Research Literature Survey

For personal research

By Xingsheng Li

Master's Student, Department of Computer Technology

2026 年 2 月 25 日

Abstract

这篇文档主要是对于个人科研工作的总结性回顾。以 Part 的形式对于各个阶段科研学习的内容进行整理。

- Part I: 2025-2026 winter break research
- Part II: Uncategorized(主要是对于项目的功能性尝试)

目录

Part I

Smart Mining Research

Chapter 1

智慧煤矿发展现状与展望：基于数字孪生与平行系统的文献综述

1.1 研究背景

煤炭作为我国的基础能源，其安全、高效与绿色的开采对于保障国家能源安全具有不可替代的战略意义。近年来，随着浅部煤炭资源的日益枯竭，煤矿开采不可避免地向深部地层延伸。深部开采伴随的高地应力、高瓦斯、高地温以及复杂的岩溶水文地质条件，使得煤与瓦斯突出、冲击地压等重特大动力灾害的防控难度急剧上升。传统的煤矿开采模式受限于高危作业环境以及粗放型的人工管理方式，已难以满足现代工业对于本质安全与生产效率的升级需求。在这一宏观背景下，煤矿智能化建设应运而生，并迅速成为应对深部开采挑战与实现煤炭工业高质量发展的核心技术支撑[1]。

早期的智慧煤矿探索主要集中在底层设备的单机自动化改造与基础信息系统的建设，旨在实现机电设备的远程启停与初步的工况参数监测[2]。然而，这种基于逻辑可编程控制器与工业组态软件的系统缺乏全局数据交互与协同规划能力。面对井下瞬息万变的地质异常与突发生产状况，孤岛式的控制单元无法形成有效的多系统联动。为了突破这一系统性瓶颈，学术界和工业界开始引入工业物联网、第五代移动通信技术、大数据分析以及新一代人工智能算法，推动矿山信息化向具备深度感知与自主决策能力的智慧化演进。

在这一演进过程中，数字孪生作为连接物理实体与虚拟数字空间的桥梁，成为构建智慧煤矿的关键共性技术[3]。该技术不仅能够对物理矿山的三维几何空间、设备运行状态及生产全流程进行高精度的全息映射，还能依托孪生计算体进行多物理场仿真推演与反向闭环控制。从单点自动化的初级阶段迈向以数字孪生为核心的全局智慧矿山，不仅是技术工具的简单迭代，更是煤炭开采范式向数据驱动与虚实互动演进的重要标志。

1.2 研究现状

当前关于智慧煤矿的研究呈现出多学科深度交叉、从宏观体系架构到微观物理执行高度协同的显著特征。现有文献的研究脉络主要可归纳为顶层设计、数据基座构建、孪生理论拓展以及底层设备协同控制等核心维度。

1.2.1 顶层设计与标准体系构建

智慧煤矿是一个涉及人、机、环、管多维复杂要素的巨系统。缺乏统一的顶层设计和标准数据接口将直接导致系统间的数据壁垒与协同障碍。为此，研究人员率先从宏观层面确立了规范化的建设框架。王国法等深入剖析了煤矿智能化标准体系的架构与建设思路，为后续各类异构智能装备的接入与多源数据的互联互通奠定了标准化基础[4]。在具体的方法论指导上，张帆等系统性地综述了智慧矿山数字孪生技术，进一步明确了矿山数字孪生的构建方法与演化机理，并指出数字孪生必须经历从几何维度的结构映射到物理维度的机理映射，再到行为维度的动态映射这一递进演化过程[5-6]，如图 1.1 为其构建的智采工作面数字孪生演化实例。鲍久圣等针对矿山环境的空间特殊性与时变复杂性，提出了矿山数字孪生模型架构，详细阐述了支撑该架构所需的高效感知网络与虚实交互关键技术[7]。邢震的最新研究综述进一步表明，面向智能矿山的数字孪生技术正加速从前期的理论架构探讨向深层次的工业现场应用落地推进[8]。

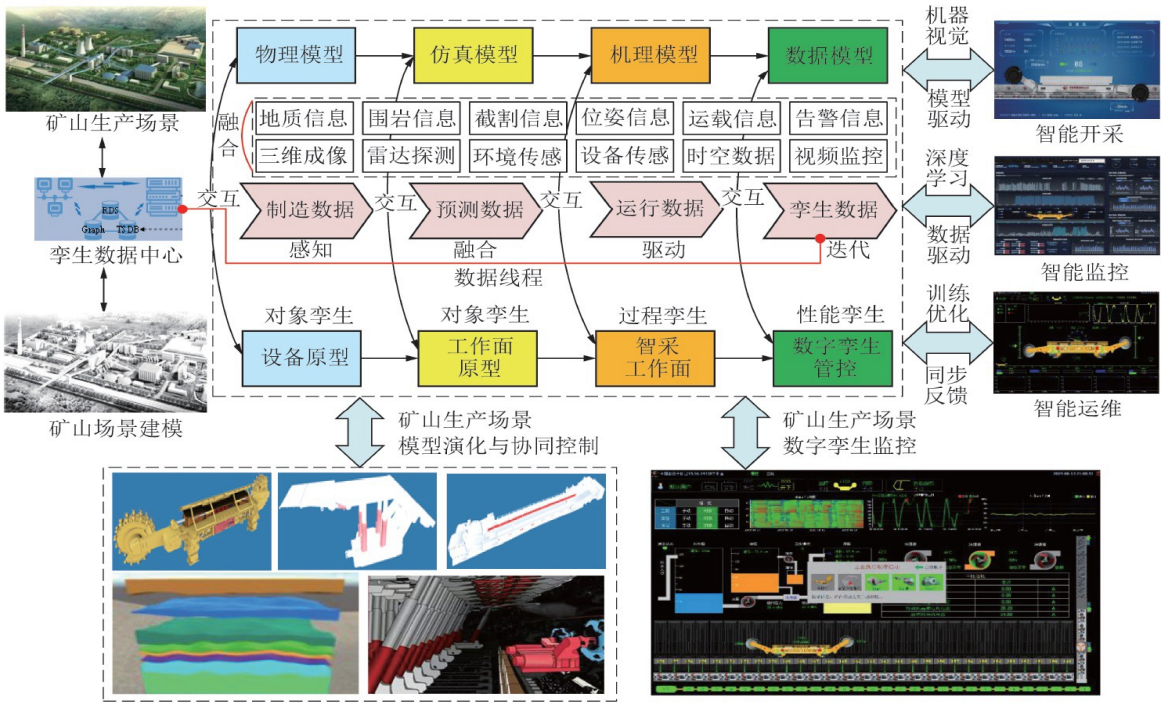

图 1.1: 矿山数字孪生演化实例

1.2.2 从数字孪生到平行矿山理论的演进

在数字孪生解决物理映射准确性的基础上，如何利用虚拟空间进行前瞻性的计算实验与决策优化，成为当前研究的前沿焦点。陈龙等创新性地将复杂系统科学中的平行系统理论与人工系统、计算实验及平行执行理念引入矿山领域，正式提出了平行矿山的概念。这一概念标志着研究视角从单纯的映射孪生迈向了具备认知与推演能力的高阶智能[9]。平行理论强调在虚拟数字空间中构建多重演化可能的人工矿山，通过大规模计算实验推演各类极端工况下的最优控制律，进而与物理矿山进行平行交互与反馈修正。这一理论在具体的采掘业务中已展现出重要的指导意义。杨健健等基于平行执行理论构建了平行掘进系统，系统性地提出了解决掘进、支护与锚固等复杂空间交叉工序智能协同控制的理论框架[10]。这种从被动状态监测向主动寻优推演的理论演变，极大地拓展了智慧煤矿的应用边界。

1.2.3 底层设备感知、协同控制与延伸应用

在底层装备执行层面，工业物联网与孪生技术的融合显著提升了重型设备的运行效能与安全可靠性。丁恩杰等系统探讨了基于物联网的矿山机械设备状态智能感知与故障诊断技术。多参量光纤传感器与高精度微机电系统惯性导航等先进感知元器件的应用，为上层孪生模型提供了高维度且低延迟的时序数据馈入[11]。针对核心的综采工作面，葛世荣等提出了数字孪生智采工作面技术架构。尤秀松等进一步聚焦采煤机、刮板输送机与液压支架，提出了数字孪生驱动的群控架构，使得复杂的重型采煤装备群体能够在虚拟模型的引导下实现多机协作与运动轨迹优化[12-13]。此外，李浩荡等将数字孪生技术与顶煤放出规律模型相协同，在提高特厚煤层资源回收率与降低混矸率方面取得了应用进展[14]。智慧煤矿的技术赋能领域也在不断拓宽。李全生等将数字孪生技术下沉应用于矿区生态环境的动态监测与修复治理，构建了生态环境数字孪生的理论内涵，深刻体现了智能化开采与绿色环境保护并重的高质量发展理念[15]。

1.3 研究不足

尽管智慧煤矿与数字孪生技术在宏观架构与局部场景应用上取得了明显进展，但对标高度复杂的井下真实生产环境，现有研究与工业实践仍面临数个亟待突破的技术局限。

首先，多源异构数据的深度融合与语义对齐能力依然有限。井下工业物联网虽然接入了海量的传感器节点，但诸如地质勘探的三维点云、机电设备的震动高频频谱与环境气体的低频浓度序列等数据往往处于异构割裂状态。现有系统多采用简单的数据堆叠拼接，缺乏统一的高维特征表示机制，难以提取出支持全局协同决策的深层交叉特征。

其次，高保真物理模型与实时算力之间存在难以调和的工程矛盾。井下开采涉及非线性岩体力学损伤、瓦斯渗流以及气固液多相流体力学的复杂耦合演化。采用传统的有限元法或离散元法构建高保真物理模型需要消耗庞大的计算资源。这种计算密集型任务导致孪生系统在面对突发顶板灾害或涌水险情时，往往产生分钟级以上的计算延迟，无法做到毫秒级的实时预警与动态阻断。受限于井下有限的通信带宽，将海量原始感知数据实时上传至地面云平台进行集中求解的模式，进一步加剧了控制时延。

再次，现有分层级联控制架构对极端地质条件的泛化能力不足。目前多数智能控制策略是基于相对理想或已知的工作面条件，通过大量人工预设阈值与专家逻辑构建的。一旦工作面遭遇断层揭露、底板剧烈突水或大面积悬顶等未见异常，确定性的逻辑规则极易失效，系统依然高度依赖人工接管与经验判断，尚未实现真正的自主适应。

最后，人工智能算法在矿山安全关键系统中的可解释性与安全约束机制尚不完善。随着深度学习在矿山视觉识别与控制参数优化中的应用增多，神经网络模型内部的黑盒属性引发了工业应用的安全隐忧。在缺乏明确数学边界证明的前提下，纯数据驱动模型输出的控制指令难以满足煤矿安全规程的严苛要求，这限制了人工智能技术在核心控制闭环中的深度应用。

1.4 研究趋势展望

面向未来，破解上述数据融合困难、计算时延过高以及模型泛化泛化性差等行业痛点，需要跨界引入人工智能科学领域的前沿范式。随着通用算力集群的普及与深度学习算法的快速迭代，智慧煤矿的发展将逐步脱离传统的规则驱动路径，向具备自监督学习与主动推理能力的智能体系过渡。其主要技术演进趋势可概括为以下五个维度的拓展。

1.4.1 多模态感知跃升：视觉语言模型赋能复杂场景认知

现有的矿山视觉感知系统大多依赖于特定任务的卷积神经网络。此类判别式模型通常只能执行诸如皮带异物识别或人员轨迹追踪等特定功能，缺乏对全局场景时空关联的理解能力。未来，视觉语言大模型将在矿山感知网络中实现深度融合与应用。这类模型能够通过自注意力机制，将矿山现场的高清视频流、微光夜视图像等视觉模态，与地质勘探文本、设备维修图纸及安全操作规程等语言模态，在统一的隐空间中进行精准的语义对齐。升级后的感知系统不仅能输出异常物体的包围框坐标，更能直接理解场景上下文的物理意义。例如，当识别到采空区瓦斯浓度梯度异常且伴有煤壁温度升高时，模型可自主结合历史多尺度防灭火资料，生成自然语言形式的详细避灾路线与通风系统调节建议。这种跨模态的推理能力将显著提升孪生系统对非结构化矿山环境的通用场景认知水平。

1.4.2 控制架构重构：端到端智能闭环与安全信封机制

在采掘装备的底层执行侧，当前的控制系统普遍采用感知、决策、路径规划到动作执行的串行模块化架构[13]。此架构在面对复杂工况时，容易因前端传感器的微小噪声引发后端控制参数的级联放大偏差。借鉴智能驾驶领域的前沿演进路线，端到端神经网络控制架构将成为矿机协同演进的重要方向。端到端模型直接摄入激光雷达点云与多维力矩传感器时序信号，通过深层残差网络提取时空耦合特征，并直接映射输出采煤机牵引速度、滚筒截割高度与液压支架推移步距等连续动作指令。这种方法省去了繁杂的中间规划逻辑，能够隐式学习重型装备与煤岩体之间的复杂博弈规律。为克服纯数据驱动带来的安全隐患，未来的端到端系统将普遍集成基于物理约束的安全信封机制。通过在神经网络的输出端引入反映物理极限与安全规程的可微惩罚函数，确保算法生成的所有控制指令严格收敛于煤矿安全边界之内。

1.4.3 演化机理重置：世界模型驱动的生成式平行推演

平行矿山理论的核心价值在于利用计算实验推演系统最优控制策略[9]。受制于微分方程求解的低效性，现有推演过程难以满足实时性要求。世界模型的引入有望为物理仿真范式带来质的飞跃。世界模型能够通过对海量历史采掘视频序列与多模态传感器信号的无监督预训练，自主内化物理世界的直观动力学规律。在实际应用中，世界模型在低维潜变量空间内以极低的算力开销，即可快速生成对矿山未来时空状态的高保真预测序列。系统无需求解纳维-斯托克斯方程或复杂岩石力学方程，即可通过神经网络快速推演未来数分钟内工作面瓦斯涌出动态或顶板位移趋势。这种生成式的计算实验使得矿山中枢大脑能够在隐空间中高效开展强化学习试错，进而将提炼出的最优参数反向映射给物理执行器，推动知识驱动与主动管控模式的真正落地[16]。

1.4.4 算力分布优化：云边协同与联邦学习破局数据孤岛

针对井下数据传输的时延瓶颈，智慧矿山的算力架构正加速向云端、边缘端与设备端高度协同的分布式网络演进。部署于采煤机防爆计算机或井下变电所的高算力边缘节点，可实现传感器数据的就地近端清洗与模型推理，保障核心急停指令的毫秒级响应。此外，联邦学习机制将在跨矿区数据利用中发挥关键作用。各矿区的边缘计算集群能够在确保底层敏感生产数据不出安全域的前提下，仅将本地模型训练产生的梯度更新量加密上传至集团云端进行全局聚合。这种去中心化的协作训练模式，在兼顾数据隐私安全的同时，打破了长期存在的信息孤岛。全局大模型得以吸纳各类复杂地质条件下的运行经验，实现泛化能力的快速迭代。

1.4.5 物理执行延伸：具身智能与仿生多智能体协同

数字孪生系统的决策推演最终需要落实在物理世界的精准交互之中。传统的矿山巡检机器人多为功能单一的履带式底盘设备，缺乏在复杂受限空间内的灵巧物理作业能力。具身智能技术的快速发展将为平行矿山提供更强大的末端执行载体。未来的井下综采工作面与掘进巷道可以广泛引入具备多模态高频感知与精细力矩反馈的仿生四足机器人与双臂灵巧作业机器人。这些具身智能体能够解析系统下发的高级语义指令，并通过多智能体强化学习机制在装备群内部进行自主路径规划与任务分配。从高危区域的超前探放水钻孔定位，到狭窄管廊内的设备带电维护与管线更换，具身机器人集群将逐步胜任高强度与高风险的体力劳动环节，进而完成数字空间向物理实体改造的最终闭环。

1.5 结论

综上所述，智慧煤矿的研究已经超越了初期的单点设备自动化范畴，全面进入了以数字孪生为核心架构、以平行理论为理论指导的深水区。客观审视当前的发展阶段，在算力资源的实时优化调度、异构数据的统一语义对齐，以及极端地质条件下的模型泛化验证等方面，工业界与学术界仍面临诸多严峻的工程与理论挑战。

应对上述挑战，跨学科融合创新是必由之路。从视觉语言大模型带来的全场景深度认知，到端到端控制架构实现的鲁棒闭环，再到世界模型赋予数字孪生的超实时预测能力，叠加边缘计算与具身智能的协同发展，新一代计算机科学技术的融合应用为智慧煤矿的升级提供了全新的解题思路。这些技术的逐步落地，将推动智慧煤矿从机械的物理状态复刻向具备自主进化能力的工业大脑演进。建立高可靠、强泛化的智能平行矿山体系，不仅是实现煤炭资源安全高效开采的必然选择，更是推动煤炭工业向低碳环保与可持续发展转型的重要基石。

References for this Part

[1] 王国法, 刘峰, 庞义辉, 等. 煤矿智能化———煤炭工业高质量发展的核心技术支撑[J/OL]. 煤炭学报, 2019(2)(2019-02-28). https://www.mtxb.com.cn/cn/article/doi/10.13225/j.cnki.jccs.2018.2041. DOI: 10.13225/j.cnki.jccs.2018.2041.

[2] 王国法, 赵国瑞, 任怀伟. 智慧煤矿与智能化开采关键核心技术分析[J/OL]. 煤炭学报, 2019(1)(2019-01-31). https://www.mtxb.com.cn/cn/article/doi/10.13225/j.cnki.jccs.2018.5034. DOI: 10.13225/j.cnki.jccs.2018.5034.

[3] 丁恩杰, 俞啸, 夏冰, 等. 矿山信息化发展及以数字孪生为核心的智慧矿山关键技术[J/OL]. 煤炭学报, 2023, 47(1): 564-578(2023-04-10). https://www.mtxb.com.cn/cn/article/id/84a0eab7-ffac-4fa3-bcef-509cb4a743fa.

[4] 王国法, 杜毅博. 煤矿智能化标准体系框架与建设思路[J/OL]. 煤炭科学技术, 2020, 48(1)(2020-01-25). https://www.mtkxjs.com.cn/cn/article/id/0d41dd60-ad22-4627-82e6-1d9ba5ed6ce0.

[5] 张帆, 葛世荣, 李闯. 智慧矿山数字孪生技术研究综述[J/OL]. 煤炭科学技术, 2020, 48(7)(2020-07-25). https://www.mtkxjs.com.cn/cn/article/id/3a06c655-8805-44cc-a5a9-53f238a7fd37.

[6] 张帆, 葛世荣. 矿山数字孪生构建方法与演化机理[J/OL]. 煤炭学报, 2023, 48(1): 511-524(2023-01-31). https://www.mtxb.com.cn/cn/article/id/a49c154d-61f7-43ee-8620-519e822cdd33.

[7] 鲍久圣, 张可琨, 王茂森, 等. 矿山数字孪生 MiDT: 模型架构、关键技术及研究展望[J/OL]. 绿色矿山, 2023, 0(01). https://www.chinacaj.net/LSKS/202301/492613.

[8] 邢震. 面向智能矿山的数字孪生技术研究进展[J/OL]. 工矿自动化, 2024, 50(3): 22-34, 41. https://doi.org/10.13272/j.issn.1671-251x.2024010079. DOI: 10.13272/j.issn.1671-251x.2024010079.

[9] 陈龙, 王晓, 杨健健, 等. 平行矿山: 从数字孪生到矿山智能[J/OL]. 自动化学报, 2021, 47(7): 1633-1645(2021-07-27). https://www.aas.net.cn/cn/article/doi/10.16383/j.aas.2021.y000001. DOI: 10.16383/j.aas.2021.y000001.

[10] 杨健健, 葛世荣, 王飞跃, 等. 平行掘进: 基于 ACP 理论的掘-支-锚智能控制理论与关键技术[J/OL]. 煤炭学报, 2021, 46(7): 2100-2111(2021-07-31). https://www.mtxb.com.cn/cn/article/id/6a44cbf3-09d6-4b27-8f28-58764e149b6e.

[11] 丁恩杰, 俞啸, 廖玉波, 等. 基于物联网的矿山机械设备状态智能感知与诊断[J/OL]. 煤炭学报, 2023, 45(6)(2023-04-10). https://www.mtxb.com.cn/cn/article/doi/10.13225/j.cnki.jccs.ZN20.0340. DOI: 10.13225/j.cnki.jccs.ZN20.0340.

[12] 葛世荣, 张帆, 王世博, 等. 数字孪生智采工作面技术架构研究[J/OL]. 煤炭学报, 2023, 45(6)(2023-04-10). https://www.mtxb.com.cn/cn/article/doi/10.13225/j.cnki.jccs.ZN20.0327. DOI: 10.13225/j.cnki.jccs.ZN20.0327.

[13] 尤秀松, 葛世荣, 郭一楠, 等. 智采工作面三机数字孪生驱动控制架构[J/OL]. 煤炭学报, 2024, 49(7): 3265-3275(2024-07-25). https://www.mtxb.com.cn/cn/article/doi/10.13225/j.cnki.jccs.2023.0684. DOI: 10.13225/j.cnki.jccs.2023.0684.

[14] 李浩荡, 董启凡, 李孝胜. 基于数字孪生与规划放煤协同的煤矿智能开采技术[J/OL]. 煤炭科学技术, 2026, 54(1): 411-423(2026-01-25). https://www.mtkxjs.com.cn/cn/article/doi/10.12438/cst.2025-0570. DOI: 10.12438/cst.2025-0570.

[15] 李全生, 刘举庆, 李军, 等. 矿山生态环境数字孪生：内涵、架构与关键技术[J/OL]. 煤炭学报, 2023, 48(10): 3859-3873(2023-10-25). https://www.mtxb.com.cn/cn/article/doi/10.13225/j.cnki.jccs.2022.1850. DOI: 10.13225/j.cnki.jccs.2022.1850.

[16] 郭一楠, 杨帆, 葛世荣, 等. 知识驱动的智采数字孪生主动管控模式[J/OL]. 煤炭学报, 2023, 48(S1): 334-344(2023-08-24). https://www.mtxb.com.cn/cn/article/doi/10.13225/j.cnki.jccs.2022.0223. DOI: 10.13225/j.cnki.jccs.2022.0223.

Part II

AI Foundations

Chapter 1

Visual Recognition Evolution

1.1 Convolutional Neural Networks

For a long time, ResNet[1] served as the backbone for most vision tasks due to its ability to solve the vanishing gradient problem in deep networks.

1.2 Vision Transformers (ViT)

Inspired by NLP success, the Vision Transformer[2] splits images into patches, treating them as sequences, challenging the dominance of CNNs.

Chapter 2

Foundations of LLMs

2.1 The Transformer Architecture

The field of Natural Language Processing was revolutionized by the introduction of the Transformer architecture[3]. This mechanism allows for parallelization tailored for modern hardware.

2.2 Scaling Laws and GPT

Recent developments have shown that scaling up model size and data volume leads to emergent abilities, as demonstrated in the GPT-4 technical report[4].

Chapter 3

Algorithmic Trading Strategies

3.1 Market Efficiency

The Efficient Market Hypothesis (EMH)[5] suggests that asset prices reflect all available information.

3.2 Machine Learning in Finance

Modern quantitative finance is moving towards ML-based approaches. As noted by Lopez de Prado[6], applying standard ML cross-validation in finance leads to overfitting due to time-series correlation.

Chapter 4

Open Source Projects

In this chapter, we track the SOTA open-source implementations for LLMs.

Llama 3

URL: https://github.com/meta-llama/llama3

Stack: `Python, PyTorch`

Description: The official Meta Llama 3 repository. It contains the model weights and inference code. State-of-the-art open weights model.

LangChain

URL: https://github.com/langchain-ai/langchain

Stack: `Python, TypeScript`

Description: Building applications with LLMs through composability. Key for RAG (Retrieval Augmented Generation) workflows.

vLLM

URL: https://github.com/vllm-project/vllm

Stack: `Python, CUDA`

Description: A high-throughput and memory-efficient inference and serving engine for LLMs, featuring PagedAttention.

Chapter 5

Quant Libraries & Tools

5.1 Backtesting Frameworks

Backtrader

URL: https://github.com/mementum/backtrader

Stack: `Python`

Description: A feature-rich Python framework for backtesting and trading. Supports multiple data feeds and brokers.

Qlib

URL: https://github.com/microsoft/qlib

Stack: `Python`

Description: An AI-oriented quantitative investment platform by Microsoft. It aims to realize the potential of AI technologies in quantitative investment.

5.2 Data Sources

AkShare

URL: https://github.com/akfamily/akshare

Stack: `Python`

Description: A purely open-source financial data interface library for Python, built for human beings! Covers stocks, futures, options, and bonds.

References for this Part

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[2] DOSOVITSKIY A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. ICLR, 2021.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[4] OpenAI. GPT-4 Technical Report[J]. arXiv preprint arXiv:2303.08774, 2023.

[5] FAMA E F. Efficient Capital Markets: A Review of Theory and Empirical Work[J]. The Journal of Finance, 1970.

[6] De PRADO M L. Advances in Financial Machine Learning[M]. Wiley, 2018.

Part III

Emerging AI Technologies

Chapter 1

World Models: Learning and Planning in Latent Spaces

1.1 Introduction

World models represent a paradigm shift in reinforcement learning and sequential decision-making, where agents learn compact latent representations of environments to enable efficient planning and generalization. Unlike traditional model-free approaches that learn policies directly from observations, world models first learn a generative model of the environment dynamics in a compressed latent space, then use this model for planning or policy learning. This separation of representation learning and control has demonstrated remarkable sample efficiency and generalization capabilities across diverse domains, from robotics to game playing.

1.2 Theoretical Foundations

The concept of world models traces back to early work on latent variable models and Bayesian filtering. Modern implementations build upon several key theoretical pillars:

1.2.1 Latent State Representation

World models learn to encode high-dimensional observations (e.g., images) into low-dimensional latent states that capture essential environmental dynamics while discarding irrelevant details. This compression enables efficient planning by operating in a much smaller state space.

1.2.2 Predictive Modeling

At the core of world models is the ability to predict future latent states given current states and actions. This predictive capability allows agents to simulate trajectories without interacting with the actual environment, enabling "imagination-based" planning.

1.2.3 Planner-Actor Separation

The world model architecture typically separates the model (which learns environment dynamics) from the planner/actor (which uses the model to make decisions). This modular design allows for different planning algorithms to be employed once the model is learned.

1.3 Key Architectures and Methods

1.3.1 Dreamer Series

The Dreamer family of algorithms has been instrumental in advancing world model research. Dreamer[1] introduced the concept of learning latent dynamics models using variational autoencoders and training policies entirely within the learned latent space. Subsequent versions, DreamerV2[2] and DreamerV3[3], improved stability, scalability, and performance across diverse environments.

1.3.2 Model-Based RL with Latent Dynamics

Other approaches include PlaNet[4], which uses a recurrent state-space model for planning, and IRIS[5], which employs autoregressive transformers for next-token prediction in latent space.

1.3.3 Generative World Models

Recent work explores generative world models that can produce diverse plausible futures. Genie[6] demonstrates how world models can be trained from internet videos without action labels, enabling foundation models for embodied AI.

1.4 Applications and Performance

1.4.1 Sample Efficiency

World models dramatically reduce the number of environment interactions required for learning. For example, DreamerV3 achieves superhuman performance on the Atari benchmark using only 100M environment steps, compared to billions required by model-free methods.

1.4.2 Generalization and Transfer

By learning generalizable environment dynamics, world models enable transfer across task variations and domain shifts. This is particularly valuable for real-world applications where environment variability is high.

1.4.3 Robotics and Embodied AI

World models are revolutionizing robotics by allowing agents to plan in learned latent spaces, reducing the need for extensive real-world trial-and-error. Applications include robotic manipulation, navigation, and autonomous driving.

1.5 Challenges and Limitations

1.5.1 Model Inaccuracy

The performance of world models depends critically on the accuracy of the learned dynamics. Inaccuracies can compound during long-horizon planning, leading to suboptimal or catastrophic decisions.

1.5.2 Computational Complexity

Training world models requires significant computational resources, particularly for high-dimensional observations. The need to maintain both encoder, dynamics model, and planner increases architectural complexity.

1.5.3 Exploration-Exploitation Trade-off

World models can suffer from confirmation bias, where the model becomes overconfident in its predictions and fails to explore regions of state space where its dynamics model

is inaccurate.

1.6 Future Directions

1.6.1 Foundation World Models

The development of large-scale, pre-trained world models that can be fine-tuned for specific tasks represents a promising direction, analogous to foundation models in language and vision.

1.6.2 Multimodal World Models

Integrating multiple sensory modalities (vision, audio, tactile) into unified world models could enable more comprehensive environment understanding.

1.6.3 Symbolic-Neural Integration

Combining neural world models with symbolic reasoning could enhance interpretability and enable more sophisticated planning capabilities.

1.6.4 Efficiency Improvements

Research into more efficient architectures, training procedures, and planning algorithms will be crucial for real-world deployment.

1.7 Conclusion

World models represent a powerful framework for sample-efficient reinforcement learning and planning. By learning compact latent representations of environment dynamics, they enable agents to reason about future outcomes and make informed decisions with minimal environment interaction. While challenges remain in model accuracy, computational efficiency, and exploration, ongoing research continues to advance the state of the art. As world models scale and integrate with other AI paradigms, they hold promise for creating more capable, generalizable, and efficient autonomous systems.

Chapter 2

Vision-Language Models: Bridging Visual and Linguistic Understanding

2.1 Introduction

Vision-Language Models (VLMs) represent a convergence of computer vision and natural language processing, enabling machines to understand and generate content that spans both visual and textual modalities. By learning joint representations of images and text, VLMs can perform tasks such as image captioning, visual question answering, cross-modal retrieval, and open-vocabulary object detection. The development of large-scale VLMs has been accelerated by the availability of web-scale image-text pairs and advances in transformer architectures.

2.2 Architectural Paradigms

2.2.1 Dual-Encoder Models

Dual-encoder architectures, exemplified by CLIP[7], use separate encoders for images and text, projecting both into a shared embedding space where similarity can be computed. This approach enables efficient zero-shot classification and cross-modal retrieval but lacks deep fusion between modalities.

2.2.2 Fusion Encoder Models

Fusion encoders, such as VisualBERT[8] and ViLT[9], process concatenated image and text inputs through a single transformer, allowing for deeper interaction between

modalities. These models excel at tasks requiring fine-grained alignment between visual and linguistic elements.

2.2.3 Generator Models

Generator architectures, including BLIP[10] and BLIP-2[11], combine vision encoders with language decoders to generate textual descriptions of images. These models are particularly effective for image captioning and visual question answering.

2.2.4 Large Multimodal Models

Recent large multimodal models, such as Flamingo[12], LLaVA[13], and GPT-4V[14], scale up VLM architectures to billions of parameters, demonstrating emergent capabilities like complex reasoning about visual content.

2.3 Training Strategies

2.3.1 Contrastive Learning

Contrastive learning objectives, as used in CLIP, train models to maximize similarity between corresponding image-text pairs while minimizing similarity between non-corresponding pairs. This approach learns rich cross-modal representations without explicit supervision.

2.3.2 Masked Language Modeling

Adapted from language modeling, masked language modeling for VLMs involves predicting masked tokens based on both visual and textual context, encouraging the model to learn grounded representations.

2.3.3 Image-Text Matching

Image-text matching objectives train models to determine whether an image and text pair correspond, improving fine-grained alignment capabilities.

2.3.4 Generative Objectives

Generative objectives train models to produce text conditioned on images, enabling capabilities like captioning and question answering.

2.4 Key Applications

2.4.1 Zero-Shot Visual Recognition

VLMs enable zero-shot classification by comparing image embeddings with text embeddings of class descriptions, eliminating the need for task-specific training data.

2.4.2 Visual Question Answering

VLMs can answer questions about images, demonstrating understanding of both visual content and linguistic queries.

2.4.3 Image Captioning

Advanced VLMs generate detailed, contextually appropriate descriptions of images, with applications in accessibility, content moderation, and creative tools.

2.4.4 Cross-Modal Retrieval

VLMs facilitate efficient search across modalities, allowing users to find images using text queries or vice versa.

2.4.5 Robotics and Embodied AI

VLMs provide robots with the ability to understand natural language instructions in visual contexts, enabling more intuitive human-robot interaction.

2.5 Challenges and Limitations

2.5.1 Modality Gap

The inherent differences between visual and linguistic representations create a "modality gap" that can limit model performance, particularly for fine-grained tasks.

2.5.2 Hallucination

VLMs sometimes generate plausible but incorrect descriptions of images, a phenomenon known as hallucination, which poses reliability concerns for real-world applications.

2.5.3 Computational Cost

Training and inference with VLMs, especially large multimodal models, require significant computational resources, limiting accessibility.

2.5.4 Bias and Fairness

VLMs can inherit and amplify biases present in training data, potentially leading to unfair or harmful outputs.

2.5.5 Evaluation Metrics

Existing evaluation metrics for VLMs often fail to capture nuanced aspects of model performance, particularly for open-ended generation tasks.

2.6 Emerging Trends

2.6.1 Efficient Fine-Tuning

Techniques like LoRA[15] and QLoRA enable efficient adaptation of large VLMs to specific domains with limited computational resources.

2.6.2 Multimodal In-Context Learning

Recent VLMs demonstrate in-context learning capabilities, allowing them to perform new tasks with few examples, similar to large language models.

2.6.3 Video-Language Models

Extension of VLM principles to video understanding, enabling temporal reasoning and long-form content understanding.

2.6.4 3D and Embodied VLMs

Integration of VLMs with 3D scene understanding and embodied AI, enabling applications in augmented reality, robotics, and autonomous systems.

2.6.5 Multilingual VLMs

Development of VLMs that understand multiple languages, improving accessibility and global applicability.

2.7 Future Directions

2.7.1 Unified Multimodal Architectures

Research toward architectures that seamlessly integrate not just vision and language, but also audio, video, and other modalities.

2.7.2 Causal Reasoning

Incorporation of causal reasoning capabilities to enable VLMs to understand cause-effect relationships in visual scenes.

2.7.3 Compositional Understanding

Improving VLMs' ability to understand complex compositional relationships between objects, attributes, and actions in images.

2.7.4 Efficiency Advances

Development of more efficient architectures, training methods, and inference techniques to democratize access to VLM capabilities.

2.7.5 Ethical Alignment

Research into techniques for aligning VLMs with human values and mitigating biases, hallucinations, and other harmful behaviors.

2.8 Conclusion

Vision-Language Models have dramatically advanced machines' ability to understand and interact with multimodal content. By bridging the gap between visual perception and linguistic understanding, VLMs enable a wide range of applications from accessibility tools to autonomous systems. While challenges remain in terms of reliability, efficiency, and fairness, ongoing research continues to push the boundaries of what's possible. As VLMs become more capable, efficient, and aligned with human values, they promise to transform how humans and machines collaborate in understanding our visual world.

Chapter 3

End-to-End Models: From Perception to Action

3.1 Introduction

End-to-end learning represents a paradigm where a single model learns to map raw sensory inputs directly to desired outputs or actions, bypassing intermediate representations and hand-engineered pipelines. This approach has revolutionized fields such as autonomous driving, robotics, speech recognition, and game playing by enabling systems to learn complex behaviors directly from data. By eliminating manual feature engineering and modular pipelines, end-to-end models can discover optimal representations and decision-making strategies that might be difficult to design manually.

3.2 Philosophical and Practical Foundations

3.2.1 The End-to-End Principle

The end-to-end principle argues that certain functions should be implemented at the endpoints of a system rather than in intermediate nodes. In machine learning, this translates to learning direct mappings from inputs to outputs, allowing the model to discover internal representations that are optimal for the task.

3.2.2 Advantages over Modular Approaches

End-to-end models offer several advantages over traditional modular systems:

- **Reduced Engineering Burden**: Eliminates need for hand-designed feature extractors and intermediate representations
- **Joint Optimization**: All components are optimized together for the final objective
- **Discovery of Novel Solutions**: Can discover strategies not envisioned by human designers
- **Adaptability**: More easily adapt to new domains or tasks through additional training

3.2.3 Historical Context

Early examples of end-to-end learning include neural networks for handwriting recognition[16] and speech recognition[17]. The approach gained prominence with the success of deep learning in image classification[18] and was later extended to sequential tasks through recurrent and attention-based architectures.

3.3 Key Application Domains

3.3.1 Autonomous Driving

End-to-end driving models, such as NVIDIA's PilotNet[19], take raw camera images as input and output steering commands directly. These models have demonstrated the ability to learn complex driving behaviors without explicit perception, planning, and control modules.

3.3.2 Robotics

In robotics, end-to-end models enable direct learning of control policies from sensory inputs. Notable examples include:

- **Visual Servoing**: Learning visuomotor policies for manipulation tasks
- **Legged Locomotion**: Training locomotion policies directly from proprioceptive sensors
- **Sim-to-Real Transfer**: Using simulation to train end-to-end policies that transfer to physical robots

3.3.3 Speech Recognition and Synthesis

Modern speech systems, such as WaveNet[20] and DeepSpeech[21], use end-to-end architectures that directly map audio waveforms to text or vice versa, eliminating the need for hand-crafted acoustic and language models.

3.3.4 Game Playing

AlphaGo[22] and subsequent systems demonstrated end-to-end learning of game-playing policies, combining perception (board state recognition) with decision-making (move selection) in a unified model.

3.3.5 Natural Language Processing

Sequence-to-sequence models[23] revolutionized machine translation by learning direct mappings between source and target language sentences, bypassing traditional pipeline components like parsing and transfer rules.

3.4 Architectural Innovations

3.4.1 Encoder-Decoder Architectures

Encoder-decoder frameworks provide a flexible template for end-to-end learning of sequence transduction tasks, with applications in machine translation, summarization, and dialogue systems.

3.4.2 Attention Mechanisms

Attention mechanisms, particularly self-attention[24], enable models to learn dynamic input-output alignments, crucial for tasks requiring variable-length context.

3.4.3 Memory-Augmented Networks

Architectures with explicit memory components, such as Neural Turing Machines[25] and Differentiable Neural Computers, extend end-to-end learning to tasks requiring reasoning and long-term information retention.

3.4.4 Multimodal Fusion

For tasks involving multiple input modalities (e.g., vision and language), end-to-end models learn to fuse information across modalities at appropriate levels of abstraction.

3.5 Training Techniques and Challenges

3.5.1 Curriculum Learning

Progressive training strategies that start with simpler versions of a task and gradually increase complexity can help end-to-end models learn complex behaviors.

3.5.2 Imitation Learning

Behavioral cloning and inverse reinforcement learning provide ways to train end-to-end policies by imitating expert demonstrations.

3.5.3 Reinforcement Learning

Policy gradient methods and Q-learning enable end-to-end learning of decision-making policies through trial-and-error interaction with environments.

3.5.4 Challenge: Sample Efficiency

End-to-end models often require large amounts of training data, particularly for tasks with high-dimensional inputs and complex objectives.

3.5.5 Challenge: Interpretability

The internal representations learned by end-to-end models can be difficult to interpret, raising concerns for safety-critical applications.

3.5.6 Challenge: Stability and Convergence

Training end-to-end models with many components can suffer from instability, vanishing/exploding gradients, and convergence issues.

3.6 Hybrid Approaches

3.6.1 Modular End-to-End Learning

Approaches that maintain some modularity while still enabling end-to-end optimization, such as neural module networks[26], offer a compromise between pure end-to-end learning and traditional pipelines.

3.6.2 Inductive Biases and Priors

Incorporating appropriate inductive biases (e.g., convolutional structure for images, recurrence for sequences) can improve sample efficiency and generalization of end-to-end models.

3.6.3 Self-Supervised Pretraining

Pretraining on auxiliary tasks can provide useful representations that accelerate and improve end-to-end learning on target tasks.

3.7 Emerging Trends

3.7.1 Few-Shot End-to-End Learning

Techniques that enable end-to-end models to learn new tasks with minimal examples, combining the flexibility of end-to-end learning with the sample efficiency of few-shot learning.

3.7.2 Neuro-Symbolic Integration

Combining neural end-to-end models with symbolic reasoning systems to enhance interpretability, reasoning capabilities, and data efficiency.

3.7.3 Causal End-to-End Learning

Incorporating causal reasoning into end-to-end models to enable more robust generalization and intervention planning.

3.7.4 Energy-Efficient End-to-End Models

Development of end-to-end architectures optimized for edge deployment, with considerations for computational efficiency, memory usage, and power consumption.

3.8 Future Directions

3.8.1 Foundation Models for End-to-End Learning

Large pre-trained models that can be adapted to various end-to-end tasks through fine-tuning or prompting, analogous to foundation models in language and vision.

3.8.2 Unified Embodied AI Models

Development of general-purpose end-to-end models that can perform diverse perception, reasoning, and action tasks across different embodiments and environments.

3.8.3 Safe and Verifiable End-to-End Systems

Research into techniques for verifying the safety and robustness of end-to-end models, particularly for critical applications like autonomous vehicles and healthcare.

3.8.4 Human-in-the-Loop End-to-End Learning

Frameworks that enable effective human collaboration with end-to-end models, including interactive training, explainability, and control.

3.9 Conclusion

End-to-end models have transformed how we approach complex learning problems by enabling direct mapping from raw inputs to desired outputs. While challenges remain in terms of sample efficiency, interpretability, and safety, ongoing architectural innovations and training techniques continue to advance the state of the art. As end-to-end learning converges with other AI paradigms like foundation models, causal reasoning, and neuro-symbolic integration, it promises to enable more capable, efficient, and trustworthy autonomous systems. The future of end-to-end learning lies not in replacing all modular approaches, but in finding optimal balances between the flexibility of end-to-end learning and the structure and interpretability of modular designs.

References for this Part

[1] HA D, SCHMIDHUBER J. World Models[J]. arXiv preprint arXiv:1803.10122, 2018.

[2] HAFNER D, LILLICRAP T, NOROUZI M, et al. DreamerV2: Mastering Atari with Discrete World Models[J]. arXiv preprint arXiv:2010.02193, 2020.

[3] HAFNER D, PASUKONIS J, BA J, et al. DreamerV3: Mastering Diverse Domains through World Models[J]. arXiv preprint arXiv:2301.04104, 2023.

[4] HAFNER D, LILLICRAP T, FISCHER I, et al. Learning Latent Dynamics for Planning from Pixels[J]. International Conference on Machine Learning, 2019: 2555-2565.

[5] MICHELI V, FLEURET F. IRIS: Implicit Reinforcement without Interaction at Scale[J]. arXiv preprint arXiv:2310.00567, 2023.

[6] Anonymous. Genie: Generative Interactive Environments[J]. arXiv preprint arXiv:2402.15391, 2024.

[7] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models From Natural Language Supervision[J]. International Conference on Machine Learning, 2021: 8748-8763.

[8] LI L H, YATSKAR M, YIN D, et al. VisualBERT: A Simple and Performant Baseline for Vision and Language[J]. arXiv preprint arXiv:1908.03557, 2019.

[9] KIM W, SON B, KIM I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision[J]. International Conference on Machine Learning, 2021: 5583-5594.

[10] LI J, LI D, XIONG C, et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation[J]. International Conference on Machine Learning, 2022: 12888-12900.

[11] LI J, LI D, SAVARESE S, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models[J]. arXiv preprint

arXiv:2301.12597, 2023.

[12] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: a Visual Language Model for Few-Shot Learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.

[13] LIU H, LI C, WU Q, et al. LLaVA: Large Language and Vision Assistant[J]. arXiv preprint arXiv:2304.08485, 2023.

[14] OpenAI. GPT-4V(ision) System Card[J]. OpenAI Technical Report, 2023.

[15] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-Rank Adaptation of Large Language Models[J]. International Conference on Learning Representations, 2021.

[16] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[17] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[18] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[19] BOJARSKI M, DEL TESTA D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[J]. arXiv preprint arXiv:1604.07316, 2016.

[20] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

[21] HANNUN A, CASE C, CASPER J, et al. Deep speech: Scaling up end-to-end speech recognition[J]. arXiv preprint arXiv:1412.5567, 2014.

[22] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[23] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[25] GRAVES A, WAYNE G, DANIHELKA I. Neural turing machines[J]. arXiv preprint arXiv:1410.5401, 2014.

[26] ANDREAS J, ROHRBACH M, DARRELL T, et al. Neural module networks[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 39-48.