中国人工智能学会
昇腾 CANN 学术基金
项目申请指南

中国人工智能学会

2023 年 11 月

1. 总则

《中国人工智能学会-昇腾 CANN 学术奖励基金》（以下简称“昇腾 CANN 基金”）是由中国人工智能学会、鹏城实验室和华为技术有限公司共同发起，致力于面向海内外高校及科研院所的 AI 学者搭建学术交流的平台，提供经费、算力、技术支持等服务，推动人工智能方法在科学领域的创新应用：由鹏城实验室通过中国算力网的充沛算力与“鹏城·脑海”国产开源大模型相关资源支持申请者基于国产人工智能生态体系开展科研探索，OpenI 启智开源社区将做为本次项目的线上开发环境，给科研工作者们提供开源开放的全流程开发体验，提前布局下一代基础模型的研究，并在大模型和科学智能两大新赛道原生生态方向，打造昇腾、昇腾 CANN 原生技术优势，推动业界在前沿技术探索与实践方面的深度合作，并支持发表基于昇腾和昇腾 CANN 的国际国内高水平会议和期刊的学术论文、孵化出原创型理论和重大算法创新成果，持续构建原生、开放的科研与人才生态。

2. 申报流程

2023 年 11 月 9 日项目正式发布申请指南，计划 2023 年 12 月 10 日前完成项目评审，并与中国人工智能学会签署项目合同。

2.1 申请条件

申请人员：高校/科研院所在职的全职教师或研究人员；

2.2 申请方式

申请者提交《项目申请书》，所有申请者均可同时申报所有课题项目，但最终只能进行一个课题项目的立项，如下附件：

申请书接收邮箱：xsjljj@caai.cn

2.3 项目评审

该项目依托中国人工智能学会运作，由技术管理委员会负责监督计划的实施以及项目的评审。委员会评审时主要考虑：

1) 申请项目的作用、意义、创新性、可行性；
2) 申请者（及团队）的学术水平、科研能力，针对部分申请者或者团队，会根据实际情况安排面试；
3) 申请者研究经历和申请项目的相关性。

经过委员会确认授予资助的研究项目需签署合同生效。

3. 申报主题

昇腾 CANN（Compute Architecture for Neural Networks）是华为针对 AI 场景推出的异构计算架构，对上支持多种 AI 框架，对下服务 AI 处理器与编程，发挥承上启下的关键作

用，是提升昇腾 AI 处理器计算效率的关键平台。同时针对多样化应用场景，提供高效易用的编程接口，支持用户快速构建基于昇腾平台的 AI 应用和业务。

Ascend C 是 CANN 针对算子开发场景推出的编程语言，原生支持 C 和 C++标准规范，最大化匹配用户开发习惯；通过多层接口抽象、自动并行计算、孪生调试等关键技术，极大提高算子开发效率，助力 AI 开发者低成本完成算子开发和模型调优部署。

昇腾社区 CANN 官方主页：https://www.hiascend.com/software/cann

昇腾社区 Ascend C 官方主页：https://www.hiascend.com/zh/ascend-c

本期昇腾 CANN 基金设置 3 大方向共 7 个具体子课题。

各方向和课题的**具体内容、验收标准**如下：

3.1 方向 1：高性能算子算法研究

3.1.1 基于 CANN 平台的高性能 MatMul 算法研究

课题概述：Matmul 是 AI 网络中常用的一种算子，在学术界存在很多相关的算法研究。推荐结合昇腾软硬件架构进行算法研究创新。实现在昇腾服务器执行时，在任何形状/数据类型/格式上都有较高的性能。

验收标准：

（1） 原型代码：基于 Ascend C 算子编程语言开发的算子实现，且代码在主流社区开源（必选 OpenI 启智社区）

（2） 设计文档：提供相关设计方案，详细阐述该研究成果的理论基础、实现原理

（3） 技术指标 1：内核执行性能随形状的变化呈线性（或近似线性）变化

（4） 技术指标 2：内核的平均 cube 利用率达 90%以上

（5） 技术指标 3：支持多种数据类型（如 Int8/FP16/FP32 等）

（6） 论文要求（可选）：首发顶会顶刊论文（原则上要求 CCF-B 类及以上，且非 Findings 论文、非 short paper；首发论文指的是：直接基于昇腾软硬件达成实验结果，并进行开源）

3.1.2 基于 CANN 平台的高性能 Conv 算法研究

课题概述：Conv2d 是 AI 网络中常用的一种算子，在学术界存在很多相关的算法研究。推荐结合昇腾软硬件架构进行算法研究创新。实现在昇腾服务器执行时，在任何形状/数据类型/格式上都有较高的性能。

验收标准：

（1） 原型代码：基于 Ascend C 算子编程语言开发的算子实现，且代码在主流社区开源（必选 OpenI 启智社区）

（2） 设计文档：提供相关设计方案，详细阐述该研究成果的理论基础、实现原理

（3） 技术指标 1：内核执行性能随形状的变化呈线性（或近似线性）变化

（4） 技术指标 2：内核的平均 cube 利用率达 90%以上

（5） 技术指标 3：支持多种数据类型（如 Int8/FP16/FP32 等）

（6） 论文要求（可选）：首发顶会顶刊论文（原则上要求 CCF-B 类及以上，且非 Findings

论文、非 short paper；首发论文指的是：直接基于昇腾软硬件达成实验结果，并进行开源）

3.1.3 基于 CANN 平台的高性能融合 Kernel 研究及优化实现

课题概述：AI 模型参数和计算量越来越大和复杂，需要更高的性能来支撑推理和训练。不停发展的模型结构和新的算法，带来大量的模型调优工作。因此诉求对模型的自动调优，获得更高的性能。通过融合多个算子，使用硬件内部多级缓存消除中间结果在 DDR/HBM 的搬移，同时减少 kernel launch 以及增加并发，可以极大的提升性能，参考 pytorch-Inductor，有 1.69x 提升。

验收标准：

（1） 原型代码：基于昇腾 CANN 平台的代码实现，且代码在主流社区开源（必选 OpenI 启智社区）

（2） 设计文档：提供相关设计方案，详细阐述该研究成果的理论基础、实现原理

（3） 技术指标 1：相比融合前，模型整体内存减小 30%，各 kernel 满足硬件缓存限制

（4） 技术指标 2：基于鲲鹏 ARM 平台，自动融合时间优于或持平 pytorch-inductor

（5） 技术指标 3：模型以 GEIR 或 Inductor IR 做输入，尽量基于 GEIR 扩展和实现代码生成、tiling 策略生成

（6） 论文要求（可选）：首发顶会顶刊论文（原则上要求 CCF-B 类及以上，且非 Findings 论文、非 short paper；首发论文指的是：直接基于昇腾软硬件达成实验结果，并进行开源）

3.2 方向 2：高效率集群通信算法研究

3.2.1 基于 CANN 平台的大规模集群 AlltoAll 算法研究

课题概述：大模型集群训练场景，随着模型规模的扩大，AlltoAll 耗时已经成为性能瓶颈，在 Torus 拓扑中多租户资源分配存在非对称 Torus 和 Mesh 拓扑等情况，AlltoAll 性能面临更大挑战。因此，需针对 Torus 拓扑探索性能高且适应性强的 AlltoAll 算法。

验收标准：

（1） 原型代码：基于昇腾 CANN 平台的代码实现，且代码在主流社区开源（必选 OpenI 启智社区）

（2） 设计文档：提供相关设计方案，详细阐述该研究成果的理论基础、实现原理

（3） 技术指标 1：在各场景下，算法性能超越现有算法：

$$\text{target} < \frac{1}{2}\lfloor\frac{n_1}{2}\rfloor\lceil\frac{n_1}{2}\rceil\prod_{i=2}^{d} n_i \frac{m}{b}$$

其中，m为任意两个processor之间传输的数据量，b为链路带宽，各维度节点数为$n_1 \times n_2 \times \ldots \times n_d$，假设$n_1$为维度节点数的最大值$\max(n_i)$

（4） 技术指标 2：在各场景下，算法步骤不超过现有算法：

$$\text{target} \leq \left\lfloor\frac{n_{max}}{2}\right\rfloor * d$$

其中，d为维度数，n_{max}为各维度节点数的最大值$\max(n_i)$

（5） 论文要求（可选）：首发顶会顶刊论文（原则上要求 CCF-B 类及以上，且非 Findings 论文、非 short paper；首发论文指的是：直接基于昇腾软硬件达成实验结果，并进行开源）

3.3 方向 3：大模型优化技术研究

3.3.1 基于 CANN 平台的亲和的 Self-Attention 算法研究

课题概述：在大模型大热的当下，随着模型参数量的不断增加，对模型的执行性能和内存都提出了挑战。当前业界针对 SelfAttention 的优化算法 FlashAttention 是结合 GPU 硬件架构提出的优化，本课题希望基于昇腾软硬件进行相关算法的研究创新。训练框架无限制。

验收标准：

（1） 原型代码：基于昇腾 CANN 平台的代码实现，且代码在主流社区开源（必选 OpenI 启智社区）

（2） 设计文档：提供相关设计方案，详细阐述该研究成果的理论基础、实现原理

（3） 技术指标 1：基于昇腾软硬件平台，内存资源消耗和性能达成 FlashAttention-2 同等水平

（4） 论文要求：首发顶会顶刊论文（原则上要求 CCF-B 类及以上，且非 Findings 论文、非 short paper；首发论文指的是：直接基于昇腾软硬件达成实验结果，并进行开源）

3.3.2 基于 CANN 平台的大语言模型（LLM）的长序列技术研究

课题概述：大语言模型（Large Language Model，LLM）能够通过读取较长的上下文理解并检索多个文档甚至一本书的信息。最近，许多 LLM 的序列长度已扩展至 32k 甚至 256k。序列长度的扩展给 LLM 的训练和推理带来巨大挑战。本课题希望基于昇腾软硬件进行原创性的 LLM 长序列、超长序列算法研究。能够研究实现一种模型结构或者算法，解决训练与推理过程中，长序列（超长序列）场景下内存占用高的问题。训练框架无限制。

验收标准：

（1） 原型代码：基于昇腾 CANN 平台的代码实现，且代码在主流社区开源（必选 OpenI 启智社区）

（2） 设计文档：提供相关设计方案，详细阐述该研究成果的理论基础、实现原理

（3） 技术指标 1：基于昇腾软硬件平台，16k/32k 序列长度下，内存资源消耗和计算性能优于业界同等算力平台

（4） 论文要求：首发顶会顶刊论文（原则上要求 CCF-B 类及以上，且非 Findings 论文、非 short paper；首发论文指的是：直接基于昇腾软硬件达成实验结果，并进行开源）

3.3.3 基于 CANN 平台开展业界流行大模型算法的应用研究

课题概述：开放性课题，希望基于昇腾软硬件平台进行原创性的前沿科学、交叉学科等领域的创新应用研究。能够挑战发表较高级别论文。训练框架无限制。

验收标准：

（1） 原型代码：基于昇腾 CANN 平台的代码实现，且代码在主流社区开源（必选 OpenI 启智社区）

（2） 设计文档：提供相关设计方案，详细阐述该研究成果的理论基础、实现原理

（3） 论文要求：挑战 Nature/Science 论文，或者学术上同等难度级别

4. 交付成果及知识产权

每个课题项目的交付成果请见申报主题中的验收标准要求。

昇腾 CANN 基金项目交付成果包含论文、代码的知识产权权利归属申请方所有，具体细节以中国人工智能学会与申请方签署的项目合同为准。

昇腾 CANN 基金项目最终解释权归昇腾 CANN 基金技术管理委员会所有。

注：如申请者选择“鹏城·脑海”大模型进行探索，经过项目组遴选后，鹏城实验室将为申请者提供额外的技术、算力、数据等方面的保障，全面支持申请者基于国产软硬件计算平台与人工智能大模型技术体系，进行前沿探索。