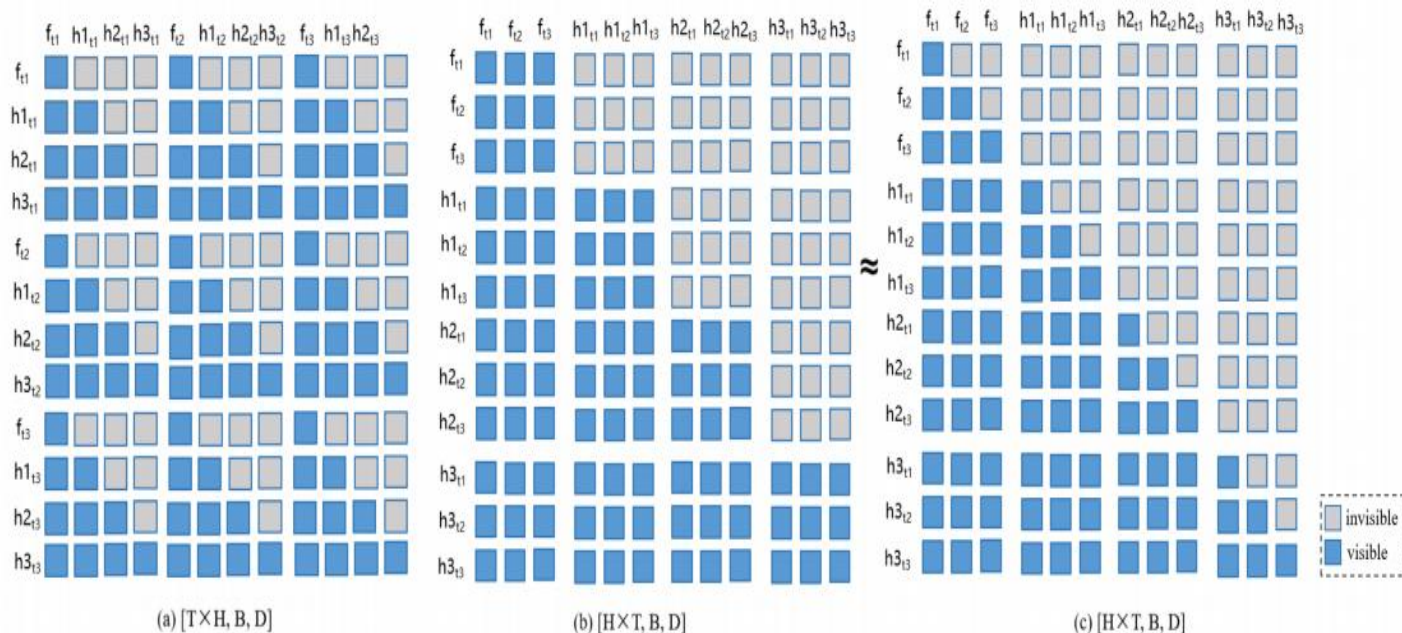


Figure 2: Different attention mechanisms via different flatten manners. a) flatten to $T \times H$ length for attention; b) flatten to $H \times T$ length for attention; c) the approximation of b).

Two-dimensional Attention Mechanism:

$$\mathbf{f} \in \mathbb{R}^{T \times B \times D} \longrightarrow \mathbf{M} \in \mathbb{R}^{H \times T \times B \times D}$$

T is the frame numbers related with the utterance length, B refers to the batch size, D denotes the feature dimension, H refers the numbers of hidden layers to be predicted plus the original feature. $[H \times T, B, D]$

Distillation Loss

$$\begin{aligned} \mathcal{L}^{(l)} &= \mathcal{L}_{\ell_1}^{(l)} + \lambda \mathcal{L}_{\cos}^{(l)} \\ &= \sum_{t=1}^T \left[\frac{1}{D} \left\| \mathbf{h}_t^{(l)} - \hat{\mathbf{h}}_t^{(l)} \right\|_1 - \lambda \log \sigma \left(\cos \left(\mathbf{h}_t^{(l)}, \hat{\mathbf{h}}_t^{(l)} \right) \right) \right], \end{aligned} \quad (1)$$

体系化AI OS

体系化人工智能(Holistic AI)主要研究对人工智能技术进行体系化重构所需的理论、技术、机制、范式和框架,其主要特征为AI服务大闭环、AI能力原子化重构、网络原生AI及安全可信AI。体系化人工智能依托泛在的网络和AI算力,在开放环境中实现对AI能力进行灵活且高效的配置、调度、训练和部署,以满足日益丰富的数智化业务需求,同时确保AI业务可信可控安全



体系化AI 原子能力

语音

面向智能助手、业务质检、客户服务等领域,构建全面先进的智能语音算法体系

编排

图像

面向图像、视频的多种类、多样化应用...

NLP

面向人机交互、文本内容分析场景,构...

网络智能化

面向智能助手、业务质检、客户服务等...

The flowchart details the atomic capabilities for speech processing, starting from '起点' (Start) and ending at '终点' (End). Key steps include: 语音特征提取, 语音端点检测 (VAD) -能量, 非语音检测-嘟嘟音, 非语音检测-彩铃音, 输入归一化-语音, 格式转换-FFmpeg, 格式转换-sox, 音频压缩算法, 语音端点检测 (VAD) -模型, 人物画像-性别, 声纹识别, 语音识别通用层-hybrid model, 语音识别通用层-v1, 语音识别通用层接口, 语音识别通用层-v2, 语音识别通用层-v3, Rich Transcription, 语音情感识别, 角色分割, 语音识别适配层-语言模型, 语音识别适配层-热词列表, 语音识别适配 (one-best\N-best\lattice\FSM), 语音纠错.

核心科学问题

- 原子化



- 端到端优化

- 基于大模型的调度体系

组合学习的训练方法与挑战

搜索空间巨大： 层级搜索、免训练 (training free)

端到端闭环数据稀疏： 无监督

参数量和内存消耗大： 适配器、蒸馏、剪枝

接口复杂： 维度一致、梯度连续

组合学习的训练方法与挑战 -1

端到端闭环数据稀疏

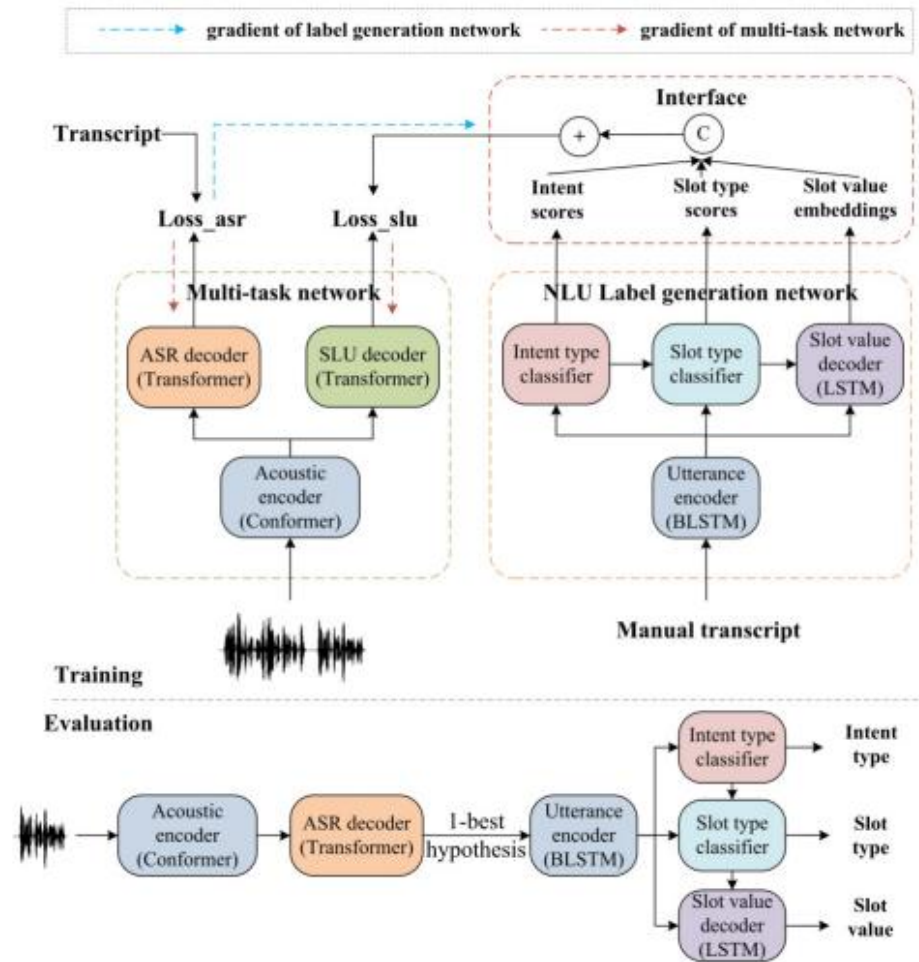


Figure 1: The proposed MAXL based SLU model.

Table 1: The interfaces between the two networks and their attributes.

	Fixed length	Gradient flow
List	×	×
Sequence	✓	×
NER tag	✓	×
Softmax	×	✓
Sum of softmax	✓	✓
Append intent and slot types	✓	✓

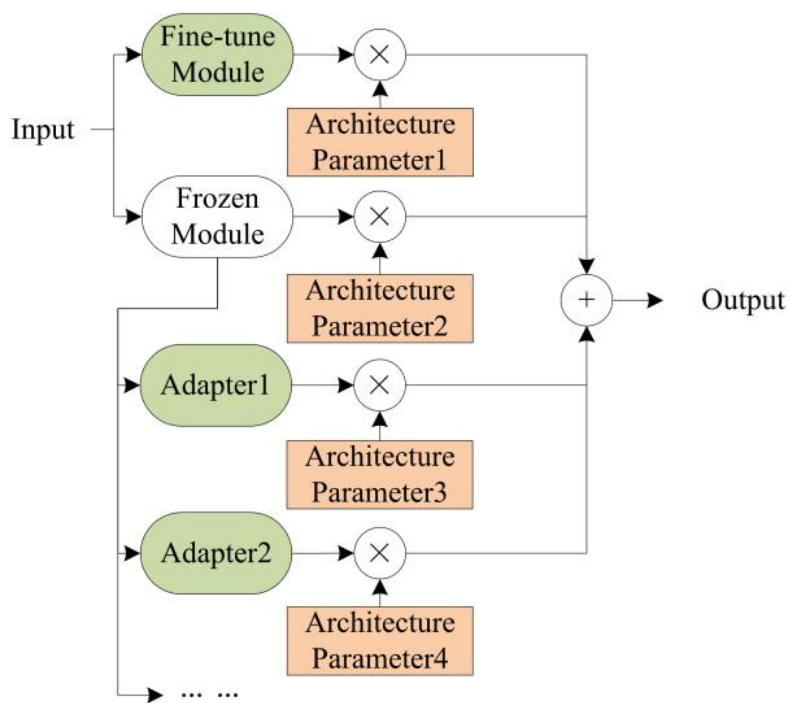
Table 2: The performance comparison between the proposal and baselines.

	CER	F1-score
Baseline	28.76	52.00
End-to-end	27.75	53.12
Fine-tuned	21.45	56.33
Proposal(MAXL)	21.21	58.27
Proposal(First-order)	21.15	59.64
Multi-task	21.30	59.25

"Meta Auxiliary Learning for Low-resource Spoken Language Understanding", Yingying Gao, Junlan Feng*, Chao Deng, Shilei Zhang. Interspeech 2022

组合学习的训练方法与挑战 -2

参数量大



(b) Nas based Fine-tuning and Adapter (NFA) Module

Figure 1: The architecture of NA and NFA module. Only the green and orange boxes are updated during training, in which the green parts denotes trainable network parameters and the orange parts are architecture parameters.

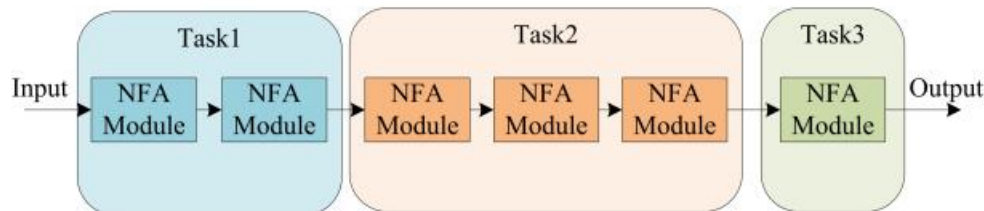


Figure 2: The illustration of an adaptive training framework based on NFA for cascaded multi-task training.

Table 2: The performance of full fine-tuning and the proposed NFA

Adapt Method	Train Param.	Selected Param.	NLU CER
Full Fine-tuning	189.25M	189.25M	12.42%
BA+fine-tuning NLU	14.15M	14.15M	14.13%
GA+fine-tuning NLU	14.17M	14.17M	18.56%
NFA	81.39M	65.11M	14.57%
+parameter control	81.39M	16.48M	12.81%
+two stage	81.39M	16.48M	12.32%

"Cascaded Multi-task Adaptive Learning Based on Neural Architecture Search", Yingying Gao, Shilei Zhang, Zihao Cui, Chao Deng, Junlan Feng*. Interspeech 2023

白盒模型的联合及优化 - 1

Fuse Multiple Models into one target model

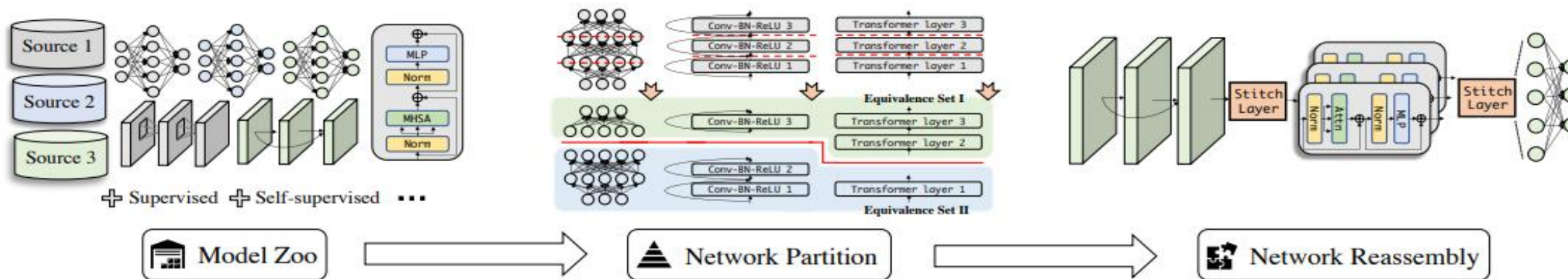


Figure 1: Overall workflow of DeRy. It partitions pre-trained models into equivalent sets of neural blocks and then reassemble them for downstream transfer. Both steps are optimized through solving constrained programs.

- 多个神经网络层形成一个功能块
- 功能相似网络：输入相似时，输出相似
- 将一个网络分成多个功能块，相似的功能块形成一个集合，这个集合称为：等同网络块集合

"Deep Model Reassembly", Xingyi Yang, etc. NeurIPS 2022

白盒模型的联合优化 -2

Stitch Multiple Big Models into one target model

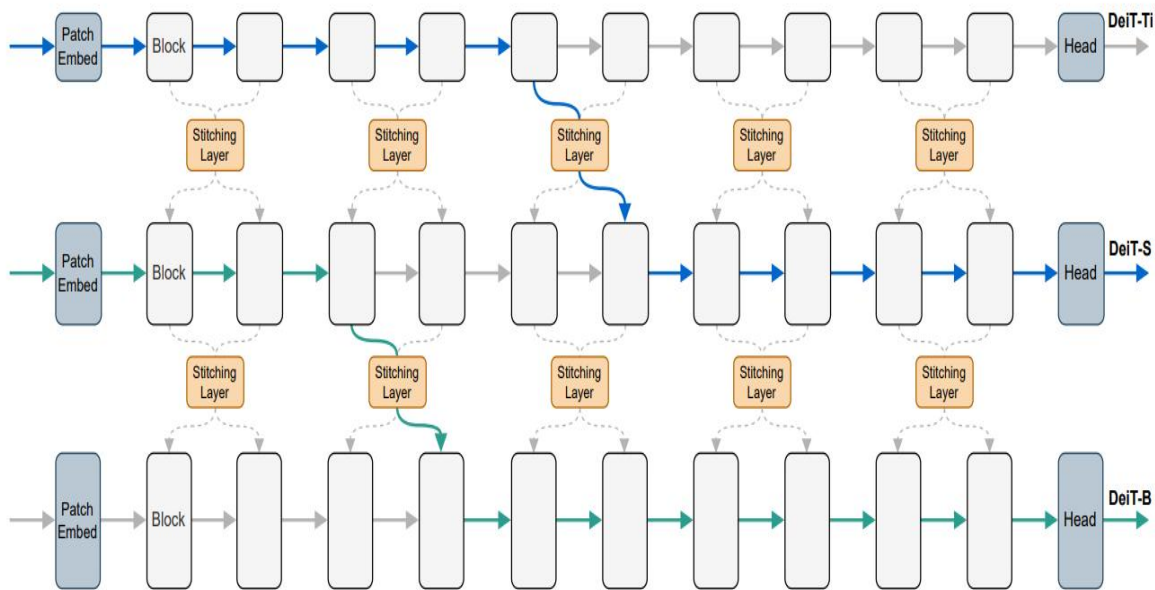


Figure 3. Illustration of the proposed **Stitchable Neural Network**, where three pretrained variants of DeITs are connected with simple stitching layers (1×1 convolutions). We share the same stitching layer among neighboring blocks (e.g., 2 blocks with a stride of 2 in this example) between two models. Apart from the basic anchor models, we obtain many sub-networks (stitches) by stitching the nearest pairs of anchors in complexity, e.g., DeiT-Ti and DeiT-S (the blue line), DeiT-S and DeiT-B (the green line). Best viewed in color.

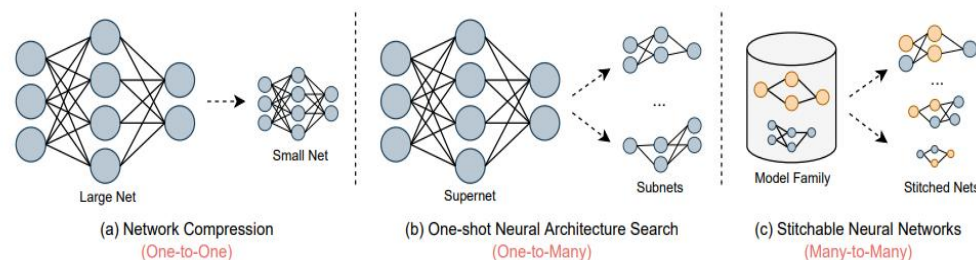
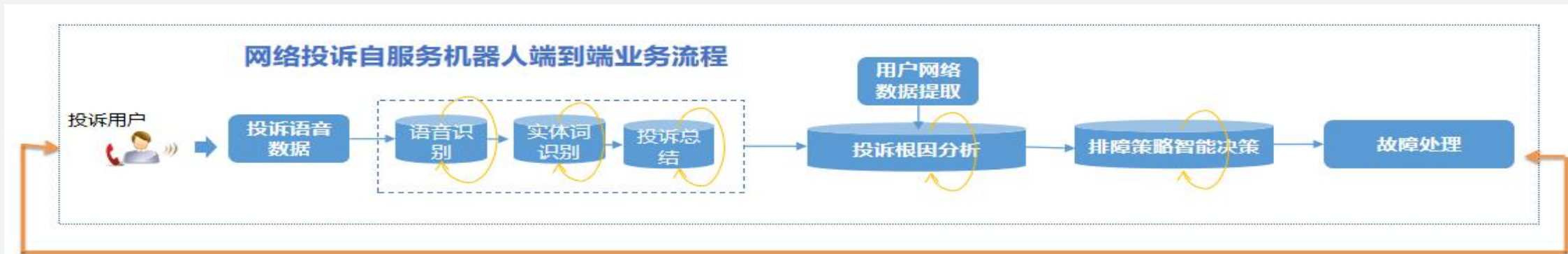


Figure 1. Compared with previous scalable deep learning frameworks. (a) Network compression shrinks a large network into a small one by techniques such as pruning, quantization and knowledge distillation, etc., which is a one-to-one mapping. (b) One-shot neural architecture search first trains a supernet that supports diverse architectural settings and then specializes a subnet given the target resource constraint during deployment, which is a case of one-to-many. (c) Our proposed Stitchable Neural Network directly stitches the off-the-rack family of pretrained models and quickly obtains new networks for efficient model design and deployment in a novel many-to-many paradigm.

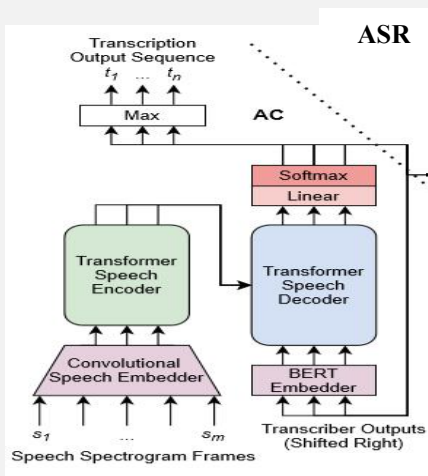
“Stichable Neural Networks”, Zizheng Pan Jianfei Cai Bohan Zhuang, Archive-2023

黑盒模型的联合优化: Holistic Neural Network (Holi-NN)

网络问题投诉 级联优化

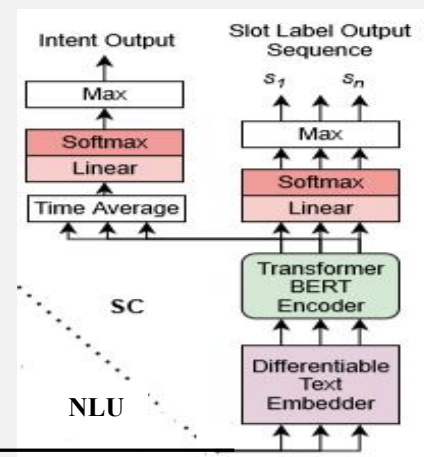


语音识别+自然语言理解 级联优化【12】



Interfaces

- Top-K:** Token Embeddings, softmax values
- Matrix multiply:** softmax output * matrix
- Gumbel softmax:** smooth distribution



黑盒模型联合优化: Holi-NN with Small Models

Fuse Multiple Models into one target model

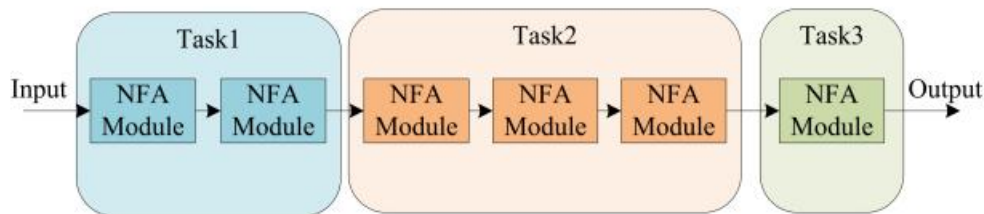
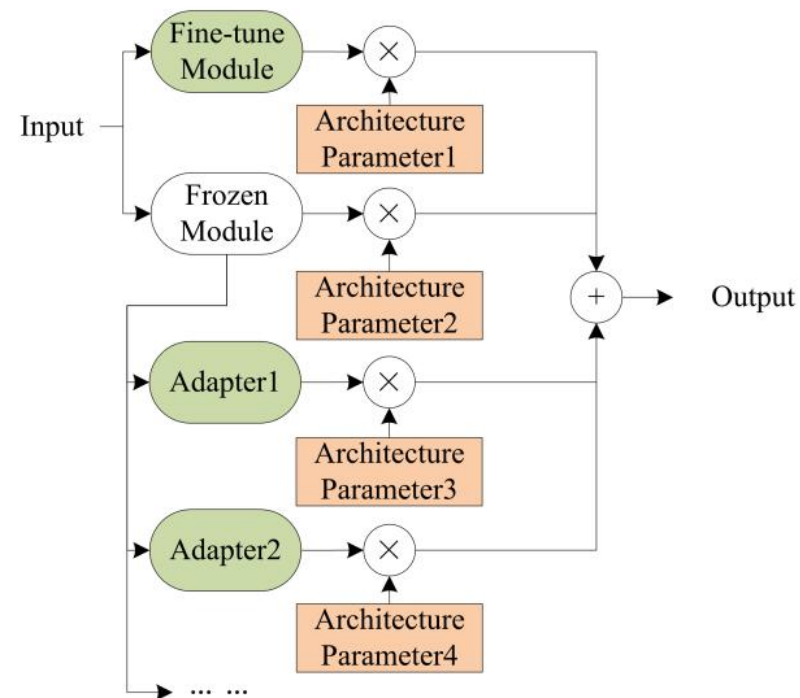


Figure 2: The illustration of an adaptive training framework based on NFA for cascaded multi-task training.



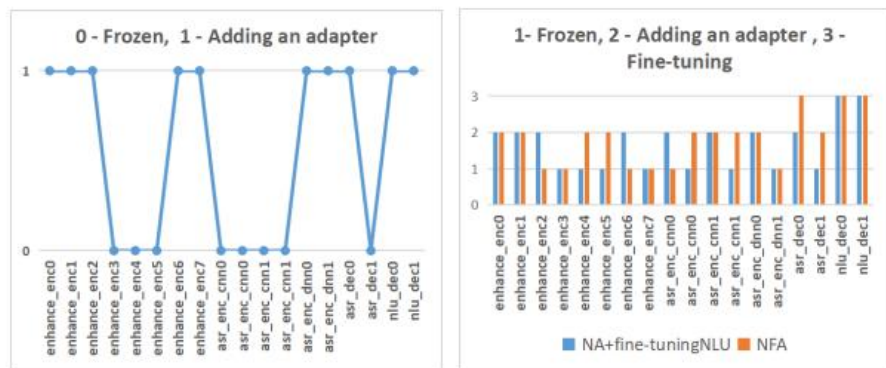
(b) Nas based Fine-tuning and Adapter (NFA) Module

Figure 1: The architecture of NA and NFA module. Only the green and orange boxes are updated during training, in which the green parts denotes trainable network parameters and the orange parts are architecture parameters.

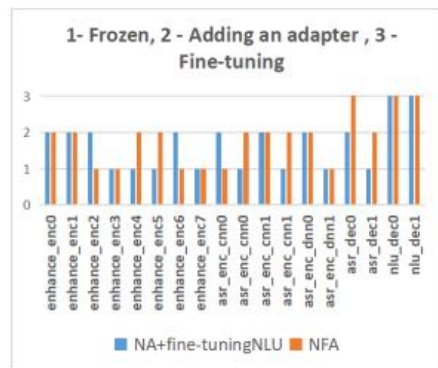
"Cascaded Multi-task Adaptive Learning Based on Neural Architecture Search", Y.Gao, Shilei Zhang, Zihao Cui, Chao Deng, Junlan Feng*. Interspeech 2023

黑盒模型联合优化: Holi-NN with 语音增强+语音识别+自然语言理解

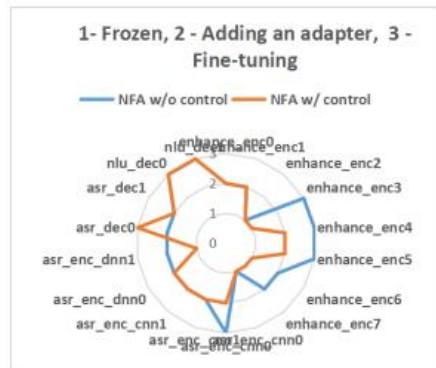
Cascade three models - speech enhancement , ASR, NLU - with Bottleneck Adapter



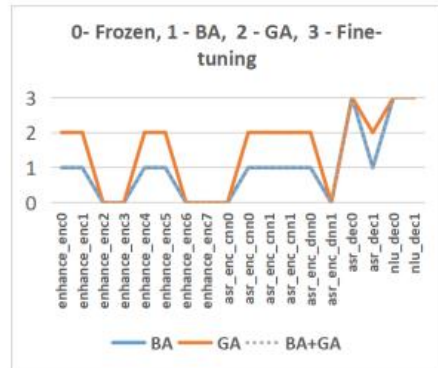
(a) NA



(b) NA+fine-tuningNLU and NFA



(c) NFA with and without parameter control



(d) NFA with different adapters

Figure 3: The searched architecture of the cascaded SE-ASR-NLU framework based on (a) NA, (b) NA+fine-tuningNLU and the proposed NFA, (c) NFA with and without parameter control and (d) NFA with different adapters.

Table 2: The performance of full fine-tuning and the proposed NFA

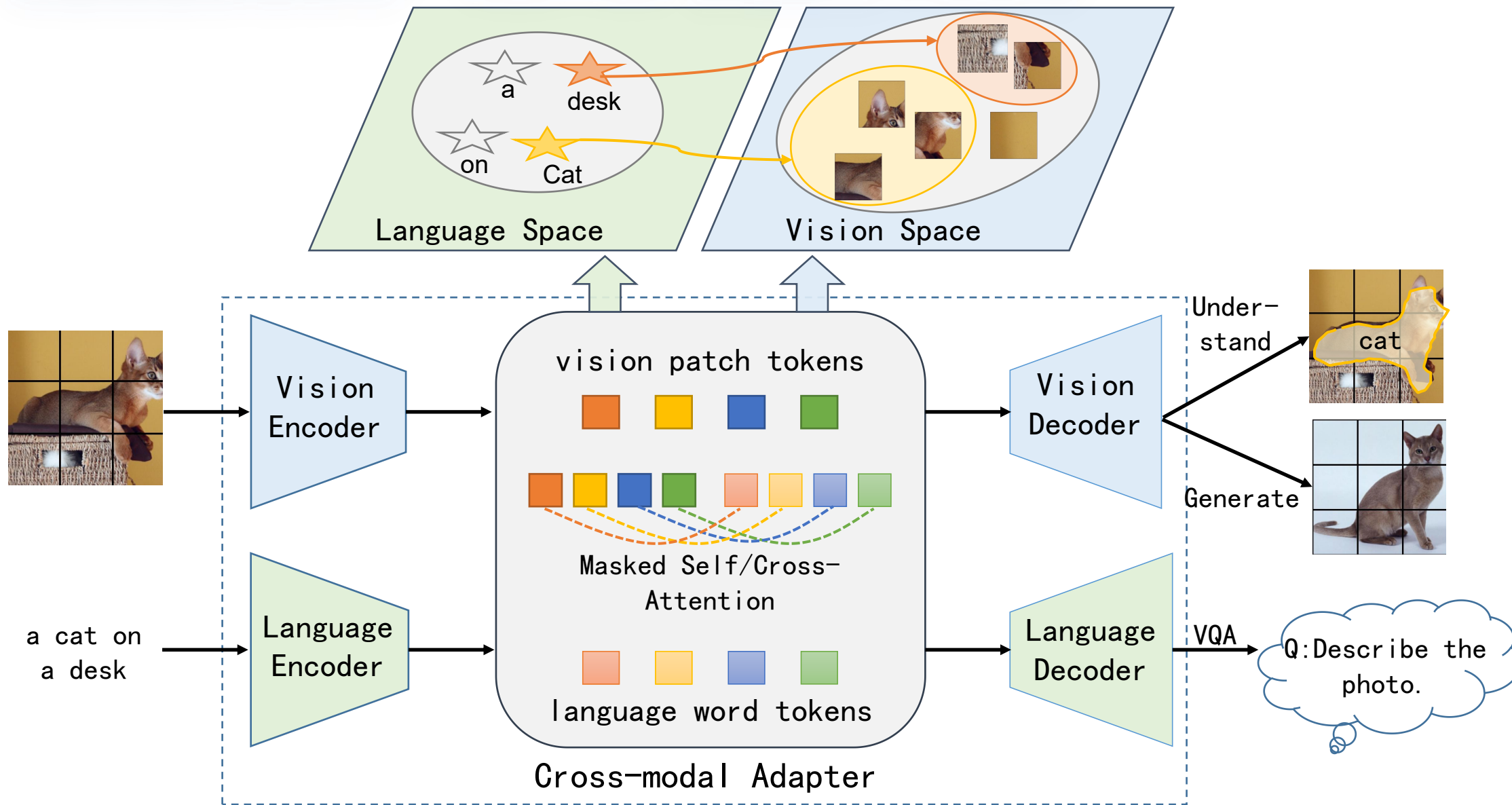
Adapt Method	Train Param.	Selected Param.	NLU CER
Full Fine-tuning	189.25M	189.25M	12.42%
BA+fine-tuning NLU	14.15M	14.15M	14.13%
GA+fine-tuning NLU	14.17M	14.17M	18.56%
NFA	81.39M	65.11M	14.57%
+parameter control	81.39M	16.48M	12.81%
+two stage	81.39M	16.48M	12.32%

Y. Qian, X. Gong, and H. Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child asr," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, 2022.

"Cascaded Multi-task Adaptive Learning Based on Neural Architecture Search", Y.Gao, Shilei Zhang, Zihao Cui, Chao Deng, Junlan Feng*. Interspeech 2023

黑盒模型联合优化: Holi-NN 多模态上下文学习



核心科学问题

- 原子化
- 端到端优化



- 基于大模型的调度体系

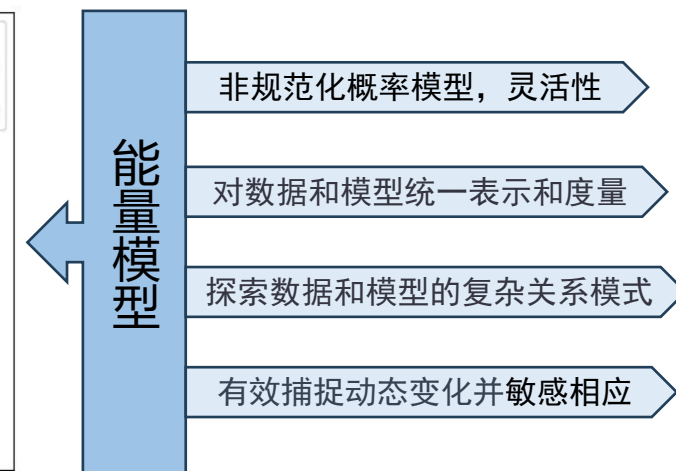
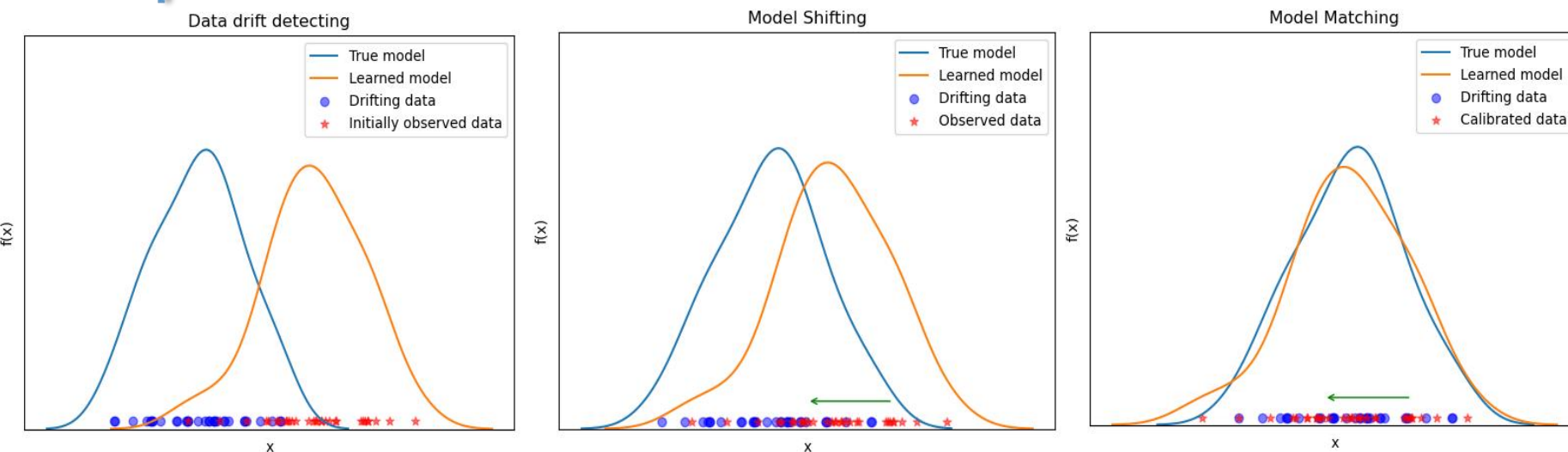
安全可靠之开放动态环境动态评估



体系化人工智能：将人工智能技术从单点应用向系统集成转变，形成具有自主学习、自主决策、自主协同等特征的人工智能系统。

体系化人工智能服务的重要基础：AI数据、模型、能力、业务的安全可信

安全可信：在开放动态环境下，保证人工智能系统的可控可靠、透明可释、隐私保护、明确责任和多元包容




欢迎使用算网智脑 HAI-OSS


基于体系化AI(HAI)的智能化服务运营

 进行业务咨询

例：“海量数据需要长期存储，要求成本低，安全可靠，不需要实时访问，有什么解决方案吗？”

 获取业务建议

例：“已经开通了安全帽检测服务，明天有新工程项目启动，怎么提升服务容量？”

 获取部署方案

例：“需要部署病毒防护标准版的安全服务，客户接入方式为移动云，如何部署能做到时延优先？”

算网智脑提供算网业务的简单入口，根据用户输入的业务请求，智能引导并发掘用户底层需求，并根据已经感知的算网资源和业务指标信息，实时计算和分析，提供业务部署和调整建议

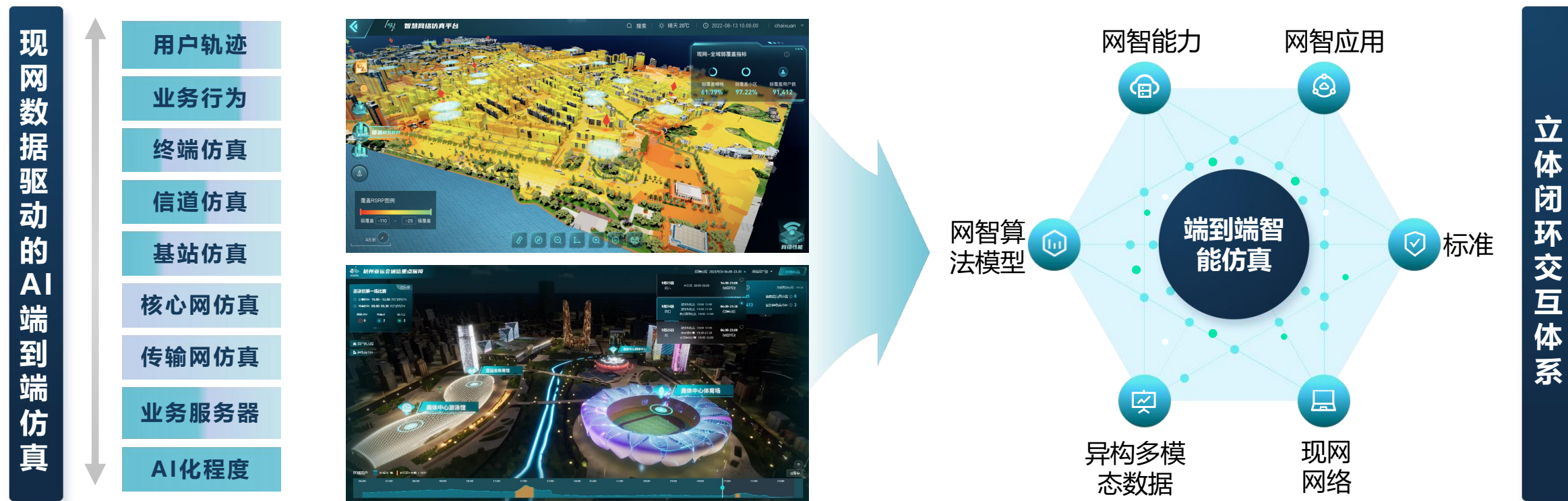
AI

请提出您的需求

回到首页

网络的智能化仿真技术

针对主流通信网络仿真系统难以精准模拟现网的难题，提出首个融合常规网络仿真、AI仿真和物理空间孪生的仿真框架，攻关多项AI仿真技术，使能仿真系统更贴近现网，基于该框架成功申请并构建“智慧网络国家新一代人工智能开放创新平台”。



支持10余种场景仿真，孵化应用31省落地，为亚运通信网络服务保驾护航



于都县长征纪念馆/毛主席...

赣州市于都县长征纪念馆/毛主席旧居附近约2平方公里的居民区，有**个建筑物，平均高度*米，密集程度*

 5G小区	 常住人口	 任务
70个	10万	27个

范围 (赣州市于都县长征纪念馆/毛主席旧居附近)
约2平方公里的居民区

进入场景 >