AI 模型优化技术专利

分析研究报告

上海市知识产权服务中心

（上海市知识产权发展研究中心）

2025 年 09 月

目录

声 明 1

摘 要 2

第一章 概述 4

1.1 研究背景与研究目的 4

1.2 AI 模型优化技术产业发展状态及政策环境 5

1.3 研究思路 10

1.4 检索策略与检索结果 10

第二章 AI 模型优化专利整体发展态势 12

2.1 专利申请趋势分析 12

2.2 区域分布及法律状态分析 13

2.3 权利人分布分析 15

2.4 技术分布分析 17

2.5 专利市场化运营 18

2.6 本章小结 19

第三章 AI 模型优化技术主要权利人分析 21

3.1 国外主要权利人 21

3.2 国内主要权利人 35

3.3 权利人对比分析 50

3.4 本章小结 52

第四章 AI 模型优化特定技术专题分析 54

4.1 AI 模型结构优化 54

4.2 AI 模型压缩 57

4.3 AI 模型参数优化 61

4.4 知识蒸馏 65

4.5 本章小结 68

声 明

1.本报告的分析基准日为 2025 年 5 月 25 日，一切分析结论均基于此基准日之前能够获取的信息与资料而得出。对于基准日之后由于法律、技术、经济环境等发生变化，可能由此引发不同的分析结论。

2.本报告形成过程中对于数据的检索、筛选、标引、分析等均依据本报告确定的分析“标的”或“事实”（分析“标的”或“事实”主要依据需求方提供的特定技术信息以及市场信息为准）及当前所适用的专利法而作出，但并不意味分析机构能够保证上述分析“标的”或“事实”的准确性与全面性，也不意味着分析机构能够保证上述检索、筛选、标引、分析等处理过程的准确性与全面性。

3.本报告在依据确定的分析“标的”或“事实”而进行创新启示建议时，不能保证也不能排除其运用于研发、生产与市场的有效性，且不能保证也不能排除与其他公开技术或者专利权利相冲突，在需求方后续具体实施中，应当再次进行技术、法律与市场层面的可行性分析。

摘 要

本项目的研究目的是：通过 AI 模型优化技术相关专利，利用聚类分析方法进行 AI 模型优化的发展趋势、区域、布局等多角度分析。明确 AI 模型优化技术的关键专利，为国内科研人员和企业提供清晰的研究方向，助力我国在 AI 模型优化技术在关键领域的自主创新和应用突破，进一步增强我国在 AI 领域的核心竞争力。

基于上述目的，本报告收集技术关键词、分类号、专利权人及其同位语组合形成的十余组检索式在全球范围内对 AI 模型优化特定关键技术进行了专利检索，经过人工筛选和去重后得到 11305 族/15894 件相关专利,其中知识蒸馏技术 3832 族；模型结构优化 3041 族；参数优化 2993 族；模型压缩 1439 族。

基于研究发现：

（1）从趋势来看，自 2017 年 Transformer 架构问世，奠定自然语言处理（NLP）新范式，催生 BERT、GPT 等大模型，随之 AI 模型优化专利申请开始逐年走高，在这一阶段中国权利人以研发机构为代表的各大高校和以百度、腾讯等互联网企业也加入到了 AI 模型的研发当中。在可以预见的未来几年中，关于 AI 模型相关专利的专利申请也将继续维持高速增长的趋势。

（2）在 AI 模型优化技术领域内，排名前三的专利权人专利数合计超过了 1500 件。其中腾讯排名第一，有 618 件专利；其次是华为有 553 件专利申请；三星电子排名第三有 496 件专利申请。整体来看，上述权利人合计专利公开为 6467 件，占总数的一半左右，头部企业聚集效益明显。整体来看，在 AI 模型优化技术领域内，主要以中国和美国权利人为主。

（3）从 AI 模型优化技术专利布局来看，四位核心权利人各有侧重。三星电子主要关注知识蒸馏技术，例如利用训练冷凝器模型以学习预先训练的教师模型和学生模型之间的参数映射函数优化模型等；IBM 则更为重视参数优化技术，涉及超参数和正则化技术，在正则化技术上，使用三元组损失正则化对样本池细化迭代。在超参数上，通过超参数分配器，利用张量交换和虚拟加速器管理有限计算资源等；

华为、腾讯和三星电子同样关注知识蒸馏技术，其中华为主要使用教师模型的无偏数据或者是中间层和输出层的输出数据改善学生模型；而腾讯则主要利用权重矩阵或者与预训练改善知识蒸馏。

建议围绕以下方向进行改进：（1）聚焦核心技术突破：加强对知识蒸馏技术深化研究；关注模型压缩与模型结构优化；推动参数优化自动化工具开发。（2）加速产学研技术转化：建立联合实验室或者设立科技转换基金；推动国内标准生态建设。（3）优化专利布局与风险防控：推动头部企业如腾讯、华为等重点在欧美地区申请核心技术专利；关注重要参与者的失效专利，通过专利挖掘获取免费技术资源进行二次开发；联合建立专利预警机制，构建通用的 AI 模型优化技术专利数据库。（4）政策与资源支持：设立专项扶持基金；建设公共算力服务平台等。

第一章 概述

1.1 研究背景与研究目的

1.1.1 研究背景

大模型自出现至今其核心能力一直在提升，尤其是在近五年更呈现出加速创新的趋势。譬如，2022 年算法创新，Chatgpt 发布，引发 Transformer 架构的风潮迭起；2023 年数据合成、数据标注成为高质量数据集建设的热点方向；2024 年算法创新、算力运营等产业诞生；2025 年 Deepseek 通过重新设计 AI Infra，优化算法架构显著提高算力利用率，打破了算力至上的传统认知。

技术的不断发展推动了 AI 大模型在感知、认知、决策、学习、执行等多个方向的应用，并且因其在各种任务中的出色表现而引起了学术界和工业界广泛的关注。AI 大模型的快速发展带来了智能化升级的核心支撑。其中，通用大模型侧重发展通识能力，垂直大模型侧重发展专业能力，模型赋能行业有效提升了效率、降低成本及优化决策过程。从产业规模看，全球人工智能快速增长。2023 年全球人工智能市场收入达 5381 亿美元，同比增长 18.5%，到 2026 年市场规模将达 9000 亿美元。从竞争格局来看，全球人工智能呈现“中美主导”的格局，尤其是 Deepseek 的出现成为了特朗普制定人工智能政策的重要参考，预示着 AI 大模型进入 G2 竞争时代。目前可以预知的手段，主要有限制英伟达芯片的对华出口，并加强新加坡等转运国的调查，同时控制中国大模型的应用，Deepseek 已被爱尔兰、意大利等美国追随国下架。

AI 大模型技术创新一直围绕算力、数据、算法三个核心要素动态循环。在算力、数据被美国限制的不利条件下，中国如何通过算法模型优化逐渐成为提升 AI 大模型性能的重要策略。通过对 AI 大模型优化相关技术的分析研究，可以深入了解该领域的技术发展趋势和竞争态势，为国内相关产业的发展提供技术指导和决策依据，帮助国内企业更好地把握市场机遇，提升产业竞争力，引导和扶持国内 AI 大模型相关产业健康发展。

1.1.2 研究目的

本报告以专利文献为主要研究对象，通过综合 AI 模型优化技术特征关键字及其同位语、国际技术分类号、主要专利权利主体及其同位语、申请人所在区域等方式，获取全球的专利数据。利用聚类分析方法进行 AI 模型优化的发展趋势、区域、布局等多角度分析。以专利分析为明线，产业分析为暗线。明确 AI 模型优化技术的关键专利和技术路线图，为国内科研人员和企业提供清晰的研究方向，提升企业在该领域的技术水平和创新能力的同时，助力我国在 AI 模型优化技术在关键领域的自主创新和应用突破，进一步增强我国在 AI 领域的核心竞争力。

1.2 AI 模型优化技术产业发展状态及政策环境

1.2.1 AI 模型优化技术概述

模型优化不仅涉及提升算法准确性的策略，还包括降低计算资源消耗及提升应用场景适应性的多种方法。一方面，合理的参数调整可以有效减少训练时间并提高预测精度；另一方面，性能优越的结构设计能够使得算法在多种条件下展现出更好的适应能力。有效的模型优化策略不仅限于算法本身，还涉及到多种因素，包括数据质量、特征选择、训练过程等。通过对这些方面进行系统性分析与调整，可以最大限度地发挥出算法的潜力。以下分别对主要的 AI 模型优化技术进行简要说明。

（一）模型结构优化：

优化模型结构不仅可以提高算法的准确性，还能增强模型的泛化能力，使其在面对新数据时仍能保持良好的表现。一种常见的方法是使用层次化结构，通过将模型划分为多个层级，可以更加高效地捕捉数据中的特征。例如，卷积神经网络（CNN）通过多个卷积层和池化层组合，能够有效提取图像信息，而循环神经网络（RNN）则适用于处理时序数据。此外，采用残差网络（ResNet）等技术也可以有效提升模型深度而不引起过拟合。残差连接通过引入捷径连接，使得深层网络能够更好地传递梯度信息，从而促进训练速度和准确性。

另一种常见的方法是剪枝与量化，由于深度学习网络模型从卷积层到全连接层存在着大量冗余参数，大量神经元激活值趋近于 0，将这些神经元去除后可以表现

出通用的模型表达能力，这种情况被称为过参数化，而对应的技术称为模型剪枝。通过去除冗余节点或降低权重精度来减小模型规模，提高推理速度。主要应用在边缘计算或资源有限的环境中，因为这类环境对计算效率和存储能力有较高要求。

（二）模型压缩：

模型压缩，就是在尽可能不改变模型效果的条件下，通过减少模型的体积，使得模型在使用的时候有更快的速度。常见的模型压缩主要有涉及模型量化、模型共享权重等。其中，模型量化即将网络的权值、激活值等从高精度转换为低精度的操作过程，例如将 32 位浮点数转化为 8 位整型数 int8，压缩到 int8 意味着内存的节省，节约了 3/4，同时也提升了计算效率，因为在 GPU 这样的硬件上，低位的浮点计算速度会远远高于高位浮点计算速度。而模型共享权重则是指模型在构建的过程中是否有些局部的信息在全局是多次出现并重复使用。而重复结构意味着卷积核在计算权重的时候会有部分权重的相似性，这些相似性能进行全局可共享。那么通过聚类的方式挖掘出这些可以共享的权重系数，并且以类别的方式让它们共享一些权重，就可以实现模型的压缩。

（三）参数优化：

在深度学习和人工智能领域，通过参数调整进行模型的性能优化是一个持续的研究课题。超参数化模型与正则化是其中两个关键的优化手段，它们对于提升 AI 模型的性能起着至关重要的作用。

首先，超参数是模型架构之外，需要通过经验或搜索来确定的参数。它们通常在模型训练之前设定，而不是通过学习过程得到。常见的超参数包括学习率、批量大小、迭代次数、优化器类型等。超参数的设置直接影响到模型的收敛速度、准确性和泛化能力。适当的超参数调整可以使模型更快地收敛到最优解，提高模型的性能。譬如，学习率是最为关键的，它决定了模型每次更新权重时的步伐。学习率过高可能导致模型震荡，无法收敛；学习率过低则可能导致训练过程缓慢，收敛速度慢。目前常见的超参数调整方法主要有经验设定、网络搜索、随机搜索和贝叶斯优化四种。

而正则化则是一种防止模型过拟合的技术，通过在损失函数中添加正则化项来约束模型参数，从而降低模型复杂度。常见的正则化方法主要有 L1 正则化（Lasso）：通过在损失函数中添加绝对值项，使模型参数向零值逼近；L2 正则化（Ridge）：通过在损失函数中添加平方项，使模型参数向较小值逼近；Dropout：在训练过程中随机丢弃部分神经元，降低模型复杂度。

在实际应用中，超参数化模型与正则化通常结合使用，以获得更好的模型性能。

（四）知识蒸馏：

大模型往往是单个复杂网络或者若干网络的集合，拥有良好的性能和泛化能力，而小模型因为网络规模较小，表达能力有限。利用大模型学习到的知识去指导小模型训练，使得小模型具有与大模型相当的性能，但参数数量大幅降低，从而可以实现模型压缩与加速，就是知识蒸馏与迁移学习在模型优化中的应用。其核心思想是一旦复杂网络模型训练完成，便可以用另一种训练方法从复杂模型中提取出来更小的模型，因此知识蒸馏框架通常包含了一个大模型和一个小模型。

1.2.2 AI 模型产业发展情况

人工智能大模型的全球化发展正以前所未有的速度重塑技术、经济与社会结构。大模型是当前科技发展的制高点，也是中美科技竞争的焦点。

其中美国企业 OpenAI 作为本轮人工智能革命的领军者，其早期研发团队是 Transformer 架构首创者和 Scaling Law 的坚定信奉者及实践先驱。自 2022 年 11 月推出现象级产品 ChatGPT 以来，该企业通过持续突破不断重塑行业认知，展现通用人工智能的早期雏形。其技术迭代在 2024 年迎来爆发期，视频生成模型 Sora 突破动态语义理解瓶颈，开启 AI 内容创作新纪元；下半年连续推出 o 系列推理模型，通过分层认知架构实现复杂决策推理；2025 年初推出支持复杂研究任务的 Deep Research 功能，发布的 GPT-4.5 更号称在认知维度实现飞跃，其万亿级参数模型不仅构建起迄今为止人类知识最完整的数字镜像，更在情感计算与共情交互领域取得突破。除了 OpenAI，美国的 Anthropic、Google、Meta 以及 xAI 等企业都在大模型

领域持续发力，逐渐形成了“OpenAI 领跑，Anthropic、Google 和 xAI 追赶，Meta 开源"的竞争态势。

中国大模型发展经历了三个主要阶段。初期探索阶段(2000 年代初至 2010 年代初)主要集中在基础 AI 理论和小规模应用的研究，缺乏大规模数据和高性能计算资源。随着 2012 年深度学习的兴起，进入了快速发展阶段(2010 年代初至 2020 年初)，期间国内大企业如阿里巴巴、百度、腾讯和华为开始大力投资 AI 技术，推动了大模型技术的发展和应用。到了成熟与应用阶段(2020 年初至今)，中国在自然语言处理和计算机视觉等领域的大模型技术已达到国际先进水平，模型广泛应用于商业、医疗、法律等多个行业。其中 2023 年是关键转折点，各大科技公司和初创企业开始推出自己的大语言模型，主要参与者包括百度、阿里、字节、腾讯、商汤、华为等。

1.2.3 AI 模型技术产业发展政策环境

近年来，中国 AI 大模型行业受到各级政府的高度重视和国家产业政策的重点支持。国家陆续出台了多项政策，鼓励 AI 大模型行业发展与创新，《信息化标准建设行动计划(2024—2027 年)》等产业政策为 AI 大模型行业的发展提供了明确、广阔的市场前景，为企业提供了良好的生产经营环境。各项支持政策已推动 AI 大模型的广泛应用，预期会推动 AI 大模型市场。具体情况列示如下：

表 1-1 大模型发展政策汇总

政策名称	颁布日期	颁布主体	政策要点
《人工智能生成合成内容标识方法》	2025 年 3 月	工信部	要求平台严格审核 AI 生成内容标识，市场监督总局发布《网络安全技术人工智能生成合成内容标识方法》国家标准，规范显隐双标识管理，《办法》将在 2025 年 9 月 1 日实施
《中小企业数字化赋能专项行动方案（2025-2027 年）》	2024 年 12 月	工信部、财政部、金融股监督局	鼓励龙头企业、交易机构、平台企业、数据服务企业等经营主体建设公共数据集、行业数据集，为中小企业提供用于人工智能模型训练的高质量数据。建设一批适用于中小企业的垂直行业大模型，强化中小企业大模型技术供给。
《信息化标准建设行动计划（2024-2027 年）》	2024 年 5 月	中央网信办、市场监督总局、工业和信	布局新兴技术领域标准，完善人工智能标准，强化通用性、基础性、伦理性、安全、隐私等标准研制，加快推进大模型、生成式人工智能标准研

		息化部	制。
《关于深化智慧城市发展发展推进城市全域数字化转型的指导意见》	2024 年 5 月	国家发展改革委等部门	AI 大模型可在智慧城市建设的智能决策、数据分析等方面发挥关键作用。
《加快数字人才培育支撑数字经济发展行动方案（2024-2026 年）》	2024 年 4 月	人力资源社会保障部等多部门	为 AI 大模型产业发展提供人才支持。
《“数据要素 x”三年行动计划（2024-2026 年）》	2024 年 1 月	国家数据局等 17 个部门	进一步明确利用数据要素推动各领域发展，其中包含大模型开发的数据支持等相关内容，助力大模型产业发展以及在多领域的应用。
《上海市推动人工智能大模型创新发展若干措施（2023-2025）》	2023 年 11 月	上海市经信委等五部门	实施大模型创新扶持计划，支持引进高水平创新企业，打造具有国际竞争力的大模型，鼓励形成数据飞轮，加速模型迭代，对重大成果予以专项奖励，支持前沿研究和研发下一代模型架构及训练方法，在市级专项中重点支持大模型创新。
《生成式人工智能服务管理暂行办法》	2023 年 7 月	国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局	实行包含审慎和分类分级监管，推动生成式人工智能基础设施和公共训练数据资源平台建设，鼓励基础技术自主创新，明确训练数据处理活动和数据标注等要求，对提供者和使用者的违法违规行为处置做出规定。
中央财经委会议	2023 年 5 月	中央财经委	提出“要把握人工智能等科技革命浪潮”，明确人工智能在新科技革命浪潮中的地位，引导各方重视人工智能发展。
中共中央政治局会议	2023 年 4 月	中共中央政治局	提出“要重视通用人工智能发展，营造创新生态，重视防范风险”，从宏观层面为通用人工智能指明方向，强调创新与风险防范。
《关于支持建设新一代人工智能示范应用场景的通知》	2022 年 8 月	科技部	面向世界科技前沿、经济主战场、国家重大需求、人民生命健康，发挥人工智能赋能经济社会发展的作用，围绕构建全链条、全过程的人工智能行业应用生态，支持一批基础较好的人工智能应用场景，加强研发上下游配合与新技术集成，打造可复制、可推广的标杆型示范应用场景，首批支持建设上个示范应用场景。
《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	2022 年 7 月	科技部等六部门	以促进人工智能与实体经济深度融合为主线，强化主体培育、加大应用示范等，加速人工智能技术攻关。围绕高端高效智能经济、安全便捷智能社会建设等打造重大场景，提升企业、高效院所等的场景创新能力，推动人工智能场景开放，鼓励发布场景清单、举办场景活动等，以场景创新

			推动人工智能技术升级和产业增长。

1.3 研究思路

1.3.1 分析方法

本项目采用专利文献研究、产业研究相结合的方式进行研究。

专利文献研究通过综合 AI 模型优化的技术特征关键字及其同位语、国际技术分类号、主要专利权利主体及其同位语、申请人所在区域等方式，通过构建检索式在专业专利数据库中获取全球的专利数据。主要采用聚类分析方法，开展趋势、区域、布局和核心专利等分析。

1.3.2 分析内容

本报告分析内容具体而言，从宏观整体态势到中观关键参与者，再到微观技术分析及应用分析的脉络展开分析。

首先，通过分析全球范围内 AI 模型优化相关技术的整体专利申请趋势、技术区域分布、专利主体类型、专利同族及引证等关键信息，揭示该技术的发展脉络与未来走向。

其次，报告聚焦核心权利人，进行具体分析，包括申请趋势、区域分布、法律状态、技术发展动态，来剖析其技术布局与创新实力，助力企业精准把握市场竞争格局。

同时，报告针对 AI 模型剪枝、蒸馏等核心技术进行专题探讨，通过解读关键专利方式，了解不同技术问题背景下各创新主体是如何进行技术攻关，以此来推动技术深化与创新，进一步为研发提供创新思路。

1.4 检索策略与检索结果

1.4.1 检索策略

本项目研究对象为：AI 模型优化技术相关专利

（1）检索地区的选择：根据本项目的实施方法和实施路径，确定本项目的专利数据分析范围为：全球。

（2）检索方式的选择：采用 IPC 分类号，辅以主要权利人等权利人进行检索。

（3）分类号的确认：

a）尝试性检索

利用 AI 模型优化技术相关市场主体（谷歌、微软、英伟达等）作为关键词，在专利网站和数据平台上进行检索，观察并记录检索结果，进而形成 AI 模型优化相关 IPC 分类号表初稿。

b）修正检索

根据与 AI 模型优化产业的相关性对 IPC 分类号和关键词表的初稿进行修正，并进行梳理分类，进而形成最终 IPC 分类号表及最终关键词表。

c）最终检索

使用最终确定的分类号确认表及关键词表，结合重点专利权人进行组合检索

1.4.2 检索结果

本报告检索范围：截止 2025 年 5 月 25 日，AI 模型优化特定技术在全球公开/公告之专利申请。

本报告检索结果：基于本项目确定的关键字词组制定检索(附件一)，检索到两万余族相关专利，经过人工筛选和去重后得到 11305 族/15894 件相关专利，其中知识蒸馏技术 3832 族；模型结构优化 3041 族；参数优化 2993 族；模型压缩 1439 族。

第二章 AI 模型优化专利整体发展态势

本章基于人工筛选所得的 15894 件与项目技术密切相关的 AI 模型优化技术相关的专利进行统计分析，以宏观观察 AI 模型优化技术，获取总体布局信息，重点研究了专利申请趋势、技术区域部署分析、专利权人排名和研发能力比较、主要专利权人分析、专利发明人分析、专利同族数分析及被引证数分析。

2.1 专利申请趋势分析

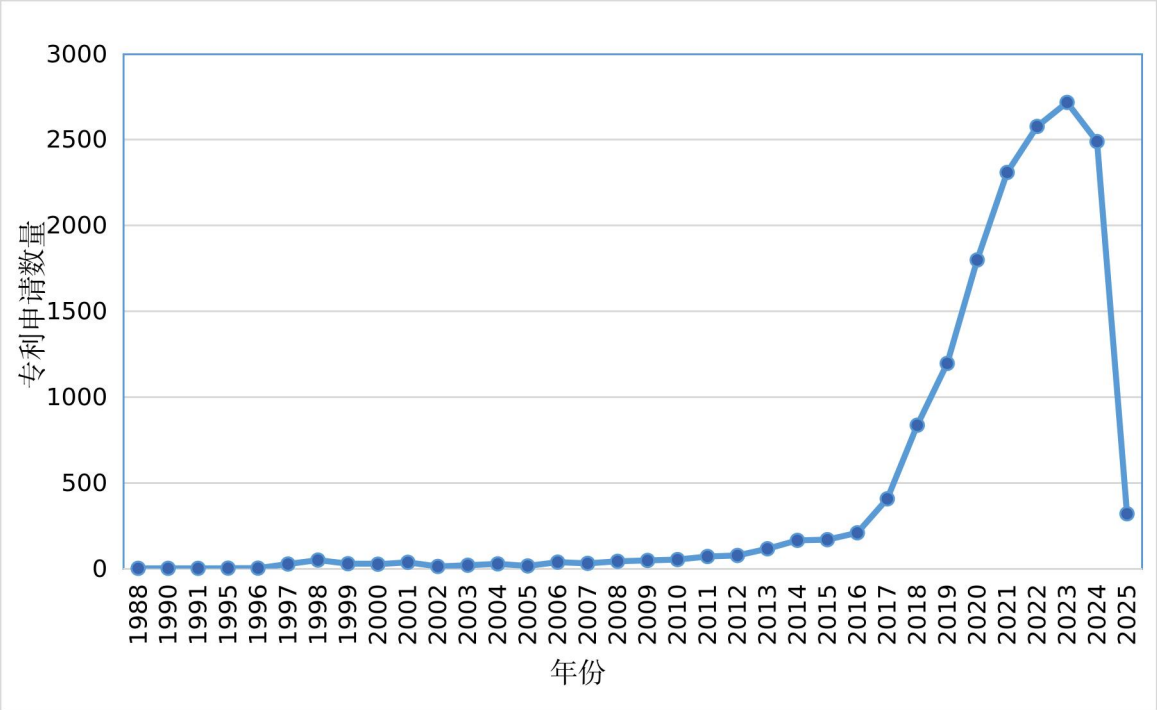


图 2-1 专利申请趋势

图 2-1 显示了 AI 模型优化技术在时间轴上的专利申请量发展态势。可见，专利总体申请量大致经历了两个阶段，即萌芽期（1988 年～2015 年），快速发展期（2016～至今）。

（1）萌芽期（1988—2015 年）

标志着 AI 成为独立学科的概念最早可以追溯到 1956 年达特茅斯会议召开。约翰麦·卡锡、马文·明斯基等学者首次提出“人工智能”术语。而从专利申请来看，AI 模型优化的专利申请最早可以追溯到 20 世纪 90 年代左右。早期由于技术突破困难和成本高昂，AI 模型优化专利的申请人主要为 IBM 以及 RICHO CORP 等国际巨

头，从专利技术来看多涉及机器学习和神经网络训练，譬如 RICHO CORP 最早在 1995 年申请了公开号为 US5636326A 的专利，涉及一种神经网络剪枝方法。在最初的十余年间，AI 模型优化技术专利申请数量一直维持低位，年均专利申请量均在 10 件左右。

（2）快速发展期（2016～至今）

2016 年 DeepMind 的 AlphaGo 击败围棋世界冠军李世石，强化学习与蒙特卡洛树搜索结合展现出通用智能潜力。深度学习开始从实验室走向大众生活。尤其是 2017 年 Transformer 架构问世，奠定自然语言处理（NLP）新范式，催生 BERT、GPT 等大模型。结合专利申请来看，自 2016 年开始 AI 模型优化专利申请开始逐年走高。在这一阶段中国权利人以研发机构为代表的各大高校和以百度、腾讯等互联网企业也加入到了 AI 模型的研发当中。

在最近的五年内，AI 模型优化专利申请开始了飞速发展，到 2023 年达到峰值 2700 余件，主要得益于生成式 AI 技术的商业化落地。2020 年 OpenAI 发布 GPT-3 以来，大模型性能逐年上升，成本逐年降低。在 2024 年 DeepSeek-R1 与 Janus-Pro 突破视觉-语言联合推理，再一次推动机器人、医疗诊断等领域发展。在可以预见的未来几年中，关于 AI 模型相关领域的专利申请也将继续维持高速增长的趋势。

2.2 区域分布及法律状态分析

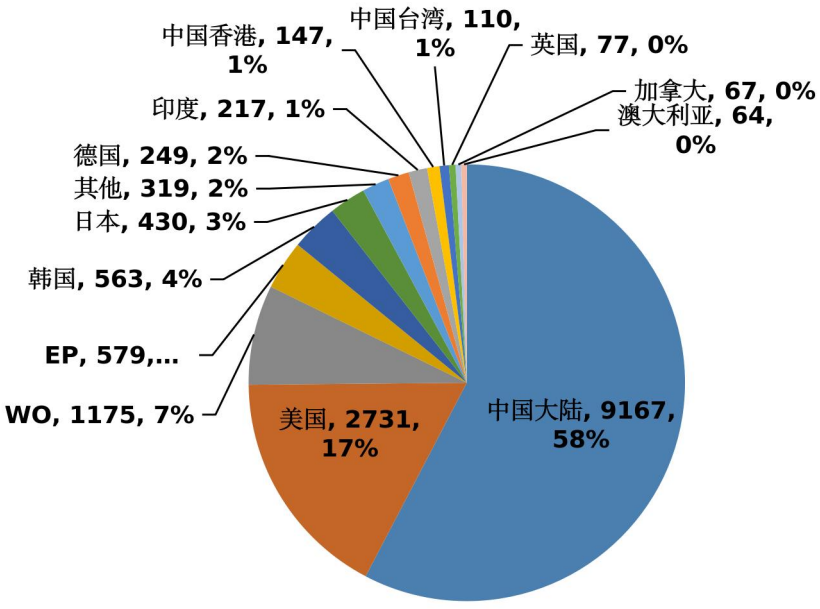


图 2-2 AI 模型优化专利区域分布

图 2-2 显示了 AI 模型优化专利的地域分布情况。从区域分布来看，目前的 AI 模型优化专利主要集中在中国大陆，合计公开 9167 件，占比 58%。其次是美国，申请量为 2731 件，占比接近 17%。整体来看，中国大陆和美国是 AI 模型优化的主要布局区域。除这两个地区外，欧洲、韩国、日本等区域是 AI 模型优化技术次一级布局区域。

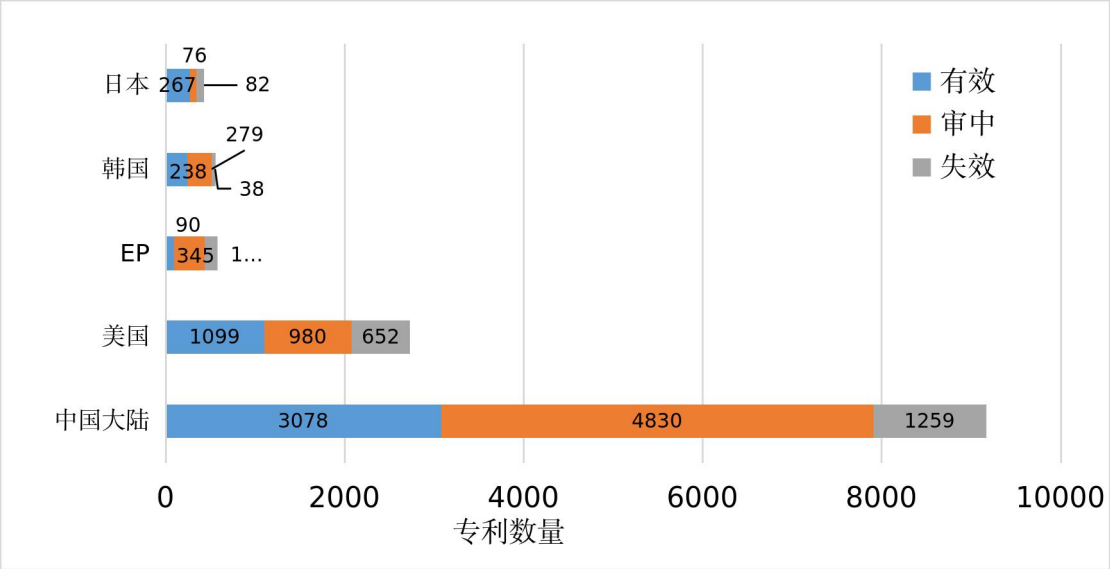


图 2-3 AI 模型优化技术简单法律状态分布图

结合专利法律状态来看，中国大陆地区模型优化专利技术处于快速发展阶段，专利在审数量超过了授权发明。而美国则与中国大陆地区不同，有效专利数量超过了在审专利，可以预见在未来几年内，中国大陆的有效 AI 模型优化专利也将快速增加。除中美外，日本、韩国和欧洲也各有不同。其中日本在审专利较少，技术活力不足；欧洲（EP）有效专利较少，主要为其 AI 模型优化专利多为德国专利公开；韩国相较于日本和欧洲在 AI 模型优化领域的技术积累和技术发展活力都较强。

2.3 权利人分布分析

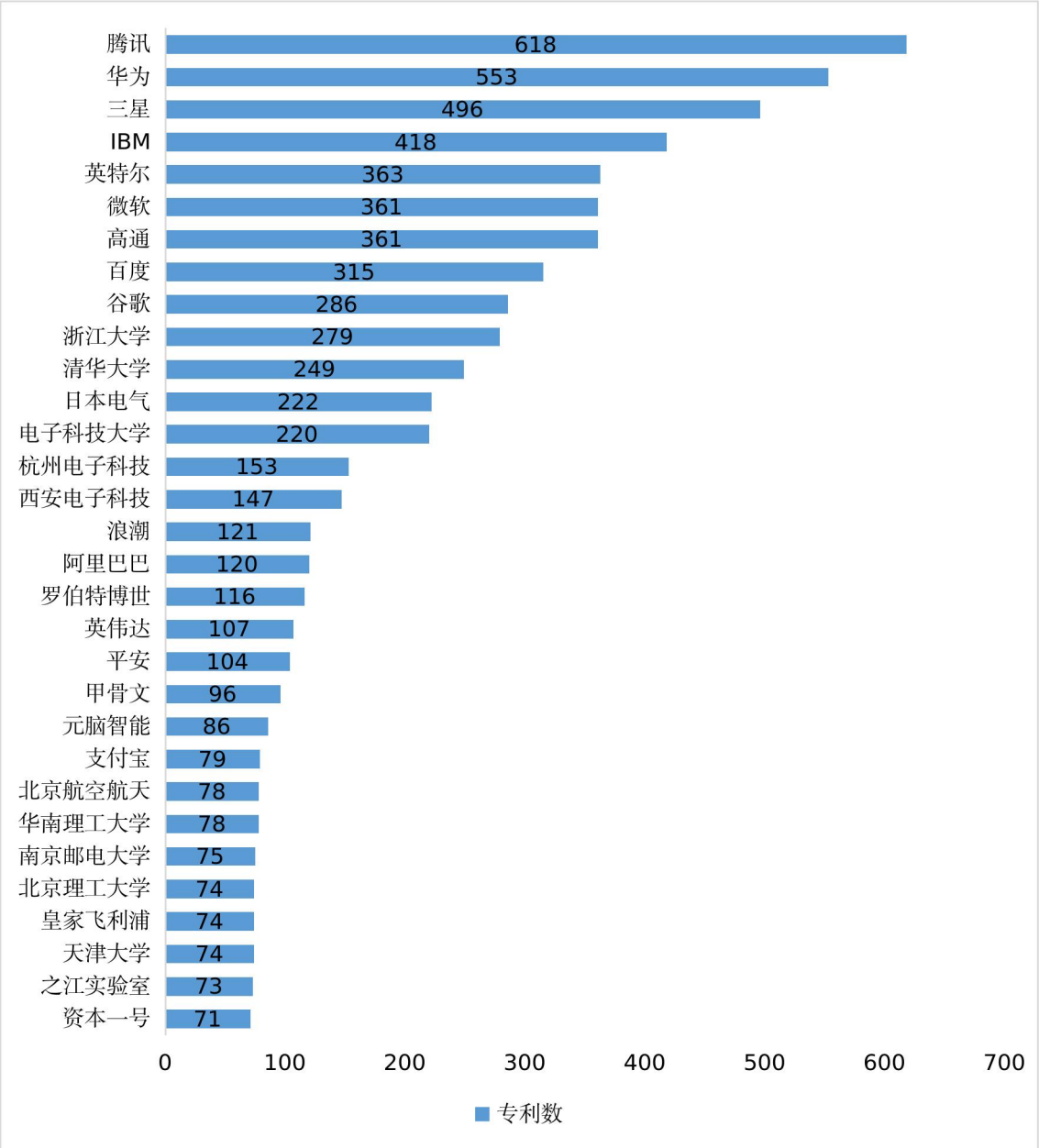


图 2-4 AI 模型优化技术主要专利权人

图 2-4 是 AI 模型优化技术专利公开数量在 70 件以上的专利权人情况。通过对专利权人的分析，可以更清楚 AI 模型优化技术的专利研发格局。

从图中数据来看，排名前三的专利权人专利数合计超过了 1500 件。其中腾讯排名第一，有 618 件专利；其次是华为有 553 件专利申请；三星电子排名第三有 496 件专利申请。整体来看，上述权利人合计专利公开为 6467 件，占总数的一半左右，头部企业聚集效益明显。

在 AI 模型优化技术领域内，主要以中国和美国权利人为主。从中美权利人来看，美国专利权人以企业为主，譬如：IBM、英特尔和微软等，表明其在 AI 模型优化技术在商业化上已取得一定成果。而中国专利权人则主要以大学和研发机构为主。以大学来看譬如浙江大学、清华大学等院校；以研发机构来看譬如之江实验室等研究机构。此外才是腾讯和华为等企业。表明中国的 AI 模型优化技术处于研发和商业化的转向阶段。

表 2-1 进一步展示了专利申请量排名前二十的专利权人的专利申请概况，从专利布局区域来看，中美是主要权利人的核心布局市场。其中，美国权利人对专利保护范围更广，在多个国家进行布局。而中国权利人主要在中美两国进行专利布局。从专利有效数量来看，腾讯、三星、百度以及英特尔等企业有效专利数量较多。

表 2-1 主要专利权人概况

排名	申请人	国家/地区	总申请量	有效专利数	区域分布	TOP 布局国家
1	腾讯	中国	618	286	11	CN、HK、US、EP、JP
2	华为	中国	553	71	19	CN、US、EP
3	三星	韩国	496	116	14	US、KR、CN、EP
4	IBM	美国	418	123	16	US、JP、CN、UK、、DE
5	英特尔	美国	363	118	22	US、CN、EP、DE、TW
6	高通	美国	361	87	34	US、CN、EP、IN、KR
7	微软	美国	361	88	14	US、EP、CN、IN
8	百度	中国	315	151	7	CN、US、JP、KR
9	谷歌	美国	286	86	12	US、CN、EP、IN、DE
10	浙江大学	中国	279	97	5	CN、US
11	清华大学	中国	249	84	5	CN、US
12	日本电气	日本	222	96	8	US、JP
13	电子科技大学	中国	220	66	3	CN
14	杭州电子科技大学	中国	153	52	3	CN
15	西安电子科技大学	中国	147	55	3	CN
16	浪潮	中国	121	34	2	CN
17	阿里巴巴	中国	120	38	8	CN、HK、US

18	罗伯特博世	德国	116	22	8	DE、CN、US
19	英伟达	美国	107	16	6	US、CN、DE
20	平安	中国	104	35	4	CN

2.4 技术分布分析

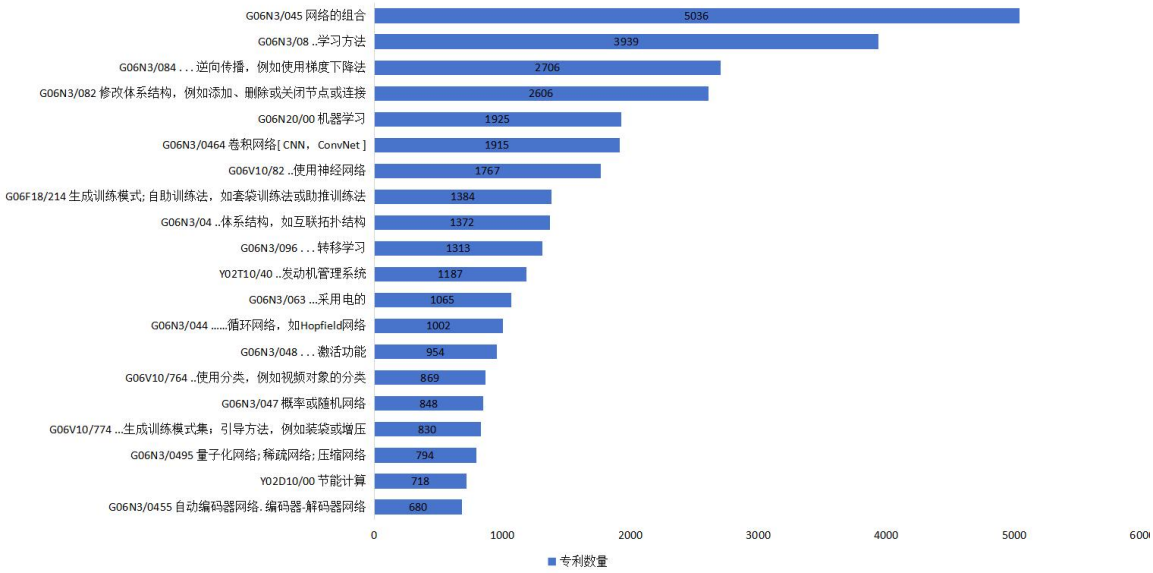


图 2-5 AI 模型优化技术 CPC 分布分析

图 2-5 显示了 AI 模型优化技术的 CPC 分布情况。从专利数量来看，G06N3/045（网络的组合）专利申请量最多，合计 5036 件；其次为 G06N3/08（学习方法）次之排名第二，专利数量为 3939 件；G06N3/084（逆向传播），合计 2706 件。整体来看，AI 模型优化专利集中在 G06N3（基于生物模型的计算装置）大类。

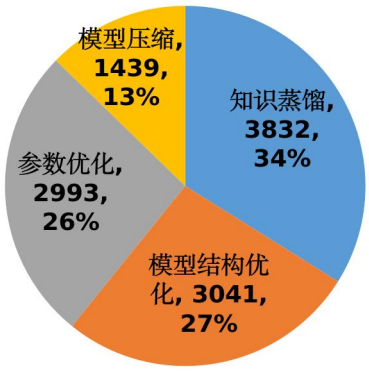


图 2-6 技术分布分析

图 2-6 显示了 AI 模型优化的技术分布情况。从标引来看，AI 模型优化技术主要分布在四个方向，其中知识蒸馏技术合计 3832 族，占比 34%；模型结构优化技术合计 3041 族，占比 27%；参数优化技术合计 2993 族，占比 26%；此外，模型压缩技术合计 1439 族，占比 13%。可见知识蒸馏是目前模型优化技术较为主要的布局方向。

2.5 专利市场化运营

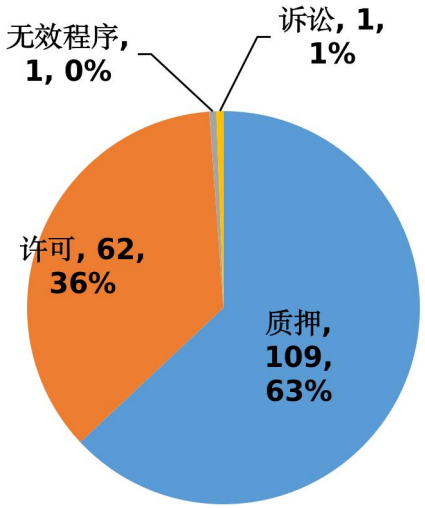


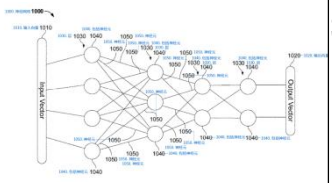
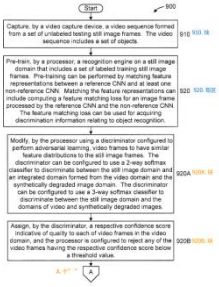
图 2-7 AI 模型优化专利市场化运营情况

图 2-7 显示了 AI 模型优化技术相关专利市场化运营情况。从类别上来看，主要涉及四种，分别是专利质押、许可、诉讼以及无效程序。

从数量上看，109 件专利进行了质押，占专利市场运营的 63%，这些专利中多数多美国权利人质押，如尼尔森、格罗方德等。专利许可则合计 62 件，占专利市场运营的 36%，除 IBM 等美国权利人外，中国权利人之间许可主要以大学如清华大学等院校为主。

专利诉讼和专利无效程序则分别为 1 件。其中专利诉讼为日本电气公开号为 US10706336B2 的美国专利，涉及图像视频领域 CNN 知识蒸馏；而专利无效程序为 COMET ML INC 公司公开号为 US11650968B2 的美国专利，涉及神经网络领域的超参数技术。

表 2-2 专利市场化运营

技术分类	功效	公开（公告）号	摘要附图	标题	市场运营情况
神经网络领域超参数	提高训练效率	US11650968B2		Systems and methods for predictive early stopping in neural network training	无效程序
CNN 知识蒸馏	提高识别准确率	US10706336B2		Recognition in unlabeled videos with domain adversarial learning and knowledge distillation	专利诉讼

2.6 本章小结

AI 模型优化技术的发展以 2016 年为界，分为萌芽期（1988 年～2015 年）和快速发展期（2016～至今）。

从发展来看，AI 模型优化的专利申请最早可以追溯到 20 世纪 90 年代左右，早期由于技术突破困难和成本高昂，AI 模型优化专利的申请人主要为 IBM 以及 RICHO CORP 等国际巨头，在最初的十余年间，AI 模型优化技术专利申请数量一直维持低位，年均专利申请量均在 10 件左右。2017 年 Transformer 架构问世，奠定自然语言处理（NLP）新范式，催生 BERT、GPT 等大模型，随之 AI 模型优化专利申请开始逐年走高，在这一阶段中国权利人以研发机构为代表的各大高校和以百度、腾讯等互联网企业也加入到了 AI 模型的研发当中。在可以预见的未来几年中，关于 AI 模型相关专利的专利申请也将继续维持高速增长的趋势。

从专利区域分布来看，目前的 AI 模型优化专利主要集中在中国大陆，合计公开 9167 件，占比 58%。其次是美国，申请量为 2731 件，占比接近 17%。整体来看，中国大陆和美国是 AI 模型优化的主要布局区域。除这两个地区外，欧洲、韩国、日本等区域是 AI 模型优化技术次一级布局区域。结合专利法律状态来看，未来几年内中国大陆的有效 AI 模型优化专利也将快速增加，而美国专利在审专利相对中国较少。

在 AI 模型优化技术领域内，排名前三的专利权人专利数合计超过了 1500 件。其中腾讯排名第一，有 618 件专利；其次是华为有 553 件专利申请；三星电子排名第三有 496 件专利申请。整体来看，上述权利人合计专利公开为 6467 件，占总数的一半左右，头部企业聚集效益明显。在 AI 模型优化技术领域内，主要以中国和美国权利人为主。从中美权利人来看，美国专利权人以企业为主，譬如：IBM、英特尔和微软等，表明其在 AI 模型优化技术在商业化上已取得一定成果。而中国专利权人则主要以大学和研发机构为主。以大学来看譬如浙江大学、清华大学等院校；以研发机构来看譬如之江实验室等研究机构。此外才是腾讯和华为等企业。表明中国的 AI 模型优化技术处于研发和商业化的转向阶段。

第三章 AI 模型优化技术主要权利人分析

权利人是专利申请的主体，也是技术发展的主要推动力量，重要权利人更是研发和创新的领先力量和领先指标，是市场竞争的主体。通过对重点权利人的研究和分析，可以掌握整个市场竞争格局，获取相关公司发展战略、专利布局特征、专利战略和研发方向等。

基于前一章对国内外重要权利人专利申请量的排名情况。特分别选取两位国内外重要权利人，具体分析包括公司概况、专利申请趋势分析、专利申请区域分析、法律状态分析和技术发展路线分析等，以期获得其在 AI 模型优化技术领域的整体发展态势和技术发展路径。其中，国外主要权利人选取的是三星、IBM 和微软这三家专利申请总量较多的企业。而国内则选取腾讯、华为和百度。

3.1 国外主要权利人

3.1.1 三星电子

3.1.1.1 公司简介

三星电子（Samsung Electronics），成立于 1969 年，主要生产消费电子产品和移动通信设备。三星电子的产品范围广泛，包括智能手机、平板电脑、电视、电脑、家电、半导体和显示器等。

三星是将 AI 大模型作为其重点投资领域，持续增加在 AI 算法研究、芯片设计、软件开发和人才引进上的资金。在全球（包括韩国、美国、加拿大、英国、俄罗斯等）设立了多个 AI 研究中心，专注于机器学习、深度学习、计算机视觉、自然语言处理等和应用研究。

三星电子对 AI 的投入是全栈式、深度融入其核心业务的。其发展策略清晰，通过自研 NPU、AI 加速芯片等为 AI 计算提供强大的底层支持。将 AI 尤其是生成式 AI 深度集成到消费电子产品中，其中以 Galaxy AI 是典范。利用 Bixby 和 SmartThings，打造跨设备的 AI 互联体验。近年，其大力投入自研大模型 Samsung

Gauss，以掌握核心 AI 技术，同时在半导体、显示等自身优势制造领域利用 AI 提质增效。

综上所述，三星电子将 AI 视为其未来增长的基石，正通过大规模、系统性的投入，力图在 AI 硬件、软件平台、终端应用和基础研究等多个层面建立全球领先优势，巩固其在消费电子、半导体等领域的霸主地位，并开拓新的增长空间。

3.1.1.2 专利申请趋势分析

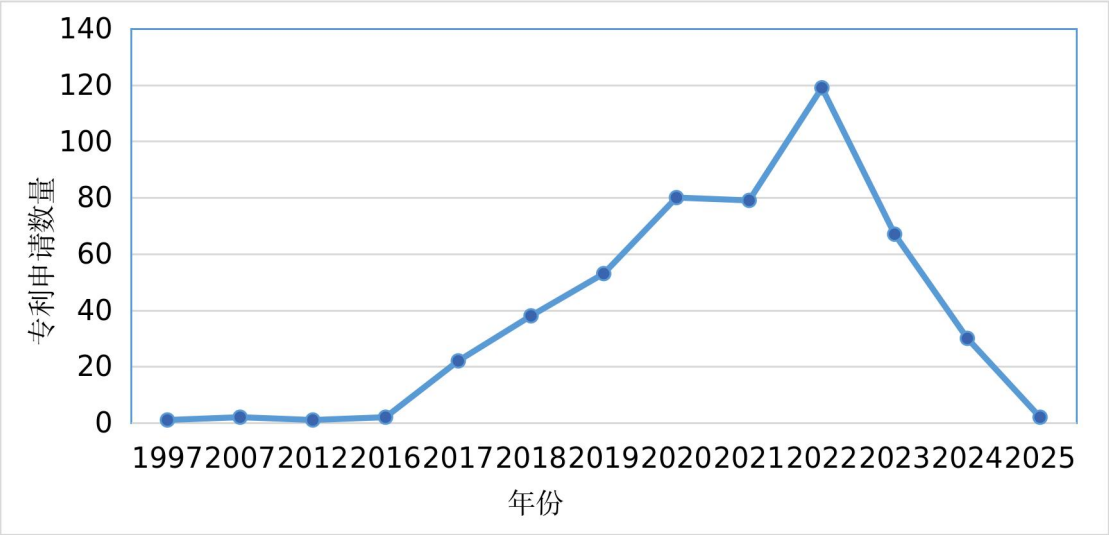


图 3-1 三星电子专利申请趋势

三星电子在全球共申请 AI 模型优化技术相关专利合计 496 件。为了进一步研究三星电子专利申请的发展情况，以下对其全球专利申请数据按时间进行统计分析。

三星电子在 AI 模型优化技术领域自 1997 年开始申请专利，其中首件 AI 模型优化技术相关专利公开号为 US6002851A，涉及一种多处理器节点算法技术。从发展趋势来看，在 2017 年之前三星电子对 AI 技术研发投入较少，专利申请也维持在低位，自 2017 年起才逐步加大了对 AI 模型优化技术的专利布局力度，年均申请由最开始的 20 件左右，逐步增长为 50 件上下，到 2022 年达到峰值 119 件。

3.1.1.3 专利区域和法律状态分析

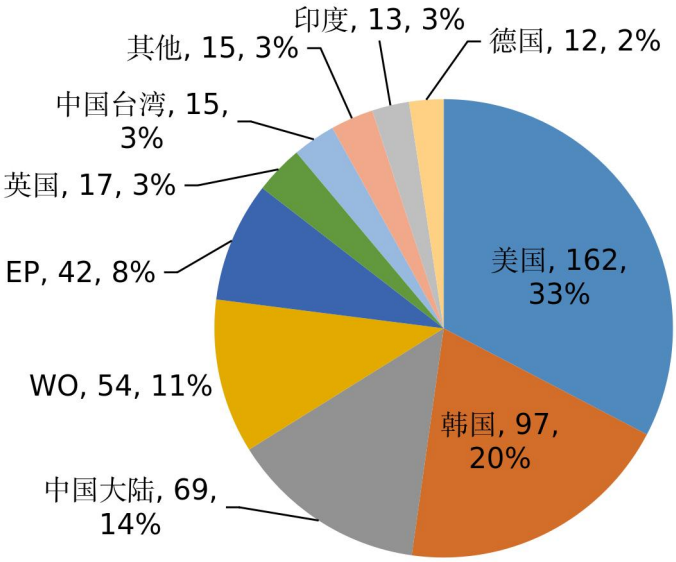


图 3-2 专利区域分布

图 3-2 显示了三星电子的 AI 模型优化技术专利区域分布及相关法律状态。隆基绿能在全球 14 个国家/地区/组织进行了专利申请。其中在美国的专利申请数量排在第一位，合计 162 件，占比 33%；其次为韩国，专利申请为 97 件，占比 20%；中国大陆次之，为 69 件，占比 14%。此外，在欧洲、英国、中国台湾等也有部分专利公开。

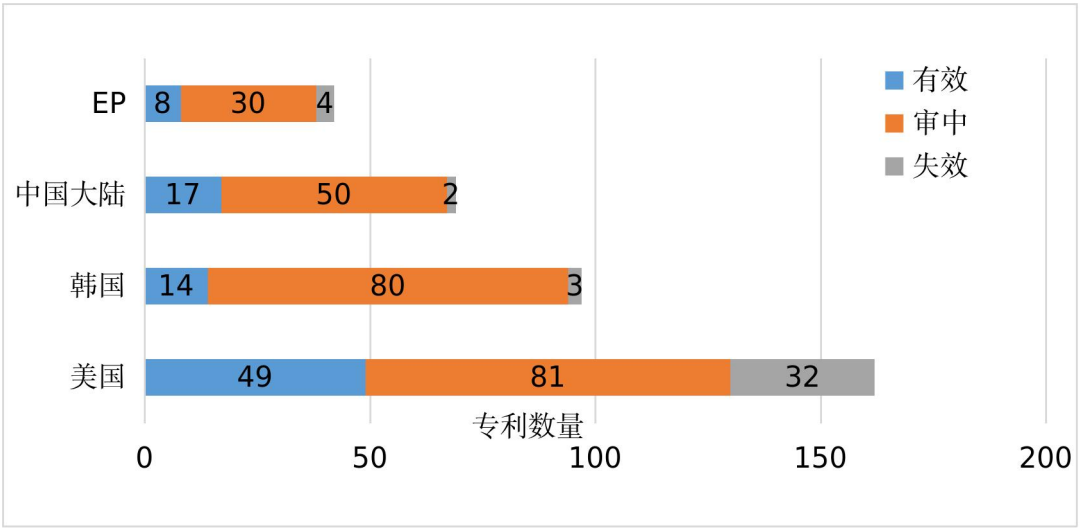


图 3-3 主要国家布局专利法律状态

结合专利法律状态来看，三星电子专利申请较晚，在主要国家的有效专利布局较少，其中美国有效专利最多也仅为 49 件，但由于其近些年加大了再 AI 领域的投入，因此在主要国家其在审专利均为有效专利的 2 倍左右，技术活跃度强。

3.1.1.4 主要发明人分析

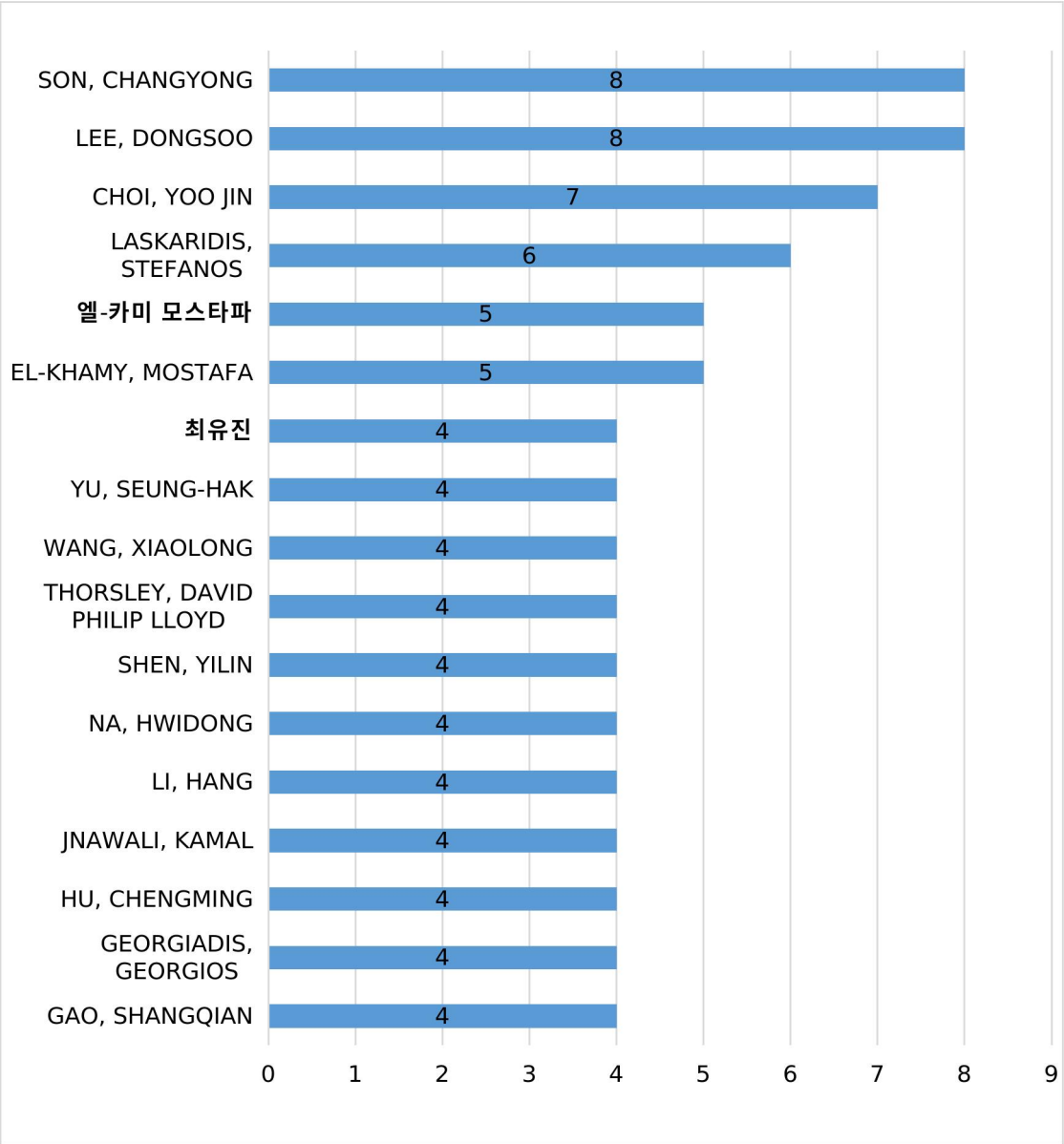


图 3-4 主要发明人排名

图 3-4 是专利申请量在 4 件以上的第一发明人基本情况。整体来看，三星电子在 AI 技术领域的发明人团队较为分散。其中并列排名第一的 SON,CHANGYONG 和 LEE,DONGSOO 也仅申请了 8 件专利，其中 SON,CHANGYONG 主要涉及轻量化神经网络技术，而 LEE,DONGSOO 则涉及神经网络压缩相关技术。排名第二的

CHOI,YOO JIN 合计申请了 7 件专利，主要涉及神经网络量化和蒸馏技术。此外，排名第三的 LASKARIDIS,STEFANOS 合计申请了 6 件专利，主要涉及联邦学习模型和可变精度神经网络技术。

3.1.1.5 关键技术分析

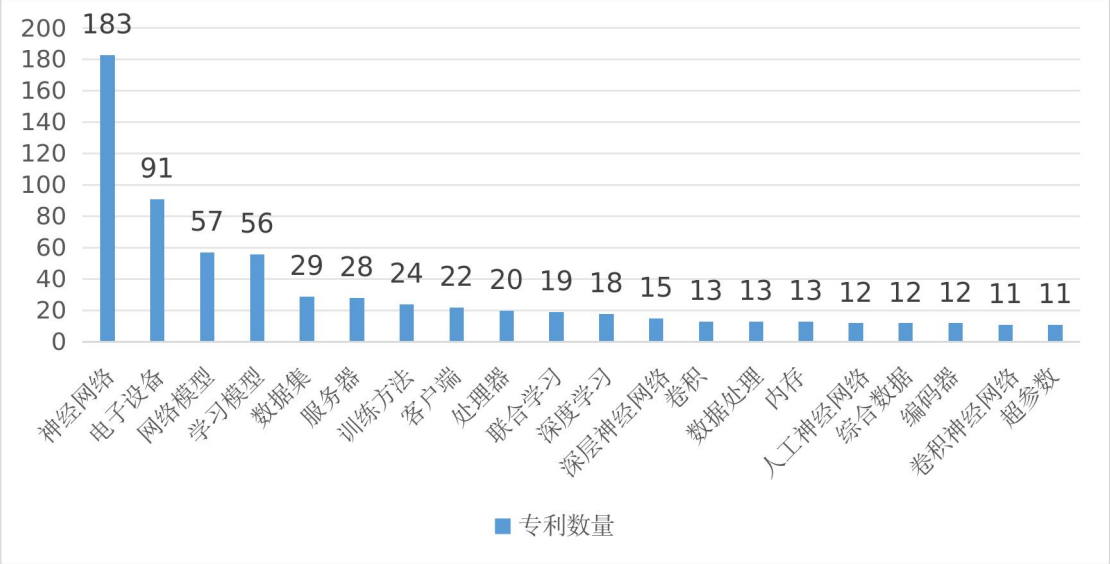


图 3-5 三星电子 AI 模型优化技术主题分布图

三星电子在 AI 模型优化技术的专利申请呈现多元布局，涵盖神经网络（187 件）、电子设备主要涉及内存数据处理等（91 件）、网络模型（57 件）等关键领域，表明其在整合 AI 模型软件及硬件方向方面投入较多，致力于提升 AI 上下游的协同处理效能。

进一步的对三星电子的以上技术进行技术归类，得到图 3-6。三星电子的 AI 模型优化技术涵盖了 4 个技术分支。其中知识蒸馏技术专利较多，为 90 族，占比 40%；其次为参数优化，合计 50 族专利，占比 22%；模型结构优化则有 46 族专利申请，占比 21%。此外，模型压缩技术相对较少，专利申请为 38 族。

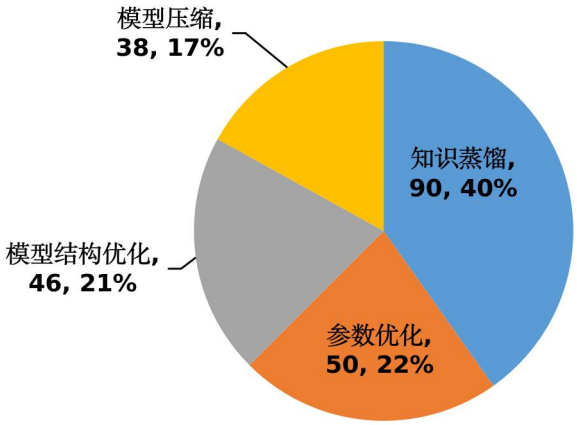


图 3-6 三星电子技术分布

3.1.2 IBM

3.1.2.1 公司简介

IBM（国际商业机器公司）是全球历史最悠久、最具影响力的科技巨头之一，成立于 1911 年。它从制表机和商用机器起家，经历了大型机、个人电脑、软件服务等多个时代转型。如今，IBM 的核心定位是混合云与人工智能解决方案提供商，专注于企业级市场。其核心竞争力在于为企业客户提供复杂 IT 环境，包括公有云、私有云和本地系统的整合、管理、安全及智能化服务。

IBM 是人工智能领域的先驱和持续推动者，早在 1997 年 IBM“深蓝”击败国际象棋冠军卡斯帕罗夫，展示计算与规则引擎的威力。IBM 于 2023 年推出 Watsonx 平台，其认知计算系统“Watson”是 IBM AI 的代名词，这是一个集成的企业级 AI 与数据平台，旨在帮助企业安全、高效地构建、部署、管理和治理 AI 应用，特别是生成式 AI 和大语言模型。它包含：Watsonx.ai、watsonx.data 以及 watsonx.governance 等。

通过 watsonx 平台，IBM 为企业提供了一个在混合云环境中规模化构建、部署和管理 AI，尤其是生成式 AI 的一站式解决方案。IBM 的 AI 战略核心在于赋能企业利用 AI 提升效率、创新和决策水平，同时确保过程的安全、透明和合规，巩固其作为企业级关键技术合作伙伴的地位。

3.1.2.2 专利申请趋势分析

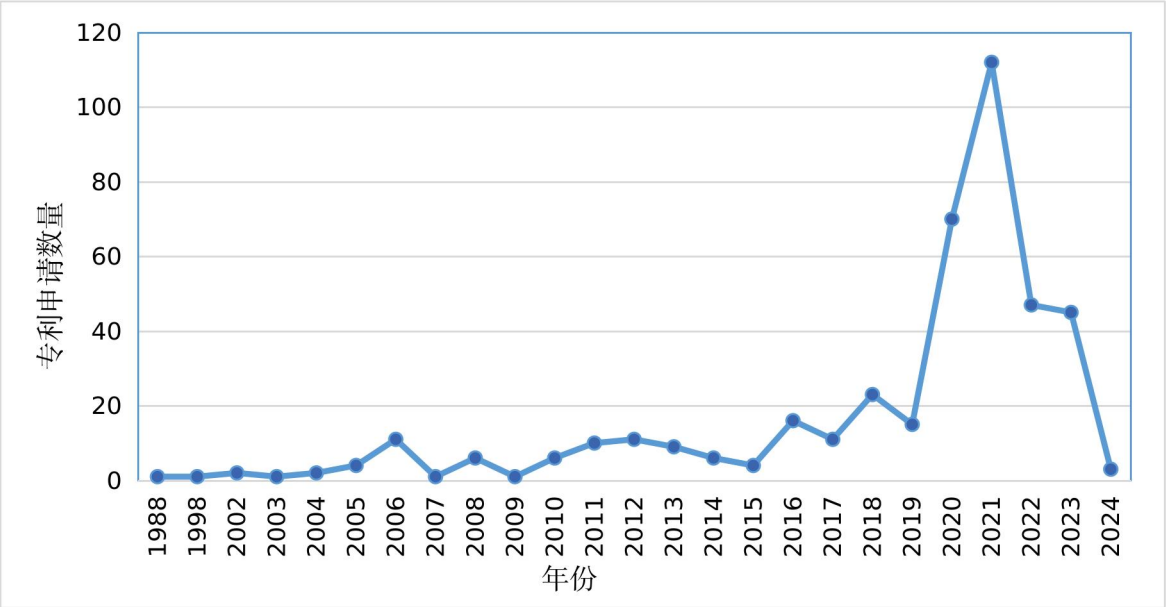


图 3-7 IBM 专利申请趋势

IBM 在全球共申请 AI 模型优化技术相关专利合计 418 件。为了进一步研究 IBM 的专利申请的发展情况，以下对其全球专利申请数据按时间进行统计分析。在 AI 模型优化领域自 1988 年开始申请专利，大致分为三个阶段，分别为 1988-2014 年的初期阶段、2015-2019 年的波动增长阶段和 2020 年以后的快速增长阶段。

技术起步初期 IBM 专利申请量一直处于较低水平，这一阶段技术的改进涉及模型数据处理，包括压缩生物模型的方法、优化输入输出的微模型算法等，在这一时间内 IBM 专利申请间隔时间长，不固定。自 2015 年开始，IBM 逐步加大了对 AI 模型的研发投入，专利逐渐开始稳步增长，维持在 10-20 件左右，涉及模型正则化、模型修剪等相关技术；自 2019 年开始，IBM 在 AI 模型领域的专利开始快速增长，在技术范围不断扩展的同时，其也加大了对各区域的专利布局。

整体来看，IBM 在 AI 模型领域的技术起步早，技术专利布局范围广。但近几年相对于三星专利申请相对滞后。

3.1.2.3 专利区域和法律状态分析

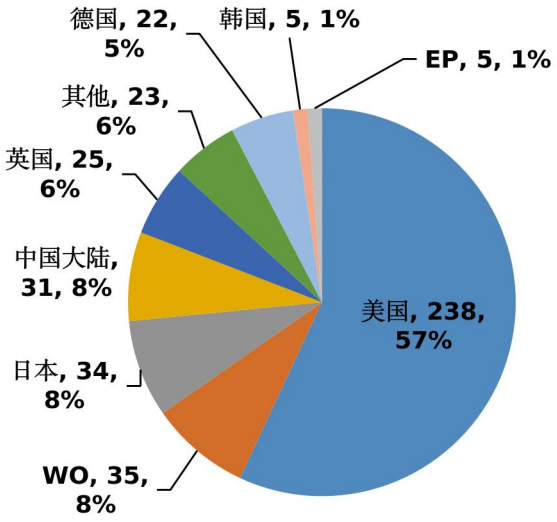


图 3-8 IBM 专利区域分布

图 3-8 显示了 IBM 的 AI 模型优化技术专利区域分布及相关法律状态。IBM 在全球 16 个国家/地区/组织进行了专利申请。其中在美国的专利申请数量排在第一位，合计 238 件，占比 57%；其次为日本和中国大陆，专利申请分别为 34 件和 31 件。此外，其在英国和德国也有部分专利布局。

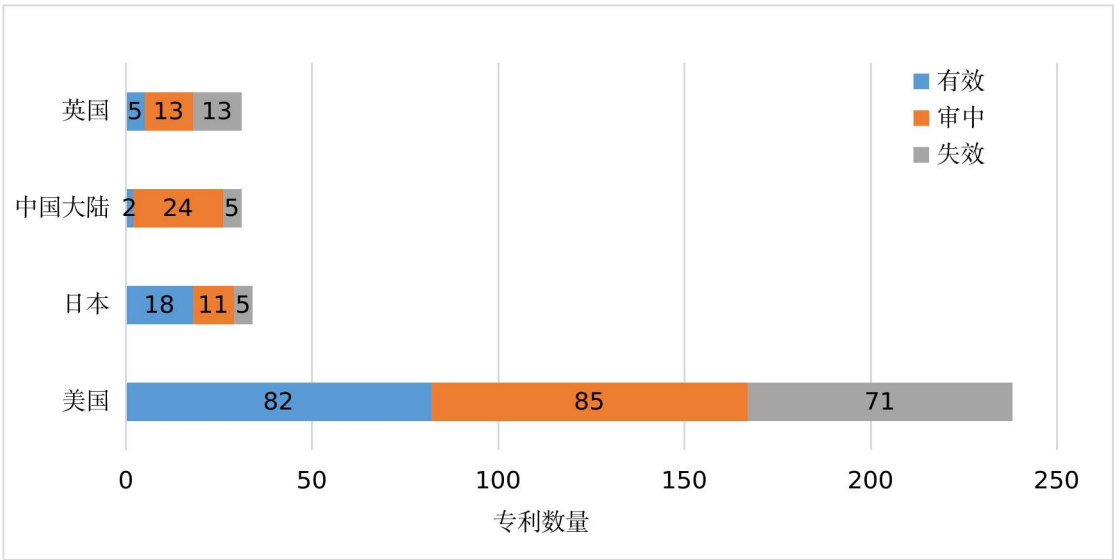


图 3-9 主要国家布局专利法律状态

结合专利法律状态来看，IBM 的有效专利主要集中在美国，有 82 件；其次为日本。从在审专利来看，IBM 的在审专利申请主要集中在美国和中国大陆，可以看出 IBM 在近两年加大了对中国大陆的技术布局。

3.1.2.4 主要发明人分析

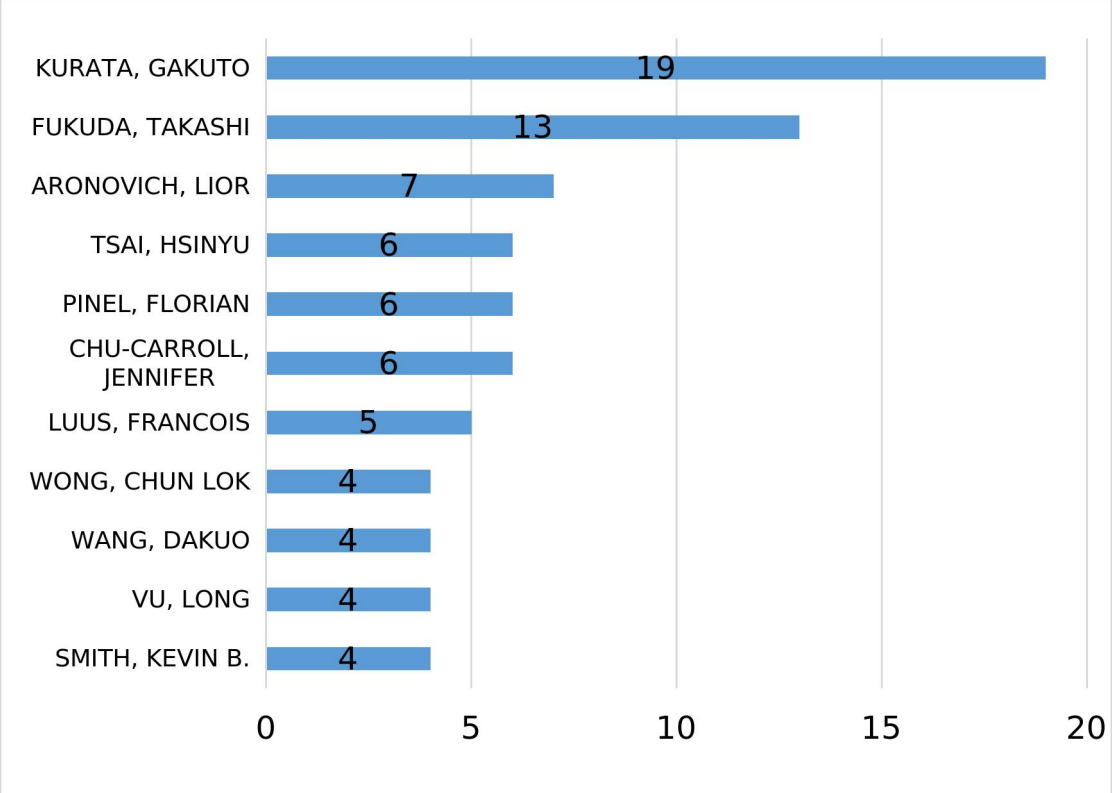


图 3-10 主要发明人专利公开

图 3-10 是 IBM 专利申请量在 4 件以上的第一发明人基本情况。其中排名第一的 KURATA,GAKUTO 合计申请了 19 件相关专利，主要涉及知识蒸馏和参数优化两个技术方向，譬如利用超参数优化词嵌入的监督训练等。排名第二的 FUKUDA,TAKASHI 合计申请了 13 件专利，主要涉及知识蒸馏技术，譬如使用无损和有损分支训练教师机器学习模型等。此外，排名第三的 ARONOVICH,LIOR 合计申请了 7 件专利，主要涉及参数优化等技术。

3.1.2.5 关键技术分析

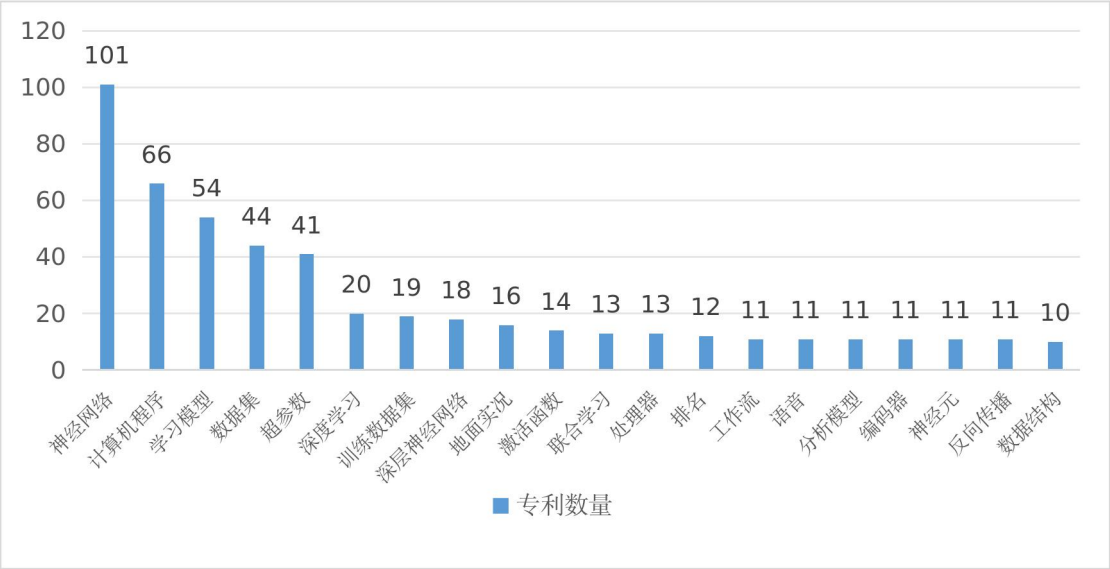


图 3-11 国际商业机器 AI 模型优化技术主题分布图

IBM 在 AI 模型优化技术的专利申请呈现多元布局，涵盖神经网络（101 件）、计算机程序（66 件）、学习模型（54 件）等关键领域，整体来看，相较于三星电子，IBM 更加偏向于数据和软件等方向的 AI 模型优化技术。

进一步的对 IBM 的以上技术进行技术归类，得到图 3-12。IBM 的 AI 模型优化技术涵盖了 4 个技术分支。其中参数优化技术专利较多，为 90 族，占比 41%；其次为模型结构优化，合计 80 族专利，占比 37%；知识蒸馏则有 35 族专利申请，占比 16%。此外，模型压缩技术相对较少，专利申请为 12 族。

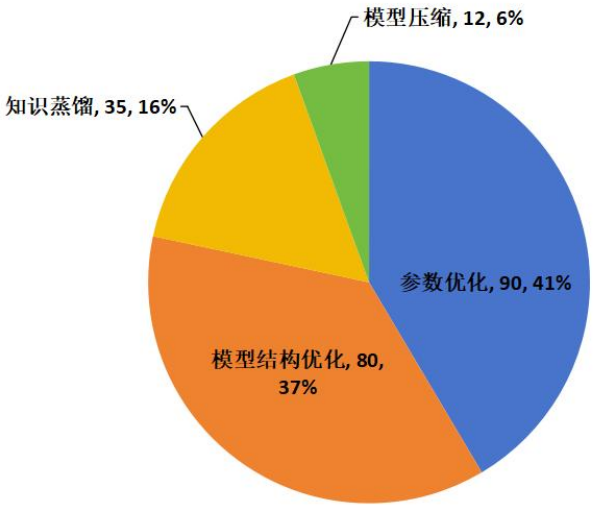


图 3-12 IBM 技术分布

3.1.3 微软

3.1.3.1 公司简介

微软（Microsoft）是全球领先的科技公司，由比尔·盖茨（Bill Gates）和保罗·艾伦（Paul Allen）于 1975 年创立，总部位于美国华盛顿州雷德蒙德。微软最初以操作系统（如 MS-DOS、Windows）和办公软件（如 Microsoft Office）闻名，业务涵盖云计算、人工智能、硬件（Surface、Xbox）、企业服务等领域。

微软是人工智能（AI）领域的领导者之一，其 AI 战略涵盖基础研究、产品整合和商业应用。Microsoft Research 是全球顶级 AI 研究机构，在自然语言处理（NLP）、计算机视觉、机器学习等领域贡献显著。2019 年起微软投资 OpenAI，并将 GPT 模型整合到微软产品如 Azure、Copilot。

3.1.3.2 专利申请趋势分析

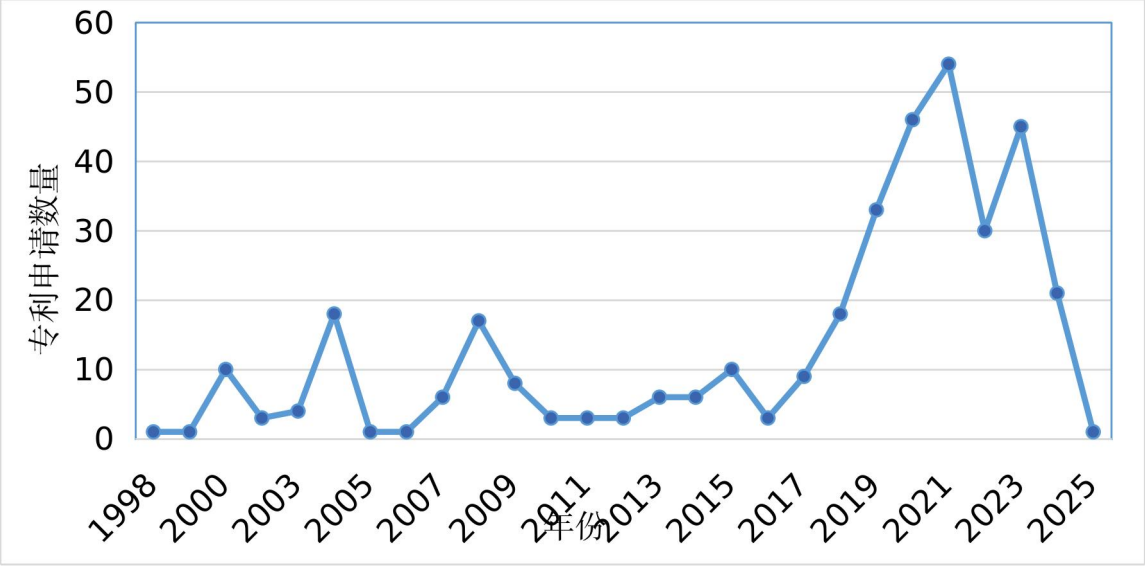


图 3-13 IBM 专利申请趋势

微软在全球共申请 AI 模型优化技术相关专利合计 361 件。为了进一步研究微软的专利申请的发展情况，以下对其全球专利申请数据按时间进行统计分析。在 AI 模型优化领域微软自 1998 年开始申请专利，大致分为两个阶段，分别为 1998-2015 年的初缓慢波动增长和 2015 年之后的快速发展阶段。

技术起步初期微软专利申请量一直处于较低水平，这一阶段技术的改进涉及模型压缩和模型优化等技术，其最早的一件专利公开号为 US6263337B1，涉及在数据库中进行聚类的算法，在这一时间内微软专利申请间隔时间长，不固定。自 2015 年开始，微软逐步加大了对 AI 模型的研发投入，专利逐渐开始稳步增长，维持在 10 件左右；自 2019 年开始，微软投资 OpenAI 后在 AI 模型领域的专利开始快速增长，在技术范围不断扩展的同时，其也加大了对各区域的专利布局。到高峰时期 2021 年，其专利申请为 54 件。

3.1.3.3 专利区域和法律状态分析

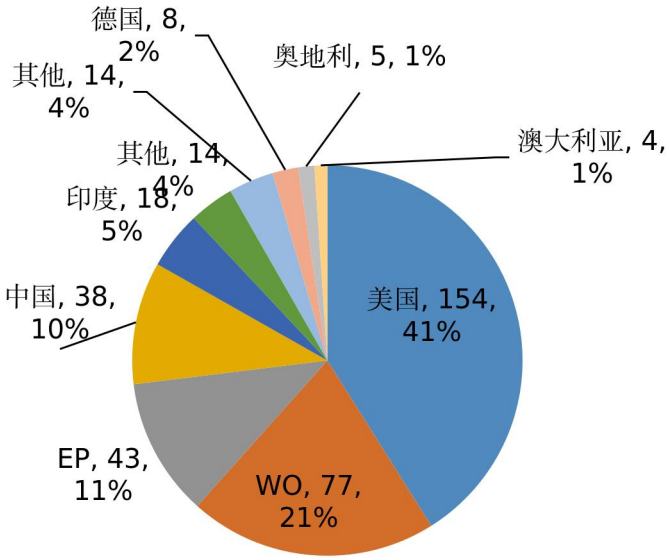


图 3-14 微软专利区域分布

图 3-14 显示了微软的 AI 模型优化技术专利区域分布及相关法律状态。微软在全球 14 个国家/地区/组织进行了专利申请。其中在美国的专利申请数量排在第一位，合计 154 件，占比 41%；其次为 EP、中国和印度，专利申请分别为 43 件、38 件和 18 件。此外，其在德国和奥地利等也有部分专利布局。

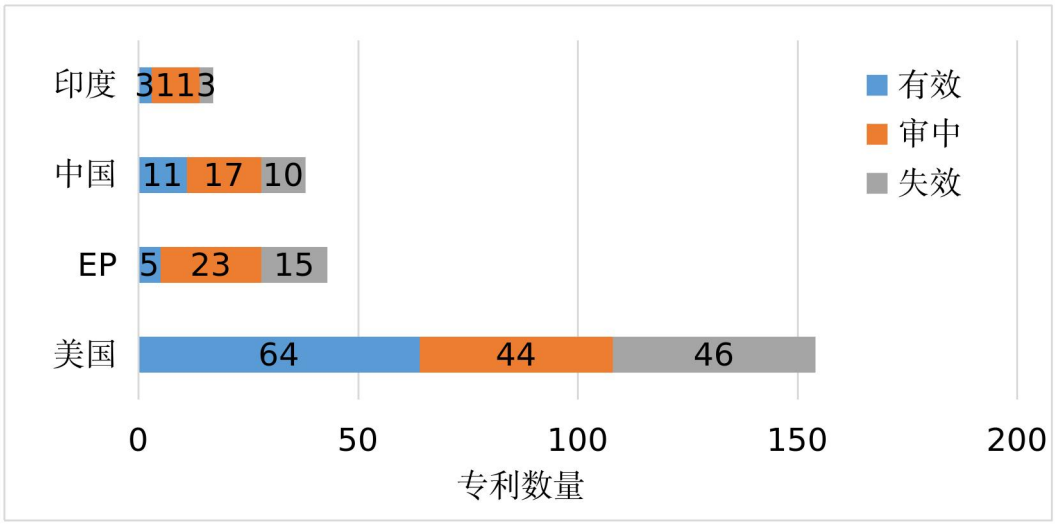


图 3-15 主要国家布局专利法律状态

结合专利法律状态来看，微软的有效专利主要集中在美国，有 64 件；其次为中国。从在审专利来看，微软的在审专利申请主要集中在美国和欧洲，可以看出微软在近两年加大了对欧洲的技术布局。

3.1.3.4 主要发明人分析

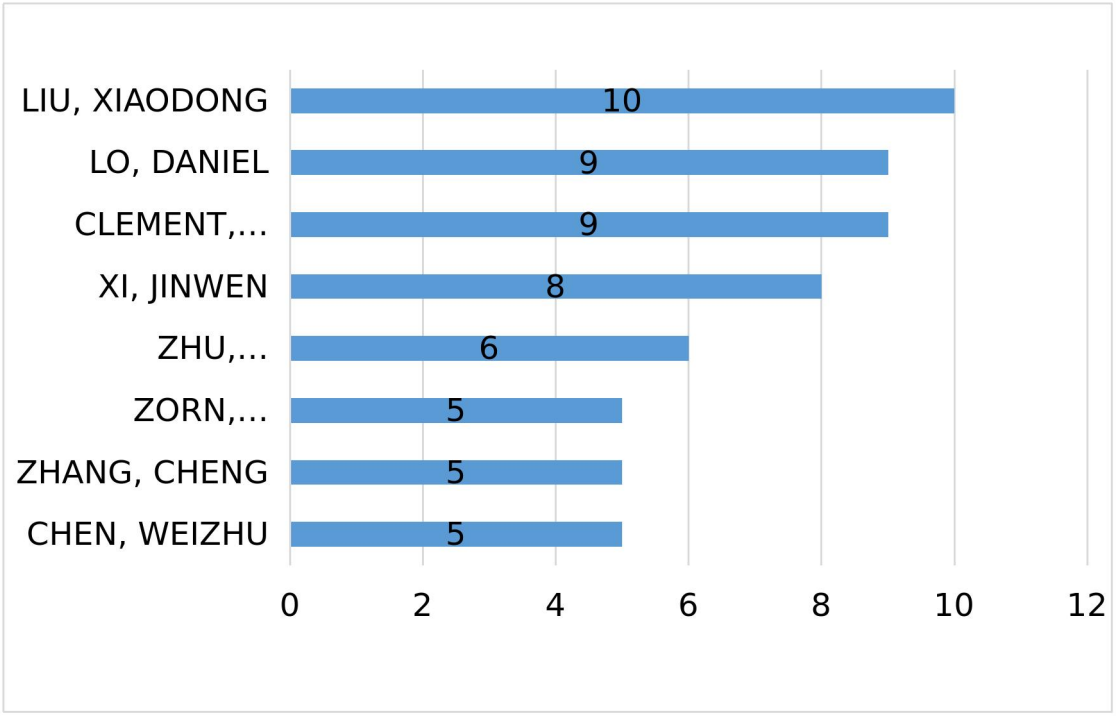


图 3-16 主要发明人专利公开

图 3-16 是微软专利申请量在 4 件以上的第一发明人基本情况。其中排名第一的 LIU,XIAODONG 合计申请了 10 件相关专利，主要涉及模型机构优化技术方向，

譬如预训练期间模型的解耦优化等。排名第二的 LO,DANIEL 合计申请了 9 件专利，主要涉及模型压缩技术，譬如使用离群值块浮点的神经网络激活压缩等。此外，并列第二的 CLEMENT,COLIN BRUCE 合计申请了 9 件专利，同样主要涉及模型压缩和知识蒸馏等技术。

3.1.3.5 关键技术分析

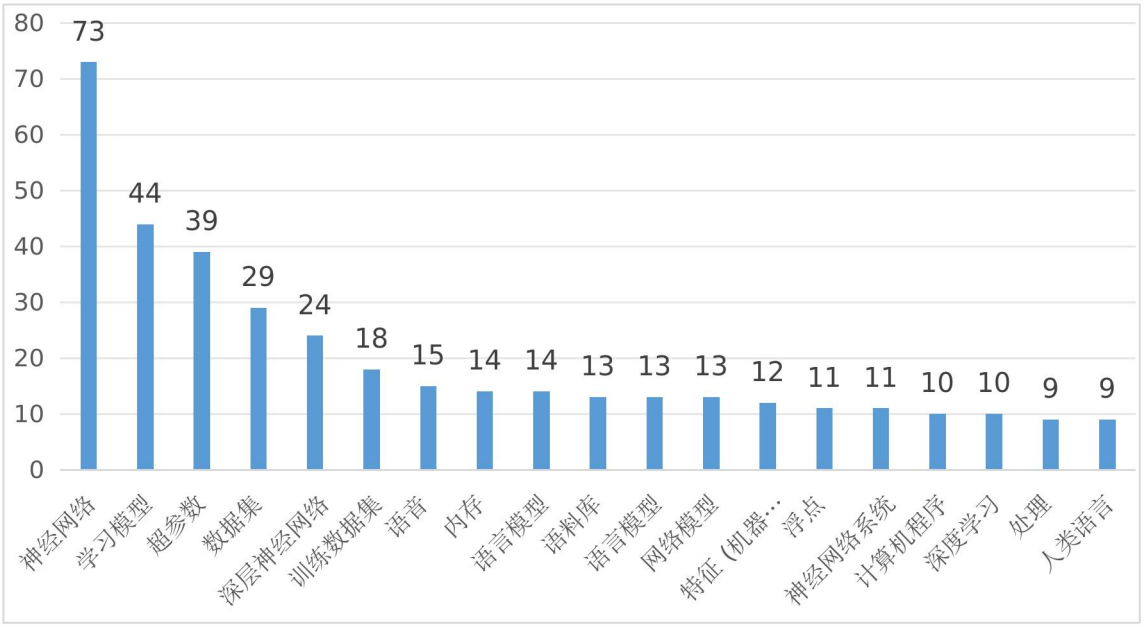


图 3-17 微软 AI 模型优化技术主题分布图

微软在 AI 模型优化技术的专利申请呈现多元布局，涵盖神经网络（73 件）、学习模型（44 件）和超参数（39 件）等关键领域，整体来看，相较于其他权利人微软更倾向于大语言模型优化。

进一步的对微软的以上技术进行技术归类，得到图 3-18。微软的 AI 模型优化技术涵盖了 4 个技术分支。其中参数优化技术专利较多，为 54 族，占比 38%；其次为模型结构优化，合计 37 族专利，占比 26%；知识蒸馏则有 29 族专利申请，占比 20%。此外，模型压缩技术相对较少，专利申请为 23 族。

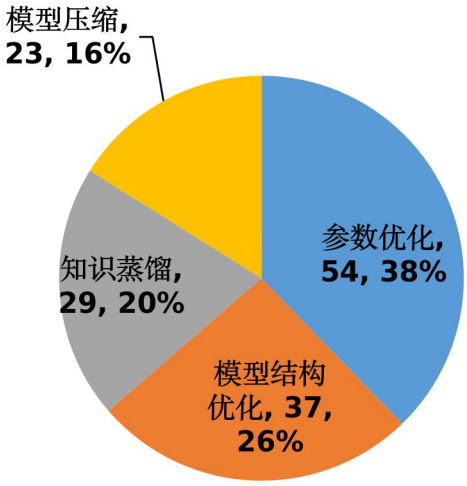


图 3- 18 微软技术分布

3.2 国内主要权利人

3.2.1 华为

3.2.1.1 公司简介

华为是全球领先的信息与通信技术（ICT）解决方案供应商，成立于 1987 年，总部位于中国深圳。公司业务涵盖电信网络、企业业务、消费者业务及云计算等领域，致力于通过创新技术推动行业数字化转型。截至 2025 年，华为已服务全球 170 多个国家和地区，构建了覆盖端、边、云的全栈技术能力。

华为自 2018 年起全面投入 AI 研发，形成了以昇腾芯片、MindSpore 框架、盘古大模型为核心的全栈技术体系。其昇腾芯片型号包括 310、910 和 920 等。预计在 2025 年下半年量产的昇腾 920 采用 7nm 工艺，单芯片算力达 900TOPS（INT8），能效比是英伟达 H20 的 20 倍。其 MindSpore 大模型框架支持端边云全场景，兼容 PyTorch/TensorFlow 生态。其盘古大模型系列包含：盘古 NLP 中文首个千亿级 NLP 大模型；盘古 CV 为 30 亿参数图像模型，可以实现工业质检、遥感图像分析高精度识别；盘古 3.0 升级多模态能力，覆盖金融、医疗、矿山等 10+行业场景。

其 AI 战略聚焦“行业智能化”，通过开放场景、构建算力基础设施、优化数据质量三大路径，推动 AI 与千行百业深度融合。截至 2024 年，华为昇腾 AI 集群

已部署于全国 25+城市人工智能计算中心，支撑超过 5000 家企业和科研机构 AI 研发。

3.2.1.2 专利申请趋势分析

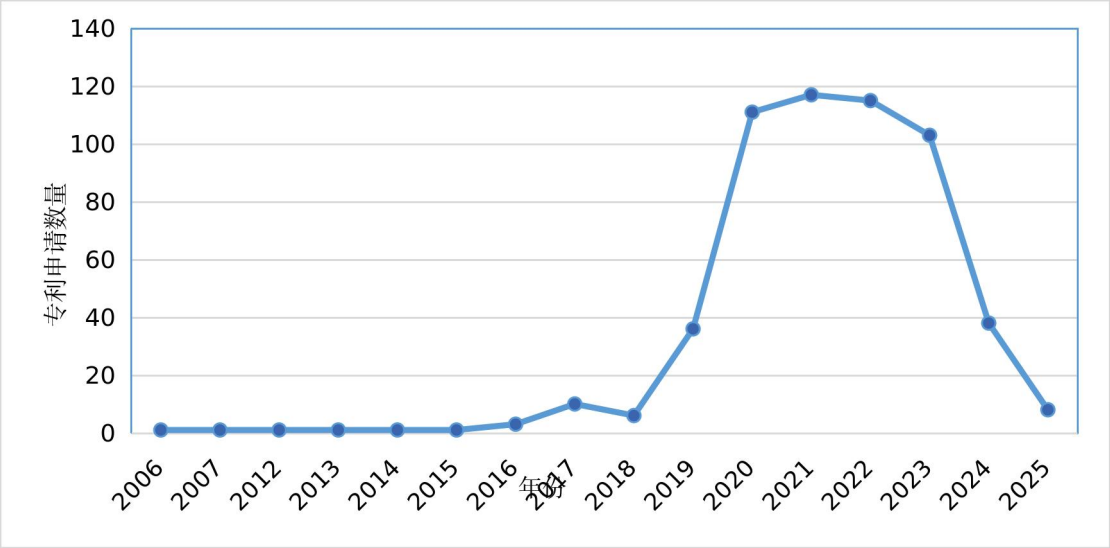


图 3-19 华为专利申请趋势

华为在全球共申请 AI 模型优化技术相关专利合计 553 件。为了进一步研究华为的专利申请发展情况，以下对其全球专利申请数据按时间进行统计分析。华为在 AI 模型领域自 2006 年开始申请专利，大致分为两个阶段，分别为 2006-2018 年的初期阶段、2019-至今的申请快速增长阶段。

在 2006 年-2018 年，华为在 AI 模型优化领域专利申请较少，年均专利申请不足 10 件，同时在该阶段华为的专利申请多涉及报文剪枝技术，核心的 AI 模型优化专利申请较少。自 2018 年起，华为逐步加大了对 AI 模型技术的研发力度，在知识蒸馏、模型压缩等多个 AI 模型优化技术方向布局了相关专利，同时基于对该技术的重视和市场化需要，华为在全球多个国家/地区均进行了相应的专利布局。这一阶段其专利数量快速增加，在 2020-2023 年均稳定在高位 100 件左右。由于专利申请的公开时滞原因，近两年专利数量有所下降。

3.2.2.3 专利区域和法律状态分析

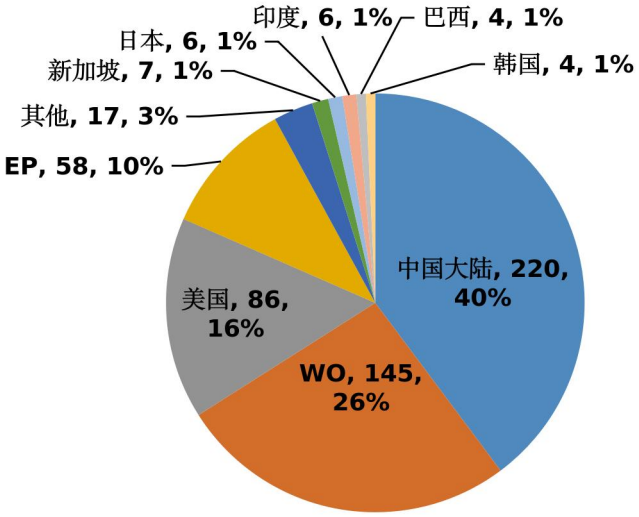


图 3-20 华为专利区域分布

图 3-20 显示了华为的 AI 模型优化技术专利区域分布状态。华为在全球 19 个国家/地区/组织进行了专利申请。其中在中国大陆的专利申请数量排在第一位，合计 220 件，占比 40%；其次为美国和 EP，专利申请分别为 86 件和 58 件。此外，其在新加坡、日本等国家也有部分专利布局。

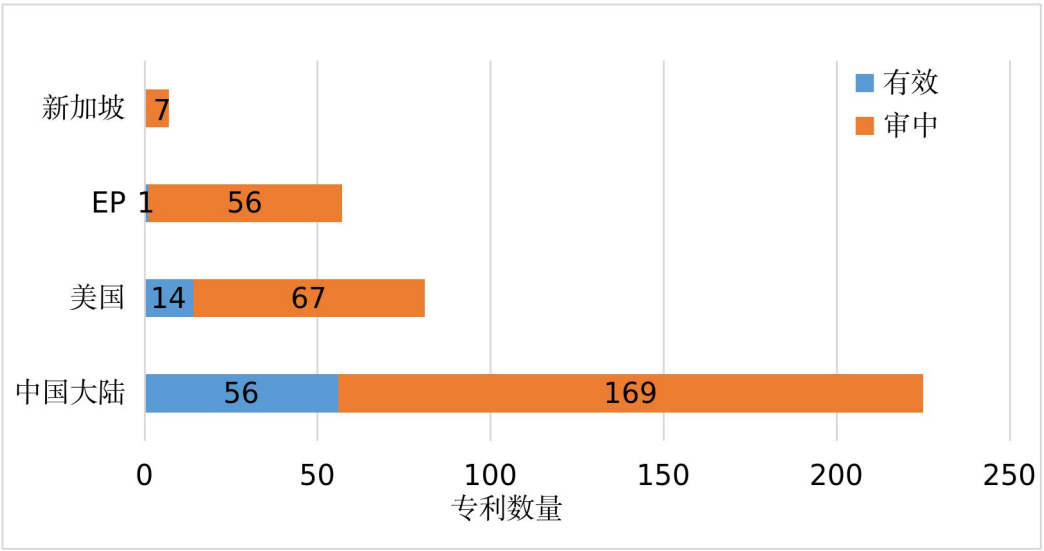


图 3-21 区域法律状态分布情况

结合专利法律状态来看，华为的有效专利主要集中在中国大陆，有 56 件。其余地区有效专利较少，尤其是欧洲和新加坡。同时，由于华为专利多为近几年申请，因此在大部分地区均为在审专利占比较多。

3.2.2.4 主要发明人分析

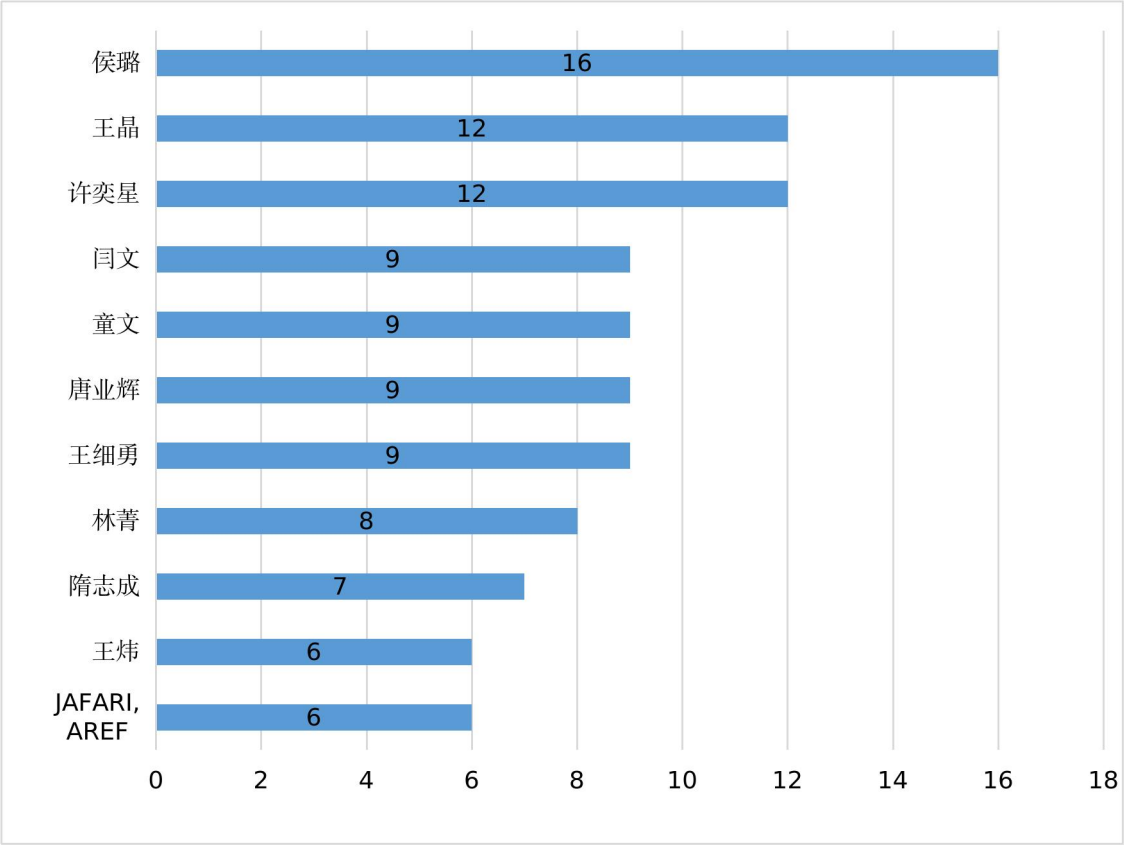


图 3- 22 华为主要发明人排名

图 3-22 是华为专利申请量在 5 件以上的第一发明人基本情况。其中排名第一的侯璐合计申请了 16 件相关专利，主要涉及知识蒸馏技术方向等。排名第二的王晶合计申请了 12 件专利，主要涉及人工智能图像处理压缩技术。排名第三的许奕星同样申请了 12 件相关专利，最有涉及神经网络的压缩和蒸馏技术。整体来看，华为的主要发明人均关注与 AI 大模型压缩和蒸馏相关技术。

3.2.2.5 关键技术分析

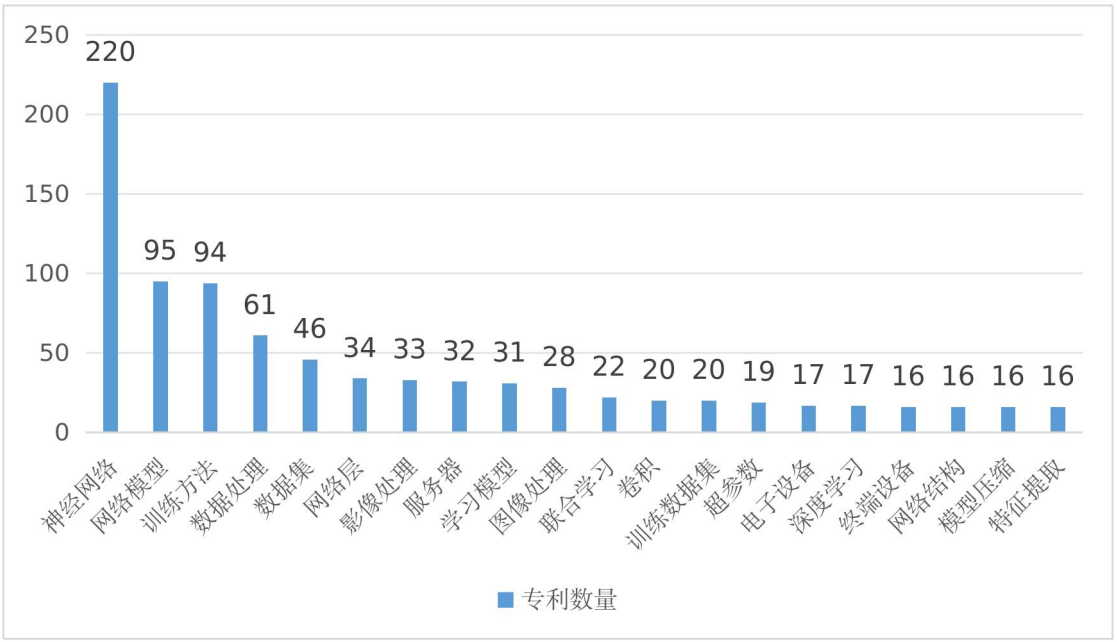


图 3-23 华为 AI 模型优化技术主题分布图

华为在 AI 模型优化技术的专利申请呈现多元布局，涵盖神经网络（220 件）、网络模型（95 件）、训练方法（94 件）等关技术分支，整体来看，华为在 AI 模型技术方向更倾向于模型训练方法、数据处理等相关技术，在应用上更侧重于图像/影像处理等实际落地场景。

进一步的对华为的以上技术进行技术归类，得到图 3-24。华为的 AI 模型优化技术涵盖了 4 个技术分支。其中知识蒸馏技术专利较多，为 101 族，占比 41%；其次为模型压缩，合计 63 族专利，占比 25%；参数优化则有 42 族专利申请，占比 17%。模型结构优化技术再次之，专利申请为 41 族。

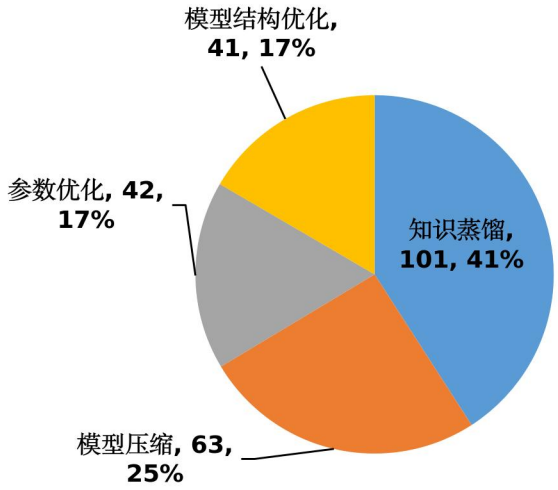


图 3-24 华为技术分布

3.2.2 腾讯

3.2.2.1 公司简介

腾讯科技（深圳）有限公司成立于 1998 年，总部位于中国深圳，是中国领先的互联网科技企业之一。公司以社交平台微信和 QQ 为核心，业务广泛布局于游戏、金融科技、云计算、人工智能等关键领域，在全球拥有庞大用户群体。腾讯致力于通过技术创新推动行业发展，积极参与 AI 等前沿技术研究与应用，旨在打造智能便捷的产品和服务。在多模态融合技术领域，腾讯的研发投入大，专利申请量国内领先，技术应用覆盖内容创作、医疗影像分析、智能驾驶等多个场景，持续探索新兴应用方向。

腾讯的混元大模型是其在多模态融合技术领域的重要成果，支持文本、图像、视频和 3D 等多种模态内容的理解与生成。模型采用混合专家模型（MoE）结构，其总体性能相比上一代提升了 50%，部分中文能力已追平 GPT-4。例如，混元大模型的图像到文本模型包含视觉变换器（ViT）用于视觉编码、视觉-语言适配器，以及采用 MoE 架构的大型语言模型。

腾讯不断探索和优化多模态融合的架构和方法。例如，通过视觉大语言模型的文本-图像融合模块，实现文字描述和图像内容的精确匹配和融合。开发了基于视

觉变换器（ViT）和 Swin Transformer 的多模态编码器，用于视觉特征提取，并通过多模态对齐和融合技术，将不同模态的特征进行有效整合。

腾讯的多模态技术在内容创作领域得到了广泛应用，如影视创作、社交内容生产、游戏资产生成和智能广告投放等。通过多模态模型的混合生成模式，实现了图生图、图生视频等创新功能，降低了创作门槛，提高了创作效率。此外，混元大模型还为腾讯视频、QQ 浏览器、微信等内部产品提供智能剪辑、智能创作等能力支持，助力内容的高效生产和推荐。

3.2.2.2 专利申请趋势分析

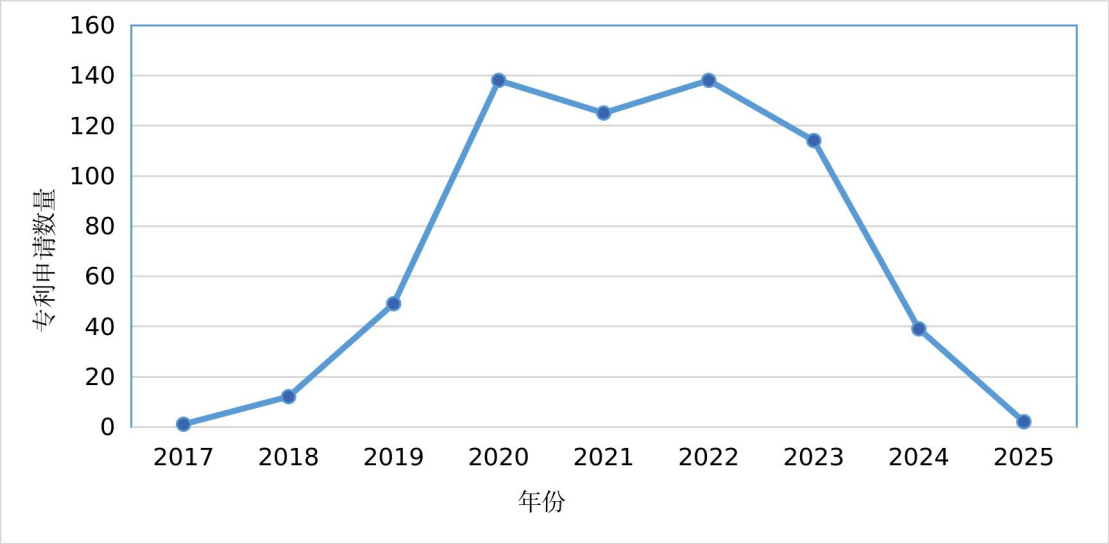


图 3-25 腾讯专利申请趋势

腾讯在全球共申请 AI 模型优化技术相关专利合计 618 件。为了进一步研究腾讯的专利申请发展情况，以下对其全球专利申请数据按时间进行统计分析。腾讯在 AI 模型优化领域自 2016 年开始申请专利，其中最早的一件专利公开号为 CN108304439B，涉及一种语义模型优化方法，利用已经优化的编码规则对另一个模型进行优化。自 2018 年开始，腾讯加大了对 AI 模型优化技术的研发投入，特别是在知识蒸馏、参数优化、以及模型优化等方向，到 2020 年-2023 年，在 AI 模型技术方向的专利申请一直维持在 100 件以上。近两年专利公开有所减少，主要是专利申请的公开时滞原因。

3.2.2.3 专利区域和法律状态分析

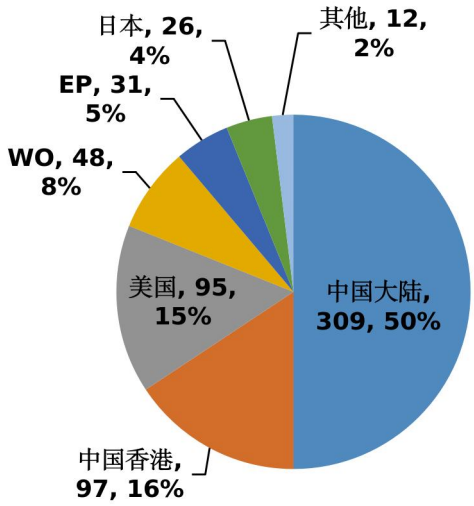


图 3-26 腾讯 AI 模型优化专利区域分布

图 3-26 显示了腾讯的 AI 模型优化技术专利区域分布及相关法律状态。腾讯在全球 11 个国家/地区/组织进行了专利申请。其中在中国大陆的专利申请数量排在第一位，合计 309 件，占比 50%；其次为中国香港和美国，专利申请分别为 97 件和 35 件。此外，其在 EP 和日本也有部分专利布局。

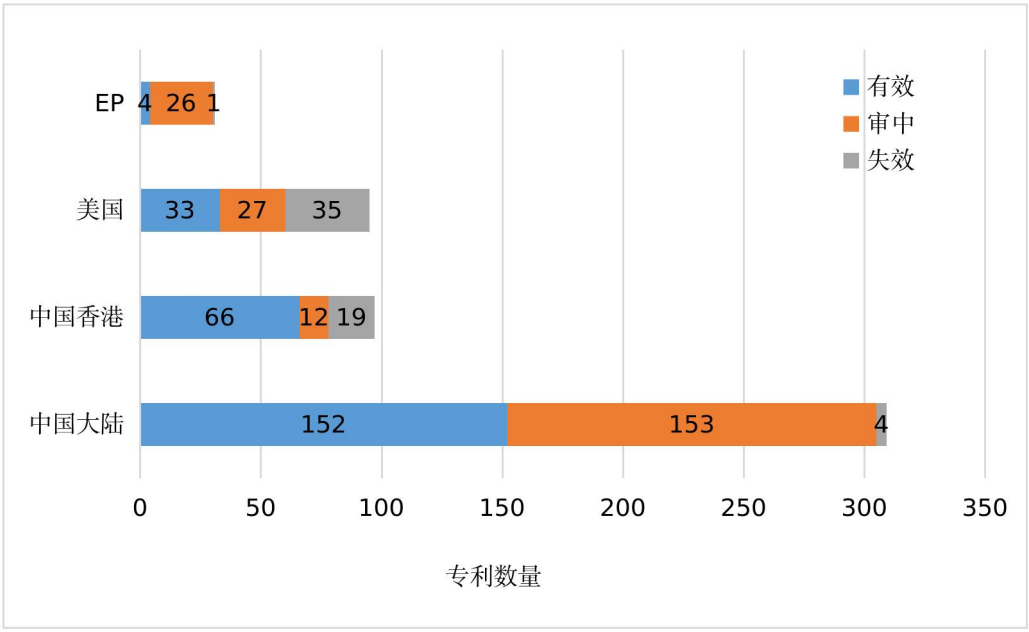


图 3-27 区域法律状态分布情况

结合专利区域法律状态来看，腾讯的有效和在审专利主要集中在中国大陆，表明腾讯仍以中国大陆为主要专利布局区域。此外，其在中国香港存在较多有效专利。从腾讯的专利布局策略来看，近几年其加大了对美国和欧洲的 AI 模型专利申请。

3.2.2.4 主要发明人分析

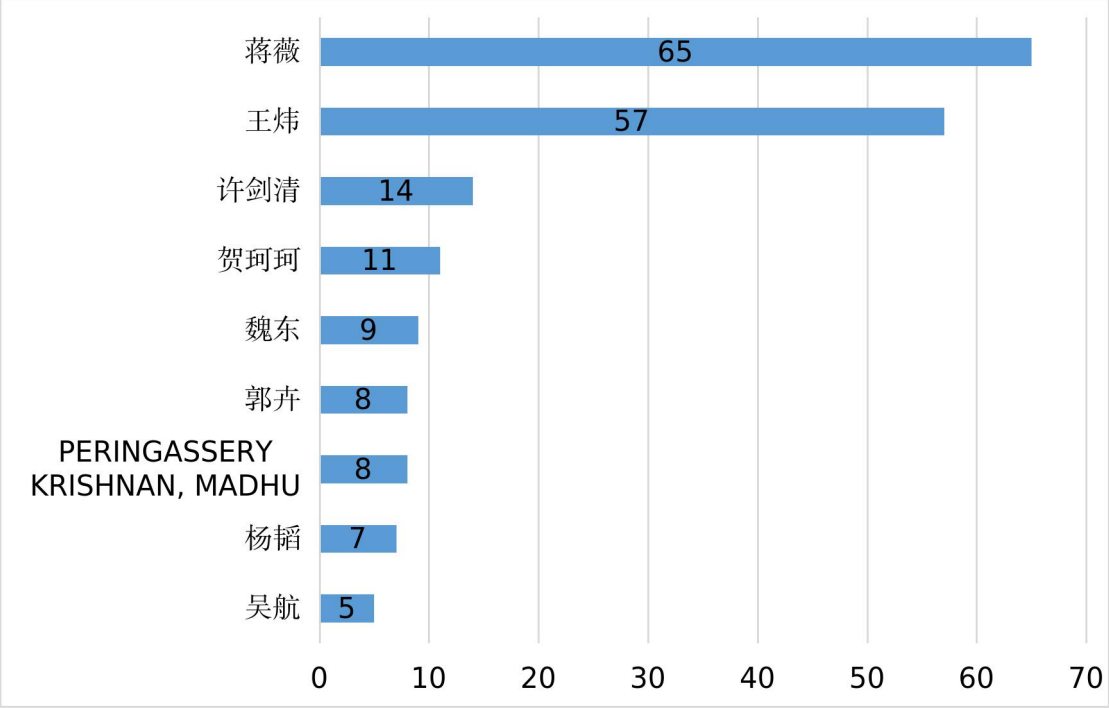


图 3- 28 腾讯主要发明人排名

图 3-28 是腾讯专利申请量在 5 件以上的第一发明人基本情况。相较于上述几家权利人，腾讯的 AI 模型优化专利发明团队较为集中，主要涉及两个核心发明人蒋薇和王炜，其中蒋薇合计有 65 件专利公开，涉及神经网络模型压缩技术；而王炜为腾讯云架构师，合计有 57 件专利公开，主要涉及神经网络矩阵乘法，块划分等相关技术。

3.2.2.5 关键技术分析

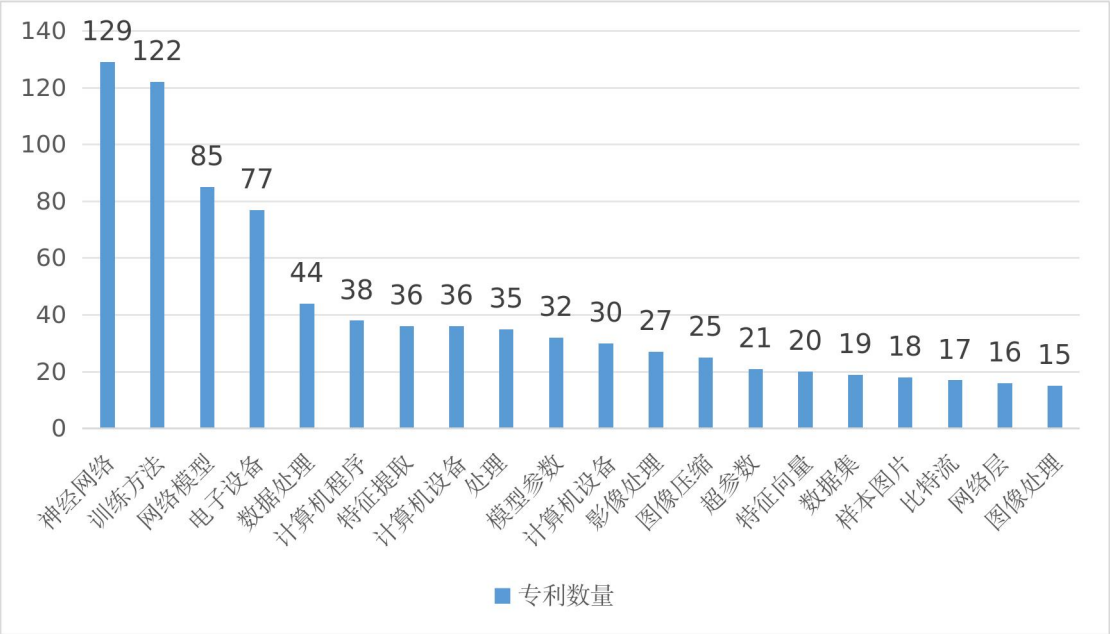


图 3-29 腾讯 AI 模型优化技术主题分布图

腾讯在 AI 模型优化技术的专利申请呈现多元布局，涵盖神经网络（129 件）、训练方法（122 件）、网络模型（85 件）等关技术分支，整体来看，腾讯在 AI 模型技术方向更倾向于特征提取、模型参数、超参数等相关技术研发，在应用上更侧重于云计算等计算程序实际落地场景。

进一步的对腾讯的以上技术进行技术归类，得到图 3-29。腾讯的 AI 模型优化技术涵盖了 4 个技术分支。其中知识蒸馏技术专利较多，为 141 族，占比 40%；其次为参数优化，合计 99 族专利，占比 28%；参数压缩则有 87 族专利申请，占比 24%。模型结构优化技术再次之，专利申请为 28 族。

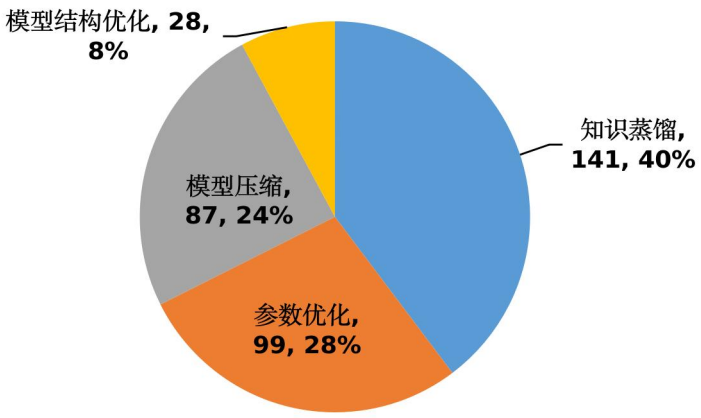


图 3-30 腾讯技术分布

3.2.3 百度

3.2.3.1 公司简介

百度是中国领先的互联网科技公司，成立于 2000 年，总部位于北京，创始人为李彦宏。百度最初以搜索引擎业务闻名，随后逐步扩展至云计算、人工智能（AI）、自动驾驶、智能硬件等多个领域。

百度是中国 AI 领域的领军企业之一，早在 2013 年就成立了深度学习研究院（IDL），并在 2017 年正式宣布“All in AI”战略，重点布局人工智能技术。其在 AI 技术方向的布局主要有：（1）百度大脑：是百度的核心 AI 技术平台，整合了深度学习、自然语言处理（NLP）、计算机视觉、语音识别等技术，为企业和开发者提供 AI 能力支持。（2）文心大模型：百度研发的文心大模型是中国领先的大语言模型之一，具备更强的多模态理解与生成能力，支持文本、图像、视频等跨模态交互。（3）百度云：百度智能云整合了 AI 能力，提供云计算、大数据分析、AI 模型训练等服务，在金融、制造、城市管理等领域落地。

3.2.3.2 专利申请趋势分析

图 3-31 百度专利申请趋势

百度在全球共申请 AI 模型优化技术相关专利合计 315 件。为了进一步研究百度的专利申请发展情况，以下对其全球专利申请数据按时间进行统计分析。百度在 AI 模型优化领域自 2011 年开始申请专利,其中最早的专利主要涉及声音模型识别。自 2017 年开始，加大了对 AI 模型优化技术的研发投入，特别是在模型训练、模型压缩等方向，到 2020 年-2023 年，在 AI 模型技术方向的专利申请一直维持在 50 件左右。近两年专利公开有所减少，主要是专利申请的公开时滞原因。

3.2.3.3 专利区域和法律状态分析

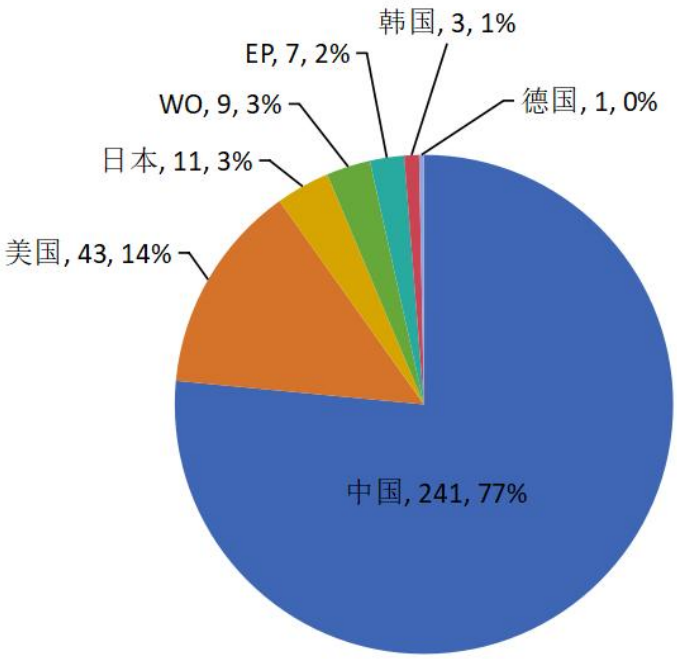


图 3-32　百度 AI 模型优化专利区域分布

图 3-32 显示了百度的 AI 模型优化技术专利区域分布及相关法律状态。百度在全球 7 个国家/地区/组织进行了专利申请。其中在中国大陆的专利申请数量排在第一位，合计 241 件，占比 77%；其次为美国和日本，专利申请分别为 43 件和 11 件。此外，其在 EP 和韩国也有部分专利布局。

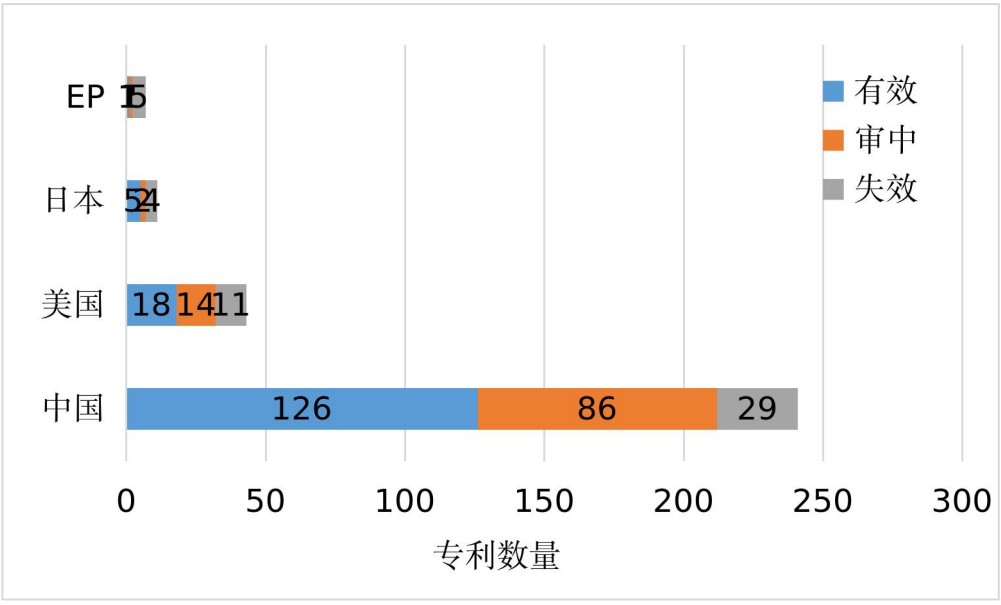


图 3-33 区域法律状态分布情况

结合专利区域法律状态来看，百度的有效和在审专利主要集中在中国大陆，表明腾讯仍以中国大陆为主要专利布局区域。此外，其在美国也存在较多有效专利和在审专利。从百度的专利布局策略来看，其主要在中美日进行专利布局。

3.2.3.4 主要发明人分析

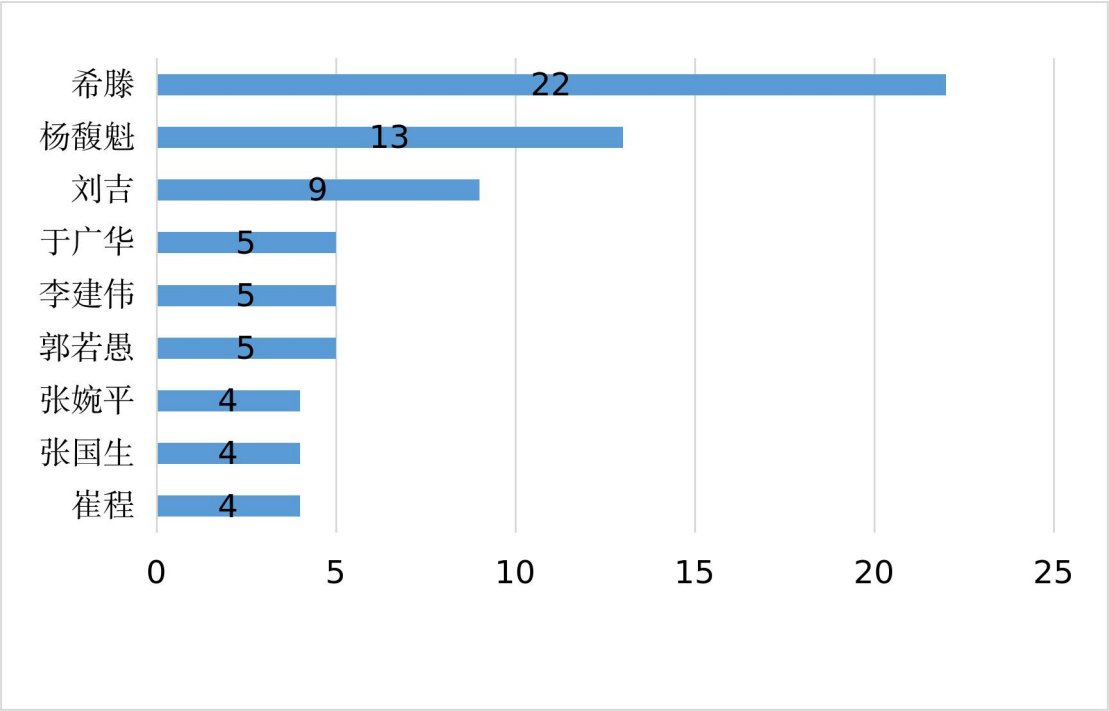


图 3-34 主要发明人

图 3-34 是百度专利申请量在 4 件以上的第一发明人基本情况。百度的 AI 模型优化专利发明团队较为集中，主要涉及三个核心发明人，其中排名第一的是希滕，合计公开专利 22 件，其主要涉及模型蒸馏、剪枝以及超参数等方向；排名第二的是杨馥魁，合计公开专利 13 件，主要涉及模型蒸馏和图形训练等方向；排名第三的刘吉，合计公开专利 9 件，主要涉及机器学习类模型优化技术。

3.2.3.5 关键技术分析

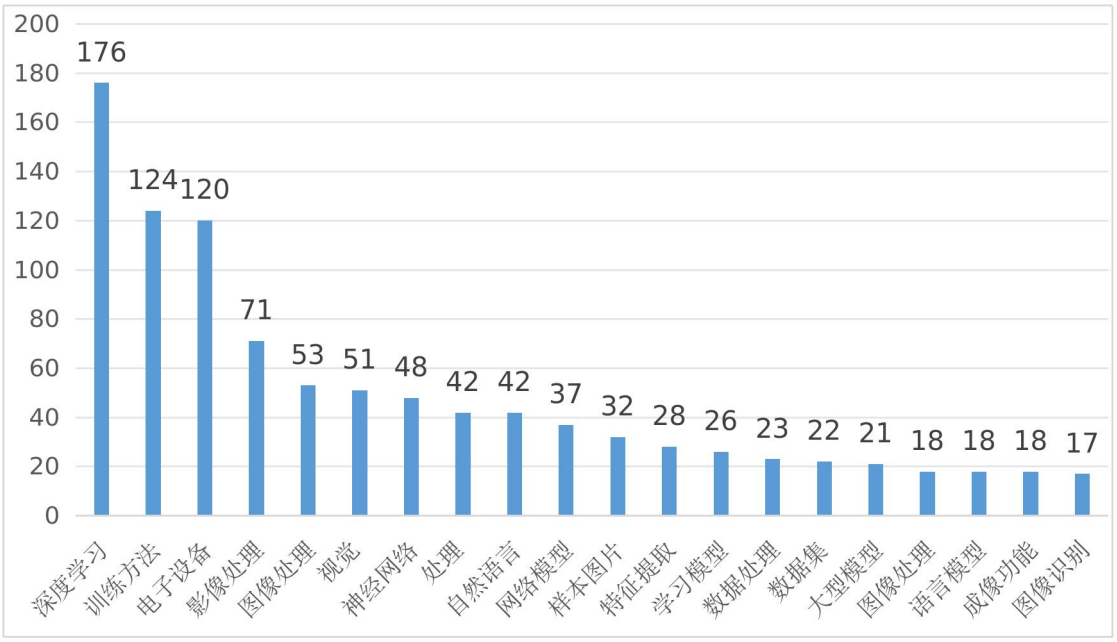


图 3-35 百度 AI 模型优化技术主题分布图

百度在 AI 模型优化技术的专利申请呈现多元布局，涵盖深度学习（176 件）、训练方法（124 件）等关技术分支，整体来看，百度在 AI 模型技术方向更倾向于深度学习等相关技术研发，尤其是在图像和语言处理方向的应用。

进一步的对百度的以上技术进行技术归类，得到图 3-36。百度的 AI 模型优化技术涵盖了 4 个技术分支。其中知识蒸馏技术专利较多，为 123 族，占比 50%；其次为参数优化，合计 48 族专利，占比 20%；模型结构优化则有 38 族专利申请，占比 15%。模型压缩技术再次之，专利申请为 37 族。

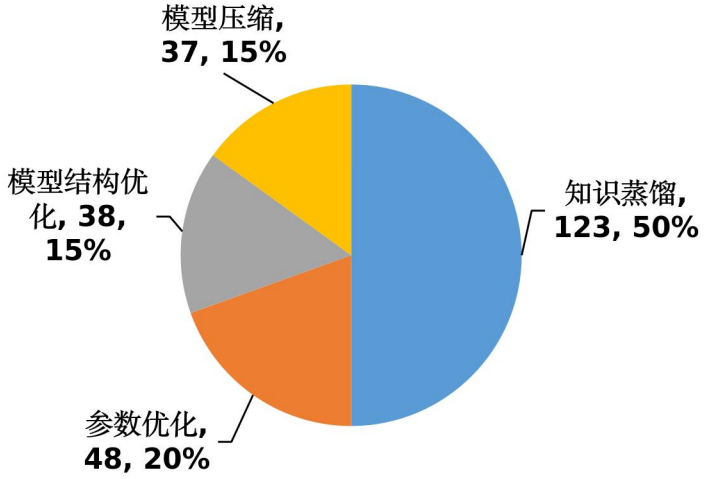


图 3-36　百度技术分布

3.3 权利人对比分析

通过对国内外重点专利权人的专利布局进行宏观分析、技术分析、代表性专利分析等，已经得到它们的专利申请量变化趋势、总体区域分布、代表性专利等具有丰富内容的信息，本节将对这些信息进行综合对比整理，以便进一步掌握目前 AI 模型优化技术领域厂商、技术、市场等要素的配置状况。

3.3.1 专利实力与目标市场比较

表 3-7 主要专利权人专利情况一览

排名	申请人	国家/地区	总申请量	有效专利数	区域分布	TOP 布局国家	最早申请时间	快速增长阶段
1	腾讯	中国	618	286	11	CN、HK、US、EP、JP	2017	2018-2023
2	华为	中国	553	71	19	CN、US、EP	2006	2019-2023
3	三星	韩国	496	116	14	US、KR、CN、EP	1997	2017-2023
4	IBM	美国	418	123	16	US、JP、CN、UK、DE	1988	2020-2023
5	微软	美国	361	88	14	US、EP、CN、IN	1998	2017-2023
6	百度	中国	315	151	7	CN、US、JP、KR	2011	2020-2023

从专利数量上看，腾讯以 618 件的总量位列第一，其次是华为和三星。结合有效专利来看，则是腾讯、百度、IBM 和三星有效专利较多。从布局区域重视度来看，主要权利人倾向性较强，其中国权利人主要以本土布局为主，三星、IBM 和微软则主要布局在美国。进一步除上述国家外，不同权利人存在些许差异，其中腾讯侧重于中国香港和日本；华为、三星、IBM 侧重于欧洲；微软侧重于印度；百度则侧重于韩国。从全球布局策略来看，除百度外，其他权利人全球化布局更广泛，均在 10 个国家/地区以上。从 AI 模型优化技术专利布局时间来看，中国权利人起步较晚，但在近几年加大了对 AI 技术研发，在专利申请上部分已超过国外权利人。

3.3.2 技术分布比较

表 3-8 主要技术布局

排名	申请人	模型蒸馏（族）	参数优化（族）	结构优化（族）	模型压缩（族）
1	腾讯	141	99	28	87
2	华为	101	42	41	63
3	三星	90	50	46	38
4	IBM	35	90	80	12
5	微软	29	54	37	23
6	百度	123	48	38	37

如表 3-8 所示，国内权利人和三星的研究重点集中在模型蒸馏方向上，其中腾讯除模型蒸馏外，主要布局在参数优化和模型压缩方向上；华为则主要集中在模型压缩上；百度主要集中在参数优化方向；三星则侧重参数优化和结构优化。IBM 和微软则主要布局在参数优化和结构优化方向。

3.3.2 技术分支重点布局技术方向比较

表 3-9 知识蒸馏技术方案

知识蒸馏	技术方案
腾讯	助理模型分别对教师和学生模型训练；特征权重矩阵计算损失函数；压缩师生网络特征
华为	不同数据集对教师和学生模型训练；调整中间层和输出层；端侧设备上训练两个模型
三星	训练冷凝器模型预先训练师生模型之间的映射函数；梯度函数确定教师模型输出向量的重要性
IBM	软标签以统一训练模型；自动蒸馏动态更新知识库
微软	教师模型和学生模型并联训练；建立实时资源博弈控制器；AMI 将任务目标和平衡目标结合
百度	特征向量输入不同模型的编码层；动态调整增广概率和目标权重

表 3-10 参数优化技术方案

参数优化	技术方案
腾讯	动态调整超参数搜索范围；采用遗传算法将神经网络超参数二进制化为基因序列；对权重系数进行量化和熵编码
华为	获取多组超参数，绘制损失函数变化地形图对超参数进行训练；选择性地将神经网络的滤波器权重量化；在二元卷输出上应用可训练缩放因子
三星	正则化空间阈权重，使用稀疏空间卷积压缩模型；量化误差减少正则化

	损失
IBM	三元组损失正则化对样本进行池化迭代；超参数调谐资源分配
微软	稀疏正则化；迭代评估和有前景的超参数提高参数调整效率；正则化对偶平均
百度	生成线性回归空间，利用训练和验证子集对正则化评估；逐步优化每一步骤的超参数组合

表 3-11 结构优化技术方案

结构优化	技术方案
腾讯	根据通过率对隐藏层进行剪枝处理；获得剪枝模型和多个输入数据，利用子模型进行预测；预训练中增加辅助参数
华为	半结构化稀疏剪枝；机器学习算法修正原始特征参数，优化模型以获得个性化新参数；
三星	零跳跃方法处理结构化剪枝；自动逐层剪枝
IBM	机器学习神经元权重剪枝；按照预测精度对模型设置置信度
微软	使用数据感知模型修剪技术，基于激活矩阵的熵值对神经网络进行修剪；基于权重的优化值选择特征嵌入大小进行训练
百度	BN 阈值搜索空间，生成剪枝策略编码生成器；剪枝掩码用于标识剪枝的自注意力头

表 3-12 模型压缩技术方案

结构压缩	技术方案
腾讯	多个参数维度以此进行压缩；四位参数张量重塑为三位参数张量；权重系数在超级块内统一以最小化损失
华为	利用预测标签和训练数据确定损失函数，对模型进行评估和更新；最大化平衡精度、压缩比和处理时间性能函数来计算最佳分解等级
三星	特征图压缩和重建；二分近似量化参数
IBM	无损和有损压缩分析冗余节点；平均边界以优化权重；识别关键参数和微调随机生成值来预训练
微软	机器学习联合训练；浮点数分离器提取尾数值和指数值；在神经网络中引入压缩块和解压缩块；
百度	通过 DDPG 算法为深度神经网络的权重张量进行分解；利用第一、二比特宽度和目标稀疏率对模型进行压缩；对单精度浮点数据的尾数进行压缩

3.4 本章小结

腾讯在 AI 模型优化专利申请量排名第一，合计申请专利达到了 618 件，在 AI 模型优化领域尤其是知识蒸馏技术上具有较强的专利控制力，和全球布局能力。华

为则排名第二，专利申请为 553 件，在 AI 模型优化技术方向的多个技术路线进行尝试，知识蒸馏、模型压缩等。三星电子排名第三，专利申请为 496 件，而 IBM 专利次之为 418 件。

从专利申请趋势来看，IBM、微软和三星电子国外权利人起步较早，均在 2000 年以前就开始进行 AI 模型优化技术研发和专利申请。而国内权利人华为和腾讯、百度最早的专利申请均在 2010 年左右，起步相对较晚。整体来看，六位权利人的专利申请高峰集中在 2017-2023 年，这得益于 Transformer 架构问世，使得 AI 大模型逐渐从实验室走向商用。

结合专利布局和专利法律状态来看，六位权利人均较为重视 AI 模型优化技术的全球化布局。但主要权利人倾向性较强，其中国内权利人主要以本土布局为主，三星、IBM 和微软则主要布局在美国。进一步除上述国家外，不同权利人存在些许差异，其中腾讯侧重于 HK 和日本；华为、三星、IBM 侧重于欧洲；微软侧重于印度；百度则侧重于韩国。从从全球布局策略来看，除百度外，其他权利人全球化布局更广泛，均在 10 个国家/地区以上。

从 AI 模型优化技术专利布局来看，三星电子主要关注知识蒸馏技术，例如利用训练冷凝器模型以学习预先训练的教师模型和学生模型之间的参数映射函数优化模型等；IBM 和微软则更为重视参数优化技术，其中 IBM 在正则化技术上，使用三元组损失正则化对样本池细化迭代。在超参数上，通过超参数分配器，利用张量交换和虚拟加速器管理有限计算资源等；而微软在正则化技术上，使用结合基于矫正方法的好处和近端随机梯度方法来实现有效的模型压缩。在超参数上，通过迭代评估和关注有前景的超参数集来提高超参数调整效率等；

华为、腾讯、百度和三星电子同样关注知识蒸馏技术，其中华为主要使用教师模型的无偏数据或者是中间层和输出层的输出数据改善学生模型；而腾讯则主要利用权重矩阵或者与预训练改善知识蒸馏；百度则主要涉及利用特征向量输入不同模型的编码层进行汇聚和处理，以及动态调整增广概率和目标权重来解决模型欠拟合的问题。

第四章 AI 模型优化特定技术专题分析

本章主要针对 AI 模型优化技术的四个主要方向进行分析，分别对模型结构优化、模型压缩、参数优化和知识蒸馏分别予以介绍，深入刨析各技术分支不同权利人相关专利。

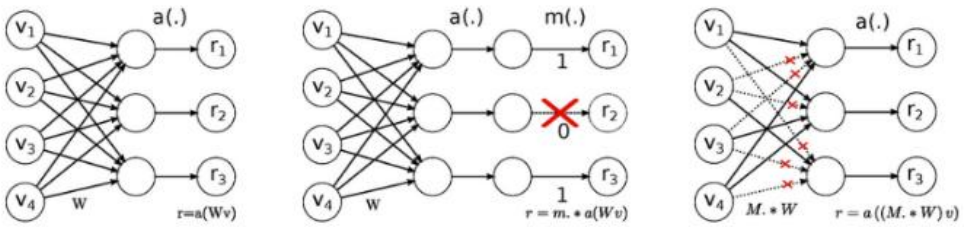
本章内容在挑选重点专利时，主要依据重点专利权人近十年重点专利及关键技术代表专利、专利同族数、同族专利被引证数，兼顾专利法律状态等进行挑选和解读。

4.1 AI 模型结构优化

4.1.1 AI 模型结构优化简介

优化模型结构不仅可以提高算法的准确性，还能增强模型的泛化能力，使其在面对新数据时仍能保持良好的表现。常见的方法是剪枝，由于深度学习网络模型从卷积层到全连接层存在着大量冗余参数，大量神经元激活值趋近于 0，将这些神经元去除后可以表现出通用的模型表达能力，这种情况被称为过参数化，而对应的技术称为模型剪枝。通过去除冗余节点或降低权重精度来减小模型规模，提高推理速度。

机器学习中，Dropout 和 DropConnect 代表着非常经典的模型剪枝技术，Dropout 中随机的将一些神经元的输出置零，这就是神经元剪枝。DropConnect 则随机的将一些神经元之间的连接置零，使得权重连接矩阵变得稀疏，这便是权重连接剪枝。它们就是最细粒度的剪枝技术，只是这个操作仅仅发生在训练中，对最终的模型不产生影响，因此没有被称为模型剪枝技术。



No-Drop Network　　DropOut Network　　DropConnect Network

图 4-1　Dropout 和 DropConnect 过程

模型剪枝不仅仅只有对神经元的剪枝和对权重连接的剪枝，可以粗分为 4 个粒度：细粒度剪枝(fine-grained)：即对连接或者神经元进行剪枝，它是粒度最小的剪枝；向量剪枝(vector-level)：它相对于细粒度剪枝粒度更大，属于对卷积核内部(intra-kernel)的剪枝；核剪枝(kernel-level)：即去除某个卷积核，它将丢弃对输入通道中对应计算通道的响应；滤波器剪枝(Filter-level)：对整个卷积核组进行剪枝，会造成推理过程中输出特征通道数的改变。

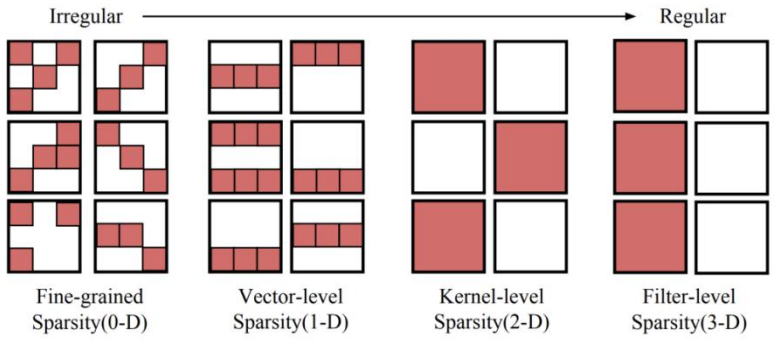
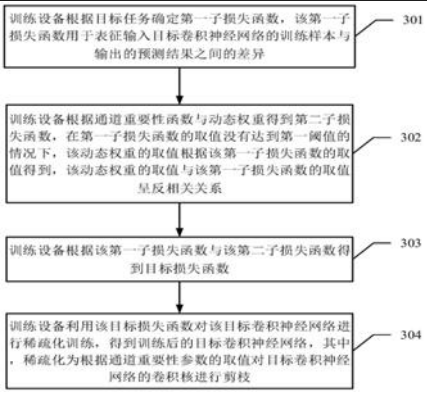


图 4-2 剪枝粒度

4.1.2 重点专利技术解读

1）结构优化相关专利解读之 CN113065636B

申请号	CN202110221926.3	申请日	2021-02-27
公开号	CN113065636B	法律状态/事件	授权
标题	一种卷积神经网络的剪枝处理方法、数据处理方法及设备		
申请(专利权)人	华为技术有限公司		
发明目的	通过构建新的目标损失函数，结合通道重要性函数和动态权重，实现卷积神经网络的灵活剪枝，解决了传统方法中模型压缩效果不佳的问题，提升了卷积神经网络在移动设备上的应用性。		
主要附图			
第一权利要求	1.一种卷积神经网络的剪枝处理方法，其特征在于，包括：		

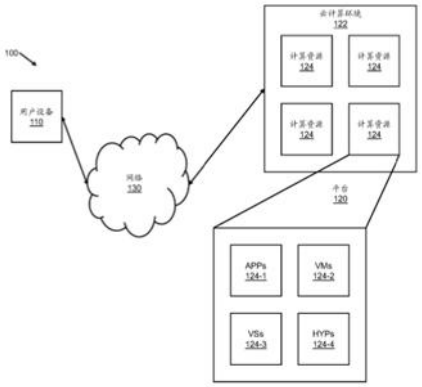
	根据目标任务确定第一子损失函数，所述第一子损失函数用于表征输入目标卷积神经网络的训练样本与输出的预测结果之间的差异，所述目标卷积神经网络为用于进行训练的卷积神经网络，所述目标卷积神经网络在针对不同训练样本的训练过程中具有不同的网络参数和损失函数，所述训练样本为训练集中的样本，所述训练样本为如下任一种：图像、视频、文本或音频； 根据通道重要性函数与动态权重得到第二子损失函数，在所述第一子损失函数的取值没有达到第一阈值的情况下，所述动态权重的取值根据所述第一子损失函数的取值得到，所述动态权重的取值与所述第一子损失函数的取值呈反相关关系； 根据所述第一子损失函数与所述第二子损失函数得到目标损失函数； 利用所述目标损失函数对所述目标卷积神经网络进行稀疏化训练，得到训练后的目标卷积神经网络，所述稀疏化为根据所述通道重要性函数的取值对所述目标卷积神经网络的卷积核进行剪枝； 将所述训练后的目标卷积神经网络部署于计算能力有限的边缘设备中的处理器上，以使得所述边缘设备中的处理器通过所述训练后的目标卷积神经网络对每个输入样本实现动态剪枝，部署有所述训练后的目标卷积神经网络的所述边缘设备用于处理图像、文本或语音相关的所述目标任务，所述目标任务至少包括分类任务、分割任务、检测任务。

发明专利 CN113065636B 主要针对卷积神经网络在移动设备上应用受限，由于模型大小大、计算复杂度高，传统的静态通道控制因子剪枝方法忽略了输入数据对网络参数和容量的不同需求，导致模型压缩效果不佳的问题进行探究。

华为技术有限公司运用了构建新的目标损失函数，通过两个子损失函数组成，第一子损失函数衡量训练样本的复杂度，第二子损失函数衡量剪枝后的子网络复杂度，动态权重根据通道重要性函数与损失函数关系，实现针对每个训练样本的灵活剪枝。能够实现针对每个训练样本的灵活剪枝，充分挖掘卷积神经网络的冗余，降低模型大小和计算复杂度，使得卷积神经网络在计算资源受限的设备上更具应用性。

2）结构优化相关专利解读之 CN114616575A

申请号	CN202180005978.5	申请日	2021-06-15
公开号	CN114616575A	法律状态/事件	实质审查
标题	用于采用微结构化权重剪枝和权重统一进行神经网络模型压缩的方法及装置		
申请(专利权)人	腾讯美国有限责任公司		
发明目的	通过微结构化权重剪枝和权重统一的方法，优化 DNN 模型，解决了		

	现有技术中模型压缩效率低的问题，实现了模型尺寸和计算速度的提升。
主要附图	
第一权利要求	1.一种神经网络模型压缩的方法，所述方法由至少一个处理器执行，并且所述方法包括： 接收输入神经网络和输入掩膜； 使用深度神经网络减少所述输入神经网络的参数，所述深度神经网络通过以下步骤进行训练： 从所述深度神经网络的、被所述输入掩膜所掩膜的输入权重的多个块中，选择待剪枝的剪枝微结构块； 基于选择的剪枝微结构块，对所述输入权重进行剪枝； 从被所述输入掩膜所掩膜的输入权重的多个块中，选择待统一的统一微结构块；以及 基于选择的统一微结构块，对已剪枝的输入权重的多个块中的一个或多个块中的多个权重进行统一，以得到所述深度神经网络的已剪枝且已统一的输入权重；以及 基于所述输入神经网络和所述深度神经网络的已剪枝且已统一的输入权重，得到具有减少的参数的输出神经网络。

发明专利CN114616575A主要针对现有技术难以有效压缩深度神经网络(DNN)模型，导致存储和计算需求增加，影响视频应用的效率的问题进行探究。

腾讯美国有限责任公司运用了采用微结构化权重剪枝和权重统一的方法，通过迭代网络再训练和微调框架，优化剪枝损失和统一损失，减少模型参数并保持预测性能，进而实现模型的压缩。能够显著减小模型尺寸，提高了计算速度，几乎不影响原始DNN模型的预测性能，适用于移动设备和芯片等场景。

4.2 AI模型压缩

4.2.1 AI模型压缩简介

模型压缩，就是在尽可能不改变模型效果的条件下，通过减少模型的体积，使得模型在使用的时候有更快的速度。常见的模型压缩主要有模型量化等。模型量化

即将网络的权值、激活值等从高精度转换为低精度的操作过程，例如将 32 位浮点数转化成 8 位整型数 int8，同时我们期望转换后的模型准确率与转化前相近。

模型量化可以带来几方面的优势：（1）更小的模型尺寸。以 8bit 量化为例，与 32bit 浮点数相比，我们可以将模型的体积降低为原来的四分之一，这对于模型的存储和更新来说都更有优势。（2） 更低的功耗。移动 8bit 数据与移动 32bit 浮点型数据相比，前者比后者高 4 倍的效率，而在一定程度上内存的使用量与功耗是成正比的。（3）更快的计算速度。相对于浮点数，大多数处理器都支持 8bit 数据的更快处理，如果是二值量化，则更有优势。

下图展示的是一个二值权重和激活值矩阵的运算，卷积过程中的乘加都可以转换为异或操作，并行程度更高，运算速度因此也更快。

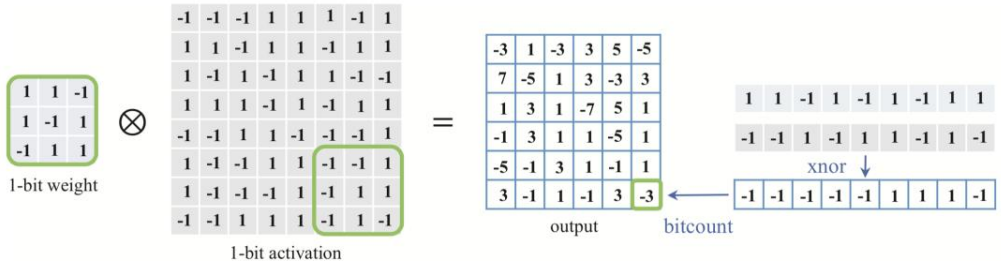
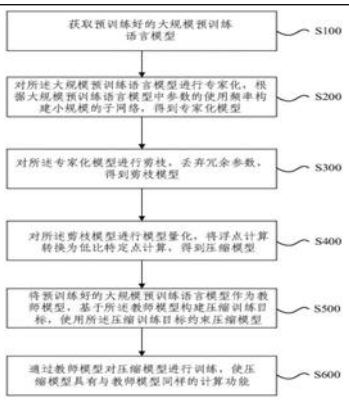


图 4-3 二值权重和激活值矩阵的运算

4.2.2 重点专利技术解读

1）模型压缩相关专利解读之 CN114742036B

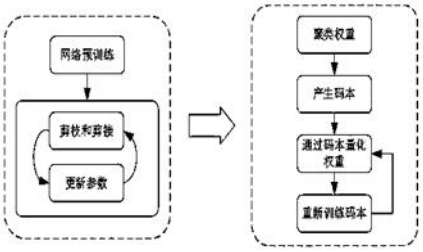
申请号	CN202210283763.6	申请日	2022-03-21
公开号	CN114742036B	法律状态/事件	授权
标题	一种预训练语言模型的组合式模型压缩方法及系统		
申请(专利权)人	清华大学		
发明目的	通过对大规模预训练语言模型进行模型专家化、剪枝和量化，结合教师模型进行监督训练，解决了预训练语言模型计算资源占用多、成本高的问题，实现了模型的高效压缩和性能保留，适用于多种自然语言处理任务。		

主要附图	
第一权利要求	1.一种预训练语言模型的组合式模型压缩方法，其特征在于，所述方法包括： 获取预训练好的大规模预训练语言模型，给句子中的每个词赋予一个静态向量，然后与句子的上下文词进行交互，最后就得到了词在句子中的变化的词向量； 对所述大规模预训练语言模型进行专家化，根据大规模预训练语言模型中参数的使用频率构建小规模的子网络，得到专家化模型； 对所述专家化模型进行剪枝，丢弃冗余参数，得到剪枝模型； 对所述剪枝模型进行模型量化，将浮点计算转换为低比特定点计算，得到压缩模型； 将预训练好的大规模预训练语言模型作为教师模型，基于所述教师模型构建压缩训练目标，使用所述压缩训练目标约束压缩模型； 通过教师模型对压缩模型进行训练，使压缩模型具有与教师模型同样的计算功能； 其中，对所述大规模预训练语言模型进行专家化，根据大规模预训练语言模型中参数的使用频率构建小规模的子网络，得到专家化模型，具体包括： 每个输入大规模预训练模型的数据在进行计算时，使用不同参数参与计算； 统计大规模预训练模型中不同参数共同被使用的频率； 基于共同使用的频率对参数进行划分，共同使用频率高的参数划分在一起，组成小规模的子网络； 根据划分好的子网络，构建路由网络，对于不同的输入数据选择不同的子网络，得到专家化模型； 通过教师模型对压缩模型进行训练，使压缩模型具有与教师模型同样的计算功能，以得到目标压缩模型，具体包括： 将所述教师模型的输出结果作为一个新的训练信号； 基于所述教师模型输出的新的训练信号对压缩模型进行训练，以形成对压缩模型的监督； 所述压缩模型结合新的训练信号形成的监督，拟合教师模型的输出结果，训练完成后压缩模型具有教师模型的计算功能。

发明专利CN114742036B主要针对大规模预训练语言模型在内存有限设备中部署时计算资源占用多、成本高，现有语言模型压缩方法主要针对特定任务，且人工设计知识蒸馏策略困难，导致模型微调费时费力，计算成本高的问题进行探究。

清华大学运用了采用预训练语言模型的组合式模型压缩方法，包括模型专家化、剪枝和量化，通过构建小规模子网络、丢弃冗余参数、将浮点计算转换为低比特计算，并利用教师模型进行监督训练，形成压缩模型，能够显著提高了模型的数据处理效率，最大限度保留了原始模型的性能，提升了模型的实用性和灵活性，适用于多种下游自然语言处理任务。

2）模型压缩相关专利解读之CN110210618A

申请号	CN201910427397.5	申请日	2019-05-22
公开号	CN110210618A	法律状态/事件	驳回
标题	动态修剪深度神经网络权重和权重共享的压缩方法		
申请(专利权)人	东南大学		
发明目的	通过动态修剪和权重共享的方法，结合K-Means量化和L1正则化，解决了深度神经网络在高压缩率下精度稳定的问题，实现了对AlexNet网络的52倍压缩，适用于移动端应用。		
主要附图			
第一权利要求	1.动态修剪深度神经网络权重和权重共享的压缩方法，其特征在于：所述的方法包括如下步骤： (1)将深度神经网络模型进行预训练； (2)根据权值重要性进行网络修剪和网络剪接； (3)将修剪和剪接后的模型参数进行一次更新； (4)迭代层重复步骤(2)和(3)，完成网络的权重修剪操作； (5)初始化K-Means质心； (6)确定量化阈值； (7)微调量化后的网络，完成网络权重共享操作。		

发明专利CN110210618A主要针对现有深度神经网络的压缩方法在高压缩率下难以保证模型精度的稳定，且对计算资源和存储的需求仍未得到有效满足的问题进行探究。

东南大学运用了采用动态修剪深度神经网络权重和权重共享的压缩方法，通过剪枝和剪接操作减少冗余参数，并在修剪后进行 K-Means 量化，结合 L1 正则化以提高模型精度和学习效率，能够在不损失模型精度的情况下，显著减少深度神经网络的参数冗余，降低存储内存需求，并将 AlexNet 网络压缩了 52 倍，适用于移动端应用。

4.3 AI 模型参数优化

4.3.1 AI 模型参数优化简介

在深度学习和人工智能领域，通过参数调整进行模型的性能优化是一个持续的研究课题。超参数化模型与正则化是其中两个关键的优化手段，它们对于提升 AI 模型的性能起着至关重要的作用。

首先，超参数是模型架构之外，需要通过经验或搜索来确定的参数。它们通常在模型训练之前设定，而不是通过学习过程得到。常见的超参数包括学习率、批量大小、迭代次数、优化器类型等。超参数的设置直接影响到模型的收敛速度、准确性和泛化能力。适当的超参数调整可以使模型更快地收敛到最优解，提高模型的性能。譬如，学习率是最为关键的，它决定了模型每次更新权重时的步伐。学习率过高可能导致模型震荡，无法收敛；学习率过低则可能导致训练过程缓慢，收敛速度慢。目前常见的超参数调整方法主要有经验设定、网络搜索、随机搜索和贝叶斯优化四种。

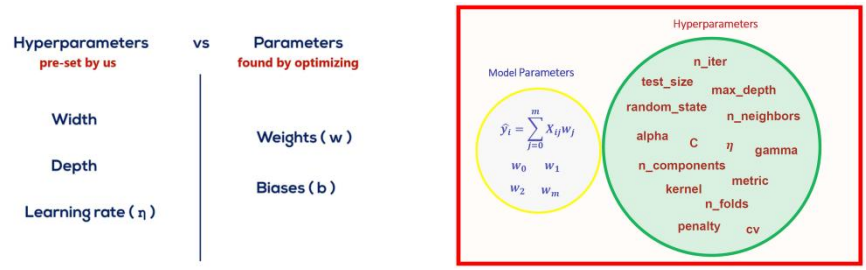


图 4-4 模型参数和超参数区别

而正则化则是一种防止模型过拟合的技术，通过在损失函数中添加正则化项来约束模型参数，从而降低模型复杂度。常见的正则化方法主要有 L1 正则化（Lasso）：通过在损失函数中添加绝对值项，使模型参数向零值逼近；L2 正则化（Ridge）：

通过在损失函数中添加平方项，使模型参数向较小值逼近；Dropout：在训练过程中随机丢弃部分神经元，降低模型复杂度。

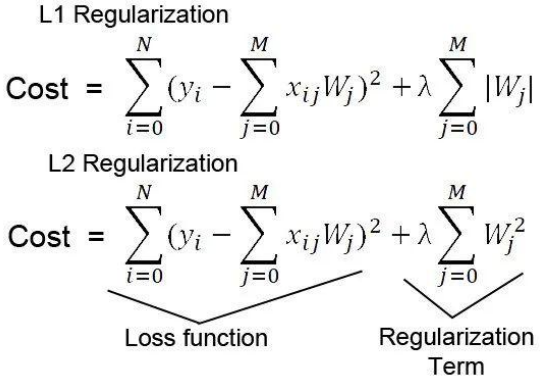
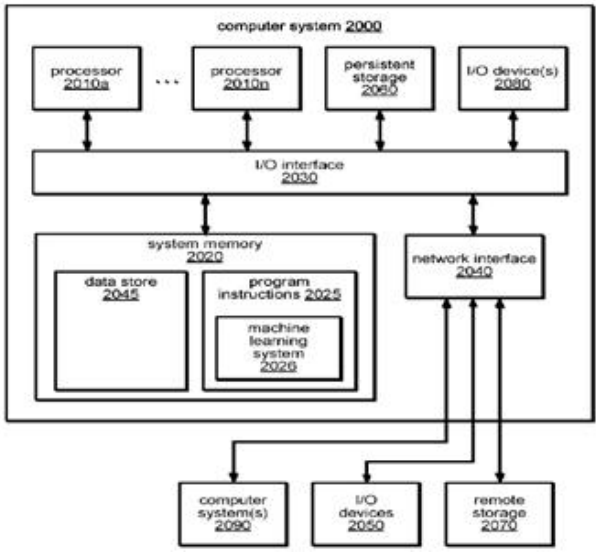


图 4-5 L1 和 L2 正则化

4.3.2 重点专利技术解读

1）参数优化相关专利解读之 US20230237336A1

申请号	US18/157755	申请日	2023-01-20
公开号	US20230237336A1	法律状态/事件	实质审查
标题	Self-Pruning Neural Networks with Regularized Auxiliary Variables		
申请(专利权)人	甲骨文国际公司		
发明目的	带有辅助参数的自剪枝神经网络解决了深度学习框架中减少不必要参数的挑战,通过动态剪枝神经元实现高效计算和高质量模型,从而克服现有技术的局限性。		
主要附图			
第一权利要求	1. A method comprising: training a neural network comprising a plurality of neuron layers, the training comprising performing, for a training batch of a plurality of training batches:		

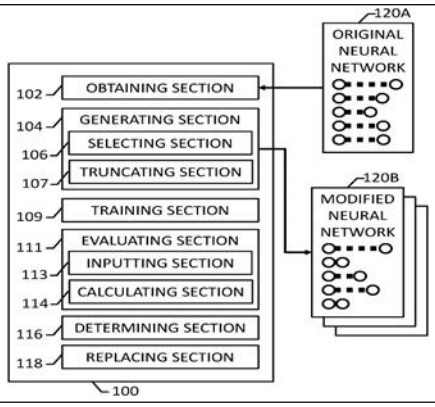
	training respective layers of the plurality of layers according to the training batch, wherein individual ones of the respective layers respectively comprise one or more neurons individually comprising a set of auxiliary parameters and a set of weighting factors, and wherein training individual ones of the respective layers updates the respective sets of auxiliary parameters and the respective sets of weighting factors of the individual ones of the one or more neurons in the respective layers; identifying one or more neurons for deletion according to the respective sets of auxiliary parameters and a regularization penalty for the neural network; and deleting the identified one or more neurons from the neural network prior to completion of the training batch.

发明专利 US20230237336A1 主要针对深度学习框架依赖于大型、过度参数化的神经网络,这些网络的训练和部署成本高昂,需要大量的计算资源和专业知识,而 L0 正则化限制了参数的数量,尽管有其理论动机,但在实践中很难实现的问题进行探究。

甲骨文国际公司运用了自剪枝神经网络的实现,在训练过程中使用辅助参数来识别和删除神经元,允许动态剪枝并实现 L0 正则化,从而减少网络参数数量并提高计算效率。这种方法可以减少网络参数,从而实现更高效的计算和高质量的模型性能,而无需完全重新训练,同时保持预测准确性并减少资源需求。

2）参数优化相关专利解读之 US10902311B2

申请号	US15/392041	申请日	2016-12-28
公开号	US10902311B2	法律状态/事件	未缴年费 | 期限延长
标题	Regularization of neural networks		
申请(专利权)人	国际商业机器公司		
发明目的	通过截断 FIFO 队列长度并评估其影响,该方法解决了具有 FIFO 队列的神经网络中的泛化问题,提高了准确性并防止过度拟合,特别是在动态玻尔兹曼机中。		

主要附图	
第一权利要求	1. A computer-implemented method comprising: obtaining an original neural network having a plurality of first-in-first-out (FIFO) queues, each FIFO queue located between a pair of nodes among a plurality of nodes of the original neural network; generating at least one modified neural network, the modified neural network being equivalent to the original neural network with a modified length of at least one FIFO queue, wherein the generating comprises: selecting the at least one FIFO queue among the plurality of FIFO queues in the original neural network to modify, wherein the selecting is based on a random variable; truncating the at least one FIFO queue to a designated length, wherein the designated length is greater than zero; and determining that each of the selected at least one FIFO queues have been truncated to the designated length; training each neural network among the original neural network and the at least one modified neural network; evaluating each neural network among the original neural network and the at least one modified neural network; and determining which neural network among the original neural network and the at least one modified neural network is most accurate, based on the evaluation.

发明专利 US10902311B2 主要针对现有的正则化方法对于具有先进先出(FIFO)队列的神经网络来说是无效的,特别是在动态玻尔兹曼机等深层架构中,当使用有限的数据进行训练时,会导致泛化问题和过度拟合问题的问题进行探究。

国际商业机器公司运用了一种生成修改神经网络的方法,通过将随机 FIFO 队列的长度截断为零,评估其与原始网络相比的准确性,如果更准确则替换它们,从而防止过度拟合并提高泛化性能。这种方法能够使用 FIFO 队列有效地规范神经网络,提高其预测准确性并减少内存所需的计算资源,特别是在时间序列的无监督学习中。

4.4 知识蒸馏

4.4.1 知识蒸馏简介

知识蒸馏就是把一个大的教师模型的知识萃取出来，浓缩到一个小的学生模型，可以理解为一个大的教师神经网络把他的知识教给小的学生网络，这里有一个知识的迁移过程，从教师网络迁移到了学生网络身上，教师网络一般是比较臃肿，所以教师网络把知识教给学生网络，学生网络是一个比较小的网络，这样就可以用学生网络去做一些轻量化网络做的事情。

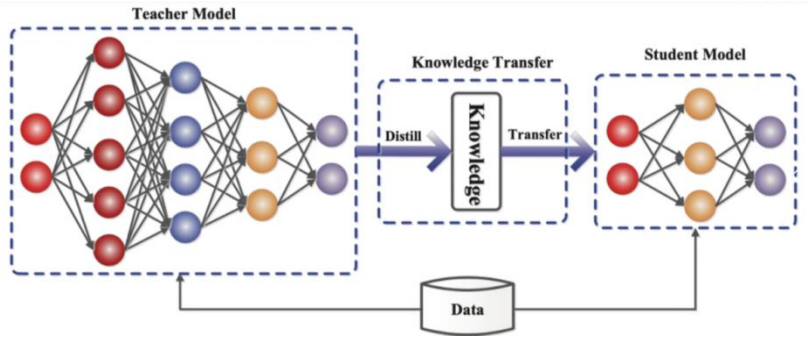


图 4-6 知识蒸馏过程

知识蒸馏使用的是 Teacher—Student 模型，其中 teacher 是知识的输出者，student 是知识的接受者。知识蒸馏的过程分为 2 个阶段:

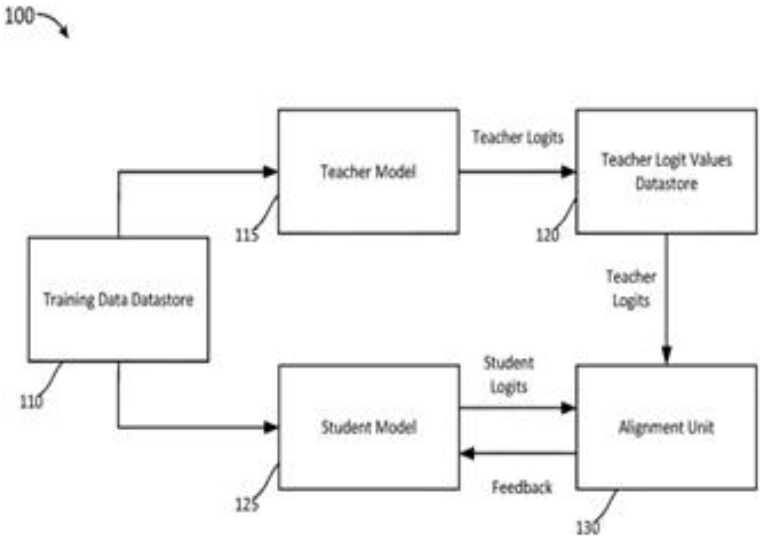
1）原始模型训练: 训练"Teacher 模型", 简称为 Net-T，它的特点是模型相对复杂，也可以由多个分别训练的模型集成而成。我们对"Teacher 模型"不作任何关于模型架构、参数量、是否集成方面的限制，唯一的要求就是，对于输入 X, 其都能输出 Y，其中 Y 经过 softmax 的映射，输出值对应相应类别的概率值。

2）精简模型训练: 训练"Student 模型", 简称为 Net-S，它是参数量较小、模型结构相对简单的单模型。同样的，对于输入 X，其都能输出 Y，Y 经过 softmax 映射后同样能输出对应相应类别的概率值。

4.4.2 重点专利技术解读

1）知识蒸馏相关专利解读之 US12307750B2

申请号	US17/837636	申请日	2022-06-10
公开号	US12307750B2	法律状态/事件	授权 | 期限延长
标题	Scalable knowledge distillation techniques for machine learning		

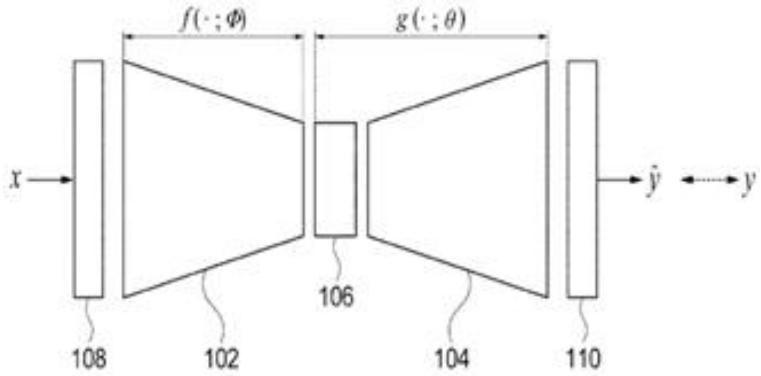
申请(专利权)人	微软技术许可有限责任公司
发明目的	通过在迭代知识蒸馏过程中在记忆中并行培训教师和学生模型,解决了当前方法的资源密集型限制,从而实现了有效有效的模型蒸馏,并减少了内存和 I/O 要求。
主要附图	
第一权利要求	1. A data processing system comprising: a processor; and a machine−readable storage medium storing executable instructions that, when executed, cause the processor to perform operations comprising: instantiating an instance of a teacher model in a memory of the data processing system; instantiating an instance of a student model in the memory of the data processing system; dividing training data into a plurality of batches of samples, wherein a size of the plurality of batches of samples is based at least in part on an amount of the memory available for training the student model after instantiating the instance of the teacher model and the instance of the student model; and training the student model to replicate performance of the teacher model using an iterative knowledge distillation process in which the teacher model and the student model are trained in parallel during each iteration of the iterative knowledge distillation process by: obtaining a respective batch of training data from the plurality of batches of samples in the memory; training the teacher model using each of the samples in the respective batch of training data; training the student model using each of the samples in the respective batch of training data; evaluating performance of the student model compared with the performance of the teacher model; and providing feedback to student model to adjust behavior of the student model based on the performance of

	the student model.

发明专利 US12307750B2 主要针对由于需要大量的内存和计算资源,尤其是在训练较小的模型以模仿较大的预训练的模型时,当前的知识蒸馏方法是资源密集型的,并且通常需要编写和读取大量数据才能持续存储的问题进行探究。

微软技术许可有限责任公司在实施迭代知识蒸馏过程,在计算环境的记忆中,教师和学生模型都经过并联培训,从而消除了编写教师模型逻辑以持续存储并减少 I/O 资源消耗的需求，这种方法大大减少了内存和 I/O 资源需求,从而有效地实施了知识蒸馏,同时保持类似于教师模型的预测性能,从而在资源有限的设备上进行部署。

2）知识蒸馏相关专利解读之 US20220383072A1

申请号	US17/826433	申请日	2022-05-27
公开号	US20220383072A1	法律状态/事件	实质审查
标题	Knowledge distillation method based on regression task and computing device for executing the method		
申请(专利权)人	三星 SDS 株式会社		
发明目的	用于回归任务的新颖知识蒸馏方法通过使用具有异常值过滤和交叉注意分数的两步学习过程来解决性能下降问题,从而提高学生模型性能的质量和稳定性。		
主要附图			
第一权利要求	1. A knowledge distillation method based on a regression task that is executed in a computing device including: one or more processors; and a memory storing one or more programs executed by the one or more processors, the method comprising: training a first deep neural network model that is a teacher model including a first encoder and a first decoder; constructing a second deep neural network model that is a student model including a second encoder and a second decoder according to a preset constraint of a target system and initializing parameters of the second deep neural network model;		

	performing a first knowledge distillation operation on the second encoder of the second deep neural network model based on the trained first deep neural network model; and performing a second knowledge distillation operation on the second encoder and the second decoder of the second deep neural network model based on the trained first deep neural network model.

发明专利 US20220383072A1 主要针对由于标签噪声和异常值,现有的深度神经网络(DNN)知识蒸馏方法对于回归任务无效,导致学生模型的性能下降的问题进行探究。

三星 SDS 株式会社运用了一种用于涉及两步学习过程的回归任务的新颖知识蒸馏方法,其中具有第一个编码器和解码器的教师模型使用第二个编码器和解码器训练学生模型,结合离群值过滤和交叉注意分数以最小化特征和预测 距离损失函数。这种方法通过过滤异常值和稳定性能来提高知识蒸馏的质量,确保学生模型在回归任务中保持稳健性和准确性。

4.5 本章小结

本章围绕 AI 模型优化的四个主要方向展开分析，包括模型结构优化、模型压缩、参数优化和知识蒸馏，并结合重点专利技术进行深入解读。

模型结构优化剪枝是主线，从传统剪枝（Dropout/DropConnect）发展到动态剪枝、结构化剪枝（Filter/Kernel 级）、联合剪枝量化等。譬如华为（CN113065636B）公开的动态剪枝方法，通过目标损失函数实现灵活剪枝，适用于移动设备；英伟达（CN111210016A）：针对元素级别操作的剪枝技术，保持准确性等。

AI 模型压缩通过量化等技术减少模型体积和计算需求，提升运行效率。量化可将 32 位浮点数转为 8 位整型，显著降低资源消耗。譬如清华大学（CN114742036B）：组合式压缩方法，结合专家化和量化，保留模型性能；微软（CN114341882A）：无损指数和有损尾数压缩，优化训练效率；英特尔（US11887001B2）：通过稀疏化减少参数密度，适用于嵌入式设备。

AI 模型参数优化通过超参数调整和正则化技术优化模型性能。超参数（如学习率）影响收敛速度，正则化（如 L1/L2）防止过拟合。譬如 IBM（US10902311B2）：正则化方法提升模型泛化能力；微软 US12198054B2：OBPROX-SG 算法等。

知识蒸馏将大模型（教师模型）的知识迁移到小模型（学生模型），实现轻量化部署。譬如微软 US12307750B2 的内存内并行蒸馏；华为 CN115699029A：对抗样本增强蒸馏等。