AI与通信融合

vivo通信研究院
2023年10月

目 录

01

第一章

引 言

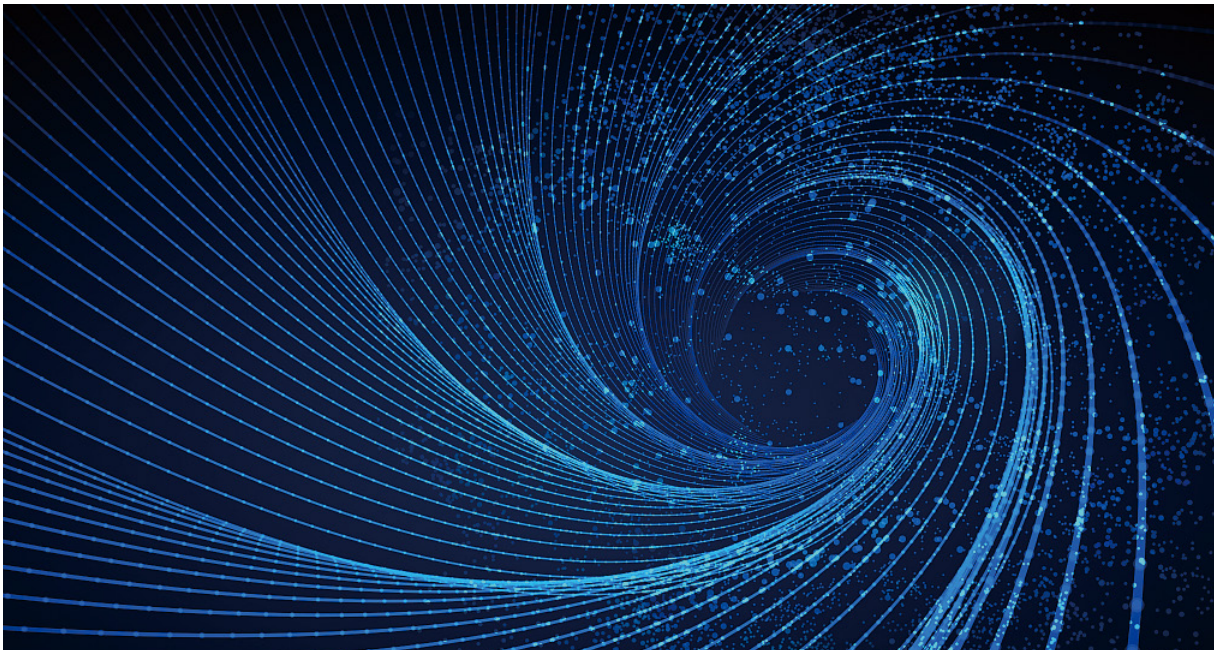
2020年和2022年，vivo通信研究院先后发布了《数字生活2030+》、《6G愿景、需求与挑战》和《6G服务，能力与使能技术》三本白皮书，分别对6G的美好愿景和使能技术进行了畅想与阐述。近几年，人工智能（Artificial intelligence，AI）在移动通信领域的应用如雨后春笋般涌现，AI与移动通信的融合有望推动未来通信范式的演进和网络架构的变革。对此，本白皮书详细地阐述了AI赋能6G的用例场景，并提出多维度的AI与通信融合的设计原则，期望为行业6G AI的发展和落地添砖加瓦。

02

第二章

AI与通信融合的驱动力与用例

2.1 AI与通信融合的驱动力

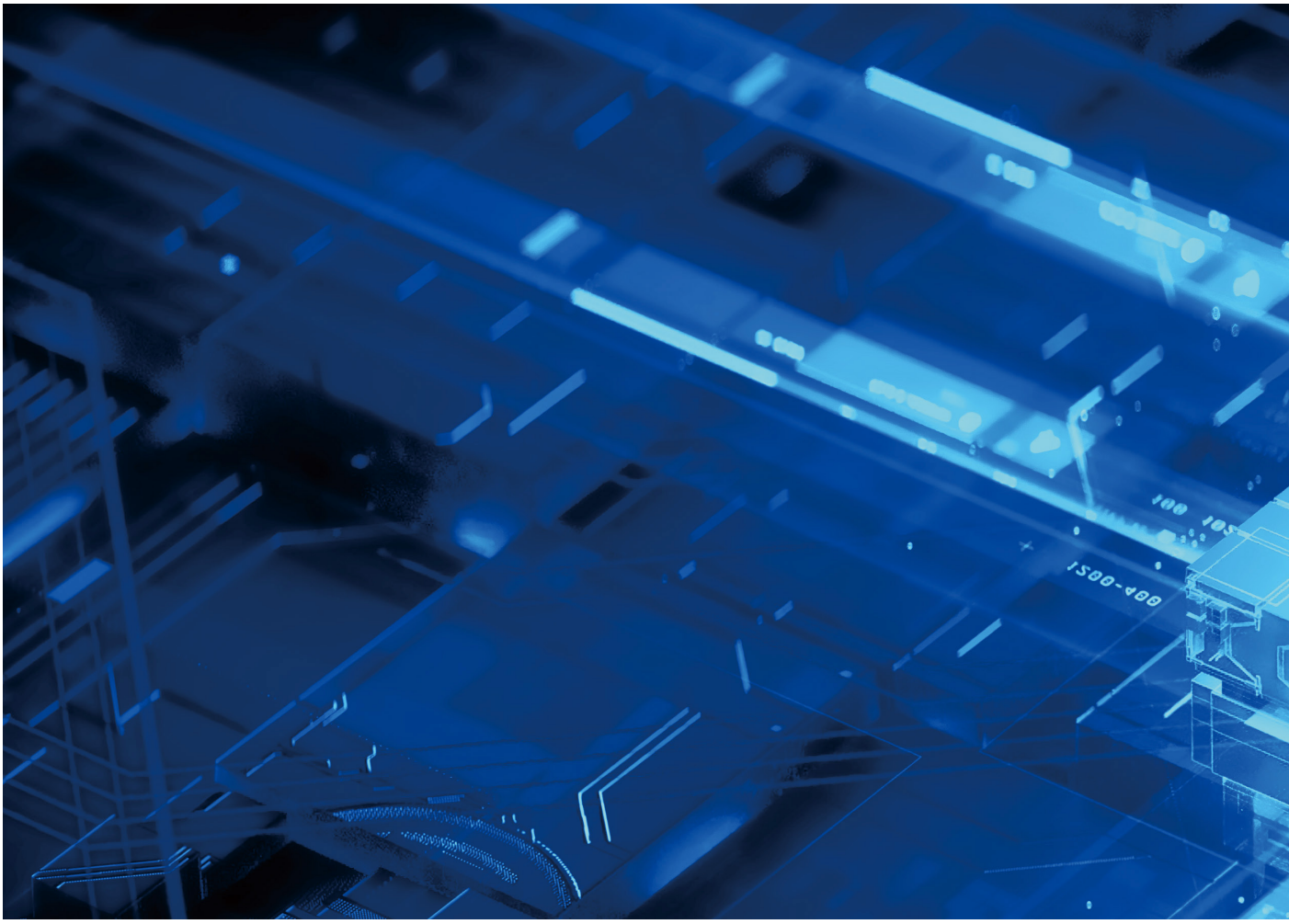


近十年，以神经网络为代表的AI技术飞速发展，正在掀起新一轮的技术革命。AI是一项数据驱动的技术，通过神经网络等工具可以从大量数据中提取特征，进一步完成判断、分类、预测、决策、内容生成等动作。目前，AI已经成功解决了一系列从前难以处理的问题，在计算机领域的图像识别、自然语言处理，机器人领域的运动控制、轨迹规划等多个方向获得了巨大的成功[1]。

与此同时，移动通信系统也在持续高速发展中，不断地向更大吞吐、更低时延、更高可靠性、更大连接数、更高频谱利用率等方向演进。随着移动通信指标要求的逐渐提高，在有些技术方向和新场景中，传统方法遇到了瓶颈。对此，AI有望提供更加高效的解决方案。

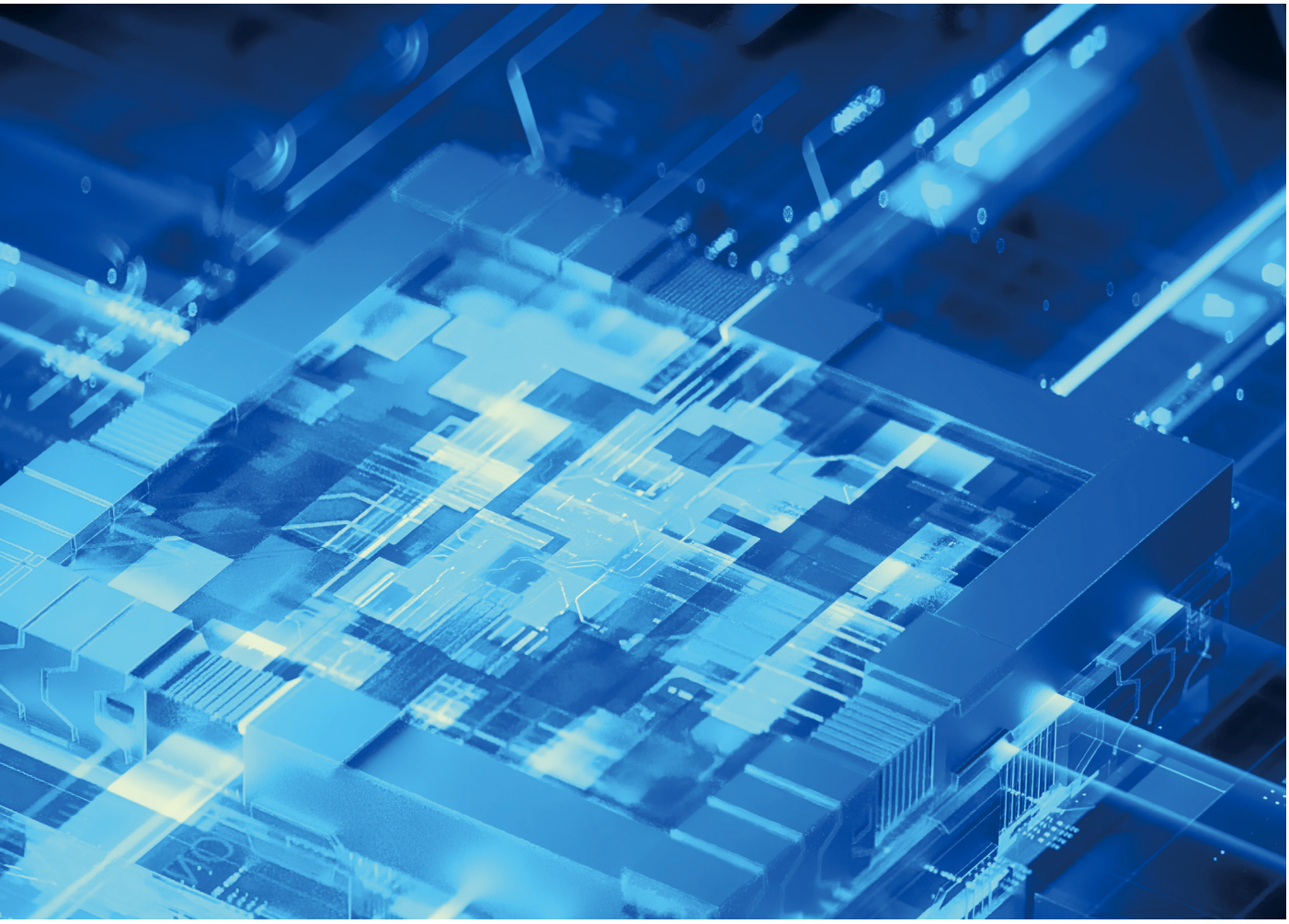
基于学界和产业界的研究，我们总结了AI在移动通信系统中主要发挥作用的方向。我们认为，AI在以下三个方向或场景中可以发挥重要价值：

（1）通信系统中传统方法不易提取的隐含关系、特征或知识（尤其是局部的）对特定功能有显著影响的场景，如实际信道对信号的接收性能的影响、信道的变化特性、用户位置与无线信道的关系、功率放大器对信号的影响、业务流量的变化特性等。而AI可以通过神经网络等工具从大量移动通信数据中提取隐含的关系、特征或知识等，更准确地完成复杂问题的建模，从而提升通信系统性能。



（2）通信系统中传统方法在期望时间内不易获得闭式解或没有显式闭式解的问题，如全局的无线资源分配、多用户配对、覆盖优化、容量优化等。这一类问题在传统方案中一般以优化问题或组合遍历的形式存在，由于其解法复杂度很高所以一直是通信系统中的一类痛点。AI可以通过数据驱动或模型驱动的方式，映射出输入信息（包括状态、条件、历史结果等）与潜在解决方案的关系，直接给出对应问题的解或近似解，从而降低通信系统复杂度。

（3） 通信系统中多个相关模块联合优化的问题。目前的通信中不同功能模块的优化是独立进行的，联合优化多个相关的功能模块难度较大，如跨层优化、 多个多输入多输出 （Multiple Input Multiple Output, MIMO）相关信号处理模块的联合优化、 信源信道联合编码， 均衡译码联合设计等。AI 可以将多个相关的功能模块建模为一个神经网络，将复杂的多模块关联问题转换为简单的数据拟合或回归问题，得到接近全局最优的解决方案。



上述AI作为一种工具辅助移动网络的模式被称为AI for Network (AI4NET) 或对内AI服务。此外，未来移动网络也将为大量的AI业务提供相关支撑。随着AI在各类垂直行业的深耕，有越来越多的行业需要使用AI业务。其中很多AI业务需要移动网络提供支撑，对移动网络的传输速率、时延、容量、覆盖等方面提出了更高的需求。比如，医疗行业中AI辅助诊疗诊断需要传输高清的检查结果并进行低时延的交互；自动驾驶、无人机操控中需要大量传感器数据的回传和指令信息的实时下发；智慧工厂中机器人的动作、轨迹信息需要低时延高可靠地完成传输，以进行高效的指挥调度。因此，需要通过对移动网络进行合理设计、对通信、计算、存储资源进行合理分配来为AI业务提供保障。这种移动网络对AI提供支撑的模式被称为Network for AI (NET4AI) 或对外AI服务。需要指出的是，本白皮书的重点是对AI4NET进行研讨。

AI在移动通信系统中有非常多的应用方向，在提高系统性能和降低复杂度方面都有很大的发挥空间。下面我们通过多个具体的用例来呈现AI在移动通信中的潜在价值。

2.2 通信与AI融合的用例

首先，我们介绍AI在通信系统中传统方法不易提取的隐含关系、特征或知识方面的应用。其代表性用例有基于AI的信道状态信息（Channel State Information, CSI）反馈增强、信道估计、波束预测、定位增强、选网、信令风暴预测、用户移动性优化、射频非线性抑制等。

基于AI的CSI反馈增强

基于AI的CSI反馈增强可进一步细分为基于AI的CSI压缩与基于AI的CSI预测两个子用例。

基于AI的CSI压缩通过AI技术实现多维信道信息的压缩与重构，从而减少空口的CSI反馈开销或提高网络侧CSI的恢复精度。其主流方案是基于编码器-解码器神经网络结构实现的：部署于终端侧的编码器神经网络对CSI进行压缩，生成用于反馈的比特流；部署于网络侧的解码器神经网络对接收到的反馈比特流进行解码，从而重构出CSI。一般来讲，用于实现CSI压缩的编码器和用于实现CSI重构的解码器需要配对使用，即某个解码器仅能有效重构出对应的一个或多个编码器所压缩的CSI。图2-1给出了基于AI的CSI压缩相对于传统基于Release 16 Type II码本的CSI压缩（非AI）的频谱效率增益，具体仿真参数详见参考文献[2]。从图中可以看出，基于AI的信道反馈可以在相同反馈开销下获得约10%的频谱效率增益。

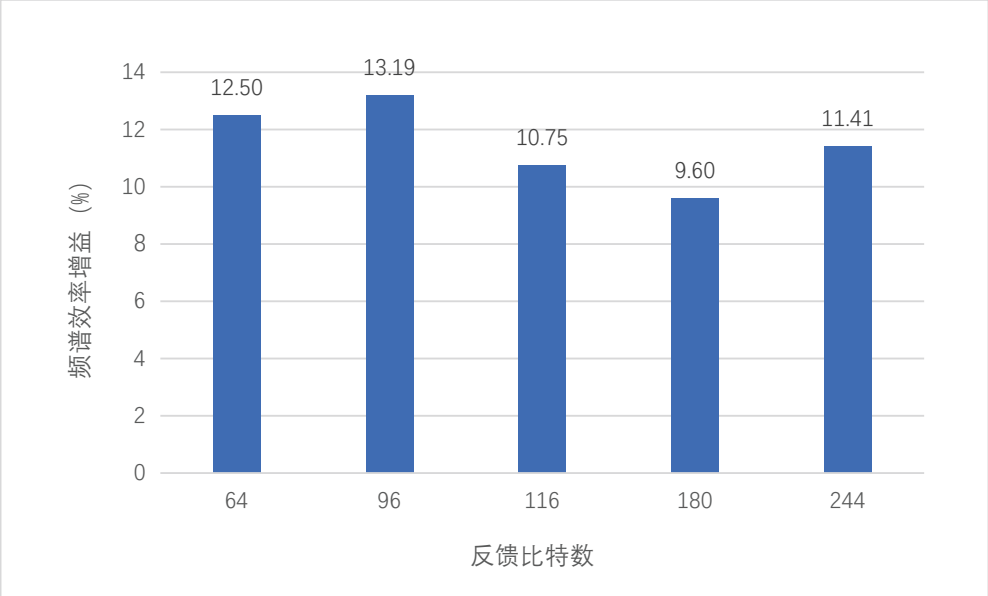


图2-1. 基于AI的CSI压缩相对于传统基于Release 16 Type II码本的CSI压缩（非AI）的频谱效率增益

CSI预测通过AI技术提取历史CSI中的时间变化规律，预测出未来时刻的CSI进行反馈，进而对CSI反馈过期问题进行补偿。其基本方案是以多个历史CSI作为神经网络的输入，再通过神经网络获得未来指定时刻的CSI。图2-2描述了在不同的资源利用率（Resource Utilization, RU）下基于AI的CSI相比于不做预测的方案和使用自回归（非AI）预测的方案可以获得的平均吞吐量增益，具体仿真参数详见文献[3]。可以看出，RU较高时基于AI的CSI预测可以获得显著的增益。

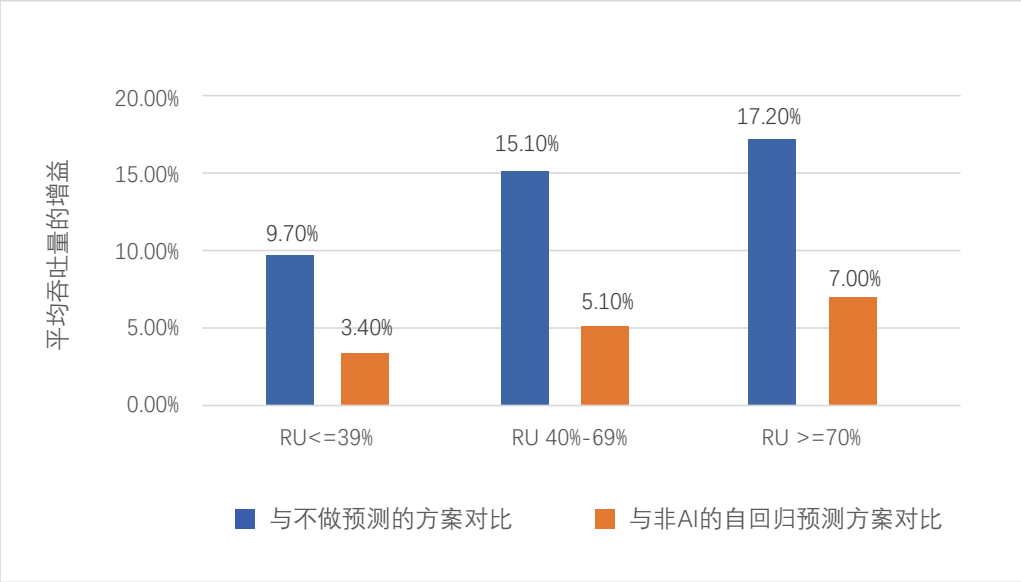


图2-2. 基于AI的CSI预测的平均吞吐量增益

基于AI的信道估计

移动通信系统需要使用参考信号进行信道估计。随着基站和终端的天线数日益增长，参考信号的开销也越来越大。对此，可以使用AI技术挖掘不同传输资源（时间、频率、空间等维度）上信道的关联关系，设计低开销、高精度的信道估计方法。

以解调参考信号（De-Modulation Reference Signal, DMRS）为例，可以将DMRS资源处的信道估计结果作为AI模型的输入，所有资源上的信道估计作为AI模型的输出来实现信道估计（如图2-3所示）。由于不同传输资源上信道的关联关系一般具有非线性特性，因此AI相比于传统插值方法可以实现更高的估计精度或降低参考信号开销。图2-4中，对基于AI的DMRS信道估计进行了验证，具体仿真参数详见文献[4]。图中蓝线表示频域50%开销（6/12）下传统线性最小均方误差算法实现的DMRS估计，其实现需要追踪参考信号（Tracking Reference Signal, TRS）的估计结果作为辅助信息；红线表示不同频域开销下使用基于AI的DMRS信道估计获得的性能（不依赖TRS）。可以看出，基于AI的方案，在没有TRS辅助的条件下，使用更低的参考信号开销实现了更高的估计精度和更低的误块率。在IMT-2020(5G)推进组于2021年主办的第二届无线通信AI大赛中，vivo承办了基于AI的信道估计赛道。651支参赛团队通过2个多月的角逐，设计了精巧的模型，显著提高了信道估计的精度，将AI与通信的融合又推进了一小步。

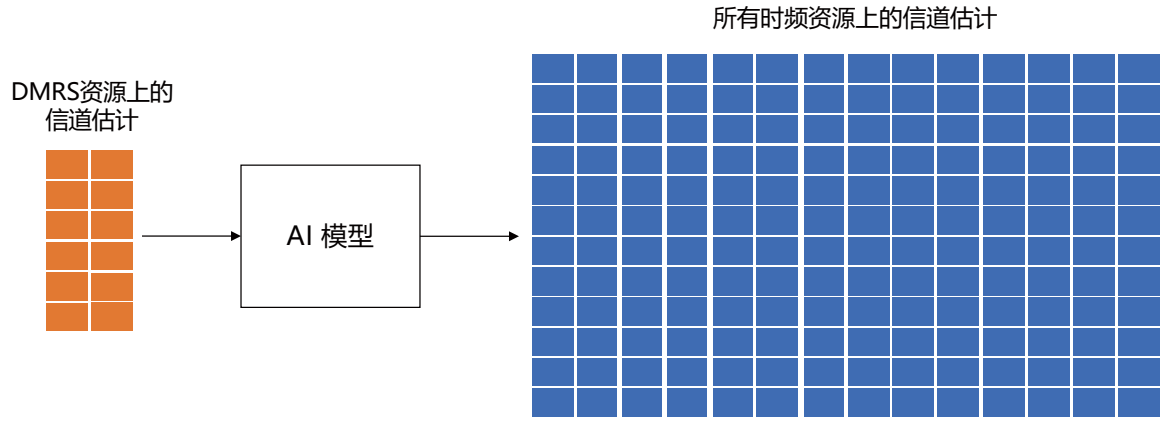


图2-3. 基于AI的DMRS信道估计示意图

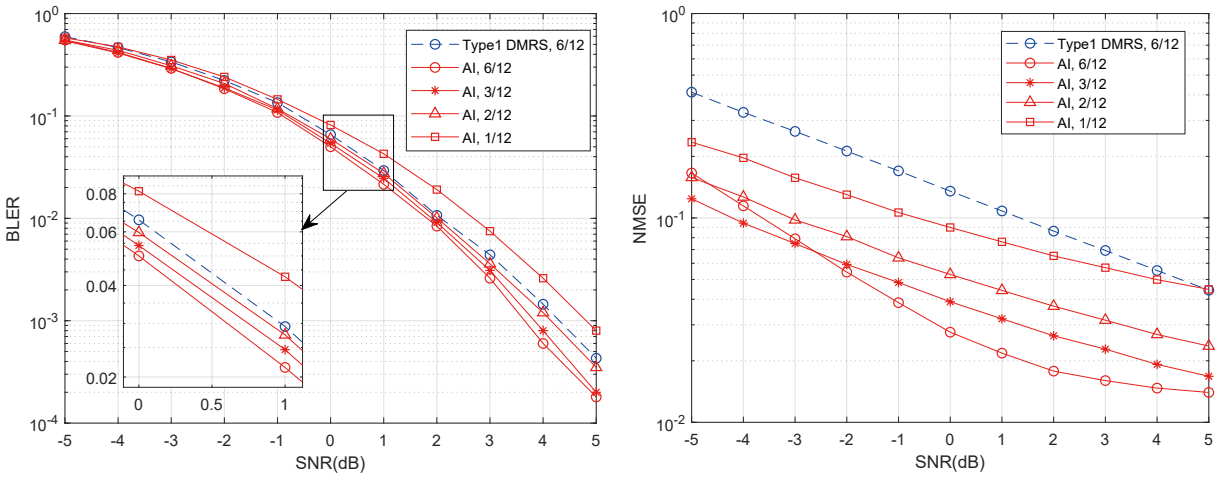


图2-4. 链路级信道中基于AI的DMRS信道估计性能：左图为误块率（Block Error Rate, BLER），右图为信道估计的归一化均方误差（Normalized Mean Square Error, NMSE）

此外，针对信道状态信息参考信号（Channel State Information Reference Signal, CSI-RS）的信道估计，可以考虑用近似消息传递（Approximate Message Passing, AMP）等压缩感知技术降低空域参考信号开销，即用稀疏的部分端口参考信号估计出全部端口的信道信息。但AMP算法的痛点是无法显式地获得最优感知矩阵和迭代估计算法中的收缩函数。对此，如图2-5所示，可以将参考信号经过感知矩阵的过程和AMP算法的迭代求解过程用模型驱动的思想展开成两个神经网络，其中第一个神经网络用于确定最优的感知矩阵，第二个神经网络用于重构AMP迭代算法。通过大量数据进行端到端的监督训练，即可确定最优的感知矩阵和收缩函数，提高基于压缩感知的信道估计的精度。

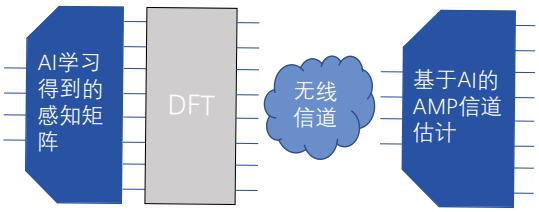


图2-5. 模型驱动的低开销CSI-RS信道估计示意图

基于AI的波束预测

波束预测包括空间域波束预测和时间域波束预测。空间域波束预测使用少量波束集合B的测量结果来预测完整集合A的波束信息，其中波束集合B可以是波束集合A的子集，或者波束集合B与波束集合A不相同（例如，波束集合B可以是宽波束，波束集合A是窄波束）。时间域波束预测的方法与空间域波束预测基本一致，但神经网络的输入包括多个历史时刻上的测量结果，神经网络的输出包括多个未来时刻上的预测结果。波束预测的核心思想是在减少波束测量开销的同时提高波束预测的准确率并降低波束预测的复杂度。因此，辅助信息也可能用于神经网络的输入从而进一步提高波束预测的准确性。

我们对基于AI的波束预测进行了验证，具体仿真参数详见文献[5]。在集合B是集合A的1/8的波束测量资源时，基于AI的方案相对比与非AI方案的Top-1波束预测准确率从12.5%提升到60%左右。在集合B是集合A的1/4的波束测量资源时，基于AI的方案相对比与非AI方案的Top-1波束预测准确率从25%提升到80%左右。可见，在使用相同的波束测量资源时，基于AI的方案的性能优于非AI方案。

基于AI的定位增强

基于AI的定位是通过AI技术建立信道测量信息与终端位置坐标之间的映射关系，根据模型输出类型可进一步划分为直接AI定位和AI辅助定位两个子用例。其中直接AI定位是指AI模型根据终端与多个发射/接收点（Transmission/Reception Point, TRP）的信道测量信息直接估计终端位置坐标；AI辅助定位是指AI模型首先根据终端与多个TRP的信道测量信息估计中间特征量，如信号到达时间（Time of Arrival, TOA），再结合多个TRP的坐标及中间特征量进一步估计终端的位置坐标。表2-1对比了在室内工厂非直视（Non Line of Sight, NLOS）径场景下不同定位方法的定位精度，具体仿真参数详见[6]。相比基于首径测量的下行到达时延差（Downlink Time Difference of Arrival, DL-TDOA）的传统定位方法，基于AI的定位可以获得显著的定位精度提升。

表2-1. 多种定位方法的定位精度

定位方法	测量信息	模型输出	定位精度（90%终端）
DL-TDOA	首径时延信息	位置坐标	32.12m
直接AI定位	信道脉冲响应	位置坐标	0.99m
AI辅助定位	信道脉冲响应	TOA	0.73m

基于AI的射频非线性抑制

正交频分复用（Orthogonal Frequency Division Multiplexing, OFDM）系统具有抗干扰能力强、抗衰弱能力强、频谱利用率高等优点，是4G和5G的基础波形。但是，OFDM具有较高的峰均功率比(Peak-to-Average Power Ratio, PAPR)，会使得功率放大器（Power Amplifier, PA）进入饱和区产生非线性失真。为了改善系统的线性与效率，需要对OFDM信号的PAPR进行抑制。近些年，深度学习启发了研究人员用数据驱动或模型驱动的方法来对抗PA非线性的影响。如图2-6所示，在降低PAPR的子载波预留（TR，Tone Reserve）算法中，使用AI技术可以获得OFDM信号和最优削峰信号之间的隐含关系。在推理阶段，当新的OFDM信号输入时，AI会给出与之对应的最优削峰信号。我们通过仿真验证发现，在预留子载波占总子载波25%的情况下，AI辅助的TR算法相比于传统TR算法可以降低PAPR约3dB。同理，在相同的PA功率回退下，基于AI的TR算法可以实现更低的误差矢量幅度（Error Vector Magnitude, EVM）[7]。

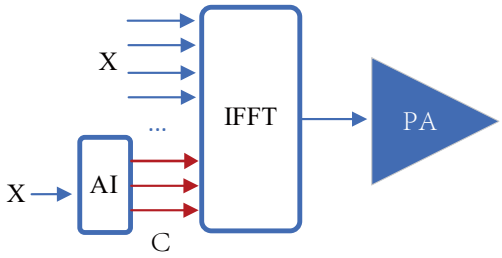


图2-6. 基于AI的TR技术示意图

类似地，在数字预失真（Digital Pre-Distortion，DPD）技术中，通过对大量的PA前信号和PA后信号的学习，AI可以得出PA对信号进行非线性变化的规律，进而得到与PA模型匹配的DPD模型。如图2-7所示，通过神经网络对输入到PA之前的信号进行DPD，可抵消掉PA对原始信号产生的非线性失真。同样，我们通过仿真验证发现，基于AI的DPD相比于传统的DPD可以降低约5%的EVM [7]。

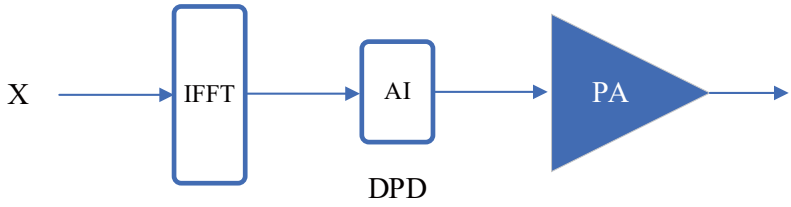


图2-7. 基于AI的DPD技术示意图

基于AI的用户移动性优化

移动性管理是移动通信系统的重要机制，能够为用户提供连续的业务体验。传统的移动性管理基于先配置，再试错后调优，想要实现几乎零失败率的切换，难度很高。因此，考虑借助AI来优化移动性管理。基于AI的移动性优化，主要借助用户轨迹预测，结合候选小区的业务负荷预测，为终端选择合适的目标小区。其目标是最小化终端的切换失败概率，并减少非期望事件发生（如切换到错误小区）。对于基于AI的移动性优化，模型推理所需的典型输入包括终端无线资源管理（Radio Resource Management, RRM）测量和终端位置。若模型推理功能仅部署在网络侧，则终端需频繁上报测量数据作为模型输入，会带来大量信令开销。此外，上报终端的实时位置会引入隐私风险，用户也可能因为隐私考虑而关闭位置信息上报许可。与网络侧基于AI的移动性优化相比，引入终端推理的移动性优化可以实现更实时/更细颗粒度的预测，如RRM测量预测、目标小区预测、以及非期望事件预测。

在图2-8中我们以RRM预测为例，介绍一种基于AI的移动性优化方案。终端在T_0时刻首次满足切换事件，此时终端可基于AI预测预定义的触发时间（Time To Trigger，TTT）内多个时刻的服务小区和相邻小区的信号质量来判断是否触发测量上报。若TTT内的RRM测量预测结果均满足上报条件，则在T_0时刻立即触发测量上报，从而避免因TTT时间内信号质量过差导致终端无法接收切换命令。此外，终端可上报更长时间的小区信号质量预测结果，供基站在切换判决中参考，以避免在随机接入到目标小区过程中出现切换失败或切换后短时间内发生无线链路失败或乒乓等切换异常事件。

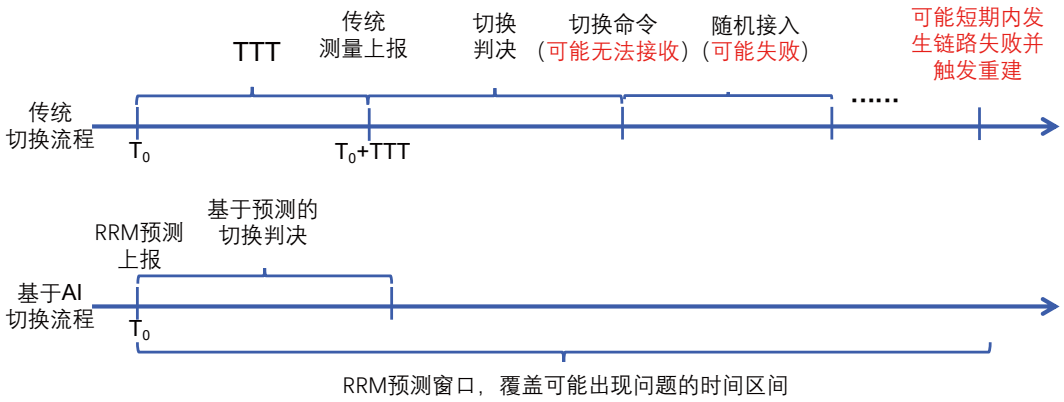


图2-8. 基于AI的DPD技术示意图

基于上述方案，在密集城区场景下仿真得到图2-9中的评估结果。相比于传统切换方案，基于AI的切换方案可显著降低切换失败概率和乒乓切换概率。

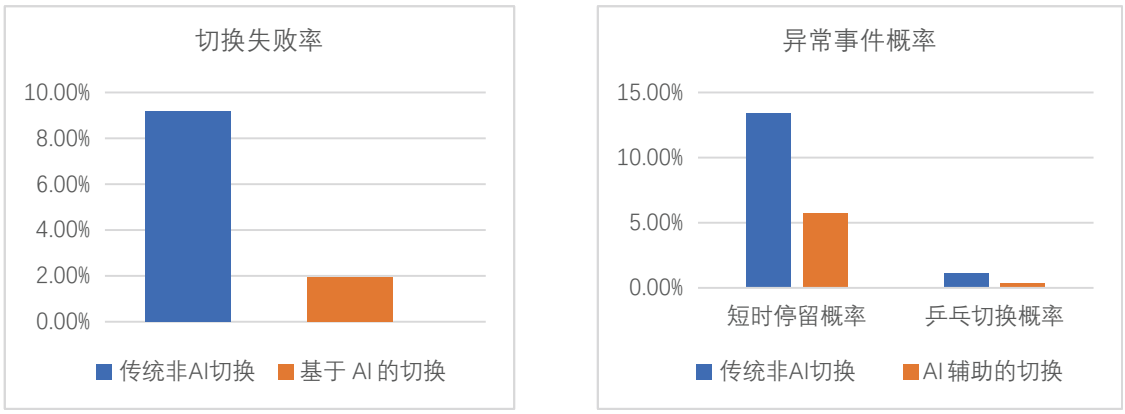


图2-9. 传统切换/基于AI切换的切换失败率（左）和异常事件概率（右）

基于AI的选网

移动通信系统的部署存在多种无线接入制式共存互补的场景。以目前的5G情况为例，虽然5G部署已经大规模开展，但在短期内5G新空口（New Radio, NR）无法达到全面覆盖。这种情况在6G时代可能同样存在。可以预见，4G、5G、6G等几种无线接入技术将会长期共存，多种接入制式如何协同工作的问题也随之产生。

不同的接入制式有着不同的性能。一般而言，4G的频率较低，时延较高，但网络覆盖较好；而5G频率较高，时延较低，但覆盖范围受限；在6G，这种容量、时延和覆盖之间的权衡依然会存在。从应用的角度考虑，不同的应用对于网络关键指标的要求不同。例如，增强现实（Augmented Reality, AR）和虚拟现实（Virtual Reality, VR）等对速率和时延要求高的业务，使用高频大带宽资源更加适合；而对移动性要求高且希望避免频繁切换的业务，就可以选择覆盖范围大的无线连接。因此，针对不同的应用或场景，存在根据业务类型选择适合的接入网络类型的需求。如图2-10所示，可以利用AI对终端用户和网络的历史数据进行分析，获取表征用户行为特征和网络性能特征的AI模型。基于该AI模型和现网数据，可预测出终端正在或即将使用何种业务、终端所在小区或具体位置，并预测网络的负载和性能情况等。根据这些信息，网络能够给出基于AI辅助的UE接入网络选择策略。

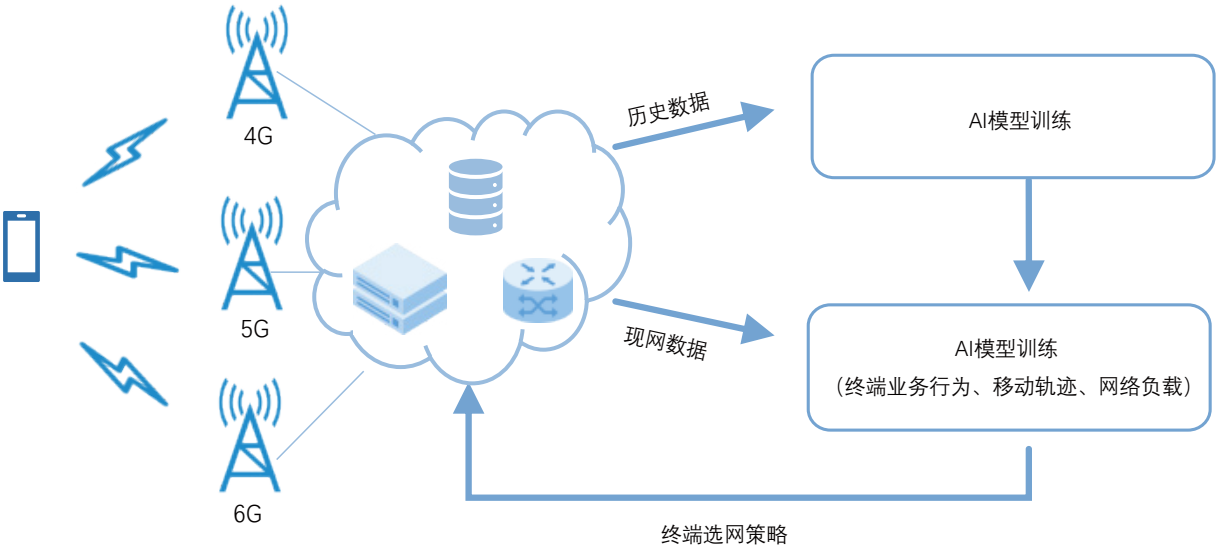


图2-10. 基于AI的智能选网示意图

基于AI的信令风暴预测

随着运营商用户数的迅速增长，产生信令风暴的风险呈显著增长态势。根据多次历史经验看来，信令风暴会导致故障影响范围大、持续时间长，对用户的个人权益和运营商的品牌形象产生较大的负面影响，因此防止网络信令风暴是运营商最为关注的难题之一。以往，运营商只能在网络运维层面后向地检测信令风暴的产生，再人为地进行根因分析和流量疏导，这一过程的时间长短往往取决于专家经验，一般可能持续几小时甚至几天。

随着AI技术在无线通信网络的运用，6G网络有望借助AI进行信令风暴的预防。如图2-11，6G网络AI实体收集正常场景和网络信令风暴场景的网络指标、行为表现、网络配置、终端行为等大量数据。AI实体挖掘这些数据指标和信令风暴发生的关联关系，从而训练出AI模型。该AI模型能够预测网络发生信令风暴的概率，甚至可以精确地预测信令风暴所产生的网络表现以及其他衍生后果。这些信息可以作为网络预警信息提前告知相应的设备或人员，甚至可以作为网络操作或配置信息的调整建议。

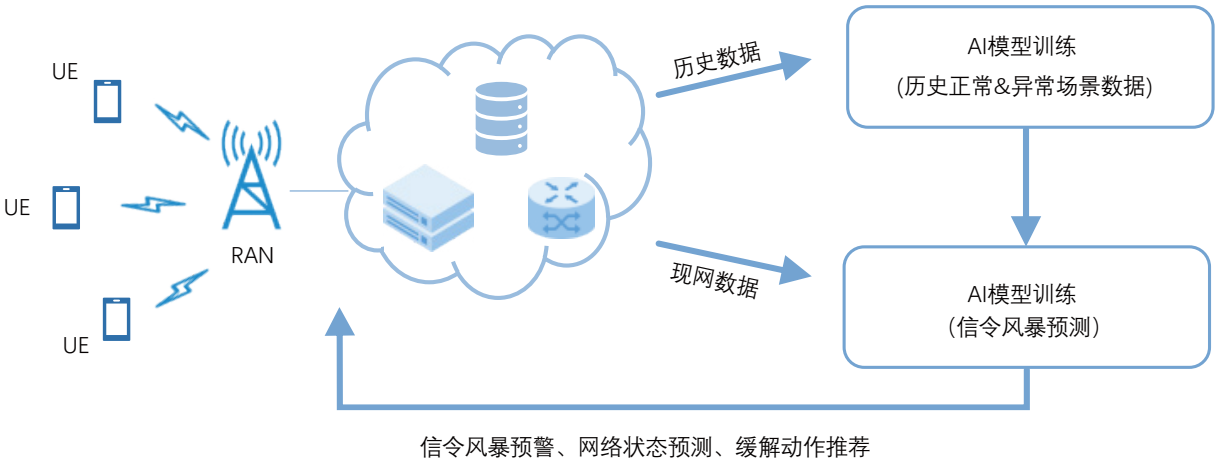


图2-11. 基于AI的信令风暴预测

接下来，我们介绍AI应用于传统方法在期望时间内不易获得闭式解或没有显式闭式解的场景。其代表性用例有基于AI的无线资源分配、网络节能和负载均衡等。

基于AI的无线资源管理

接入网中有一大类任务（用例），如用户调度、接入控制、资源分配等，是根据无线环境、负载、用户数等的变化连续调整的过程。其每一轮给出的决策不仅影响本轮的系统性能，也会影响下一轮的决策。这一类问题是强化学习的擅长领域。强化学习能够通过与系统、环境的交互，学习出系统和环境的静态与动态特征，从而自适应地实现最优策略。目前，基于强化学习的方案在频谱资源分配，动态功率分配，车联网、无人机通信中的调度，大连接通信中的接入控制、切片资源分配等诸多方面，已有较为成熟的研究成果。相比传统方案，基于强化学习的无线资源相关决策可以更好地与系统、环境适配，实现更高的系统性能和资源利用率。

以用户调度为例，6GANA于2023年主办的6G网络AI挑战赛中，vivo承办了基于AI的cell free场景无线资源调度赛道。72支队伍经过初赛和复赛两轮的角逐，设计了性能优异的AI模型，显著优化了资源分配策略，提高了调度的整体性能增益。基于AI的无线资源调度，以网络节点到各个用户节点的当前信道信息以及历史调度速率作为AI模型输入，以未来时刻的时/频/空域资源分配策略作为AI输出。图2-12中我们展示了大赛中选手设计的基于AI的调度方案和传统贪婪调度方案的调度得分。调度得分综合考量了网络整体调度速率及各个用户的调度公平性，得分越高表明网络整体吞吐量越高且用户调度公平性越好。可以看出，基于AI的方案可以实现比传统非AI的方案更好的调度。

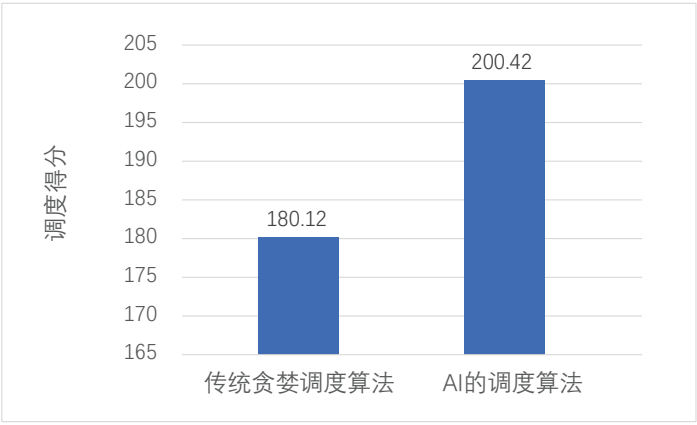


图2-12. 基于AI的调度方案和传统贪婪调度方案的调度得分

基于AI的网络节能

日益增长的通信业务需求，促使网络站点数量和网络能耗急剧增加。同时，网络运营商对网络节能的需求也日益迫切。当前，网络节能主要通过小区级别的去激活/激活实现。简而言之，当小区业务负荷低于阈值，可以去激活小区，并将业务卸载到相邻小区，以降低网络整体能耗；当相邻小区业务负荷高于阈值，可请求激活已关断小区，以避免网络拥塞影响用户体验。但是小区的关断会对周边其他小区的服务造成影响，仅以单一小区的负载情况进行决策可能会导致小区频繁改变激活状态，造成用户业务中断。而传统方法受限于计算复杂度，很难直接获得大面积小区的激活/去激活最优解，难以实现网络能量效率的最优化。如图2-13所示，基于AI的网络节能一方面通过预测终端轨迹和各节点的业务负荷情况，确定激活/去激活的候选小区。另一方面，AI还可以挖掘负载状态与全局最优的去激活/激活方案之间的关系，直接通过神经网络给出指定区域内候选小区的激活/去激活决策，平衡网络能耗与服务质量，最大化网络能量效率。类似的思路也可以应用于负载均衡，让不同类型的用户的不同业务在各频段中合理分配，最优化网络的频谱效率。

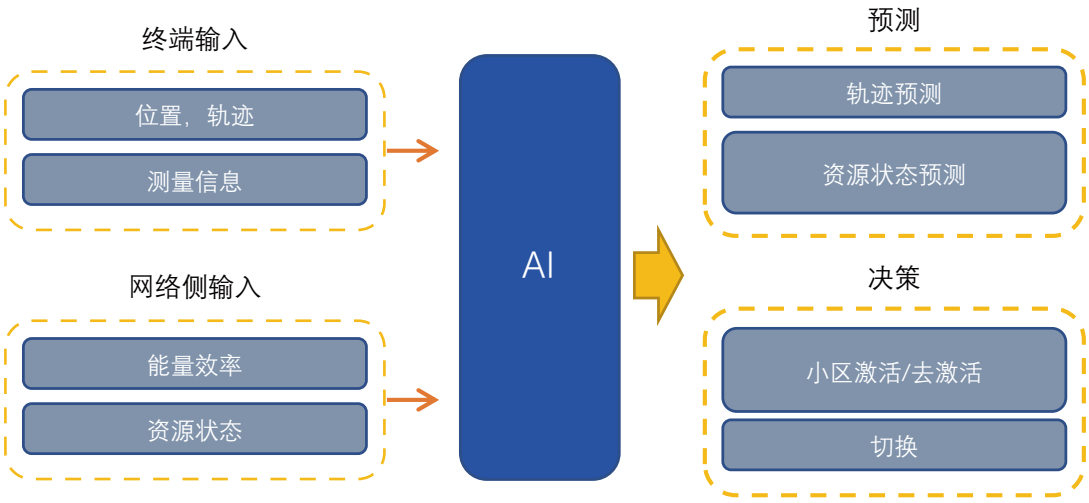


图2-13. 基于AI的网络节能示意图

最后，介绍AI应用于多模块联合优化的场景。其代表性用例有MIMO信号处理中关联功能的联合优化、信源信道联合编码、跨层优化等。

MIMO信号处理多功能联合优化

AI的一个优势是对数据背后隐含的关联关系进行挖掘。通信系统中有一些功能是相互关联的，但其具体关系无法显式地表示出来，导致联合优化的成效受限。对此，可以考虑使用AI对这类关联的多个功能进行联合优化。MIMO系统的系统容量由预编码决定，预编码是基于信道估计（时分双工系统基于信道互易性、频分双工系统基于信道信息反馈）生成，信道估计又与导频序列的设计相关。单模块的AI优化虽然可以提升每个模块的性能，但从整体系统性能来讲并不能实现全局最优。对此，可以考虑将MIMO信号处理中关联功能建模为一个联合问题，建立全局损失函数，获得最优的MIMO传输方案。在图2-14中，以两用户MIMO的CSI反馈为例，可以通过3个子神经网络分别实现CSI-RS序列的选取、信道反馈和预编码矩阵生成等3个功能。其中，第一个子网络的参数即我们通过神经网络得到的参考信号序列，第二个子神经网络生成CSI反馈比特，第三个子网络生成预编码矩阵。训练的时候将3个子网络拼接起来，用频谱效率作为全局损失函数，通过梯度下降的方式获得各个子神经网络的参数。这样得到的全局神经网络会兼顾多个子网络之间的依赖关系，实现比分别训练的方案更高的频谱效率。

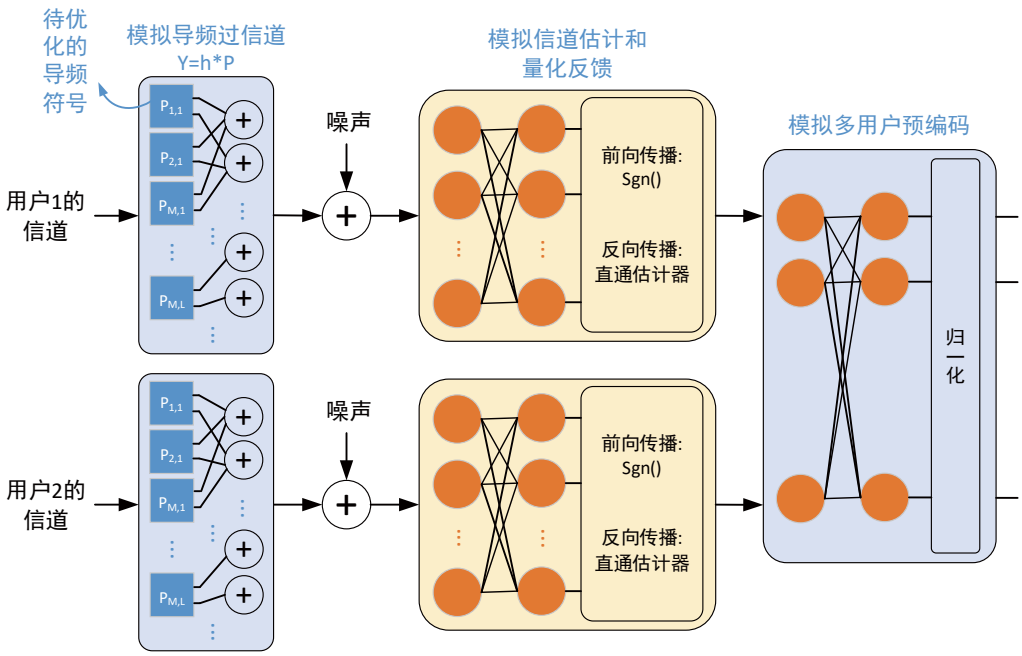


图2-14. CSI-RS序列的选取、信道反馈和预编码矩阵生成的联合优化

信源信道联合编码

传统的信源信道联合编码允许用户根据信道或网络条件改变信源编码参数，或是根据信源特性选择信道编码、调制及网络参数。其本质是研究人员使用专家知识精巧地设计信源信道编码方案。近年来，AI在图像、语音压缩等领域的发展为信源信道联合编码的设计提供了新的思路。其中，自编码器（Auto-Encoder， AE）的结构非常适合实现信源信道联合编码。以信源待传输的信息和当前的信道信息作为编码器的输入，编码器通过神经网络的正向传递获得待传输的比特流信号。接收端的解码器是与编码器成对匹配训练得到的，实现编码器的逆操作。因此，将接收到的比特流信号输入到解码器，在输出端可以恢复信源侧传输的信息。

此外，也可以基于语义通信的思路设计信源信道联合编码。其特征是首先利用神经网络提取信源的多维语义特征，再根据特定的先验知识和各维度语义的重要性对其使用不同的编码策略，从而达到更好的利用已知信道特性的目的。比如在图像和视频传输中，通过引入语义通信可能获得比传统图像压缩编码技术在相同码率下更佳的质量。此外，还可以将CSI压缩视为一种特殊的信源信道联合编码问题，借助语义通信的思路进一步提高CSI反馈的效率。具体的，发端原始的CSI对应信源信息，待反馈的CSI信息对应信源信道联合编码后的比特流，接收端再做逆变换实现CSI的重构。

跨层优化

跨层功能也有一定的潜力进行联合优化。比如，不同的流量需求和业务特征有对应的最优传输资源和传输模式，但由于不同层的功能并没有串接在一起，且不同层的优化目标也不相同，因此很难直接给出闭式解。对此，可以将频谱资源分配与流量预测/业务预测设置为一个联合任务，以历史的流量、业务、资源配置等作为输入，以优化最终性能（如吞吐量）为目标，通过AI模型来给出推荐的资源配置；类似的，物理层的自适应调制编码（Adaptive Modulation and Coding, AMC）与数据链路层的混合自动请求重传（Hybrid Automatic Repeat request, HARQ）都是为了适配传输环境而进行动态调整的功能。也可以将其视为联合任务，基于AI实现跨层联合优化。跨层优化的目标一般是网络的最终性能，需要将AI训练/推理过程与实际系统联动起来，使用监督学习很难实现。对此，强化学习是一种解决方案：可以在训练AI模型的时候将AI模型给出的推荐配置应用在系统中，并通过系统的真实反应（网络的最终性能或相关的衍生指标）作为激励/评分来指导优化AI模型。这种交互迭代优化的模型训练好之后，再在实际系统中部署应用。

随着学术界和产业界研究的深入以及AI技术和资源的普及，将会有更多的高价值用例涌现，不断提升移动通信系统的性能。比如，AI在波形设计、调制解调、信道编译码、信号检测、信号均衡等更基础的领域以及收发机联合设计、基于端到端架构的传输机制（如无导频传输，无循环前缀传输）等更加复杂的领域也有很大的潜力。

03

第三章

AI与通信融合的设计原则

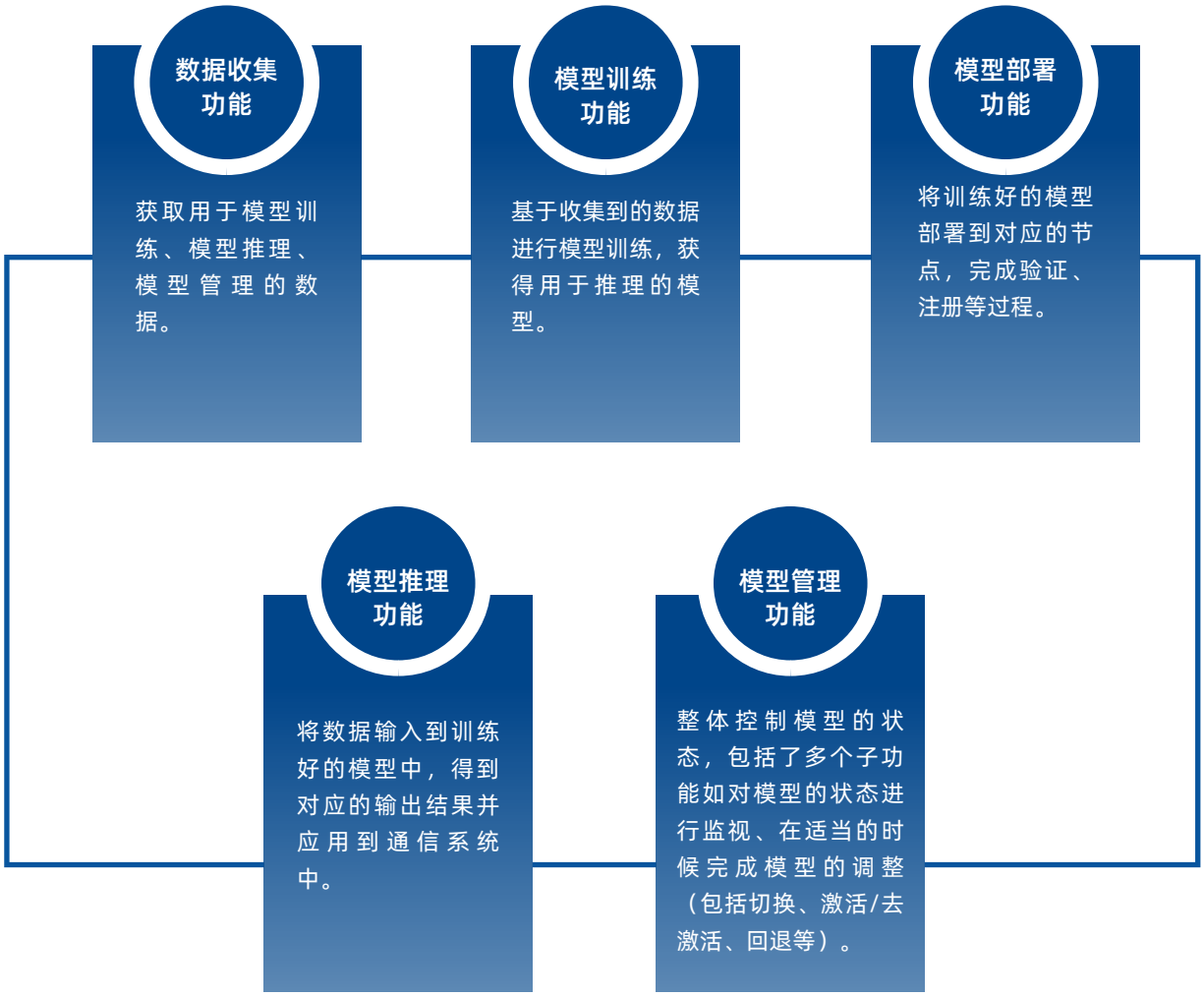
AI在6G中将是一个原生的、泛在的技术，为6G提供全面的支撑。为了在6G更好地发挥AI的作用，需要结合用例的丰富性和系统的执行效率，提出更科学、更高效、更普适的设计原则。鉴于5G已经初步探索了AI在移动通信中的应用，我们首先对AI在5G中的应用现状进行总结。其次，对智慧内生的基本逻辑进行阐述，并基于此从多个维度提出设计原则，用来实现AI与移动通信的深度融合。

将AI应用到移动通信的标准化工作已经在第三代合作伙伴计划（3rd Generation Partnership Project, 3GPP）陆续展开。目前3GPP已经有AI/ML for OAM、AI/ML for NG-RAN、Enablers for Network Automation for 5G、5G Systems Support for AI/ML-based Services和AI/ML for Air Interface等多个项目开展了AI与移动通信系统结合的标准化研究。在用例研究方面，目前，3GPP在核心网、接入网以及物理层均展开了基于AI的增强型研究。第二章介绍的用例中，CSI反馈增强、波束预测和定位增强已在RAN1进行研究；移动性优化、负载均衡和网络节能在RAN3进行了研究；智能选网在核心网进行了研究。可见，3GPP正在从各个层的视角对AI的价值进行评估。在5G阶段，AI是针对5G网络中特定的已识别问题进行逐用例、外挂式优化，是5G网络的自然延伸，对所述特定的已识别问题均有一定的提升。

基于5G的讨论，我们从智慧内生的基本逻辑出发，提出多种AI与通信融合的设计原则，为AI赋能6G搭建坚实的基础。

3.1 6G智慧内生的基本逻辑

本白皮书涉及的6G智慧内生主要从AI4NET或对内AI服务的角度进行阐述。6G智慧内生是指将AI能力原生地融入到6G网络中，在设计6G之初就将AI考虑在体系架构设计中，预留各种AI用例所需的功能、接口、能力、和信令结构，实现与移动通信网络的深度融合。本白皮书的第二章介绍了多种AI赋能6G的用例，随着AI技术的成熟和AI设备的普及，将会有更多高价值的AI用例。如果仍沿用5G中逐用例做协议设计的方式，协议的复杂度和冗余度将会大幅增加。因此，在6G中部署AI用例应以逻辑功能和逻辑节点为基准进行架构设计。因为不同用例的物理执行位置不尽相同（比如波束管理用例涉及终端或基站，定位增强用例涉及终端或基站或位置管理网元），根据物理节点设计的架构将无法适应不同的使用情况。但是，不同用例实现的逻辑功能是相同的，即所有用例都离不开数据收集、模型训练、模型部署、模型推理和模型管理等逻辑功能。因此，可以基于上述逻辑功能设计一套统一的架构。各逻辑功能的具体内容如下。



对应的，上述逻辑功能需要在一个逻辑节点或多个逻辑节点协作执行。所述逻辑节点包括数据源、训练节点、推理节点、执行节点和管理节点。其中在推理节点获得AI模型的输出信息，执行节点将AI模型的输出信息应用于通信系统的具体功能中，多数情况下推理节点和执行节点是同一个节点。图3-1给出了逻辑功能和逻辑节点之间的关联关系。

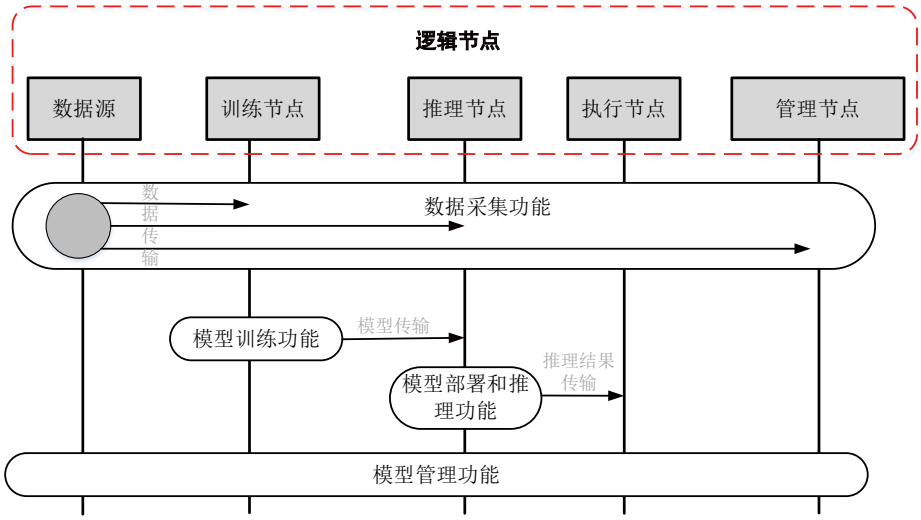
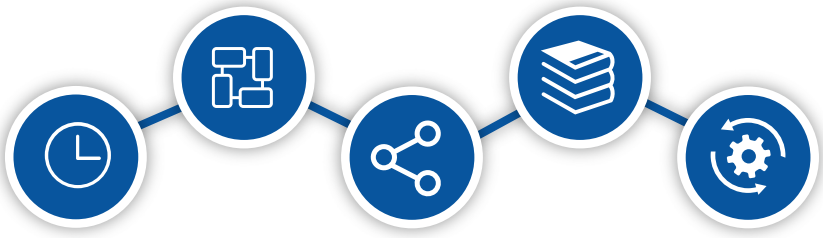


图3-1. 逻辑功能和逻辑节点之间的关联关系

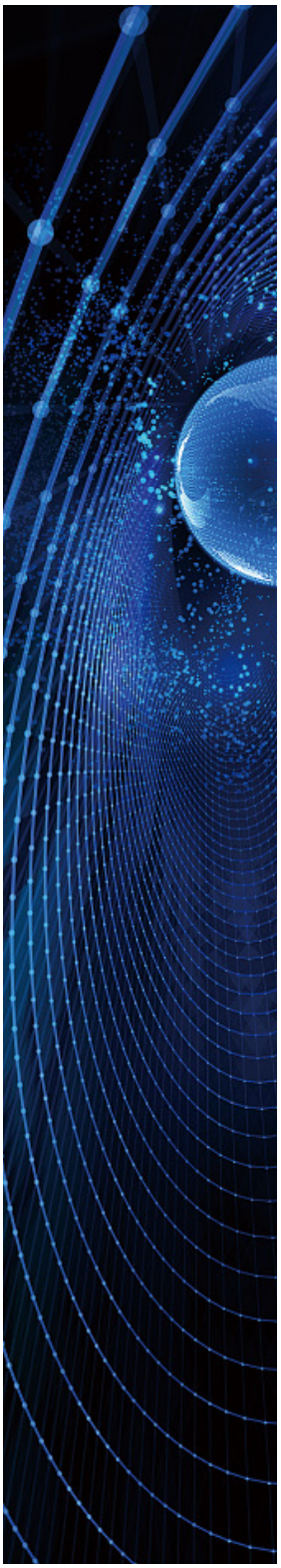
设计好一套基于逻辑功能和逻辑节点的协议框架后，在具体的用例中只需要将逻辑节点映射到实际的物理节点。实际中，存在多个逻辑节点对应到一个物理节点的情况，此时这些逻辑节点之间的交互流程将在物理节点内部实现，无需信令支持。

接下来，我们从以下五个维度探讨AI与通信融合的设计原则：

- 生命周期管理
- AI逻辑功能的分布
- AI资源的共享
- 学习架构
- 演进



3.2 内生统一的生命周期管理



在AI和通信融合的6G AI系统中，一个AI模型从无到有再到调整变化需要经历数据收集、模型训练、模型传递、模型验证、模型部署、模型推理、模型监视、模型调整等生命周期中的多种功能。生命周期管理是AI应用于移动通信中必需且特有的操作。其主要原因是AI模型是基于数据训练得到的，其有效性与数据质量、应用环境的匹配度等紧密相关。由于实际中无法通过完美的数据集训练出完美的模型，因此必然存在AI模型与应用环境失配的情况，引起模型泛化不足的问题。对此，需要对模型的生命周期进行管理。

上一节介绍的逻辑功能与逻辑节点对移动通信系统中所有的AI用例均适用，基于这些逻辑节点和逻辑功能可以实现一套统一的模型生命周期管理。对AI模型的生命周期管理过程进行拆分后，可以发现，模型的生命周期管理是一个闭环过程。如图3-2所示，这个闭环过程还可以进一步分为大、小两个闭环过程。其中大闭环的侧重点是获取一套全新的模型，可通过数据收集-模型训练-模型传输-模型注册-模型推理-模型监视来实现。小闭环的侧重点是进行模型调整，主要通过数据收集-模型调整-模型传输-模型注册-模型推理-模型监视来实现。但数据收集和模型传递并不是小闭环必需的步骤，取决于具体的模型调整方案。可以看出，大、小两个闭环有一定的重叠区，即模型注册、推理、监视等跟模型执行紧密相关的过程。另外，大闭环中的数据收集一般会涉及大规模的数据收集，以离线收集为主；而小闭环的数据收集一般涉及模型微调所需要的小规模数据收集，以在线收集为主。

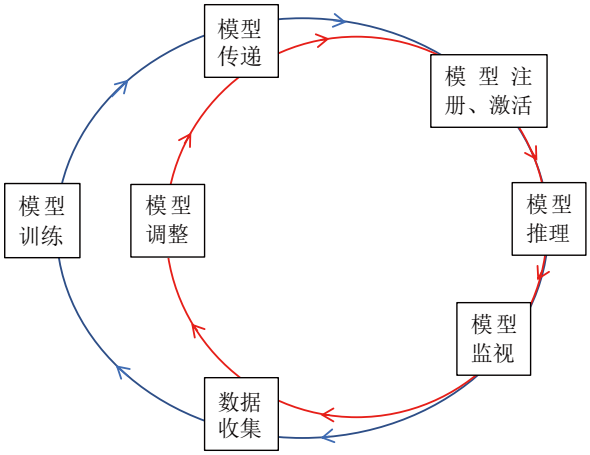


图3-2. 模型生命周期中的闭环

实际中，所有的用例都可以基于这样一套包含大、小闭环的生命周期管理来运行。在模型大版本更新时执行大闭环，而随着用户的移动、环境的变化需要做模型调整时通过小闭环来解决。

统一的生命周期管理中一项关键的功能是数据收集，如何基于统一的方案实现数据收集是一个关键。我们发现，5G中已定义的用户面和数据面均无法满足面向AI的数据收集：

1. 目前控制面单次交互承载量是9000字节[8]，无法满足大数据的传输需求。
2. 用户面只能在终端和用户面功能（User Plane Function, UPF）之间进行数据传输，例如终端通过数据面进行数据传输，则基站无法获得数据。此外，用户面的收费政策会影响用户的偏好，因此用户面的数据收集可能还需要征得用户的同意。
3. 利用控制面和用户面进行数据收集，存在相同或类似数据的重复收集的问题。

为了解决上述问题，我们建议6G借助数据面进行统一的数据收集，如图3-3所示。数据面的数据交互具有一点对多点、多点对一点、多点对多点，以及数据终结点在核心网、无线接入网或终端的多样性，因此数据面可以高效灵活地支持多样化的数据交互需求。数据面中的数据控制功能可以根据所收集数据的服务质量和数据分级等信息，负责在数据提供方和数据请求方之间建立、修改和释放用于数据传输的通道。其中数据提供方即前面介绍的逻辑节点中的数据源，数据请求方可以是逻辑节点中的训练节点、推理节点和管理节点等。此外，数据面的数据控制功能还包含数据收集管理功能，对所收到的数据请求进行处理。所述处理包括合并相同的数据收集请求，并生成数据收集控制信息来指导数据提供方。这样就可以避免数据重复收集的问题，实现高效统一的数据收集。

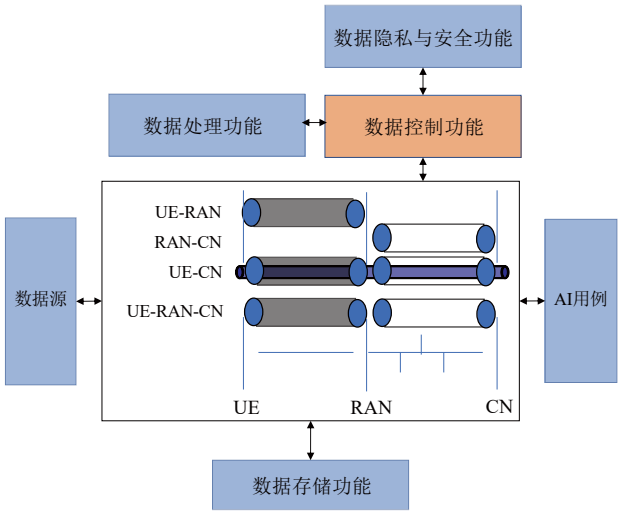


图3-3. 基于数据面为AI用例收集数据

3.3 6G系统中AI逻辑功能的分布

传统通信系统中，算力主要集中在网络侧，主要用于通信业务的计算处理。随着AI行业的迅速发展，以图形处理器（Graphics Processing Unit, GPU）、神经网络处理器（Neural Network Processing Unit, NPU）、张量处理器（Tensor Processing Unit, TPU）等智能处理器为代表的新型算力不断地涌现，其成本、能效、算力水平等都有了飞跃性提升。同时，在未来移动通信系统中，几乎所有网元都需要借助AI进行增强。因此，如图3-4所示，每个网元都会有自己专属的用例和跨网元节点的协作用例，且不同AI功能也会分布在各个网元节点中。具体的，单节点用例可以进一步划分为终端用例、基站用例、核心网用例以及网管用例。跨节点用例可以划分为终端-基站用例，终端-核心网用例、基站-核心网用例以及网管-基站用例等。

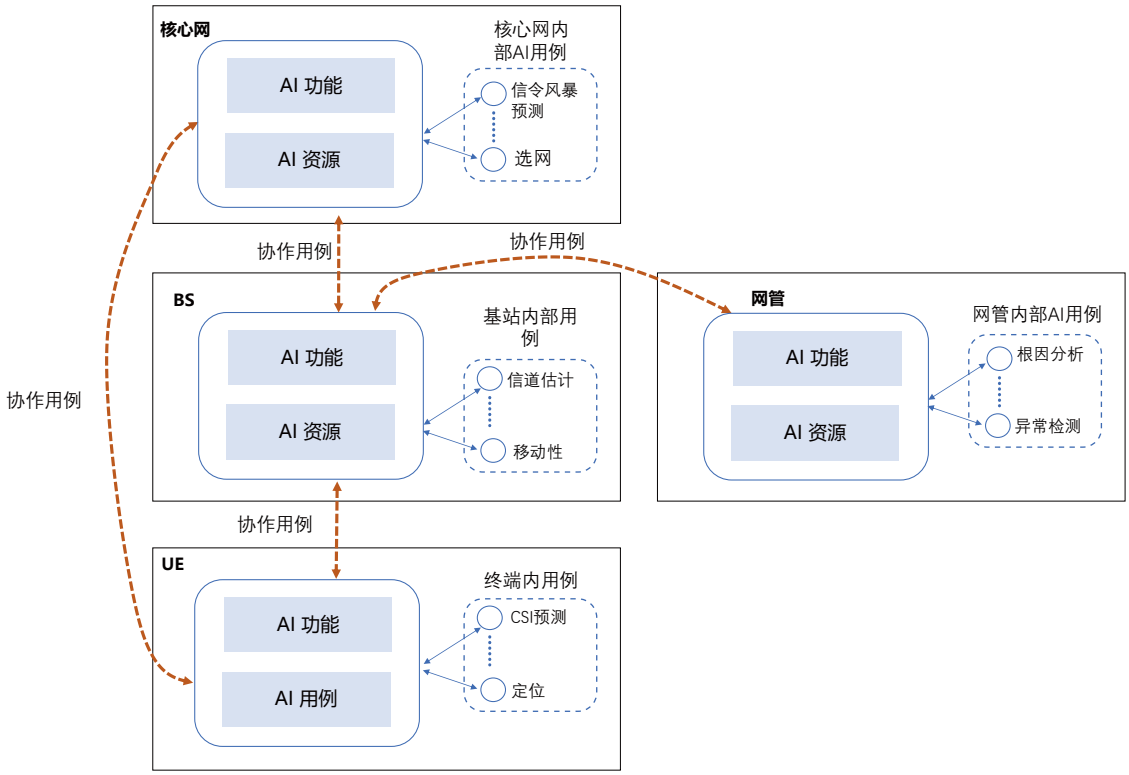


图3-4. 6G 网络中AI功能与用例的分布

将AI功能部署到实际物理节点时，不同的AI功能将遵从不同的原则。

模型训练功能的分布

模型训练是6G AI中非常关键的一环。5G阶段的研究中，AI模型的训练以集中式训练为主，需要将数据汇聚到一起后再进行训练。这种方式对数据隐私和数据安全考虑较少。

数据隐私

每个网元的数据中都直接或间接蕴含着该网元及其环境的信息，合理应用可以提升通信系统的价值，但也存在被恶意使用风险。同时，数据是数据源节点的资产。因此，各个网元的数据一定程度上存在数据不出域和保护隐私的要求。集中式的AI架构需要将数据传输到一个集中资源池中进行处理，与数据隐私和数据资产保护的需求相悖。

数据安全

集中式训练架构需要有集中式数据存储，数据库如果被网络攻击攻破，将会导致海量数据的泄露。

因此需要尽量让数据留在本地进行训练。但这又面临本地数据特征有限、数据量有限、算力有限等问题。既要满足数据隐私和安全需求，又要保障充足的训练数据和算力，分布式学习是一种解决方案。分布式学习是一类学习方法的统称，其核心思想是多个节点参与训练，数据不需要汇聚到一个集中式节点。分布式学习方法有很多种，如联邦学习，群体学习，分割学习等。从表3-1总结的不同分布式学习架构的特点可以看出，不同的分布式学习架构有各自的训练模式、拓扑结构、传输内容和适用场景。因此，需要设计一套统一的分布式框架去支持多种分布式学习方法，在具体业务中选择适用的学习方法来提供服务。

表3-1. 多种分布式学习方法的特点汇总

分布式学习方法	特点
联邦学习	最常见的分布式学习架构，一般由中央参数服务器和多个分布式的客户端节点组成，模型在客户端节点训练，在服务器汇聚再分发
群体学习	没有中央参数服务器，不需要将参数上传到中央服务器进行聚合
分割学习	中央参数服务器和分布式的客户端节点分别训练一个完整神经网络的一部分

在移动网络中搭建一个分布式学习的虚拟网络，需要选取合适的节点参与协作。选取参与协作的节点，最重要的依据是该节点的数据数量和质量。此外，也需要视节点的算力、传输能力等因素进行合理规划。

模型推理功能的分布

模型推理也需要关注数据隐私和数据安全问题。但与模型训练不同，模型推理需要的数据量小，数据隐私和安全的风险相对较小。因此，在保障数据隐私和安全的前提下，在模型推理阶段可以进行一定的卸载和协作。模型推理的一个关键需求是推理时延。如果数据源节点的算力充裕，则应尽量在数据源节点完成推理。这是一种静态的AI推理功能分布，此时推理时延主要是计算时延。但在数据源节点的算力有限时仍在数据源节点进行推理，计算时延会非常大。鉴于不同节点的AI能力以及AI业务的不同需求，网络内可用AI能力具有分布不均且动态变化的特征。对此，可以让AI能力充裕的节点（称作助力节点）协助AI能力需求大的节点（称作需求节点），完成特定AI推理任务，再将期望的信息反馈给需求节点。这是一种动态的AI推理功能分布，此时推理时延由计算时延和传输时延两部分组成，其关键是如何降低传输时延。因此，选取协作节点的原则是一个关键。选取助力节点时，需要根据节点与需求节点的距离，传输开销、算力等多重因素进行合理规划。

模型监视功能的分布

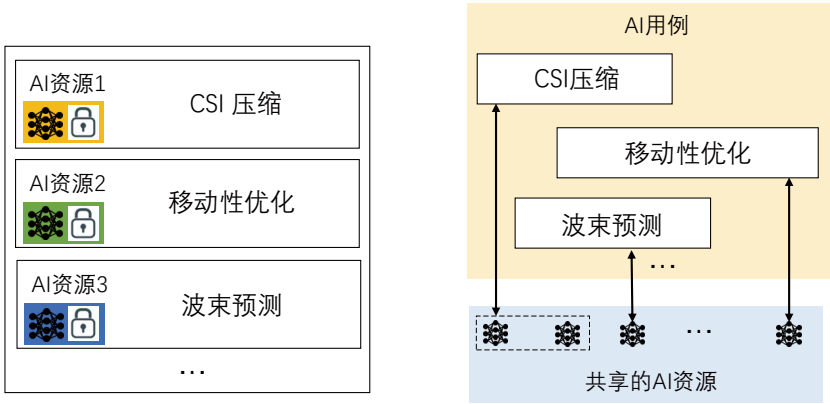
模型监视对时效性要求较高，如果不能按时执行模型状态的调整将会造成模型性能的严重下降。模型监视和决策模型状态（如模型调整、重新训练）是紧密关联的。部署模型监视的位置时，我们认为应遵从尽早决策原则。首先，尽量将计算监视指标和决策模型状态两个动作整合在同一个节点，避免传输、协商等流程导致的延迟。其次，对于基于模型输出的真值计算监视指标的场景，尽量模型监视功能部署在获取真值的节点，避免真值信息传输带来的传输延迟。

通过合理地将AI功能分布于6G网络中，可以实现更高的执行效率和资源利用率，有助于实现高效的移动通信。

3.4 AI资源与用例解耦

6G AI的用例对AI资源的需求存在一定的错峰现象，即不同的用例对AI资源的需求在时间上呈现不同的规律，不同用例的资源需求峰值的出现时间也不尽相同。对此，可通过AI资源与用例解耦，以更小的成本和代价实现更优的资源利用率和性能。AI资源包括硬件资源和软件资源。硬件资源包括GPU、NPU、TPU、应用专用集成电路（Application Specific Integrated Circuit，ASIC）等，软件资源包括AI框架、算法等。

目前，5G中AI用例的设计没有考虑到上述AI资源的共享，存在如下问题：1）AI新用例兼容性差：由于AI资源是与特定的AI用例绑定，新增的AI用例无法使用已有的AI资源。2）AI资源利用效率低：如图3-5（a）所示，通信设备同时支持了基于AI的CSI压缩，移动性优化和波束预测。但是这三个用例并不一定需要同时推理，假如在某时刻只有波束预测需要进行模型推理，剩余的两个AI用例无需推理，但其对应的2块AI资源并不能用于波束赋形的模型推理，无法有效利用AI资源。



(a) AI用例与特定AI资源绑定　　(b) 多个AI用例共享AI资源

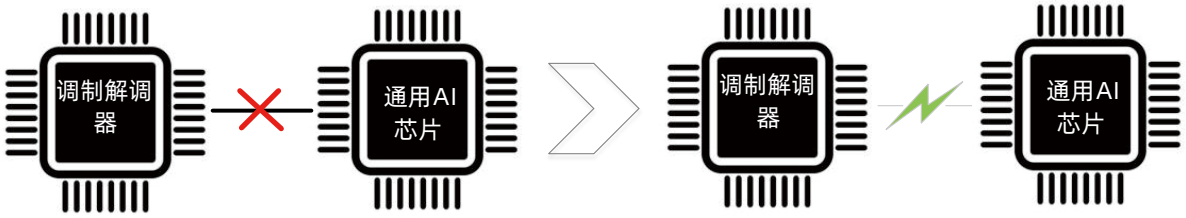
图3-5. AI用例与AI资源的关联关系

为解决以上问题，6G AI资源与AI用例应解耦，进而实现多个用例共享AI资源。当新增AI用例时，可以直接复用已有的AI资源，从而无需更新AI硬件资源，提升了AI新用例的兼容性。此外，如图3-5（b）所示，如果某一时间只有波束预测用例需要进行模型推理，则可以使用整块AI资源，缩短推理时间，提升AI资源的使用效率。多用例共享AI软件资源时，需要考虑多个用例的平台是否兼容，以及多个用例的算法软件接口是否兼容。

此外，AI资源与AI用例解耦对模型训练也大有益处。AI资源与AI用例解耦后，AI资源将变成一个通用资源。针对通用性较强的AI任务，多个模型供应商（如不同的运营商和终端厂商）可以将AI资源共享，联合训练一个模型，再分发给各自的用户。这样可以显著提高AI资源的利用率和模型训练的效率。

上述设计原则对AI通用设计的影响在于通用的资源交互信令。如果AI资源与用例绑定，则不同用例交互AI资源相关的信令是不同的。而当AI资源与用例解耦后，多个用例使用的AI资源交互信令可以是通用的。AI资源通用信令可以包括节点可用资源上报/通知，AI模型/功能资源请求，资源分配等。需要注意的是，在某两个节点之间不同用例可以使用通用的资源交互信令，但不同节点例如终端和核心网，以及终端和基站之间交互信令往往是不同的。

除了上述AI资源与用例解耦，终端侧调制解调器专用AI资源和终端通用AI资源之间的共享也值得关注。随着AI的普及，终端侧通用硬件的AI软硬件能力将大大提升，甚至会远高于调制解调器专用的AI软硬件能力。因此，可以将调制解调器内部的AI资源和通用AI芯片上的AI资源进行协调和共享，为AI用例提供服务。这两种资源的共享可能会导致不同的模型推理延迟，因此需要进一步考虑对不同用例的适用性。



（a）调制解调器和通用AI芯片之间存在壁垒　　（b）调制解调器和通用AI芯片之间AI资源共享

图3-6. 调制解调器和通用AI芯片之间的资源关系

在上述场景中可以依据AI用例的重要性、进程需求（或时序需求）和服务质量（Quality of Service, QoS）需求等来设计AI资源交互机制。比如，对系统性能影响大的用例分配更多的AI资源。如果两个功能有严格的时序关系，则进行资源共享时要优先保障时序在前的用例。同理，对于QoS质量需求高的用例也需要优先分配AI资源。例如，终端向基站申请的AI模型/功能资源请求中包含了多个用例时，由于各用例有不同的处理时延需求。因此，可以将多用例的优先级顺序，以及各用例的处理时延需求都通知给基站，以便基站根据这些要求，决定执行用例的顺序。

3.5 模型传输为基础的多种学习方法

在第二章，我们通过丰富的用例描绘了AI在移动通信中的多彩蓝图。可以从AI用例的应用环境特征来对所有的AI用例进行划分，并提取用例的共性，进而辅助智慧内生的设计。我们参考了计算机领域对AI用例的划分思路[9]，将移动通信中的AI用例按照环境的静态/动态和封闭/开放划分到四个象限中，如图3-7所示。静态是指环境的特征有限且稳定，动态是指环境的特征不固定、具有一定的未知性。封闭是指AI的输出以及后续动作不会对环境产生影响，开放是指AI的输出以及后续动作会对环境产生影响。为了便于理解，我们在图3-7的每一个象限举了一个用例作为例子。

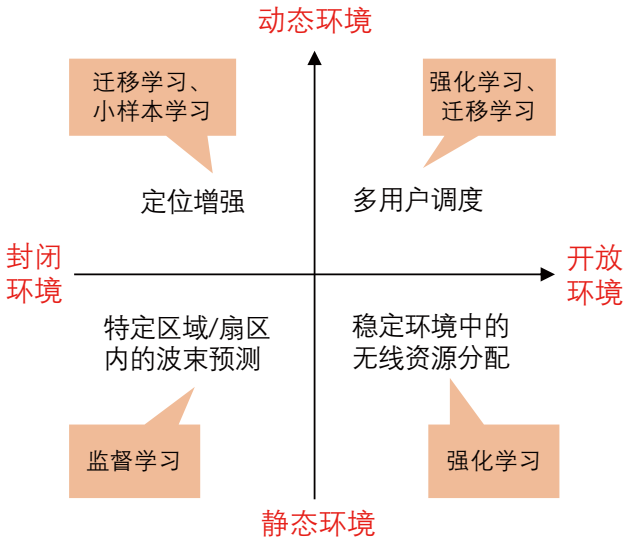


图3-7. 基于环境的用例划分

如图3-7所示，不同类型的用例一般需要采用不同的训练/学习方法。封闭-静态环境中的用例是最简单的一类用例，通过监督学习的方式即可获得泛化性较高的模型，并可以直接部署推理。封闭-动态环境中很难通过离线训练获得泛化好的模型，因此需要通过迁移学习、小样本学习等技术实现模型与环境的适配。开放-静态环境中的关键是用例与环境或系统会存在较强的交互，环境或系统会根据模型给出的结果发生变化，强化学习是解决此类问题的重要手段。而开放-动态环境是最复杂的环境，需要联合迁移学习、强化学习等才能实现优秀的推理性能。此外，网络中的数据往往分布在各个节点，同时考虑到数据隐私的需求，分布式学习也是一种非常有价值的学习架构。

综上所述，6G AI中需要支持多种学习方法来满足丰富的AI用例。在上述多种学习方法和架构中，一项共需的关键功能是模型传输。具体的，不同学习方案对模型传输的需求如下：

在联邦学习中

需要模型（完整或部分）或模型相关信息（如梯度）传输才能完成训练。

在分割学习中

为了在不同节点支持灵活的模型分割训练，也需要模型或梯度传输。

在迁移学习和小样本学习中

需要把基础模型从一个节点传输到另一个节点。

在强化学习中

如果模型更新节点和推理节点是分离的，则也需要将每轮迭代训练后将模型传输到推理节点，然后由推理节点收集状态并推理得到行为。

在监督学习中

如果训练节点和推理节点分属于不同的物理位置，则也需要模型传输。

模型是将AI引入到移动通信后产生的一种新的待传输信息，它既不属于传统移动通信系统中的业务信息，也不属于控制信息。因此，我们认为，在一开始设计6G网络时需要定义清楚用什么格式、在什么信道、通过什么流程来完成模型的传输。模型的传输格式，可以分为公开格式和专有格式。其中公开格式是指两端能够识别的模型描述格式，专有格式是指某厂商特有的私有模型描述格式。表3-2对比了两种格式的细节特征。可以看出，公开格式的主要优势在于模型更新的高度灵活性和对区域特定模型的支持。为了支持跨域模型传输，业界需要定义这样一种可共同识别的格式。

表3-2. 模型的公开格式与专有格式特点汇总

格式 / 特性	公开格式	专有格式
可否跨厂商互通模型信息	是	否
厂商依赖性	低	高
模型存在的问题	透明	不透明
模型更新和优化的灵活性	好	差*
区域级模型的支持	高效	受限**
离线优化需求	小	大
多厂商部署灵活性	好	差
模型存储开销	小	大

*公开格式允许任意拥有数据的节点更新模型；专有格式只允许能进行编译的节点更新模型
**公开格式中，每个便于获取区域级数据的节点都可以参与训练；专有模式中，能进行编译的节点不一定能获得区域级数据

我们认为，以公开格式模型传输为基础的多种学习方法将会为丰富的6G AI用例提供支撑，保障模型更好地匹配移动通信复杂多变的环境，实现更优的推理性能。

3.6 持续自演进

无线环境和系统需求随着时间都会不断变化，所以用例和模型都需要不断地演进来适配系统。这种演进，如果仍依赖于人工调参和用例选择，演进效率将会大打折扣。因此，如何将模型和用例的演进过程自动化是一个值得深入研究的问题。融合了AI的未来移动通信系统在运行过程中将不断地、自动地收集数据、提取知识、与环境和用户迭代交互，自动化地实现旧模块的更新、淘汰以及新模块的衍生，逐步搭建更高效的通信系统，称之为自演进。内生的统一生命周期管理和基于数据面的统一数据收集为AI自演进搭建了基础。按照难易程度，AI自演进可划分为L1~L3共三个层级，自演进能力逐级提升。

L1级自演进

AI模型参数自演进，如图3-8所示。L1级自演进的是在短时间内完成模型参数的自动更新，适配业务需求和部署环境的变化。该阶段的自演进主要依靠迁移学习、元学习、强化学习等在线学习实现。通过预定义一套在线的模型参数调整框架，对于特定的模型，自动地进行数据收集、在线训练、模型验证和模型参数取值替换等操作。

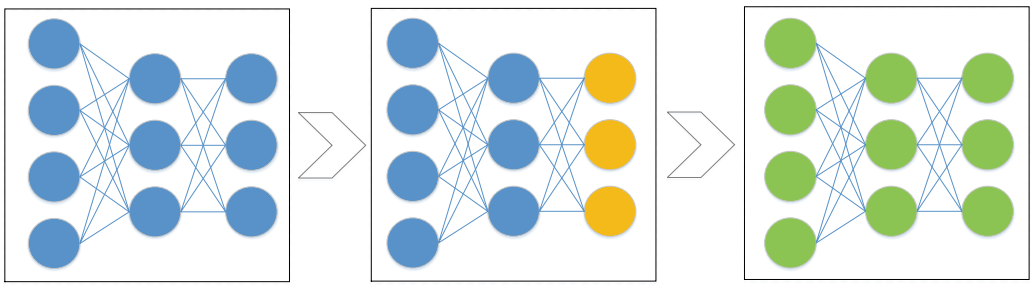


图3-8. L1级自演进示意图

L2级自演进

AI模型超参数（如输入、输出、结构）自演进，如图3-9所示。L2级自演进是L1级自演进的高阶版本，不仅可以解决参数适配问题，还可以基于实际环境中的数据自动地找出最适配该用例的模型超参数。该阶段的自演进主要依靠自动化机器学习（Automatic Machine Learning, AutoML），神经架构搜索（Neural Architecture Search, NAS）等技术来实现。通过预定义一套在线模型架构调整框架，对于特定的用例，自动地进行数据收集，在线模型搜索与训练、模型验证和模型替换等操作。

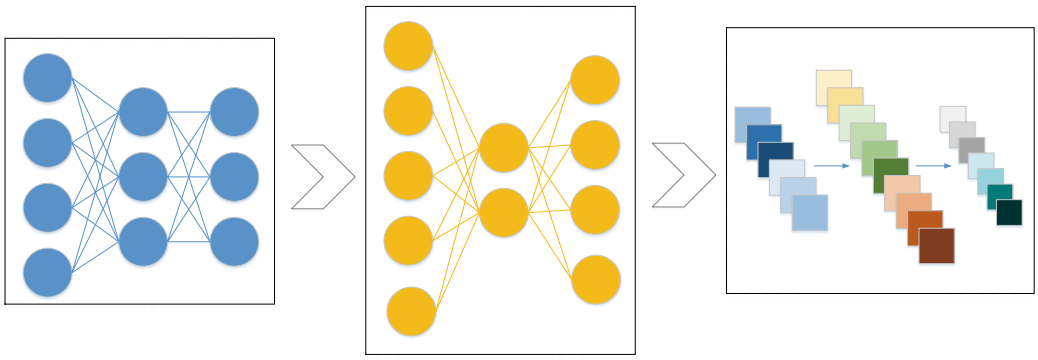


图3-9. L2级自演进示意图

L3级自演进

AI用例自演进，如图3-10所示。这一级别的自演进跳出了特定用例的束缚，可以对新用例进行探索，也可以对旧用例进行淘汰。而用例变化的过程中自然伴随着L1和L2级的自演进。L3级自演进实现的是用例自生成，需要具备灵活的数据收集，用例自主探索等高级功能以促进旧用例的更新、重组、淘汰以及新用例的诞生。

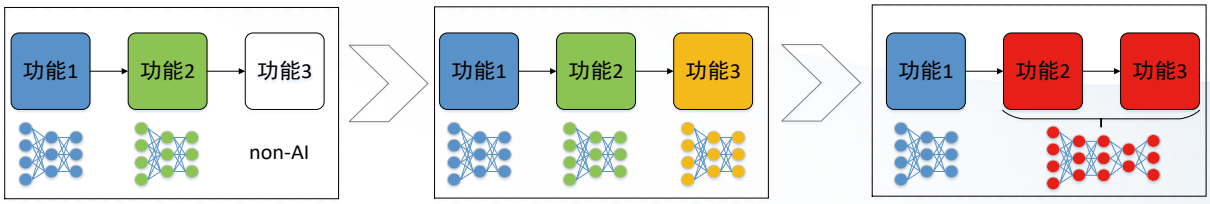


图3-10. L3级自演进示意图

从对三个自演进阶段的描述可以看出，L1级的自演进是整个自演进体系的基础。为了达到L1级自演进，需要新增演进自动触发、训练数据自动收集、模型自动训练、模型自动验证和模型参数取值自动替换等操作。其中，演进自动触发包括收集执行结果或监视结果，进而判断是否启动自演进。触发了自演进后，管理功能向训练功能发送数据收集，训练策略和验证相关的参数配置。训练数据自动收集是指数据源根据数据收集指示进行数据收集并将数据反馈给训练功能。模型自动训练是指训练功能根据管理功能配置的训练策略（包括学习方法，训练超参数，优化函数指示，训练验证数据集划分方法等）进行模型训练。模型自动验证是指训练功能根据管理功能配置的验证数据集划分方法和验证KPI确定模型性能。如果验证结果满足预设门限，则训练功能将模型参数和版本发送给推理功能，推理功能用新的模型参数进行参数取值替换，并通知训练功能参数更新成功，训练功能向管理功能反馈模型版本和验证性能。

相比于L1级自演进，L2自演进有一些新的需求：L2自演进中模型的输入输出是可调整的，进而数据收集配置也需要对应地调整；在L1自演进的在线自动训练训练中超参数，优化函数等都是固定的，而在L2自演进中，模型结构，训练超参数，训练算法，优化函数等都可调整。因此，为了达到L2级自演进，需要将L1级自演进已有的训练数据自动收集和模型自动训练策略变得更灵活，还需再增加模型结构和参数自动更新等操作。其中，模型自动训练策略的灵活化可通过多种配的模型尝试或AutoML、NAS等技术实现。模型结构和参数自动更新是指训练功能向推理功能发送更新的模型结构、参数和版本，推理功能完成模型更新后向训练功能反馈模型更新成功。

为了达到L3级自演进，相比于L2级自演进需要新增的功能包括更灵活的数据收集和更灵活的模型部署等操作。在L2级自演进的配置中，数据收集配置是与已有的特定用例对应的。然而，在L3自演进中可以是为当前非AI功能收集数据训练AI模型，或者是为当前2个AI用例融合后的功能收集数据训练AI模型，因此需要更为灵活的数据收集配置。灵活的模型部署是指支持将某功能从非AI替换为AI，或者AI替换为非AI，或者将多个独立功能替换为融合AI功能。

综上所述，自演进的实现复杂度从L1至L3逐级增加。我们认为，L1级自演进在6G实现的可能性最大。L2级自演进的实现取决于AutoML技术的成熟度和算力的增长速度，也有可能在6G实现。L3级自演进需要灵活的数据收集和交互接口配置，属于更长远的愿景。

04

第四章

结　束　语

AI与通信的融合涉及空口、网络、协议、算法等多个维度，并将深刻影响感知、计算和控制等场景或功能，有望推动未来通信范式的演进和网络架构的变革，对未来移动通信技术的研究具有重要意义。

AI与通信融合的用例丰富多彩，不仅涉及了移动通信网络的多个层，还具有多种协作模式，在提高移动通信系统性能和降低复杂度方面均有很高的价值。随着学术界和产业界对AI与通信融合的深入研究以及AI资源的普及，将会有更多的高价值用例涌现。

为了支持如此多的用例，需要在6G设计之初就基于多种用例的共性特征和需求，设定合理的设计原则，打造通用性强、部署灵活、资源利用率高、且支持多种新型学习架构和具备演进能力的6G AI架构。6G将是一个AI用例更加丰富、AI能力灵活普适的智能系统。

参考文献

[1] vivo，“6G服务，能力与使能技术白皮书”，2022.7

[2] 3GPP, R1-2203550, “Evaluation on AI/ML for CSI feedback enhancement”, vivo, RAN1 #109e, 2022.5

[3] 3GPP, R1-2304471, “Evaluation on AI/ML for CSI feedback enhancement”, vivo, RAN1 #113, 2023.5.

[4] vivo, “基于AI的DMRS信道估计-多种信道验证”, IMT-2030_WX_AI, 202304

[5] 3GPP, R1-2306742, “Evaluation on AIML for beam management”, vivo, RAN1 #114, 2023.8

[6] 3GPP, R1-2306744, “Evaluation on AI/ML for positioning accuracy enhancement”, vivo, RAN1 #114, 2023.8

[7] vivo，“基于AI的对抗PA非线性方法”, IMT-2030_WX_AI, 202305

[8] 3GPP, TS38.323, “Packet Data Convergence Protocol (PDCP) specification (v17.5.0)”, 2023.6

[9] 刘嘉，人工智能的认知神经基础论坛，2022北京智源大会，2022.6

缩略语

英文缩写	英文全称	中文
3GPP	3rd Generation Partnership Project	第三代合作伙伴计划
4G	The fourth generation mobile communication systems	第四代移动通信系统
5G	The fifth generation mobile communication systems	第五代移动通信系统
6G	The sixth generation mobile communication systems	第六代移动通信系统
AI4NET	AI for Network	为网络服务的AI
AMC	Adaptive Modulation and Coding	自适应调制编码
AMP	Approximate Message Passing	近似消息传递
AR	Augmented Reality	增强现实
ASIC	Application Specific Integrated Circuit	应用专用集成电路
AutoML	Automatic Machine Learning	自动化机器学习
BLER	Block Error Rate	误块率
CN	Core Network	核心网
CSI	Channel State Information	信道状态信息
CSI-RS	Channel State Information Reference Signal	信道状态信息参考信号
DL-TDO	Downlink Time Difference of Arrival	下行到达时延差
DMRS	De-Modulation Reference Signal	解调参考信号
DPD	Digital Pre-Distortion	数字预失真
EVM	Error Vector Magnitude	误差矢量幅度
GPU	Graphics Processing Unit	图形处理器
HARQ	Hybrid Automatic Repeat request	混合自动请求重传
LTE	Long Term Evolution	长期演进
MIMO	Multiple Input Multiple Output	多输入多输出
NAS	Neural Architecture Search	神经架构搜索
NET4AI	Network for AI	为AI服务的网络

英文缩写	英文全称	中文
NLOS	Non line of Sight	非直视
NMSE	Normalized Mean Square Error	归一化均方误差
NPU	Neural Network Processing Unit	神经网络处理器
NR	New Radio	新空口
OFDM	Orthogonal Frequency Division Multiplexing	正交频分复用
PA	Power Amplifier	功率放大器
PAPR	Peak-to-Average Power Ratio	峰均功率比
QoS	Quality of Service	服务质量
RAN	Radio Access Network	接入网
RRM	Radio Resource Management	无线资源管理
RU	Resource Utilization	资源利用率
TOA	Time of Arrival	到达时间
TPU	Tensor Processing Unit	张量处理器
TR	Tone Reserve	预留子载波
TRP	Transmission/Reception point	发射/接收点
TRS	Tracking Reference Signal	追踪参考信号
TTT	Time-To-Trigger	触发时间
UE	User Equipment	终端
UPF	User Plane Function	用户面功能
VR	Virtual Reality	虚拟现实