中国科学: 信息科学 2018 年 第 48 卷 第 12 期: 1589–1602
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

基于 AI 的 5G 技术 —— 研究方向与范例

尤肖虎[1], 张川[1*], 谈晓思[1], 金石[1], 邬贺铨[2]

1. 东南大学移动通信国家重点实验室, 南京 210096
2. 中国工程院信息与电子工程学部, 北京 100088
* 通信作者. E-mail: chzhang@seu.edu.cn

收稿日期: 2018–07–05; 接受日期: 2018–08–14; 网络出版日期: 2018–11–27

国家自然科学基金 (批准号: 61501116, 61521061) 资助项目

摘要 第 5 代移动通信 (5G) 技术将为移动互联网的快速发展提供无所不在的基础性业务能力, 在满足未来 10 年移动互联网流量增加 1000 倍发展需求的同时, 为全行业、全生态提供万物互联的基础网络技术. 相对已有的移动通信技术, 5G 技术适用面更为广泛, 系统设计也更为复杂. 重新复兴的人工智能 (AI) 技术为 5G 系统的设计与优化提供了一种超越传统理念与性能的可能性. 本文在概述 5G 移动通信关键技术的基础上, 梳理了 AI 技术在 5G 系统设计与优化方面富有发展前景的若干发展方向, 并给出了有关 5G 网络优化、资源最优分配、5G 物理层统一加速运算以及端到端物理层联合优化等若干典型范例.

关键词 5G 移动通信, AI 技术, 网络优化, 资源分配, 统一加速, 端到端联合优化

1 引言

5G 是面向 2020 年之后发展需求的新一代移动通信系统, 其主要目标可概括为 “增强宽带、万物互联”. 5G 应用被划分为 3 个典型的场景: 增强型的移动宽带 (enhanced mobile broadband, eMBB)、海量连接的机器通信 (massive machine-type communications, mMTC), 以及高可靠、低时延的物联网应用 (ultra-reliable and low-latency communications, URLLC), 并规定了多个维度的关键技术指标需求, 包括: 峰值速率、频谱效率、时间延迟、网络可靠性、连接密度, 及用户体验速率等. 大规模天线阵列、密集网络、新型波形复用与信道编译码 (如极化码), 以及毫米波接入将成为 5G 无线网络的核心关键技术 [1~4]. 5G 另一项富有前景的关键技术是网络虚拟化与切片技术, 其基本思想是将 5G 网络构建在云计算平台上, 通过计算资源的隔离、动态调配与迁移, 实现网络资源的灵活调配, 以适应未来 5G 极为丰富的应用场景 [5].

引用格式: 尤肖虎, 张川, 谈晓思, 等. 基于 AI 的 5G 技术 —— 研究方向与范例. 中国科学: 信息科学, 2018, 48: 1589–1602, doi: 10.1360/N112018-00174
You X H, Zhang C, Tan X S, et al. AI for 5G: research directions and paradigms (in Chinese). Sci Sin Inform, 2018, 48: 1589–1602, doi: 10.1360/N112018-00174

5G 技术标准正处于不断完善与成熟的过程中 [6,7]. 国际标准化组织 3GPP 于 2017 年 12 月公布了第一个 5G 技术标准, 支持非独立组网 (non-standalone, NSA) 与 eMBB 功能 [8]. 2018 年 06 月 14 日, 3GPP 批准了 5G 独立组网 (standalone, SA) 技术标准, 5G 自此进入了产业全面冲刺的新阶段 [9]. 在 4G 技术基础上, 5G 新无线接口 (5G new radio, 5G NR) 的改进包括: (1) 对已有的多输入多输出 (multiple-input multiple-output, MIMO) 技术进行了增强, 引入了大规模天线阵列技术; (2) 对 OFDM 时隙结构和时频资源块 (resource block, RB) 划分方案进行了补充, 提出了更为灵活的空中接口技术; (3) 预计下一个 5G 标准版本将引入非正交多用户接入 (non-orthogonal multiple access, NOMA) 技术, 以支撑广域覆盖的中低速率物联网应用; (4) 沿用了前期的分布式无线网络构架 [10], 将无线网络功能单元划分为分布式单元 (distributed units, DU) 和中心单元 (central units, CU), 并引入了基于云计算的网络虚拟化与切片技术.

5G 技术应用范围的扩展使得其系统设计和优化更为复杂. 传统移动通信系统的优化目标主要体现在对系统传输速率和移动性能力的支持方面. 5G NR 将其应用特性的支持能力进一步扩展至: 时间延迟、网络可靠性、连接密度, 及用户体验速率等多个关键性能指标 (key performance indicator, KPI) 方面. 5G NR 系统设计需要在这些 KPI 之间进行折中与优化 [11]. 同时灵活空中接口、网络虚拟化与切片技术的引入, 极大地增加了系统设计的复杂性, 并为 5G 网络运维和优化带来了极大的挑战. 可喜的是, 人工智能 (artificial intelligence, AI) 技术为 5G 系统的设计与优化提供了一种超越传统理念与性能的可能性, 已成为业界重点关注的研究方向. 3GPP, ITU 等组织均提出了 5G 与 AI 相结合的研究项目 [12,13].

AI 技术诞生于 20 世纪中叶, 几经沉浮, 近年来借助于现代计算和数据存储技术的迅猛发展而再次复兴. AI 技术涵盖遗传算法 [14] 和人工神经网络 [15,16] 两大类, 其本身是一种普适性的机器学习技术. 凡是给定场景涉及到了数据的统计、推断、拟合、优化, 及聚类, AI 均能找到其典型应用. 根据训练数据的标签 (label) 或对应关系是否已知来区分, AI 学习算法可以粗略地分为监督学习和非监督学习两类. 而增强学习算法既不是一般意义上的非监督学习算法, 也不是监督学习算法, 因此又自成一类. 以上 3 类学习算法的特点和典型范例如下:

- **监督学习.** 在监督学习中, 每个训练数据组 (data pair) 都是由一个输入对象和一个期望的输出值组成的, 其目标是习得输入和输出数据的一种函数关系, 并依据该函数关系推断其他输入数据可能的输出值. 监督学习的一种典型范例即为图 1 所示的深度神经网络 (deep neural networks, DNN) 的训练. 该训练方法通过一组先验的数据对 (data pairs) 对多层人工神经网络节点间的加权系数进行离线训练. 当训练收敛后, 该分层人工神经网络可以实现对新数据的辨识与推断.
- **非监督学习.** 非监督学习的训练数据是无标签的数据. 通过非监督学习, 我们试图找到这些数据中隐藏的结构. 自组织映射 (self-organizing map, SOM) 网络的训练就用到了非监督学习方法. 在 SOM 中, 如图 2, 任意维度的无标签训练数据被输入一个人工神经网络, 转换为低维度 (通常为二维) 的离散映射 (map), 并通过非监督学习算法对权重向量有选择性地微调, 以拓扑有序的方式自适应地执行这种变换.
- **增强学习.** 增强学习的典型例子是: 可在线处理的增强学习方法 (见图 3). 它基于智能实体 (agent) 与环境 (environment) 之间的动态交互. 当智能实体感知到环境信息后, 依据自己采取动作 (action) 所可能带来的奖赏 (reward) 或惩罚 (penalty), 确定下一步动作, 并进一步观察环境的反应, 循环往复, 直至收敛至某一稳态目标.

两种常见的学习方法如下:

- **反向传播学习算法.** 反向传播 (backpropagation, BP) 学习算法是分层人工神经网络中最为经典

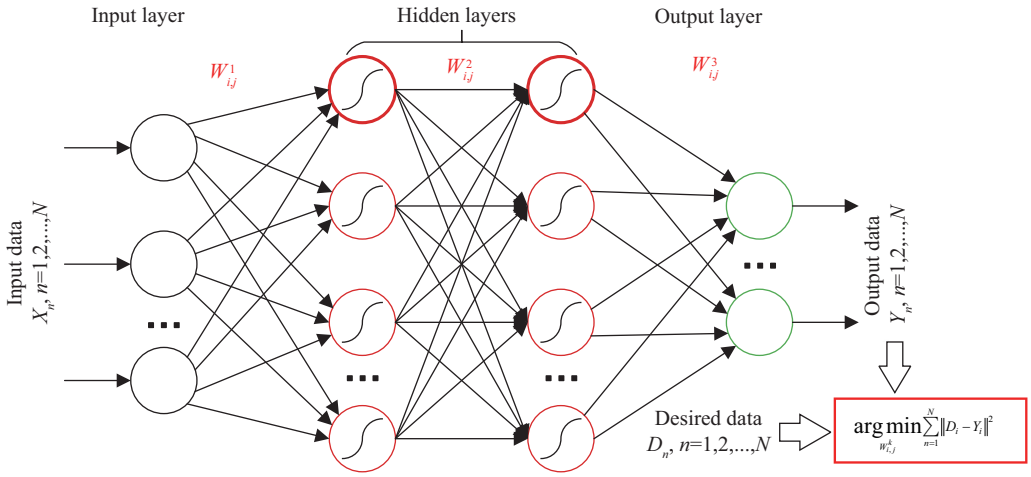

图 1 (网络版彩图) 监督学习示例: 多层人工神经网络学习

Figure 1 (Color online) Example of supervised learning: learning in deep neural network

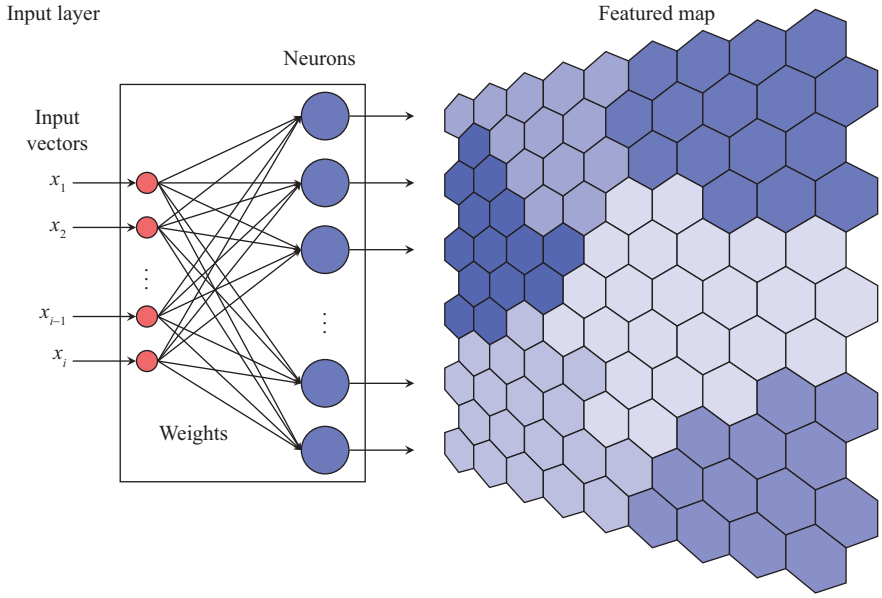

图 2 (网络版彩图) 非监督学习示例: 自组织映射网络的学习

Figure 2 (Color online) Example of unsupervised learning: self organizing map

的训练算法 [15], 是最速下降优化算法的一种具体形式. 其目标是通过迭代, 调整节点间的加权系数, 使得分层人工神经网络的输出逐步接近已知的输出. 有关 BP 算法的动态参数优化及加速实现形式, 可参见作者早期的有关论文 [16, 17]. 有关分层人工神经网络拓扑结构的选择, 以及如何避免陷入局部最优, 可参见作者早期的论文 [18]. 而今, BP 学习算法被广泛用于训练深度神经网络 (DNN), 并取得了良好的效果. 拥有两层或更多隐藏层 (hidden layer) 的神经网络均可被称为 DNN. 卷积神经网络 (convolutional neural networks, CNN) 是一类常见的前馈 DNN, 其隐藏层包含: 卷积层、池化层、全连接层, 及归一化层. CNN 也可以使用反向传播算法进行训练, 并能在图像和语音识别等方面给出比其他 DNN 更好的结果.

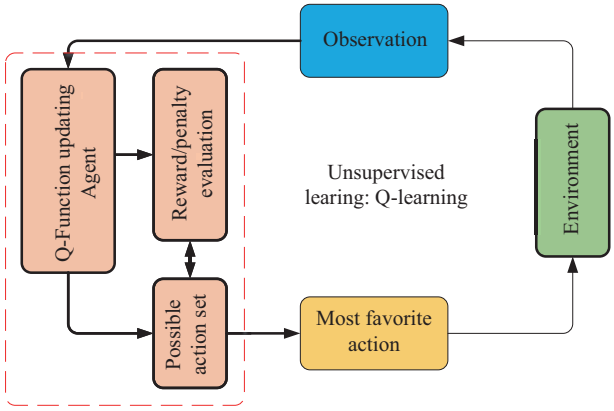

图 3　(网络版彩图) 增强学习示例: Q 学习算法

Figure 3　(Color online) Example of reinforcement learning: Q-learning

• **Q 学习算法.** Q 学习算法又称为 Bellman 算法 [19], 是增强学习最为经典的算法. 其基本思路是: 选择某一函数 (Q 函数), 作为衡量一个智能实体执行某种动作的代价 (奖励或惩罚) 函数; 该智能实体根据所处的环境, 对所有可能的动作进行 Q 函数评估, 并从中选择出奖励成份最大的动作, 并加以行动. Bellman 给出了 Q 函数常见的迭代更新形式, 从而使该智能实体的动作、环境的变化和 Q 函数的调整能以在线的方式实现. Q 算法的迭代收敛性证明可参见文献 [20].

传统数据科学的 AI 机器学习算法 (例如线性模型 (linear models)、决策树 (descision tree)、k 均值聚类 (k-means clustering) 等) 日臻成熟并已经部分投入商用. 而近年来, 深度学习方法 (例如 DNN、CNN、增强学习等) 迅速发展, 并在认知技术等领域取得了重大突破, 受到了前所未有的关注. 与此同时, 元学习 (meta-learning) 等深度学习算法的新分支也在不断开拓中, 提出了 “学会学习” 等崭新的概念. 例如, 文献 [21] 提出的未知模型元学习法 (model-agnostic meta-learning, MAML) 不会对模型的形式做任何假设, 也没有为元学习引入额外的参数, 极易应用于许多领域, 包括分类、回归和增强学习等. 近期 AI 技术的发展在文献 [22~24] 中有详细的总结. 这些新技术使深度学习方法适用于更广泛、更复杂的场景, 也为 AI 在通信等领域中的应用制造了新的机遇.

2　AI 应用于 5G 系统的发展方向

作为普适性的机器学习技术, AI 可以广泛地应用于 5G 系统设计和优化的各个环节, 大体上涉及 3 类技术问题: 组合优化问题、检测问题, 及估计问题.

• **组合优化问题.** 5G NR 的资源分配问题是一个典型的组合优化问题. 它需要从资源池中穷举出一组最优的资源配置方式, 并据此将资源分配给网络覆盖范围内的多个用户, 最大化资源利用效率.

• **检测问题.** 5G 通信最优接收机的设计就是一个典型的检测问题, 其目标是对接收信号进行辨识, 确定对应的发射信号, 并使检测错误概率最低.

• **估计问题.** 5G 通信的信道参数的准确估计是实现系统相干接收的必要条件. 它需要根据 5G 系统所发送的导频信号 (事先确知), 估计出无线信号传播从发射端到接收端所历经的信道畸变.

AI 技术在 5G 系统中的应用已有大量的文献 [25~37] 可供参考, 但并非所有的研究均有潜在的发展生命力. 首先, 移动通信系统本身经过长期的发展, 已经拥有了较为完备的经典设计和处理方法. 大量实践已经证明, 这些经典方法在工程上极为有效, 且易于实现. 其次, 移动通信系统通常存在性能界

(如 Shannon 容量限), 现有的方法经过精心设计已经可以充分逼近上述性能界. 例如, 文献 [38] 中给出了逼近 MIMO 容量界的一种设计方法, 其只需对功率最优注水分配问题进行简单的迭代, 便可使 MIMO 的性能充分逼近 Shannon 容量界. 这意味着, 即使采用先进的 AI 学习技术, 也无法超越这些经典算法. 第三, AI 学习本身也有明显的局限性, 无论是 BP 算法还是 Q 算法均存在训练的收敛时间问题. 能否满足移动通信系统实时处理的需求, 需要进行较为充分的评估. 最后, 与经典的方法相比, AI 学习算法的计算复杂度通常较高, 如果不能带来性能上的明显提升, 其本身显然不具备足够的竞争力.

这里, 并非否定 AI 技术在 5G 系统设计与优化中的潜在价值. 相反, 5G 系统中存在大量传统方法难以建模、求解, 或高效实现的问题, 为 AI 技术在 5G 中的有效应用提供了可能. 同时, 一些新的 AI 算法正在不断发展中, 为 AI 在 5G 中的应用提供了新的机遇. 因此, 有必要对 AI 技术在 5G 系统设计与优化中的典型问题进行梳理, 从而确定其在 5G 系统中更有潜在应用价值的研究方向.

为此, 本文将 AI 在 5G 中的应用问题划分为以下 4 种类型: 无法建模问题、难以求解问题、统一模式高效实现问题, 及最优检测与估计问题. 我们将看到: 对于前两类问题, 由于缺乏有效的传统解决方案, AI 技术通常更具应用潜力. 而对于后两类问题, AI 技术相对于传统的解决方案是否在性能或实现上具备更强的竞争力, 则视其具体解决方案而定. 4 类问题具体分析如下:

- **无法建模问题.** 移动通信系统的网络优化涵盖一类难以统一建模的技术问题, 如: 覆盖问题、干扰问题、邻区选择, 及越区切换问题等, 其运维通常更多地依赖于工程人员的实践经验. 相比之下, 5G 系统涉及的应用场景更为综合, KPI 种类更多, 特别是 5G NR 中基于大规模天线阵列的密集波束应用 [39], 需要引入更高维度的优化参数, 对网络运维带来了更为艰巨的挑战. 5G NR 另一个难以建模的问题是 KPI 优化. 如前所述, 5G NR 的 KPI 涉及峰值速率、频谱效率、时间延迟、网络可靠性、连接密度, 及用户体验速率等多个维度, 这些指标往往是相互依赖或相互矛盾的 [11], 因而难以建立全局性的优化模型.
- **难以求解问题.** 5G NR 涉及一系列资源分配问题 [39,40], 包括: 小区间时频资源块分配、正交导频资源分配、波束分配、大规模 MIMO 多用户聚类, 及无线网络虚拟化资源池调配等. 这些问题的模型优化目标是使得整个无线网络的吞吐率最大化, 并满足一定程度的用户服务比例公平性. 其最优解求解问题通常属于 NP-hard 类型的组合优化, 对应的计算复杂度随系统规模的增加而指数增长. 传统的解决方法一般将此类问题进行静态分割, 从而以较低的计算量获取次优的解决方案. AI 技术则为此类问题的解决提供了可能的技术途径.
- **统一模式高效实现问题.** 5G NR 涉及一些基本功能模块的级联组合. 以 5G NR 的物理层为例, 涉及大规模 MIMO 多用户空时处理、NOMA 信号检测, 及 LDPC 码和 Polar 码信道编译码等功能模块. 上述每个功能模块算法各不相同, 但理论上均可以单独采用 AI 学习技术逐一加以解决 [25~33]. 这启发我们, 可用统一的、基于 AI 技术的软硬件方案解决 5G NR 物理层所有的关键功能模块 [41], 从而简化系统的设计流程、加速工程实现的进程、提高物理层实现的可配置性, 并最终降低系统实现成本、提高实现效率.
- **最优检测与估计问题.** 将 AI 应用于 5G 系统的直观想法是, 用人工神经网络来取代传统发射机和接收机的基本功能模块. 如前所述, 基于人工神经网络训练的方法, 其无线传输性能最多也只能接近 Shannon 容量界; 但相对于经典的优化设计方法, 其计算量可能更为庞大, 且训练所需的收敛时间也会限制其实际应用. 另一个更值得探讨的研究方向是, 使用 AI 技术进行跨层联合优化 [42], 包括: 物理层与媒体控制层的联合优化 [43]、信源与信道的联合优化 [44], 及算法设计与硬件实现的联合优化等 [45], 这些均是传统方法所难以解决的.

3 AI 应用于 5G 系统的典型范例

本节将具体介绍 AI 应用于 5G 系统的 4 个典型范例: 网络自组织与自优化、时频资源最优分配、5G 通用加速器, 及 5G 物理层端到端优化.

3.1 网络自组织与自优化

自组织网络 (self organizing network, SON) 已被 3GPP 列为 LTE 网络优化关键技术. 相比于传统无线通信, 5G 应用场景更为复杂, 网络优化与管理更为艰难, 因此对 SON 的技术需求将更为强烈. SON 包括了网络自配置、自优化, 及自愈合 3 项功能, 旨在淡化传统人工干预, 实现网络规划、网络配置, 及网络优化的高度自动化, 以节省运营成本, 降低人为故障. 文献 [46~48] 对 AI 技术在 SON 中的应用进行了总结, 涉及基站自主参数配置、动态规划、迁移学习, 网络故障的自动检测与定位, 及网络参数的自动优化等, 所采用的 AI 方法包括人工神经网络学习、蚁群优化, 及遗传算法等.

下面以文献 [34] 提出的自动故障分析为例, 介绍 AI 技术在 SON 中的具体应用. 为了实现 LTE 网络自动故障诊断, AI 技术需要克服两大问题: (1) 现有大量数据的 KPI 种类多, 又缺乏已知的故障标签, 难以进行简单的归类诊断; (2) 鉴于人工诊断的成本较高和能力有限, 需要尽量减少人工参与. 因此, 研究者结合监督和无监督两种学习方法, 提出了基于 AI 的故障诊断系统 [34]. 诊断分为以下几步:

步骤 1. 利用图 2 所示的无监督的自组织映射 (SOM) 算法实现对无标签高维度数据的初步分类. 多种类的 KPI 指标带来了高维的历史数据. SOM 作为人工神经网络, 通过训练能将任意维的输入数据在输出层映射成二维神经元网络. 神经元的拓扑结构即代表了原数据的分布情况: 越相近的神经元, 其映射的原数据越接近. 这样就实现了高维数据的低维表示和初步聚类.

步骤 2. 完成 SOM 训练之后, 我们再对 SOM 建立的神经元进行一次无监督的聚类. 因为神经元之间的欧氏距离即表示其映射数据之间的差异, 所以基于欧式距离的沃德 (Ward) 聚类算法即可实现对神经元的聚类.

步骤 3. 经过以上两个步骤, 数据已经被分为几个大类. 此时, 再引入专家对分好类的数据进行故障分析, 有监督地贴上故障标签.

以上 3 个步骤完成了故障诊断系统的设计, 建立了一套自动的诊断流程, 如图 4 所示. 此后产生的新数据在输入该系统之后, 将先由 SOM 定位到最接近的神经元, 再由该神经元的类别标签判断其是否故障以及原因. 在诊断一定数量的新数据之后, 以上 3 个步骤可以再次被执行用以验证和更新系统. 文献 [34] 的仿真结果表明, 即便在主要使用无监督学习进行构建, 并且人工参与量极低的情况下, 上述自动故障诊断系统仍能达到非常高的诊断准确率.

3.2 时频资源最优分配

相比于 4G LTE-A, 5G NR 将面临更为复杂的 OFDM 时频资源块 (RB) 分配问题, 以适应 5G 3 种典型的应用场景. 图 5 给出了一个典型的多小区、多用户下行链路 RB 分配示意图. 其中, 同一小区内不同用户的 RB 分配是正交的, 系统整体干扰主要取决于相邻小区用户 RB 的分配方案. 假设每个用户的信息容量可在信干比 (SIR) 测量值的基础上得出, 则系统 RB 最优分配的目标是使所有用户的信息容量之和最大化. 这是一个典型的 NP-hard 组合优化问题, 所需的计算量与覆盖范围内移动用户数的阶乘成正比.

以基于 Q 学习算法的应用为例. 假设某一智能实体负责上述移动用户的 RB 分配, 则该智能实体的动作可以遵循以下原则对用户 RB 进行更新: (1) 在同一小区内, 选择 SIR 较好的空闲 RB 分配给

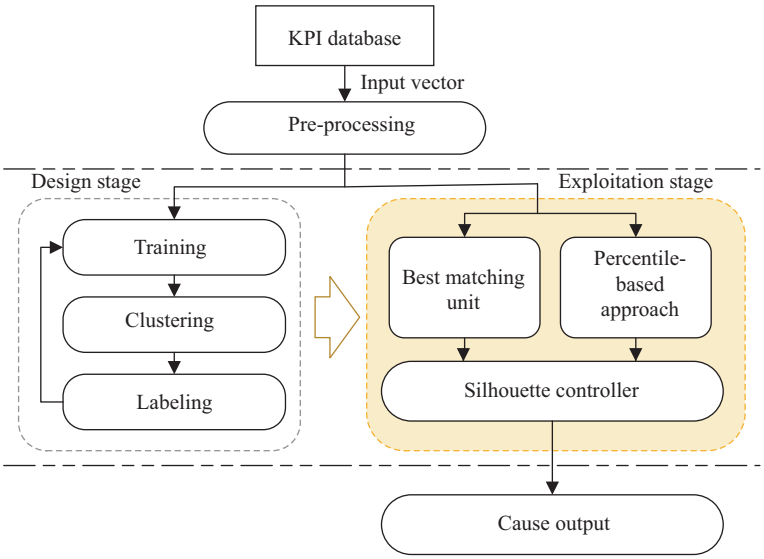

图 4 (网络版彩图) 自动故障诊断系统流程 [34]

Figure 4 (Color online) Automatic root cause analysis [34]

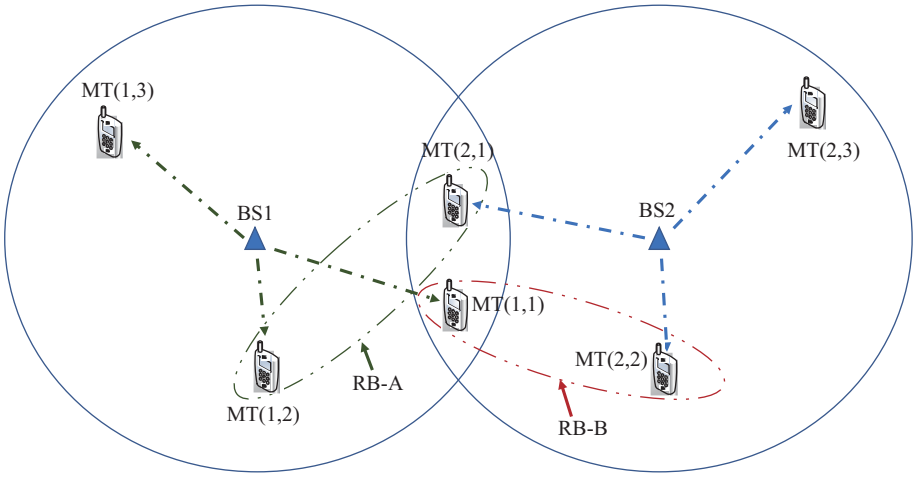

图 5 (网络版彩图) 多小区多用户动态无线资源分配

Figure 5 (Color online) Dynamic resource allocation for multi-cell and multi-user systems

用户; (2) 不断更新本小区 SIR 最差用户的 RB, 以寻求更好的系统性能; (3) 对于同一 RB, 把本小区 SIR 最差的用户与邻小区 SIR 最好的用户进行配对或分簇, 如图 3 所示. 前两个原则易于理解, 而第 3 个原则旨在避免位置相近, 且处于小区边缘的用户被分配相同的 RB. 此时, 相邻基站无论如何调整发射功率, 这些用户均无法同时获得正常工作所需的 SIR.

智能实体在动作集合确定后, 以所有用户的信息容量之和最大化为准则, 选择当前最优的动作对 RB 进行调整, 并按照 Bellman 公式对 Q 函数进行实时更新 [19]. 如此迭代操作, 直至 Q 函数趋于稳定.

上述迭代过程还应与用户的功率最优分配相结合. 文献 [49] 基于博弈论框架, 给出了多小区用户采用相同 RB 时的最优功率分配方法. 如果系统需要进一步考虑用户 QoS 的比例公平性, 可引入

Lagrange 乘子法, 构造相应的动作评估准则和 Q 函数. 文献 [50, 51] 对 AI 在 5G 资源分配中的一些应用进行了总结和展望. 而文献 [52] 也初步提出了增强学习方法在 5G 新框架中的网络切片技术方面的新应用.

3.3 5G 通用加速器

相比于 4G, 5G 的基带处理需要考虑更多的模块, 例如, 大规模 MIMO 检测、NOMA 检测, 及 Polar 码译码等, 这会使硬件面积增加, 实现架构不规律. 可以注意到, 尽管 5G 基带模块众多, 但所有功能均可以用基于因子图的置信传播算法实现 [53~57]. 针对特定的基带功能, 置信传播算法只需确定变量符号集、变量间关系等参数, 而保持其余的部分不变. 因此可以用一个基于置信传播的、参数可配置的通用加速器实现整个基带功能.

尽管基于置信传播的算法可以实现 5G 通用加速器, 但受其性能限制, 置信传播算法在一些场景下仍然无法满足要求. 为此我们尝试在置信传播通用加速器的基础上, 实现基于 AI 的 5G 通用加速器. 我们可以通过以下两种方式, 将一个置信传播算法改进为一个性能更好的 AI 算法.

方法一. 将置信传播算法改为深度神经网络 (DNN) 算法. 方法如下: (1) 将置信传播算法的因子图复制多次, 并按照原有的方式连接为一个深度神经网络, 复制次数等于置信传播算法的迭代次数. (2) 对 DNN 进行训练. 文献 [30] 提出的基于 DNN 的 Polar 码译码器, 文献 [58] 提出的基于 DNN 的 MIMO 检测器等是方法一在基带模块上实现的范例.

方法二. 将置信传播算法改为卷积神经网络 (CNN) 算法. 方法如下: (1) 将置信传播算法的因子图节点排列在图片上, 其中每个像素代表一个节点. 像素相邻意味着其对应的节点在因子图中相连. (2) 用连接得到的图像对 CNN 进行训练. 文献 [35] 提出的 BP-CNN 信道译码器应用了方法二.

AI 技术中的神经网络算法具有高度的自适应性与可靠性. 通过将基带算法转化为神经网络能够处理的问题, 我们受益于以下两方面: (1) 系统性能的提升; (2) 硬件架构的统一. 观察神经网络的算法, 我们发现 CNN 核心操作是卷积运算, 而 DNN 核心操作为二维矩阵乘法运算. 二维脉动阵列可以同时完成上述两种运算. 因此利用同一脉动阵列, 只需对输入数据进行合理调度, 即可同时实现 CNN 与 DNN 的功能, 从而实现基于 AI 的 5G 通用加速器. 文献 [28] 综合研究了神经网络的高效量化策略以及硬件实现. 一种二维脉动阵列架构如图 6 所示, 可以看出其高度的规整性和可扩展性.

文献 [31] 指出, 在由信道编码、信道、信道均衡器, 及译码器组成的系统中 (如图 7), 接收机的均衡器与译码器分别可用 CNN 和 DNN 实现. 对应的 AI 加速器有两种实现方案. (1) 硬件消耗优先的设计: 我们可将整个接收机折叠为一个通用处理器. 通用处理器首先工作于均衡器模式 (CNN), 输入为来自信道的信息, 输出被保存下来; 然后工作于译码器模式 (DNN), 将刚保存的结果作为输入, 输出最后的译码结果. (2) 吞吐速率优先的设计: 接收机由两个通用处理器组成流水线, 两个处理器分别工作在均衡器和译码器模式. 可以看出, 相比于传统实现, 通用处理器可带来更多硬件设计自由度, 以更好地满足不同系统要求.

3.4 5G 物理层端到端优化

AI 算法在物理层若干模块上成功实现了功能优化, 例如, 文献 [36] 提出的基于神经网络的调制模式识别, 文献 [30] 提出的 DNN 极化码译码器, 及文献 [35] 提出的基于 DNN 的 MIMO 检测算法等. 在两个或多个物理层模块的联合优化问题中, AI 算法也得到了成功的应用, 例如文献 [31] 提出的基于神经网络的信道均衡和信道译码的联合优化. 文献 [37, 59] 均对 AI 在物理层各模块的应用进行了较

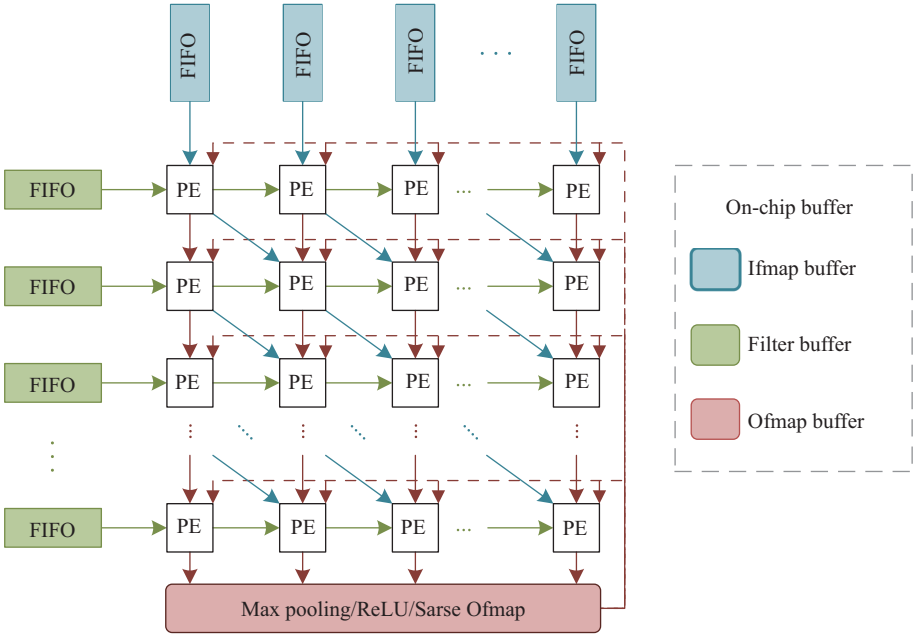

图 6 (网络版彩图) 神经网络的二维脉动阵列硬件架构 [28]

Figure 6 (Color online) Systolic array hardware structure for neural networks [28]

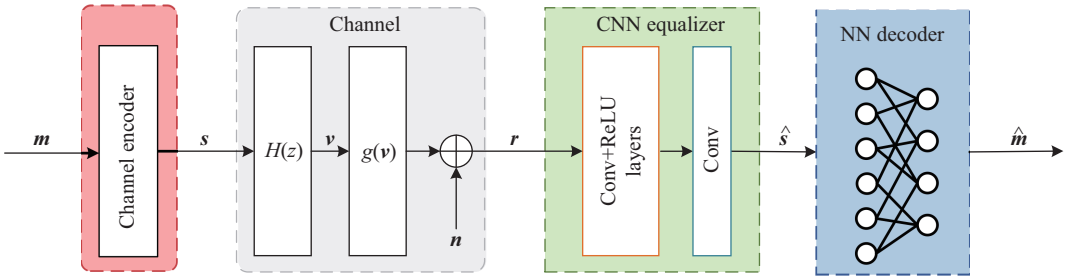

图 7 (网络版彩图) 包含均衡器与译码器的接收机示意 [31]

Figure 7 (Color online) Architecture of a receiver including neural network equalizer and decoder [31]

全面的总结. 但单个模块的优化并无法保证整个物理层端到端通信的整体优化 [37], 而端到端通信的实现中, 多个基于迭代算法的 AI 模块的拼接反而会带来更高的训练和计算复杂度. 因此, 我们需要一种对物理层端到端的联合优化方法.

文献 [59] 提出将物理层通信看作一个端到端的信号重构问题, 并应用自编码器概念来表示物理层通信过程, 进行端到端通信的联合优化. 自编码器是一种无监督深度学习算法, 属于神经网络, 通过学习输入信息的压缩形式来进行压缩信息的重构. 在利用自编码器构建的端到端通信模型中, 编码、调制、信道均衡等物理层模块, 被简单表示为发射端、信道和接收端 3 个模块: 发射端和接收端都分别表示为全连接的 DNN, 其中发射端连接一个归一化层来确保输出值符合物理约束, 接收端则连接一个 softmax 激活函数层, 最后输出一组概率向量来决定接收到的信息. 两者中间的 AWGN 信道则用神经网络的一个噪声层 (noise layer) 表示, 从而将通信系统表示为结构如图 8 所示的大型自编码器. 该自编码器基于端到端的误比特率 (BER) 或误块率 (BLER) 表现进行训练, 完成训练的自编码器即可基

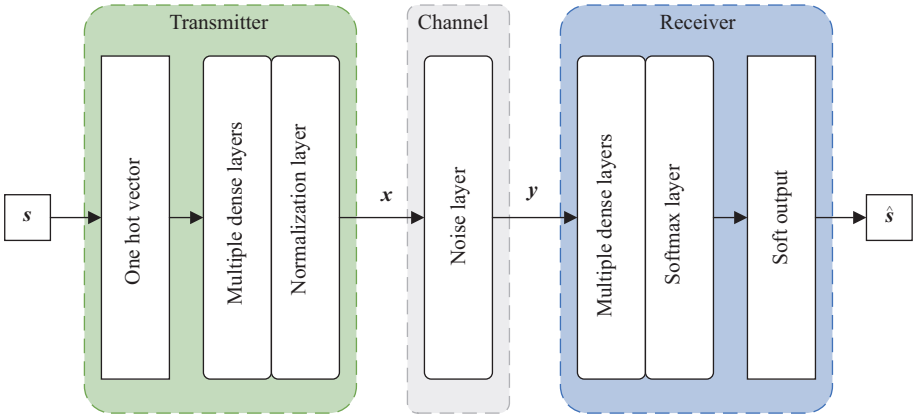

图 8 (网络版彩图) 用自编码器构建端到端通信模型的简单结构 [37]

Figure 8 (Color online) A simple autoencoder for an end-to-end communication system [37]

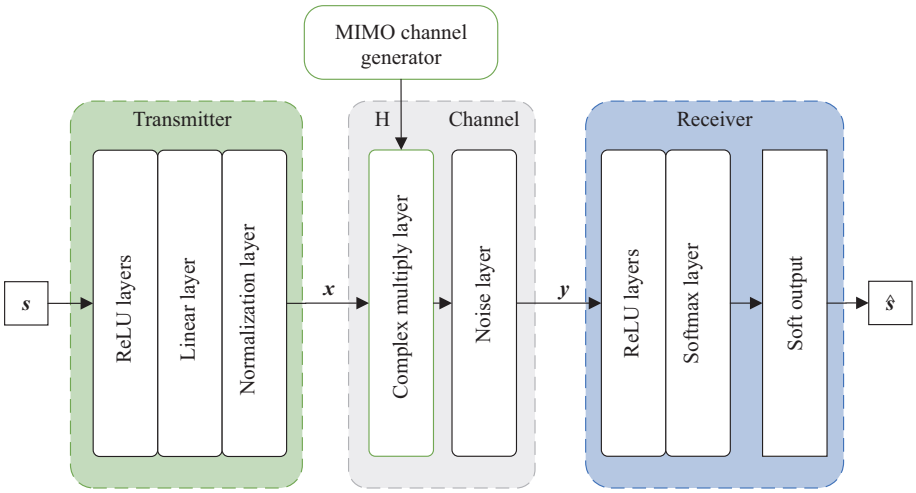

图 9 (网络版彩图) 用自编码器构建的 MIMO 端到端通信模型结构 [37]

Figure 9 (Color online) A general MIMO channel autoencoder architecture [37]

于接收信号对传输信号进行重构.

自编码器方法不基于任何经典的编码、检测方法, 而是将整个端到端通信构筑为一个用于信息重构的 DNN 并进行联合优化. 在难以建立准确模型的复杂现实环境中, 自编码器方法不采用经典模型, 而是利用大量数据的支撑和机器学习算法的强大功能来 “学习” 复杂的信道状况, 从而优化整个通信过程. 同时, 它也有效避免了多种模块拼接而产生的多层网络所带来的高复杂度和大计算量. 文献 [59] 将自编码器模型推广到干扰信道的多用户通信模型上, 而文献 [60] 则将这种自编码器优化方法推广到 MIMO 上, 通过增加信道矩阵相关模块, 形成了如图 9 所示的 MIMO 自编码器通信模型. 文献 [59,60] 的实验结果均显示自编码器方法建立了一种可用于不同 CSI 信息, 天线数等情况下的统一物理层框架, 并且在更低的计算复杂度下可通过 “学习” 得到比经典方法更低的误比特率. 而上述端到端优化策略可以利用上文所述的 “基于 AI 的 5G 通用加速器” 加以高效实现.

4 结束语

5G 突破了传统移动通信系统的应用范畴, 在大幅提升传统移动互联网服务能力的同时, 将应用的触角渗透至各行各业的物联网应用, 从而演变成为支撑全社会、全行业运行的基础性互联网络. 以统一的技术框架支撑极度差异化的繁杂应用, 5G 技术的发展正面临着前所未有的挑战. AI 技术的新一轮复兴及迅猛发展, 为应对上述挑战并超越传统移动通信设计理念与性能提供了潜在的可能性.

AI 技术在 5G 移动通信系统中应用, 已有大量的研究文献可供借鉴. 本文并非试图全面地总结该领域已有研究成果, 而是希望厘清 AI 技术在 5G 系统中最有发展生命力的研究方向, 并通过在这些方向上的进一步努力, 使 5G 系统性能与实现的便利性可以显著超越传统移动通信系统. 基于本文所给出的若干典型应用范例, 我们有理由期待上述努力在不远的未来取得显著的成效.

参考文献

1 You X H, Pan Z W, Gao X Q, et al. The 5G mobile communication: the development trends and its emerging key techniques (in Chinese). Sci Sin Inform, 2014, 44: 551–563

2 Li L M, Wang D M, Niu X K, et al. mmWave communications for 5G: implementation challenges and advances. Sci China Inf Sci, 2018, 61: 021301

3 Wang C X, Wu S B, Bai L, et al. Recent advances and future challenges for massive MIMO channel measurements and models. Sci China Inf Sci, 2016, 59: 021301

4 Zhang J H, Tang P, Tian L, et al. 6–100 GHz research progress and challenges from a channel perspective for fifth generation (5G) and future wireless communication. Sci China Inf Sci, 2017, 60: 080301

5 Tao X F, Han Y, Xu X D, et al. Recent advances and future challenges for mobile network virtualization. Sci China Inf Sci, 2017, 60: 040301

6 3GPP. Way forward on the overall 5G-NR eMBB. Workplan RP-170741. 2017. ftp://ftp.3gpp.org/TSG RAN/TSG RAN/TSGR 75/Docs/RP-170741.zip

7 3GPP. Study on new radio access technology: radio access architecture and interfaces (release 14). TR38.801, v14.0. 2017. http://www.3gpp.org/ftp/Specs/archive/38 series/38.801/38801-e00.zip

8 ITU-R. Minimum requirements related to technical performance for IMT2020 radio interface(s). Report ITU-R M.2410-0. 2017. https://www.itu.int/pub/R-REP-M.2410-2017

9 3GPP. LTE Enhancements and 5G Normative Work. Release-15. 2018. http://www.3gpp.org/release-15

10 You X H, Wang D M, Sheng B, et al. Cooperative distributed antenna systems for mobile communications. IEEE Wirel Commun, 2010, 17: 35–43

11 Yang W J, Wang M, Zhang J J, et al. Narrowband wireless access for low-power massive internet of things: a bandwidth perspective. IEEE Wirel Commun, 2017, 24: 138–145

12 ITU-T. LS/o on the results of the 1st meeting of the ITU-T focus group on ma- chine learning for future networks including 5G (FG ML5G). FG ML5G-0-004. 2018. http://www.3gpp.org/ftp/tsg sa/WG1 Serv/TSGS1 82 Dubrovnik/Docs/S1-181271.zip

13 3GPP. 5G system network data analytics services stage 3. TS 29.520 (CT3). 2018. http://www.etsi.org/deliver/etsi ts/129500 129599/129520/15.00.00 60/ts 129520v150000p.pdf

14 Whitley D. A genetic algorithm tutorial. Stat Comput, 1994, 4: 65–85

15 Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw, 2015, 61: 85–117

16 You X H, Chen G A, Cheng S X. Dynamic learning rate optimization of the backpropagation algorithm. IEEE Trans Neural Netw, 1995, 6: 669–677

17 You X H. Can backpropagation error surface not have local minima. IEEE Trans Neural Netw, 1992, 3: 1019–1021

18 Yu X H, Chen G A. Efficient backpropagation learning using optimal learning rate and momentum. Neural Netw, 1997, 10: 517–527

19 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey. J Artif Intell Res, 1996, 4: 237–285

20 Watkins C J C H, Dayan P. Q-learning. Mach Learn, 1992, 8: 279–292

21 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. 2017. ArXiv: 1703.03400
22 Wu J X, Gao B B, Wei X S, et al. Resource-constrained deep learning: challenges and practices. Sci Sin Inform, 2018, 48: 501–510
23 Zhou Z H. Machine learning: recent progress in China and beyond. China Sci Rev, 2018, 5: 20
24 Zhong Y X. Artificial intelligence: concept, approach and opportunity. Chin Sci Bull, 2017, 62: 2473
25 Gatherer A. Machine learning Modems: how ML will change how we specify and design next generation communication systems. IEEE ComSoc Tech News, 2018. https://www.comsoc.org/ctn/machine-learning-modems-how-ml-will-change-how-we-specify-and-design-next-generation
26 Yang C, Xu W H, Zhang Z C, et al. A channel-blind detection for SCMA based on image processing techniques. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2018. 1–5
27 Zhang C, Xu W H. Neural networks: efficient implementations and applications. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 1029–1032
28 Xu W H, You X H, Zhang C. Efficient deep convolutional neural networks accelerator without multiplication and retraining. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018. 1–5
29 Xu W H, Wang Z F, You X H, et al. Efficient fast convolution architectures for convolutional neural network. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 904–907
30 Xu W H, Wu Z Z, Ueng Y L, et al. Improved polar decoder based on deep learning. In: Proceedings of IEEE International Workshop on Signal Processing Systems (SiPS), 2017. 1–6
31 Xu W H, Zhong Z W, Be'ery Y, et al. Joint neural network equalizer and decoder. In: Proceedings of IEEE International Symposium on Wireless Communication Systems (ISWCS), 2018. 1–6
32 Xu W H, Be'ery Y, You X H, et al. Polar decoding on sparse graphs with deep learning. In: Proceedings of Asilomar Conference on Signals, Systems, and Computers (Asilomar), 2018. 1–6
33 Xu W H, You X H, Zhang C. Using Fermat number transform to accelerate convolutional neural network. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 1033–1036
34 Gómez-Andrades A, Munoz P, Serrano I, et al. Automatic root cause analysis for LTE networks based on unsupervised techniques. IEEE Trans Veh Technol, 2016, 65: 2369–2386
35 Liang F, Shen C, Wu F. An iterative BP-CNN architecture for channel decoding. IEEE J Sel Top Signal Process, 2018, 12: 144–159
36 Lv X Z, Wei P, Xiao X C. Automatic identification of digital modulation signals using high order cumulants. Electronic Warfare, 2004, 6: 1
37 Wang T Q, Wen C K, Wang H Q, et al. Deep learning for wireless physical layer: opportunities and challenges. China Commun, 2017, 14: 92–111
38 Gao X Q, Jiang B, Li X, et al. Statistical eigenmode transmission over jointly correlated MIMO channels. IEEE Trans Inform Theor, 2009, 55: 3735–3750
39 Wang D M, Zhang Y, Wei H, et al. An overview of transmission theory and techniques of large-scale antenna systems for 5G wireless communications. Sci China Inf Sci, 2016, 59: 081301
40 Gesbert D, Hanly S, Huang H, et al. Multi-cell MIMO cooperative networks: a new look at interference. IEEE J Sel Areas Commun, 2010, 28: 1380–1408
41 Jing S S, Yu A L, Liang X, et al. Uniform belief propagation processor for massive MIMO detection and GF (2^n) LDPC decoding. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 961–964
42 Gandhi V S, Maheswaran B. A cross layer design for performance enhancements in LTE-A system. In: Proceedings of IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016. 905–909
43 Kuen J, Kong X F, Wang G, et al. DelugeNets: deep networks with efficient and flexible cross-layer information inflows. In: Proceedings of IEEE International Conference on Computer Vision Workshop (ICCVW), 2017. 958–966
44 Farsad N, Rao M, Goldsmith A. Deep learning for joint source-channel coding of text. 2018. ArXiv: 1802.06832
45 Xu X W, Ding Y K, Hu S X, et al. Scaling for edge inference of deep neural networks. Nat Electron, 2018, 1: 216–222
46 Wang X F, Li X H, Leung V C M. Artificial intelligence-based techniques for emerging heterogeneous network: state

of the arts, opportunities, and challenges. IEEE Access, 2015, 3: 1379–1391
47 Klaine P V, Imran M A, Onireti O, et al. A survey of machine learning techniques applied to self-organizing cellular networks. IEEE Commun Surv Tut, 2017, 19: 2392–2431
48 Pèrez-Romero J, Sallent O, Ferrús R, et al. Knowledge-based 5G radio access network planning and optimization. In: Proceedings of IEEE International Symposium on Wireless Communication Systems (ISWCS), 2016. 359–365
49 Wang J H, Guan W, Huang Y M, et al. Distributed optimization of hierarchical small cell networks: a GNEP framework. IEEE J Sel Areas Commun, 2017, 35: 249–264
50 Bogale T E, Wang X, Le L B. Machine intelligence techniques for next-generation context-aware wireless networks. 2018. ArXiv: 1801.04223
51 Li R, Zhao Z, Zhou X, et al. Intelligent 5G: when cellular networks meet artificial intelligence. IEEE Wirel Commun, 2017, 24: 175–183
52 Zhao Z, Li R, Sun Q, et al. Deep reinforcement learning for network slicing. 2018. ArXiv: 1805.06591
53 Ren Y R, Zhang C, Liu X, et al. Efficient early termination schemes for belief-propagation decoding of polar codes. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2015. 1–4
54 Fossorier M P C, Mihaljevic M, Imai H. Reduced complexity iterative decoding of low-density parity check codes based on belief propagation. IEEE Trans Commun, 1999, 47: 673–680
55 Yang J M, Song W Q, Zhang S Q, et al. Low-complexity belief propagation detection for correlated large-scale MIMO systems. J Sign Process Syst, 2018, 90: 585–599
56 Liu L, Yuen C, Guan Y L, et al. Gaussian message passing iterative detection for MIMO-NOMA systems with massive access. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), 2016. 1–6
57 Yang J M, Zhang C, Zhou H Y, et al. Pipelined belief propagation polar decoders. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2016. 413–416
58 Tan X S, Xu W H, Be'ery Y, et al. Improving massive MIMO belief propagation detector with deep neural network. 2018. ArXiv: 1804.01002
59 O'Shea T, Hoydis J. An introduction to deep learning for the physical layer. IEEE Trans Cogn Commun Netw, 2017, 3: 563–575
60 O'Shea T J, Erpek T, Clancy T C. Deep learning based MIMO communications. 2017. ArXiv: 1707.07980

AI for 5G: research directions and paradigms

Xiaohu YOU[1], Chuan ZHANG[1*], Xiaosi TAN[1], Shi JIN[1] & Hequan WU[2]

1. *National Mobile Communications Research Laboratory, Southeast University, Nanjing* 210096*, China*;
2. *Chinese Academy of Engineering, Beijing* 100088*, China*
* Corresponding author. E-mail: chzhang@seu.edu.cn

Abstract Fifth-generation wireless communication (5G) technologies not only fulfill the requirement of 1000 times increase of Internet traffic in the next decade but also offer the underlying technologies to the entire industry and ecology for the Internet of everything. Compared with the existing mobile communication technologies, 5G technologies are more widely applicable and have more complicated corresponding system design. In order to better balance the complexity and performance, artificial intelligence (AI) technologies have been considered for 5G. Typical and potential research directions to which AI can make promising contributions need to be identified, evaluated, and investigated. To this end, this overview paper first combs through several promising research directions of AI for 5G, based on the understanding of the key aspects of 5G technologies. Furthermore, the paper devotes itself in providing design paradigms including 5G network optimization, optimal resource allocation, 5G physical layer unified acceleration, and end-to-end physical layer joint optimization.

Keywords 5G mobile communication, AI techniques, network optimization, resource allocation, unified acceleration, end-to-end joint optimization

Xiaohu YOU obtained B.S., M.S., and Ph.D. degrees in electrical engineering from Nanjing Institute of Technology, Nanjing, China, in 1982, 1985, and 1989, respectively. From 1987 to 1989, he was a lecturer at Nanjing Institute of Technology. From 1990 to the present time, he has been with Southeast University, first as an associate professor and later as a professor. His research interests include mobile communications, adaptive signal processing, and artificial neural networks, with applications to communications and biomedical engineering.

Xiaosi TAN obtained her Ph.D. degree in mathematics from Texas A&M University, College Station, Texas, USA, in 2015. She is currently a postdoctoral researcher of National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing, China. Her current research interests include emerging technologies for 5G cellular networks, including machine learning for wireless networks, massive multiple-input multiple-output and massive machine type communication communications.

Chuan ZHANG is now an associate professor of National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing, China. He obtained his B.E. degree in microelectronics and M.E. degree in very-large-scale integration (VLSI) design from Nanjing University, Nanjing, China, in 2006 and 2009, respectively. In 2012, he obtained both MSEE and Ph.D. degrees in the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities (UMN), USA. His current research interests include low-power high-speed VLSI design for digital signal processing and digital communication, bio-chemical computation and neuromorphic engineering, and quantum communication.

Hequan WU obtained his degree from the Wuhan Post and Telecommunications Institute in 1964. He has been with the China Academy of Post and Telecommunications, Ministry of Posts and Telecommunications, since 1964. He was the vice president and chief engineer of the China Academy of Telecommunications Technology from 1997 to 2003. He studied optical fiber transmission systems and broadband networks and managed a series of national research and development projects. In recent years, he has focused on the development strategy of next-generation networks and next-generation Internet as well as 3G, 4G, and 5G.