CAICT 中国信通院

No.202201

AI 框架发展白皮书

（2022 年）

中国信息通信研究院

2022年2月

版权声明

前　言

AI 助力当前经济社会步入智能经济时代。世界正在进入以新一代信息技术驱动发展的重塑时期，人工智能（AI，Artificial Intelligence）作为其中重要的使能技术，对激活实体经济具有溢出带动性很强的“头雁效应”，对构筑国家科技影响力具有举足轻重的意义。人工智能成为了全球各国新的科技热点，人工智能基础设施建设也成为重要抓手与着力点。未来十年是全球发展数字经济、迈入智能经济社会的黄金发展期，着力发展人工智能基础设施，将为我国人工智能产业发展壮大、数字经济蓬勃发展提供强大牵引力。

AI 框架是智能经济时代的操作系统。作为人工智能开发环节中的基础工具，AI 框架承担着 AI 技术生态中操作系统的角色，是 AI 学术创新与产业商业化的重要载体，助力人工智能由理论走入实践，快速进入了场景化应用时代，也是发展人工智能所必需的基础设施之一。随着重要性的不断凸显，AI 框架已经成为了人工智能产业创新的焦点之一，引起了学术界、产业界的重视。

在此背景下，白皮书致力于厘清 AI 框架的概念内涵、演进历程、技术体系与作用意义，通过梳理总结当前 AI 框架发展现状，研判 AI 框架技术发展趋势，并对 AI 框架发展提出展望与路径建议。由于 AI 框架仍处于快速发展阶段，我们对 AI 框架的认识还有待持续深化，白皮书中存在的不足之处，欢迎大家批评指正。

目 录

图 目 录

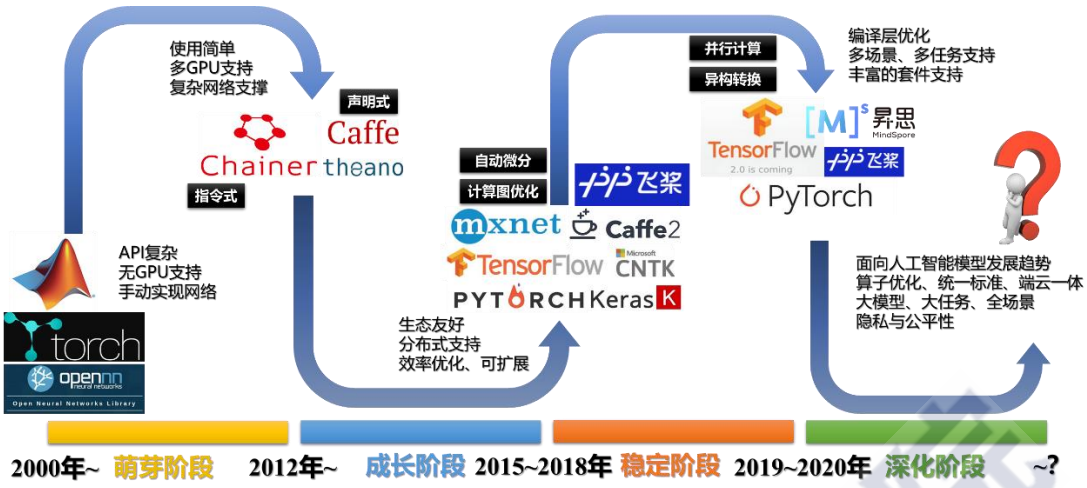
表 目 录

一、AI 框架技术持续演进，已形成较为完整的体系

AI 框架是 AI 算法模型设计、训练和验证的一套标准接口、特性库和工具包，集成了算法的封装、数据的调用以及计算资源的使用，同时面向开发者提供了开发界面和高效的执行平台，是现阶段 AI 算法开发的必备工具。当前，人工智能基础性算法理论研究创新日益活跃，深度神经网络日趋成熟，各大厂商纷纷投入到深度神经网络算法的工程实现并发力建设算法模型工具，进一步将其封装为软件框架供开发者使用，这个过程中 AI 框架（业界也称 AI 开发框架、深度学习框架等）应运而生。AI 框架负责给开发者提供构建神经网络模型的数学操作，把复杂的数学表达转换成计算机可识别的计算图，自动对神经网络进行训练，得到一个神经网络模型用于解决机器学习中分类、回归的问题，实现目标分类、语音识别等应用场景。

（一）AI 框架演进步入深化阶段

结合人工智能的发展历程和 AI 框架的技术特性来看，AI 框架的发展大致可以分为四个阶段，分别为萌芽阶段（2000 年初期）、成长阶段（2012~2014 年）、稳定阶段（2015 年~2019 年）、深化阶段（2020 年以后）。其发展脉络与人工智能，特别是神经网络技术的异峰突起有非常紧密的联系。



来源：中国信息通信研究院

图 1 AI 框架技术演进

萌芽阶段：受限于计算能力不足，这一阶段的神经网络技术影响力相对有限，因而出现了一些传统的机器学习工具来提供基本支持，也就是 AI 框架的雏形，但这些工具或者不是专门为神经网络模型开发定制的，或者 API 极其复杂对开发者并不友好，且这些工具并没有对 GPU 算力进行支持。**这一阶段的 AI 框架并不完善，开发者不得不进行大量基础的工作，例如手写反向传播、搭建网络结构、自行设计优化器等。**

成长阶段：2012 年，Alex Krizhevsky 等人提出了一种深度神经网络架构，即著名的 AlexNet，在 ImageNet 数据集上达到了最佳精度，并碾压第二名，引爆了深度神经网络的热潮。自此极大地推动了 AI 框架的发展，出现了 Caffe、Chainer 和 Theano 等具有代表性的早期 AI 框架，帮助开发者方便地建立复杂的深度神经网络模型，如 CNN、RNN、LSTM 等。不仅如此，这些框架还支持多 GPU 训

练，让开展更大、更深的模型训练成为可能。**在这一阶段，AI 框架体系已经初步形成，声明式风格和命令式风格为之后的 AI 框架趟出了两条不同的发展道路。**

稳定阶段：2015 年，何恺明等人提出的 ResNet，再次突破了图像分类的边界，在 ImageNet 数据集上的准确率再创新高，也终于凝聚了产业界和学界的共识，那就是深度学习将成为下一个重大技术趋势。在这一到两年里，Google 开源了著名的 TensorFlow 框架，它至今仍是机器学习领域最流行的 AI 框架。Caffe 的发明者加入了 Facebook（现更名为 Meta）并发布了 Caffe2；与此同时，Facebook AI 研究团队也发布了另一个流行的框架 PyTorch，该框架拓展自 Torch 框架，但使用了更流行的 Python API。微软研究院开发了 CNTK 框架。Amazon 采用了 MXNet，这是华盛顿大学、CMU 和其他机构的联合学术项目。国内的百度则率先布局了 PaddlePaddle 飞桨深度学习框架并于 2016 年发布。

TensorFlow 和 CNTK 借鉴了 Theano 的声明式编程风格，而 PyTorch 则继承了 Torch 的直观和开发者友好的命令式编程风格。Francois Chollet 几乎是独自开发了 Keras 框架，该框架提供了神经网络和构建块的更直观的高级抽象。同时各种 AI 框架不断进行迭代，为框架提供各种面向高效友好开发的核心组件，例如几乎所有 AI 框架都支持的自动微分能力，TensorFlow 提供了分布式版本的 AI 框架和支持 iOS 系统的能力，PyTorch 则在完全拥抱 Python 的基

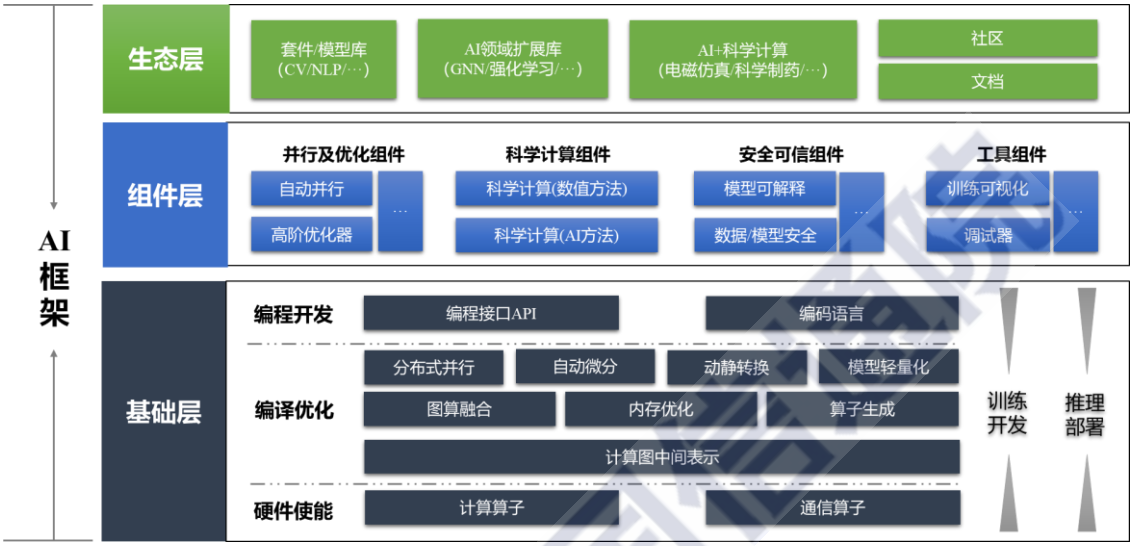
础上提供了一整套包括优化器、库函数、API 工具等支持。**AI 框架迎来了繁荣，而在不断发展的基础上，各种框架不断迭代，也被开发者自然选择。**

经过激烈的竞争后，最终形成了两大阵营，TensorFlow 和 PyTorch 双头垄断。2019 年，Chainer 团队将他们的开发工作转移到 PyTorch；Microsoft 停止了 CNTK 框架的积极开发，部分团队成员转而支持 PyTorch；Keras 被 TensorFlow 收编，并在 TensorFlow2.0 版本中成为其高级 API 之一。

深化阶段：随着人工智能的进一步发展，新的趋势不断涌现，例如超大规模模型的出现（GPT-3 等），向 AI 框架提出了更高的要求。随着人工智能应用场景的扩展以及与更多领域交叉融合进程的加快，越来越多的需求被提出，如对全场景多任务的支持、对高算力的需求等，这就要求 AI 框架最大化的实现编译优化，更好地利用算力、调动算力，充分发挥硬件资源的潜力。此外，人工智能与社会伦理的痛点问题也促使可信赖人工智能在框架层面的进步。基于以上背景，现有的流行框架都在探索下一代 AI 框架的发展方向，如 2020 年华为推出昇思 MindSpore，在全场景协同、可信赖方面有一定的突破；旷视推出天元 MegEngine，在训练推理一体化方面深度布局。**在这一阶段，AI 框架正向着全场景支持、超大规模 AI、安全可信等技术特性深化探索，不断实现新的突破。**

（二）AI 框架技术演化出三个层次

根据技术所处环节及定位，当前主流 AI 框架的核心技术可分为基础层、组件层和生态层。



来源：中国信息通信研究院

图 2 AI 框架核心技术体系

1.基础层

基础层实现 AI 框架最基础核心的功能，具体包括编程开发、编译优化以及硬件使能三个子层。编程开发层是开发者与 AI 框架互动的窗口，为开发者提供构建 AI 模型的 API 接口。编译优化层是 AI 框架的关键部分，负责完成 AI 模型的编译优化并调度硬件资源完成计算。硬件使能层是 AI 框架与 AI 算力硬件对接的通道，帮助开发者屏蔽底层硬件技术细节。

编程开发-编程接口 API：开发者通过调用编程接口来描述算法的计算过程。对于开发者来说，编程接口的易用性以及接口的表达

能力非常重要，对算法的描述会映射到计算图上。编程接口主要可以分为 3 类：一类是基于数据流图的编程接口，流行的基于数据流图的机器学习编程框架包括 TensorFlow、MXNet、Theano、Torch7 等；另一类是基于层的编程接口，如 Caffe；还有一类是基于算法的编程接口，主要用于传统机器学习算法的实现，如 Scikit-Learn。

编程开发-编码语言：人工智能应用场景众多，人工智能开发者基于不同场景选择使用的编程语言多样，完善的 AI 框架应支持多种不同的语言，例如 Python/仓颉/Julia 等。面向使用不同编程语言的开发者，AI 框架需要提供功能相同、性能可比的开发服务和技术支持。

编译优化-分布式并行：指数据流并行、模型并行、Pipeline 并行、优化器并行等策略。随着模型规模的增大，传统的数据并行无法有效处理，自动并行技术的使用将会是常态。需要将大模型切分到不同的设备上，切分就是将不同大块计算切分成小块计算，并将小块计算发送到不同的计算资源进行计算，最后将小块计算的结构进行规约合并。而切分策略寻优是很困难的，不同的切分产生的通信量差异巨大，计算利用率也很不一样，比如 Pipeline 并行往往在计算利用率方面存在较大的挑战，算子切分的并行则在通信量方面存在较大的挑战，需要 AI 框架来支持。

编译优化-自动微分：自动微分是将一个复杂的数学运算过程分解为一系列简单的基本运算，每一项基本运算都可以通过查表得出

来。自动微分有两种形式，包括前向模式(forward mode)和反向模式(reverse mode)，前向模式是在计算图前向传播的同时计算微分，反向模式需要对计算图进行一次正向计算，得出输出值，再进行反向传播。因此反向模式的内存开销要大一点，它需要保存正向传播中的中间变量值，这些变量值用于反向传播的时候计算导数。

编译优化-动静转换：静态图在定义执行前的所有操作和网络结构，并将其呈现给传感器流，在训练期间提供了更高的性能，但这样做的代价是不易于使用、不够灵活。动态图计算是即时执行的，提供了更大的灵活性和更容易的调试,但这样做的代价是性能较低。TensorFlow2.0、MindSpore 等均支持动态图和静态图的转换技术，可以实现计算效率和灵活性的平衡。

编译优化-模型轻量化：轻量化是指为满足 AI 模型尺寸小、计算复杂度低、电池耗电量低、下发更新部署灵活等要求下，AI 框架所配置的轻量化技术。一般来说，模型轻量化就是指模型压缩和加速，其中压缩重点在于减少网络参数量，加速则侧重在降低计算复杂度、提升并行能力等。算法层压缩加速主要包括结构优化（如矩阵分解、分组卷积、小卷积核等）、量化与定点化、模型剪枝、模型蒸馏等；框架层加速主要包括编译优化、缓存优化、稀疏存储和计算、NEON 指令应用、算子优化等。

编译优化-图算融合：通过自动分析和优化现有网络计算图逻辑，并结合目标硬件能力，对计算图进行计算化简和替代、算子拆分和

融合、算子特例化编译等优化，以提升设备计算资源利用率，实现对网络性能的整体优化。相比传统优化技术，图算融合具有多算子跨边界联合优化、与算子编译跨层协同、基于 Polyhedral 的算子即时编译等独特优势。另外，图算融合只需要开发者打开对应配置，整个优化过程即可自动完成，不需要网络开发人员进行其它额外感知，使得开发者可以聚焦网络算法实现。

编译优化-内存优化：由于硬件系统的内存资源有限，特别是 AI 芯片的内存资源有限，需要有高效的内存优化策略降低 AI 网络对系统内存的消耗。一般常用的内存优化技术有：静态内存复用优化和动态内存分配机制。静态内存复用优化会分析计算图的数据流关系，基于数据的内存占用大小、数据间的生命周期重叠关系，规划数据的内存复用策略，从而最小化内存占用。动态内存分配机制是在运行时创建大块内存，并按照实际算子执行过程中需要的内存进行内存切片提供，当算子执行完且相关数据的引用均已结束时，释放内存切片，从而实现内存的有效复用。

编译优化-算子生成：AI 框架会提供基础常用的算子，但是这些算子往往不能满足开发者算法不断演进的需求。因此，需要 AI 框架具备针对不同算力设备的统一算子生成和优化的能力，使得开发人员只需要编写高层编程语言（如 DSL）就可以通过 AI 框架提供的算子编译生成能力，生成高质量的底层算子，极大降低 AI 框架和硬件平台的开发和维护成本，拓展应用范围。

编译优化-中间表示：中间表示（Intermediate Representation，简称 IR）是对计算图和算子格式的定义。完备的中间表示需要支持不同硬件设备算子定义和计算图的性能优化，支持不同类型的 AI 模型网络结构的灵活表达，支持不同设备间的模型中转和迁移。

硬件接入-计算算子：在深度学习领域计算算子特指计算图中的一个函数节点，一个在张量上执行的计算操作，它接受零或多个张量作为输入，得到零或多个张量作为输出，利用梯度、散度、旋度的表达方式进行计算。

硬件接入-通信算子：用于分布式节点通信的函数节点。

2.组件层

组件层主要提供 AI 模型生命周期的可配置高阶功能组件，实现细分领域性能的优化提升，包括编译优化组件、科学计算组件、安全可信组件、工具组件等，对人工智能模型开发人员可见。

并行及优化组件-自动并行：指对自动并行技术的多样化组合支持。AI 框架支持开发者进行多种不同并行进行组合，根据需要形成混合并行策略，例如数据流并行和模型并行的组合、数据流和 Pipeline 并行的组合等，支持开发者个性化的选择自己的并行策略，以更灵活的姿态支持人工智能模型训练、应用适配。

并行及优化组件-高阶优化器：AI 框架支持多种不同的一阶/二阶优化器，能为开发者提供灵活方便的接口，例如 SGD 优化器、

SGDM 优化器、NAG 优化器、AdaGrad 优化器、AdaDelta 优化器、Adam 优化器、Nadam 优化器等。

科学计算组件-科学计算（数值方法）：人工智能发展的重要方向之一是科学计算，因此要求 AI 框架向开发者提供科学计算相关的功能支持，通过函数式编程范式为 AI+科学计算提供融合的表达方式，使得开发者以更加接近数学计算的方式进行编程，以缓解当前 AI 框架的编程接口主要面向深度神经网络设计，但是科学计算中需要大量的数学公式的表达（例如微分方程求解）的情况。

科学计算组件-科学计算（AI 方法）：针对 AI 方法直接替代数值方法取得计算结果的形式，AI 框架需要具备“AI+科学计算”统一的数据底座，将传统科学计算的输入数据（如传统科学计算软件生成的仿真数据）转换为 AI 框架的输入数据（即张量）。针对 AI 方法与数值方法配合取得计算结果形式，除了需要具备统一的数据引擎之外，AI 框架需要支持传统数值计算的方法，例如高阶微分求解、线性代数计算等，并通过计算图对传统数值方法和 AI 方法的混合计算优化，从而实现“AI+科学计算”端到端加速。

安全可信组件-AI 可解释：AI 框架需要具备三个层面的能力支持可解释人工智能。建模前的“数据可解释”，分析数据分布，找出代表性的特征，在训练时选择需要的特征进行建模。构建“可解释人工智能模型”，通过与传统机器学习（如贝叶斯概率编程）结合的方式，对人工智能结构进行补充，平衡学习结果的有效性和学习模型

的可解释性。对已构筑模型进行“解释性分析”，通过分析人工智能模型的输入、输出、中间信息的以来关系分析（如 TB-Net 的方式）及验证模型的逻辑。

安全可信组件-数据安全：人工智能领域的数据安全问题不仅仅涉及到原始数据本身的保护，还要防止通过模型推理结果反推出数据隐私关键信息。因此，AI 框架本身除了要提供数据资产保护能力，还需要通过差分隐私等方式，保护模型数据的隐私。同时，为了源头保护数据安全，AI 框架通过联邦学习等方式进行模型训练，使得数据不出端的情况下模型得到训练更新。

安全可信组件-模型安全：训练模型时样本训练不足，使得模型泛化能力不足，导致模型面对恶意样本时，无法给出正确的判断结果。为此，AI 框架首先需要提供丰富的人工智能鲁棒性检测工具，通过黑盒、白盒、灰盒测试等对抗检测技术测试人工智能模型的鲁棒性，如静态结构分析，动态路径分析等；其次，AI 框架可以通过支持网络蒸馏、对抗训练等方式帮助开发者提高模型的鲁棒性。

工具组件-训练可视化：支持训练过程可视化，可通过页面直接查看训练过程中的核心内容，包括训练标量信息、参数分布图、计算图、数据图、数据抽样等模块。

工具组件-调试器：神经网络训练中经常出现数值误差情况，如无穷大等，开发者希望分析训练无法收敛的原因。但是，由于计算被封装为黑盒，以图的方式执行，开发者很难定位其中的错误。调

试器是训练调试的工具，开发者可以在训练过程中查看图的内部结构以及节点的输入/输出，例如查看一个张量的值，查看图中的节点对应的 Python 代码等。此外，开发者还可以选择一组节点设置条件断点，实时监控节点的计算结果。

3.生态层

生态层主要面向应用服务，用以支持基于 AI 框架开发的各种人工智能模型的应用、维护和改进，对于开发人员和应用人员均可见。

套件/模型库：AI 框架应对领域通用任务提供预训练模型或者定义好的模型结构，方便开发者获取和开展人工智能模型训练和推理，如 CV、NLP 等。

AI 领域扩展库：AI 框架要能够提供丰富的领域任务支持，并为相关任务提供典型案例，从而提供更好的应用服务，如 GNN、强化学习、迁移学习等。

AI+科学计算：与 CV、NLP 等传统信息领域不同，科学计算问题的求解需要具备相对专业的领域知识。为了加速 AI+科学计算融合的研究和落地，AI 框架需要面向不同的科学计算领域（如电磁仿真、科学制药、能源、气象、生物、材料等）提供简单易用的科学计算套件，这些套件包含高质量的领域数据集、高精度的基础 AI 模型和用于前后处理的工具集合。

文档：AI 框架应提供完善的文档体系，包括但不限于框架说明文档、框架 API 文档、框架版本变更文档、框架常见问题文档、框架特性文档等。

社区：人工智能服务发展需要社区支持，AI 框架应该经营或者维护良好的社区环境，好的 AI 框架具备较好的维护性和易用性，同时 AI 框架社区中应该有代表性项目并长期支持基于该框架的项目和应用。

（三）AI 框架重要性愈加突显

AI 框架承上启下，是整个人工智能技术体系的核心。从技术体系中的功能定位看，AI 框架对下调用底层硬件计算资源，能够屏蔽底层差异并提供良好的执行性能，对上支撑 AI 应用算法模型搭建，提供算法工程化实现的标准环境，是 AI 技术体系的关键核心。除完成 AI 算法的工程实现外，AI 框架还能极大提高人工智能学习效率、强化 AI 算法模型能力，如基于 TensorFlow 的 AlphaGo 在极短时间内学习到战胜前任 AlphaGo 的技能。

AI 框架是应对智能经济时代的技术利器。大规模并行计算及智能应用是未来智能经济时代的主要特点。当前硬件计算以 CPU 为代表，软件栈主要针对串行指令进行优化。由于人工智能算法涉及大量的矩阵计算和并行数值计算，面向智能经济时代的硬件计算已经显示出从串行迁移到并行计算的趋势，未来可能以 GPU 为代表，软件栈主要针对大规模并行计算进行优化，这其中 AI **框架将成为大**

规模并行计算的关键调度者。此外，人工智能模型将主导智能经济时代各行各业细分场景，智能应用将呈现规模化、深度化等特点，而 AI 框架就是智能应用快速落地的关键支撑者。

AI 框架将成为智能经济时代的操作系统。当前互联网时代，操作系统是 IT 业的核心枢纽点，建立硬件和应用软件之间的联系，左右着数字设备的整个生态，通过与通用计算芯片的深度绑定，形成 Windows+Intel、Android/iOS+ARM 两大稳定的技术体系格局。智能经济时代，AI 框架承担着 AI 技术生态中操作系统的角色，是 AI 学术创新与产业商业化的重要载体，助力人工智能由理论走入实践，快速进入场景化应用时代。总体来说，**“AI 框架+算力芯片”的组合在一定程度上决定了人工智能产业应用的主体技术路线**，其研发能够促进生态圈关联及外围的芯片、系统、软硬件平台等产业发展，从而促进人工智能核心生态圈的建设。随着价值不断凸显，**AI 框架已经成为了人工智能产业创新的焦点之一，引起了学术界、产业界的重视**。

二、全球 AI 框架繁荣发展，多元化竞合态势渐显

（一）供给主体方面，企校贡献最活跃

科技企业与顶尖高校对 AI 框架的发展成熟贡献最为活跃。数字科技企业巨头与顶尖高校是 AI 框架发展壮大的主体维护力量，打造技术产业生态、营造学术创新氛围，是两大主体的源动力。个

人及开源组织也扮演着重要的角色，是 AI 框架创新性、公益性的重要体现。

数字科技企业巨头是 AI 框架发展壮大的核心力量。自身 AI 业务场景需求激发 AI 框架的应用，并实现 AI 框架的验证完善。国际知名数字科技巨头主导开源 AI 框架技术生态，我国数字科技企业近年来也积极布局并不断创新。Google、Meta、Microsoft、Amazon 等国外数字科技企业巨头在基础算法框架研发方面具有先发优势，依托自身 AI 业务场景以及庞大的数据资源，能够对算法框架进行有效试验验证及功能完善。在此基础之上，数字科技企业巨头将原本服务于内部业务场景的 AI 框架进行开源，为产业链下游合作伙伴提供底层 AI 核心能力，满足工业级应用需求，逐步完善整体生态，实现合作共赢。国内数字科技巨头纷纷布局推出 AI 框架，立足满足自身的 AI 应用需求外，也对外拓展服务，如华为 MindSpore、百度 PaddlePaddle、腾讯 TNN、阿里 MNN、字节跳动 BytePS 以及小米 Mace 等。

高校及科研院所是最早启动 AI 框架研发的主导力量之一，并持续发挥着积极作用。高校及科研院所拥有强大的人才资源，基于实验室科研创新需求对 AI 框架开展基础性理论研究工作，布局整体早于数字科技企业，更易实现革命性突破创新。高校最早推出的 Theano、Caffe 等开源框架能够满足学术研究需求，并对 AI 框架的整体发展起到巨大推动作用，但在大规模分布式计算等场景下的性

能不及企业推出的 AI 框架。随后，高校通过更换维护主体以持续释放作用价值。例如，MXNet 框架发起于卡内基梅隆大学，后捐赠给 Apache 基金会，现成为 Amazon AWS 最主要的 AI 框架。我国高校日渐重视 AI 框架研发，如清华大学已陆续开发出开源框架计图 Jittor、贝叶斯深度学习算法框架“珠算”等。

（二）开源生态方面，全球进入活跃期

开源本质上是一种人才、智慧的聚合，能够助推 AI 框架快速升级。茁壮的开源生态对于 AI 框架的发展至关重要。开发者通过在开源社区进行代码开源、项目托管、协作分享、沟通交流等一系列活动，实现与开源 AI 框架的紧密互动。开源社区是 AI 框架开发者必不可少的学习与交流环境，可以说开源社区在推动 AI 框架发展的过程中起着巨大的作用。开源社区的相关指标，也体现着 AI 框架在整个行业内的发展情况。对 AI 框架来说，国外最知名社区是 Microsoft 收购的开源代码托管平台 Github，国内知名社区是由 OSCHINA.NET 推出的代码托管平台 Gitee（码云）。

表 1 Github 社区中主流 AI 框架情况（2022.1）

Rank	Framework	Commits[1]	Fork[2]	Star[3]	Contributors[4]
Foreign Framework					
1	TensorFlow	124494	86300	163000	3056

[1] Commits 代表开源代码提交的次数，表征开源项目活跃度.
[2] Fork 代表代码复刻、分叉，表征开源项目被引用情况.
[3] Star 代表点赞数，表征开源项目关注度.
[4] Contributors 代表贡献者，表征开源项目贡献者规模.

Rank	Framework	Commits[1]	Fork[2]	Star[3]	Contributors[4]
2	PyTorch	43390	14800	53700	2137
3	Theano (Stop Developing)	28127	2500	9500	352
4	CNTK (Stop Developing)	16116	4400	17100	201
5	MXNet	11776	6900	19800	868
Domestic Framework					
1	MindSpore	37308	514	2700	267
2	PaddlePaddle	33753	4300	17500	524
3	MegEngine	2282	462	4100	32
4	OneFlow	7621	351	3000	99
5	Jittor	1266	235	2300	31

来源：根据 Github 社区数据整理

Github 作为业内认可度最高的开源社区，也是 AI 框架开发者最关注的代码托管平台。从 Github 指标看，国外 AI 框架方面，TensorFlow **的各项指标均高居榜首，并远超第二名，是全球目前活跃度最高、应用最广的** AI **框架**。近年来在学术领域表现亮眼的后起之秀 PyTorch 紧随其后，虽在顶会占据了主流地位，但与 TensorFlow 相比仍略逊一筹。MXNet 表现也较为亮眼，但与前两者不在同一量级。**我国主体推出的 AI 框架方面，**MindSpore **是目前活跃度最高的** AI **框架**，在贡献者方面也已集聚了一定规模使用群体。百度 PaddlePaddle 开源时间较早，在关注度方面较其他框架有一定优势。其余框架中，OneFlow 的活跃度与贡献者规模处于领先位置。

表 2 Gitee 社区中主流 AI 框架情况（2022.1）

Rank	Framework	Commits	Fork	Star	Contributors
1	MindSpore	38549	2400	6100	774
2	PaddlePaddle	32788	195	3600	561
3	OneFlow	7521	2	1	126
4	MegEngine(镜像）	2280	6	16	35
5	Jittor	1239	3	11	34

来源：根据 Gitee 社区数据整理

国内最大的开源代码托管平台 Gitee 目前主要是我国企业所主导 AI 框架进行发布交流的平台。国内知名的框架除旷视 MegEngine 尚未在社区上发布外，其他框架均有所布局，也吸引了国内的开发群体。其中，MindSpore 在 Gitee 中的各项指标都远超其他 AI 框架，是国内社区中最活跃、关注度最高、被应用最多的框架，处在我国开源生态的引领者地位。

（三）市场格局方面，双寡头持续引领

全球来看，国际主流 AI 框架由 Google、Meta 等科技巨头主导。目前以 Google、Meta、Amazon、Microsoft 等代表的互联网科技巨头，凭借自身的数据、技术和资本等优势，持续在 AI 框架生态领域发力，引领全球 AI 框架技术创新升级趋势，并逐步形成了以 Google-TensorFlow 和 Meta-PyTorch 为代表的双寡头格局。从市场占有情况看，产业界以 TensorFlow 为主，学术界以 PyTorch 为主。Github 中 Star 数表征开源项目流行度，是开源项目在产业界中市场

份额的生动体现，据表 1 数据显示，TensorFlow Star 数达到 163000，远高于排名第二的 PyTorch（53700），且 Google 于 2019 年推出 TensorFlow Enterprise，为大型企业提供 TensorFlow 的优化版本以及长期的技术支持，并与 Google Cloud 服务深度集成，持续巩固 TensorFlow 在产业界的领先地位。据 Papers With Code 数据[5]显示，2021 全年基于 PyTorch 的论文数量在所有基于 AI 框架的论文中占比高达 58.56%，远高于排名第二的 TensorFlow（12.38%），PyTorch 在学术界的领先优势在持续加强。

国内来看，双寡头并驱态势下 AI 框架市场格局向着多元发展。我国在 AI 应用方面优势显著，相当规模的 AI 应用均构筑在国际主流 AI 框架之上，从底层开源代码贡献、底层硬件适配，到中间算子研发迭代、模型库完善，以及上层算法模型构建，双寡头持续为国内 AI 应用生态输出能力。不仅如此，近两年国内厂商推出的 AI 框架市场占有率也正稳步提升。MindSpore 框架开源后获得国内外开发者的积极响应，在 Gitee 千万个开源项目中综合排名第一，成为国内最活跃的 AI 开源框架。百度飞桨 PaddlePaddle 开发者规模也在持续壮大，从 IDC 2021 年调研的 350 份中小企业开发者样本数据显示，飞桨开发者认知度占比已超 20%。

[5] https://paperswithcode.com/trends.

（四）支撑应用方面，科研与产业齐驱

1.AI 框架赋能学术科研

AI 与超级计算机的结合，使科研领域的计算能力普遍提升到一个新的高度。2021 年世界排名前 500 的超级计算机中，68.4%采用了 AI 技术进行了加速。美国橡树岭国家实验室利用 TensorFlow 在 Summit 超级计算机上训练了 1.1EFLOP/s 的极端天气预报模型，用来模拟预测气候变迁会产生的极端天气，提升了气象研究的精准度和可能性。美国劳伦斯伯克利国家实验室在基于 CPU 的高性能计算平台上，使用 TensorFlow 框架开发了大型科学应用程序 CosmoFlow，利用机器学习插件前所未有的将 TensorFlow 框架扩展到 8000 多个节点，以这种规模处理三维空间数据卷，主要应用在暗物质 N 体模拟实验中，为科学家提供了一个全新的平台来加深对宇宙的了解。

TensorFlow 被广泛应用于学术科研领域。美国航空航天局使用 TensorFlow 对开普勒任务中积累的大量数据进行分析，由于机器学习能够比人类更高效地搜索更广范围的信号，发现了一直以来忽视的开普勒-90i 行星，这一发现使开普勒-90 星系成为了目前所知除太阳系外唯一八颗行星绕一颗恒星运行的星系，取得了天体物理学领域的一项重大突破。宾夕法尼亚大学研究利用 TensorFlow 解决农业病虫害问题，通过注释大量木薯植株图像来识别和分类疾病，目前在坦桑尼亚部分地区试验应用，农民们可以通过在木薯叶子前挥动手机，快速实现病株识别，并给出最佳的方式来进行管理。雨林保

护组织 Rainforest Connection 基于 TensorFlow 开发了世界上首款可自动识别盗伐行为的可扩展、实时监控报警的热带雨林环保系统，在亚马逊雨林试验应用，通过当地的手机蜂窝网络向中央云计算服务器发送声音采样，依托 TensorFlow 来分析和审计数据，从中甄别电锯、木运卡车等与非法砍伐相关的声音，以防止人工监听遗漏。

我国框架作为后起之秀在学术科研领域已经崭露头角。基于 MindSpore 的鹏程.盘古作为全球首个发布的千亿级预训练中文大模型，模型规模高达 2000 亿参数，MindSpore 采用全自动并行训练方式支撑鹏程.盘古大模型在 4096 张 NPU 芯片上高效训练。紫东.太初是基于 MindSpore 框架构建的全球首个图文音三模态、千亿级参数预训练大模型，具备跨模态理解与跨模态生成能力。武汉大学运用 MindSpore 打造了全球首个专用深度学习遥感框架武汉.LuojiaNet，实现大规模卫星遥感影像的智能遥感解译。PaddlePaddle 联合鹏城实验室发布了鹏城-百度·文心，模型参数规模达到 2600 亿，是目前全球最大中文单体模型，在机器阅读理解、文本分类、语义相似度计算等 60 多项任务取得最好效果。此外，百度基于 PaddlePaddle 研发推出量子机器学习工具集量桨（Paddle Quantum），建立起了人工智能与量子计算之间的桥梁，可以快速实现量子神经网络的搭建与训练，同时还提供多项前沿量子应用。

2.AI 框架赋能产业应用

空客公司使用 TensorFlow 开发的模型进行异常监测，保障空间站运行安全。空客公司为哥伦布实验舱的运行及其在国际空间站上的有效载荷提供多项服务，哥伦布实验舱是欧洲航天局最大的国际空间站项目，装备有多种实验设备，能开展细胞生物学、外空生物学、流体和材料科学、人类生理学、天文学和基础物理学等多方面的实验，由多个组件组成，能够产生约 17000 个独特的遥测参数。空客使用 TensorFlow 开发的模型在数据流监控过程中进行异常检测，并实现实时报告，大大的简化了异常原因分析过程并缩短了解决时间。

生物制药龙头 Celgene 公司借助 MXNet 促进药品研究和发明。Celgene 是一家从事免疫医疗的制药企业，通过训练神经网络识别和决策带有标记细胞的显微镜图像，解决了使用经典的图像分析方法难以大规模识别和区分正常细胞和肿瘤细胞的问题。MXNet 框架对于毒理学预测尤其重要，可以无需活体患者承担风险，虚拟分析潜在药物的生物学影响。

PyTorch 帮助采矿企业 Datarock 进行基于深度学习的岩心钻探。Datarock 通过深度学习模型帮助地质学家更快地分析钻芯样品图像。传统模式下地质学家会一厘米一厘米地仔细研究这些样本，以评估矿物学和结构，工程师则会寻找诸如断层、裂缝和岩石质量等物理特征，这个过程既缓慢又容易出现人为错误。使用 Datarock 的技术，

可以将手动记录耗费的 5-6 小时缩短在半小时内，使地质学家从繁重的基础工作中解放出来。

MindSpore **在行业赋能方面成绩斐然，拥有** 300 **多个** SOTA **模型，超过** 4000 **个开源生态社区贡献者，支持超过** 5000 **个在线** AI **应用，广泛应用于工业制造、金融、能源电力、交通、医疗等行业。** MindSpore 赋能工业制造，通过 AI 技术助力降低重复劳动，华为松山湖南方工厂通过引入 MindSpore 及 AI 质检算法，将印制电路板的缺陷检测精度由 90%提升至 99.9%，并将质检人员的工作效率提升了 3 倍。基于 MindSpore 的金融解决方案在深圳、上海等地银行网点运行效果显著，有效提升潜在客户转化率，同时利用 OCR 识别技术和生物识别技术，实现企业年报、合同、保单、发票等各类文档及工单文本电子化，迅速提升工作效率。基于 MindSpore 的智能输电线路巡检方案对输电线路的设备和周界情况进行前端监控，并分析异常问题及时报警，南方电网、深圳供电局更是开辟了“以系统智能分析为主、人工判断辅”的崭新模式，使原来需要 20 天才能完成的现场巡视工作，输电监控指挥中心现在仅需 2 小时就可完成，巡检效率提高了近 80 倍。除此之外，基于 MindSpore 孵化的紫东.太初、武汉.Luojia 已从学术科研向产业应用转化，支撑央视、爱奇艺、新华社技术局、航天宏图等企业开展创新应用。

PaddlePaddle **服务企业遍布能源、金融、工业、医疗、农业等多个行业，助力千行万业智能化升级。**PaddlePaddle 赋能人民日报

“创作大脑”，覆盖了全媒体策划、采集、编辑、传播效果分析等各环节和业务场景，可以大幅提高新闻产品的生产效率，能够进行视频直播关键人物、语句识别、全网热点数据自定义监测预警、批量生成可视化大数据报告等多种智能化生产。连心医疗基于 PaddlePaddle 平台开发上线“基于 CT 影像的肺炎筛查与病情预评估 AI 系统”，已首先在湖南郴州湘南学院附属医院投入使用，可快速检测识别肺炎病灶，为病情诊断提供病灶的数量、体积、肺部占比等定量评估信息，同时辅以双肺密度分布的直方图和病灶勾画叠加显示等可视化手段，为临床医生筛查和预诊断患者肺炎病情提供定性和定量依据，提升医生诊断和评估效率。

旷视 MegEngine 充分发挥视觉领域优势，实现行业赋能。旷视为某摄像头模组企业提供的智能质检解决方案实现了产品的在线实时检测，基于 Brain++平台的私有化部署版本 MegOne，能够实时发现产品划伤、折痕、油污、破损等缺陷，缺陷检测率同比提高 90%，降低 85%以上人工成本，整体维护成本降低 10%。旷视推出供应链操作系统——河图，在电商仓库中协同 500 台机器人并发工作，将仓库效率提升了 40%。旷视为华润电力部署了园区安全管理系统，利用人脸识别、物体检测等计算机视觉算法，对变电设备周边等危险区域实现了 7*24 小时警戒，显著提升了安全管理水平。

一流科技 OneFlow 充分发挥分布式可扩展性能优势，已服务科研、政务、军工、金融等诸多行业客户。一流科技基于 OneFlow 框

架，集成大数据、云计算等组件，提供商业化产品 OF 智能云，包括人工智能开发平台 OneBrain、强化学习解决方案 OneAgent 及 AI 实训及编程平台 OneLab。其中 OneBrain 助力中关村智用研究院打造一站式人工智能开发平台，提供多种混合算力解决方案，支持资源按需扩容，该项目交付智用投入使用后，经计算，其系统算力率可提升 30%，模型训练时间相较传统方式节省 80%，整体解决了智用复杂业务场景、高算力要求和边界灵活延展要求。

（五）推广途径方面，三条路齐发并进

致力于社区生态的壮大与优化，吸引更多学术界与产业界开发者。主流 AI 框架通过繁荣开源社区生态，打造忠实的贡献者团队，从而吸引更多开发者参与生态构建。Google TensorFlow 团队基于 GitHub 开源，并逐步吸引早期开发者向贡献者转变。围绕 TensorFlow 开源社区，贡献者除了贡献 TensorFlow 高阶 API 代码外，还积极参与 TensorFlow 社区的管理、贡献 TensorFlow 延伸出来的开源项目以及传播知识和分享经验。华为推出 MindSpore 开发者扶植计划，为开发者提供优惠的云服务资源和相关的知识赋能培训资源，帮助个人开发者学习和构建基于 MindSpore 的技术能力，以获得持续职业发展。百度携手社区开发者共建生态，成立飞桨城市/高校领航团 150 个、飞桨特别兴趣小组 12 个，目前全国范围内已有 132 个城市和高校自组织社区在主动自发举办飞桨社区活动。

与高校科研院所联动，拓展高校学术科研开发者规模以及学术科研应用。高校的人才培养和开发者的发展已成为整个 AI 框架生态的重要组成部分，当前国内主流 AI 框架积极融入高校教学体系。华为与教育部联合启动建设“智能基座”产教融合协同育人基地，目前 MindSpore 课程已经在 100 多所高校开设，并积极开展计算机系统能力提升高级研修班，培养 AI 先锋教师。百度支持教育部产学合作协同育人项目，截至目前，PaddlePaddle 已累计培训了 3000 多位高校教师，并且参与编写了一系列人工智能教材。此外，主流 AI 框架也选择通过设立创新基金激励框架的创新应用。华为于 2020 年与中国人工智能学会共同发起《中国人工智能学会-华为 MindSpore 学术奖励基金》，旨在激励原创性科学研究开展，构建中国人工智能科学研究的全球影响力，累计已投入 1600 万资金，支持 120 多个项目，据 Papers With Code 数据显示，2021 年 10 月统计基于 MindSpore 的论文数量在所有基于 AI 框架的论文中占比 10%（当月排名第 2），成效显著。百度于 2020 年与中国计算机学会联合成立了“CCF-百度松果基金”，旨在为青年学者提供经费、平台、数据、技术支持等服务，推动 AI 框架在科研领域的应用。

面向产业应用提供基础设施及解决方案服务，不断吸纳下游合作伙伴。围绕产业应用，AI 框架有三种层次的赋能路径。首先是将 AI 框架融入算力基础设施，提供 AI 能力服务，如各地政府在建的和已上线运营的人工智能计算中心，重点依托我国 AI 框架构建底

层 AI 开发能力，其中 MindSpore 成为主要选择。其次是打造软硬一体化方案，将 AI 框架作为打通底层算力硬件与上层应用的通道，如 PaddlePaddle 积极与硬件厂商合作，完成适配或正在适配的芯片与 IP 型号 31 种，进一步促进软硬件联合优化、协同发展；之江实验室天枢人工智能开源平台，以 OneFlow 框架为核心，上承算法应用，下接底层硬件。另外，还可依托 AI 框架打造面向具体行业的应用平台，如华为联合合作伙伴基于 MindSpore 推出“昇腾智造”、“昇腾智城”、“昇腾智行”、“昇腾智巡”四大行业解决方案。

三、应对未来多样化挑战，AI 框架有六大技术趋势

（一）泛开发：AI 框架将注重前端便捷性与后端高效性的统一

AI 框架需要提供更全面的 API 体系以及前端语言支持转换能力，从而提升前端开发便捷性。AI 框架需要能为开发者提供完备度高、性能优异、易于理解和使用的 API 体系，TensorFlow、JAX 等相关开源项目成员组织的 Consortium for Python Data API Standards 已经在启动构建相应的标准。目前 PaddlePaddle 已经初步形成较完备的 API 体系。同时，AI 框架在产业落地应用时，需要能够与产业级开发语言（C++、C#、Java、Go 等）无缝衔接，也需要提供配套的编程接口与功能支持。从开发语言来看，众多已有的开发框架主要以 Python 语言的支持为主，Julia、Swift for TensorFlow 及仓颉等新的编程语言正尝试在 AI 框架领域构建 Python 之外的语言生态，

从目前看，尽管 Julia（科学计算）和 Swift（工业级开发应用）都有些特色，但是短期内还很难撼动 Python 在 AI 框架领域的地位。

AI 框架需要提供更为优质的动静态图转换能力，从而提升后端运行高效性。从开发者使用 AI 框架来实现模型训练和推理部署的角度看，AI 框架需要能够通过动态图的编程范式，来完成在模型训练的开发阶段的灵活易用的开发体验，以提升模型的开发效率；通过静态图的方式来实现模型部署时的高性能运行；同时，通过动态图转静态图的方式，来实现方便的部署和性能优化。目前，国际主流基本均已经实现动态图开发、静态图部署的编程范式，具备动静态图转换的能力，不过基于开发效率考虑，动态图与静态图的转换与统一需要持续迭代优化。

（二）全场景：AI 框架将支持端边云全场景跨平台设备部署

AI 模型需要适配部署到端边云全场景设备，对 AI 框架提出了多样化、复杂化、碎片化的挑战。随着云服务器、边缘设备、终端设备等人工智能硬件运算设备的不断涌现，以及各类人工智能运算库、中间表示工具以及编程框架的快速发展，人工智能软硬件生态呈现多样化发展趋势。但主流框架训练出来的模型却不能通用，学术科研项目间难以合作延伸，造成了 AI 框架的“碎片化”。目前业界并没有统一的中间表示层标准，导致各硬件厂商解决方案存在一定差异，以致应用模型迁移不畅，增加了应用部署难度。因此，**基于**

AI 框架训练出来的模型进行标准化互通将是未来的挑战。

AI 框架需要与硬件基础设施平台充分解耦，通过标准的硬件注册接口实现跨设备平台的快速部署。随着处理任务的复杂化、处理数据的密集化，跨架构的开发能力将会成为常态化的需求。AI 框架迫切需要开放一套可解耦的硬件注册接口，支持硬件厂商无需触碰框架核心代码即可完成适配，避免硬件厂商面对多种 AI 框架以及不同框架版本的适配代码进行维护。可解耦的硬件注册接口，需包括标准的硬件运行态管理、算子抽象定义、性能优化适配等接口，使得 AI 框架和硬件平台开发者遵从相同接口定义设备驱动、运行时以及算子和计算图等关键信息。除上述接口标准化外，还应该对模型的中间表示和算子进行标准化，硬件厂商只需基于同一种模型格式和同一套算子即可完成不同 AI 框架的适配，满足端-边-云不同业务场景同步适配的业务需求。

（三）超大规模：AI 框架将着力强化对超大规模 AI 的支持

超大规模 AI 成为新的深度学习范式。OpenAI 于 2020 年 5 月发布 GPT-3 模型，包含 1750 亿参数，数据集（处理前）达到 45T，在多项 NLP 任务中超越了人类水平。这种通过超大规模的模型参数及超大规模的数据集的 AI 大模型范式，实现了深度学习新的突破。产业界和学术界看到这种新型范式的潜力后纷纷入局，继 OpenAI 后，华为基于 MindSpore 框架发布了盘古大模型、智源发布了悟道

模型、阿里发布了 M6 模型、百度发布了文心模型等。超大规模 AI 正成为下一代人工智能的突破口，也是最有潜力的强人工智能技术。

超大规模 AI 需要大模型、大数据、大算力的三重支持，对 AI 框架也提出了新的挑战，可总结为“五堵墙”。一是内存墙，大模型训练过程中需要存储参数、激活、梯度、优化器状态，鹏程.盘古一个模型的训练就需要近 4TB 的内存。**二是算力墙**，以鹏程.盘古 2000 亿参数量的大模型为例，需要 3.6EFLOPS 的算力支持，这要求必须构建大规模的异构 AI 计算集群，才能满足这样的算力需求，同时算力平台要满足智能调度，来提升算力资源的利用率。**三是通信墙**，大模型并行切分到集群后，模型切片之间会产生大量通信，从而通信就成了主要的瓶颈。**四是调优墙**，在 E 级算力集群上训练一个千亿参数规模的，节点之间的通信关系非常复杂，要保证计算的正确性、性能和可用性，手动调试难以全面兼顾。**五是部署墙**，超大规模 AI 面临“大模型、小推理”的部署难题，需要对大模型进行完美压缩以适应推理侧的部署需求。

AI 框架将通过自动混合并行、全局内存管理、可视化调优以及分布式推理等核心技术支持超大规模 AI 发展。AI 框架可通过**多维度自动混合并行**，支持数据并行、模型并行、流水并行、优化器并行、子图并行等多种维度的 AI 并行计算技术，解决模型及集群的横向扩展问题，支持超大规模模型切分到大集群高效训练，并实现最优的计算通信比，进而提升算力的利用率。AI 框架可通过**全局内**

存管理及计算调度，实现 CPU 内存、NPU 内存和 NVMe 三层存储的统一管理，从而提升单卡的纵向扩展能力。超大规模 AI 的数据集、网络深度和宽度都非常大，AI 框架需要通过张量分析、图码结合等方式，快速定位出现精度异常的网络结构或者算子，提供方便快捷的**精度问题定位能力**，并通过可视化的方式记录并且分析开发者的调优路径和 AI 模型的精度收敛趋势，向开发者推荐**调优策略**，加速调优过程。此外，对于大模型的推理服务，AI 框架需要自动从分布式训练模式转换成**分布式推理**模式，并实现服务化封装，支持快速上线大模型服务。

（四）科学计算：AI 框架将进一步与科学计算深度融合交叉

传统科学计算领域亟需 AI 技术加持融合。科学计算一般以准确的数学模型为根基，以严谨的计算方法为手段，对应用领域中气候气象、能源材料、航空航天、生物医药等问题进行模拟。传统科学计算方法通过数值迭代的方式解决问题，面临着维度灾难引起的计算量指数上升的问题，导致在复杂问题或者场景中“算不起”，甚至是“算不动”。在科学计算的诸多领域仍旧存在着大量待求解的问题，因为机理不清楚，或是计算过于复杂，以至于传统算法难以求解。而人工智能则往往依赖于以神经网络为代表的具有“万能逼近”性质的数学工具从数据中挖掘规律，从而在图像处理等类型的任务上，实现超越人类水准的突破。

AI 框架提供了科学计算问题求解的新范式，推动科学计算与 AI 共同发展。AI 框架需构建 AI 与科学计算的统一加速引擎，支持传统数值计算的方法，并通过计算图对传统数值方法和 AI 方法的混合计算优化，从而实现 AI+科学计算端到端加速。**AI 框架需要强化自动微分功能**，通过改进框架自动微分机制和底层算子实现，支持高阶微分，使得 AI 框架具备表达复杂科学计算公式的能力。**AI 框架需丰富编程接口**，通过新增 Jacobian、Hessian、JVP、VJP 等接口，为 AI+科学计算提供融合的表达方式，使得开发者以更加接近数学计算的方式进行编程。**AI 框架需内置专业领域的科学计算套件**，面向不同的科学计算领域提供简单易用的科学计算套件，包含高质量的领域数据集、高精度的基础 AI 模型和用于前后处理的工具集合。MindSpore 内置 MindSpore Science 功能组件，并推出面向电子信息行业的 MindSpore Elec 套件和面向生命科学行业的 MindSpore SPONGE 套件。PaddlePaddle 通过扩展底层框架以及开发 PaddleScience 科学计算开发套件，具备求解科学计算问题的能力。

（五）安全可信：AI 框架将助力提升 AI 模型可解释性与鲁棒性

可解释性的需求增加对 AI 框架提出进阶性要求。通过对模型决策结果以人类可理解的方式呈现，有助于人们理解复杂模型内部的工作机理以及模型如何做出决策等重要问题。安全可信的 AI 框架需对模型可解释性进行支持，将黑盒的人工智能决策转化为可解

释的决策判断。这不仅能增加开发者对 AI 模型决策的理解与信任，也能帮助诊断出影响模型性能的因素，加以改进，进一步提升模型性能。目前已有部分框架开始支持可解释性的需求，比如基于 PyTorch 框架出现了 Captum 等可解释库支持，基于 TensorFlow 出现了 TF-explain 等库支持，以及同时支持 PyTorch 和 TensorFlow 的 AIX360、Alibi 等可解释库，国内则有 MindSpore 的 MindSpore XAI，以及 PaddlePaddle 的 InterpretDL。另外，已经有一些平台从可解释的角度出发对模型进行评测，例如启智社区的重明平台、瑞莱智慧平台等。

AI 框架需要提供丰富的 AI 鲁棒性检测工具，提升 AI 模型的鲁棒性。训练模型时样本训练不足，使得模型泛化能力不足；模型面对恶意样本时，无法给出正确的判断结果。AI 框架可通过支持网络蒸馏、对抗训练等方式，以及黑盒、白盒、灰盒测试等对抗检测技术，帮助开发者提高模型的鲁棒性。MindSpore 推出鲁棒性测试工具 MindSpore Armour，基于黑白盒对抗样本、自然扰动等技术提供高效的鲁棒性评测方案，帮助客户评估模型的鲁棒性性、识别模型脆弱点。PaddlePaddle 推出 PaddleSleeve 模型安全工具，完整提供了从 AI 模型鲁棒性评估测试，到模型攻击防御，再到模型鲁棒性提升的一整套能力。

（六）工程化：AI 框架将加速 AI 应用产业规模级工程化落地

AI 工程化是 AI 深度赋能实体经济的必经之路。工程化是人工智能技术从理论算法走向实践的基本路径，是在较为成熟的算法基础上，结合产业需求，形成可落地可实施且适宜规模化部署的工程方案。近年来越来越多的行业领域涌现出智能化应用，但其工程化落地情况尚不理想，目前仅有半数项目能够从 AI 原型转化为生产[6]。Gartner 于 2021 年 10 月发布 2022 年十二大重要战略技术趋势，再次将 AI 工程化确定为重要战略技术趋势之一，并预测到 2025 年，10%建立 AI 工程化最佳实践的企业从其人工智能工作中产生的价值，将至少比 90%未建立该实践的企业高出三倍。

AI 框架需要支持 AI 模型跨平台的快速迁移，通过模型自适应等技术实现开发者开发调试代价的最小化。不同应用场景中、不同任务中，设备的资源约束不同，对 AI 模型的精度产品化需求也不同。AI 框架需要针对不同场景或不同任务，权衡设备资源约束和精度要求，通过自动学习（AutoML）、模型轻量化（量化、剪枝等）、迁移学习等模型自适应技术对 AI 模型进行调优。迁移部署可以针对同一应用场景中的不同任务，或者不同应用场景的同一任务，避免了从零开始的再次开发，充分利用已有技术基础，实现快速部署，减少开发者开发的时间、人力等各方面成本，也便于 AI 产品的快

[6] 2021 年重要战略科技趋势研究报告，Gartner 发布.

速推广复用。

AI 框架将依托增量学习更灵活地面对动态数据训练需求，实现 AI 应用开发更快、成本更低。面对新增样本数据或新任务时，传统一次性的数据学习需要耗费大量的计算资源和时间进行重新学习，并且在新任务上训练时，在旧任务上的表现力通常会显著下降，出现“灾难性遗忘”缺陷。增量学习能力能够很好的解决上述问题，充分利用历史训练结果实现知识累积，显著减少后续训练时间的同时缓解遗忘缺陷，适用于数据库庞大或数据流应用场景。此外，AI 框架对端侧、边侧增量学习的支持，也能够优化轻量化部署效率，减少与云侧数据的交互，进一步提升训练性能。

应用工程化将推动 AI 框架向着精细化、多元化发展。AI 应用产业规模级工程化部署往往涉及云边端不同场景下的硬件设备，包括云服务器、移动终端以及 IoT 设备等。对于移动终端和 IoT 设备，由于硬件资源限制，云侧的模型和推理运行框架体积太大，无法直接部署，因此 AI 模型的压缩和端侧推理框架的轻量化成为移动终端和 IoT 设备上部署的关键。部分主流 AI 框架坚持训练推理一体化布局，推出面向移动终端和 IoT 设备的推理引擎组件，加速 AI 工程化，如 TensorFlow Lite、PyTorch Mobile、MindSpore Lite、Paddle Lite 等。此外还有专门为推理而设计的 AI 推理框架，如 NVIDIA TensorRT、Intel OpenVINO、腾讯优图 TNN、阿里 MNN 等。全行业 AI 应用有着丰富的 AI 推理需求，包括精度需求、易用需求、性

能需求等，随着 AI 工程化的不断发展，AI 推理框架生态将愈加繁荣。

四、AI 框架生态远未成熟，未来发展空间可观

AI 框架进入主流视野仅五六年时间，从技术演进，到开源生态、市场格局，再到应用赋能、推广辐射，AI 框架的整体生态还远未成熟。软硬件协同、开源打造、开发者推广、关键领域赋能等方面，将为 AI 框架生态成熟升级提供重要助力。

（一）从硬件适配向算子接口标准化演进

为应对人工智能软硬件生态面临多样化、复杂化、碎片化的挑战，亟需推进 AI 框架硬件适配、算子接口标准化工作。鼓励 AI 头部企业通过 AI 框架与底层 AI 芯片的适配逐步构建标准化硬件接口，驱动硬件厂商主动适配 AI 框架，从 AI 芯片主导适配向统一硬件接口主导适配转变。支持研制统一的 AI 算子接口标准，通过屏蔽不同的底层硬件架构细节，制订标准化的开发接口，为 AI 技术研究、软硬件研制和应用开发提供统一规范。**从标准工作切入，推进 AI 框架统一中间表示 IR 的标准化，加速 AI 框架形成支持跨平台快递迁移部署的能力，将为 AI 框架构筑起协同生态。**

（二）强化开源社区打造与开源氛围营造

着力开源开放，多措并举构建 AI 框架开源生态，营造创新良好的 AI 算法框架发展环境。建议遵循开源开放原则，联合建设开

源社区，引领各方积极参与贡献力量。鼓励有技术实力的企业构建开源生态，重点在开源算法框架、数据库、操作系统等关键基础领域创新突破；鼓励我国高校、企业、行业组织等产业各方融入国际开源社区生态，提升参与度与影响力；配套建设开源风险监测、开源生态监测等平台，强化开源生态治理意识。**通过构筑 AI 框架开源生态，为人工智能企业本身的技术创新、产品优化、应用拓展、人才引进提供持续的支持。**

（三）重视与高校科研院所广泛开放合作

引导学术界高校科研机构与产业界企业基于主流 AI 框架构建其 AI 应用系统，并在项目申报、科创资金申请等方面予以政策倾斜。通过支持和鼓励高校、科研机构和合作伙伴参与到主流 AI 框架的众筹开发中，通过共建联合实验室、创新中心等方式，开发和调优多个主流 AI 框架下的网络模型，持续补足算子和模型，不断优化算子和模型的精度与性能，培养一大批优秀开发者。鼓励 AI 头部企业与高校广泛开展合作，包括“人才培养、教材/教辅书籍、教学课程、技术合作、科研、大赛、项目孵化”等方面，支持高校建设融合 AI 框架的核心课程和数字教学资源，开展基于 AI 框架的理论教学、实验实训及 AI 技术合作项目。

（四）推进融入 AI 基础设施布局落地

AI 基础设施是以“数据资源、算法框架、算力资源”为核心能力

要素，以“开放平台”为主要赋能载体，能够长期提供公共普惠的智能化服务的基础设施。**鼓励 AI 框架主体通过融入人工智能计算中心、AI 公有云以及 AI 应用开放平台等，对外提供 AI 能力服务**，如各大中城市主导建设的人工智能计算中心，重点依托 MindSpore 等优质 AI 框架夯实底层 AI 开发能力。支持政府、企事业单位积极采购 AI 基础设施服务，逐步扩大 AI 框架辐射范围。

（五）支持深度赋能大模型及科学计算

支持 AI 框架主体钻研科学计算基础研究领域，通过深度赋能超大规模 AI、融入科学计算领域，实现 AI 框架的快速发展。超大规模 AI 是近两年 AI 持续变革的核心动力，而 AI 融合计算则是 AI 走入各学科领域的关键支柱，两者均是各国基础科研领域发力的新高地，也是新的科学装置。超大规模 AI 与 AI 融合计算向 AI 框架提出更高要求，从性能、准确性、时效、能耗等多维度，推动 AI 框架在技术上持续完善升级。同时，AI 框架需要主动融入超大规模 AI 与 AI 融合计算领域开展 AI 创新应用，从而为生态层套件/模型库提供丰富资源。

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：18613805918

传真：010-62304980

网址：www.caict.ac.cn