

中国知网“CNKI 研究生数智化学习平台”开通通知

为丰富我校研究生教学手段，提升研究生培养质量，研究生学院特引进中国知网“CNKI 研究生数智化学习平台”产品，该平台基于导学导研模式，紧扣研究生培养过程中关键业务、关键环节、关键要素，为研究生、导师提供优质课程资源、基本文献阅读、课题组研讨会、毕业论文指导、论文格式检查、学术规范训练等功能。请相关师生积极使用，现通知如下：

一、使用范围

2024 级全日制硕士研究生及其导师。

二、登录方式

师生选择对应身份，使用知网研学账号登录。如无知网研学账号可先用个人手机号进行免费注册。首次登录时需要验证职工号/学号、姓名，完成验证绑定后才可进入系统，免费享用机构所订功能权限。首次绑定以后，后续只需使用知网研学账号登录即可，无需再次验证。

平台地址：<https://yjs.cnki.net/hbueab>



三、不同角色作用

1.研究生端

提供体系化学习环境，打造线上线下相结合的学习模式，助力研究生学术科研能力提升。

2.导师端

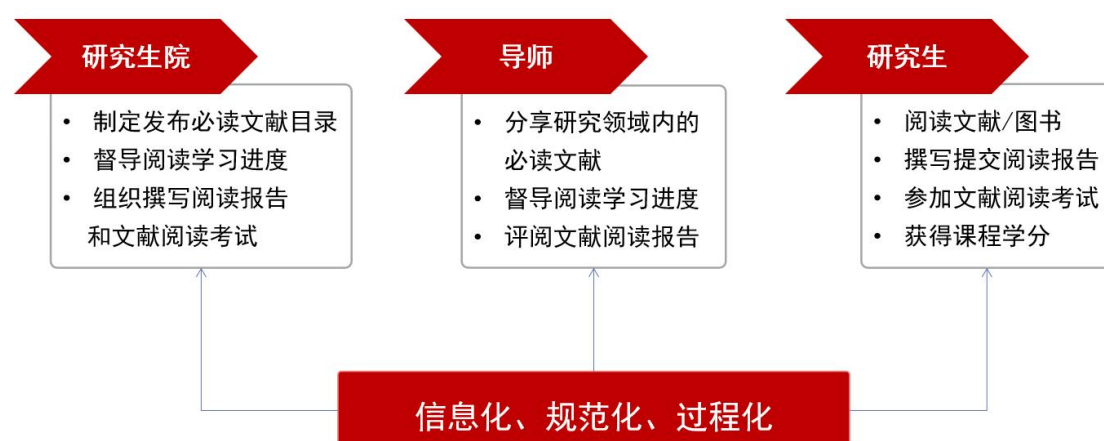
提供指导评阅的线上环境，实时了解学生文献阅读进展、论文完成情况，降低沟通成本，提高指导效率。

四、主要功能简介

平台功能包含基本文献阅读、研读学习、课程培训、创作投稿、学术写作训练、写作指导与审阅、学情数据管理、课题组研讨、AI 专题探究学习、AI 探究式阅读、AI 创作等。

1.基本文献阅读

提供 XML 数据阅读，提供文献大纲阅读功能；提供笔记、划线、摘录以及相关标签和管理等操作；提供参考文献、引证文献一键链接，提供作者、关键词、机构等知网节链接功能；提供一键查询专业词汇释义功能；提供多项在线翻译功能；提供单篇、专题、引文文献三种类型的文献矩阵功能；支持英文文献全文翻译，支持单次 200 页文献内容的翻译，提供原版式对照和逐句对照的翻译结果呈现方式，且翻译结果可导出；提供在线撰写阅读报告，支持本地上传。



2.在线创作功能

提供选题调研分析，依托大数据分析选题可行性；提供多学科开题模板，支持在线撰写和本地上传；支持在线写作，可直接生成参考文献、提供素材库、翻译、CNKI 检索；提供写作引用分析，分析论文中的引用问题；提供多种期刊的官方投稿地址和 CNKI 腾云采编平台投稿地址，可按学科导航查看，筛选核心刊；依托数据分析，分析期刊数据，推荐

投稿期刊。

学生

待办

通知

帮助

首页

基本文献阅读

学校推荐

导师推荐

自主阅读

文献阅读报告

论文写作

研究生知网团队

研究生数智化学习平台

专题

资源包

金融

智慧城市

医药

论文写作

环境保护

文博

图书馆

出版发行

国防

人力资源

心理学

新建目录

学习资料(70)

文献矩阵

AI专题探究

请输入文献标题检索

上传

AI文献矩阵

共70篇

每页显示: 10

文献标题

作者

来源

发表时间

类

A high-throughput method for dereplication and assessment of metabolite distribution in Salvia species using LC-MS/MS

上传

Whole-genome sequence of Phellinus gilvus (mulberry Sanghuang) reveals its unique medicinal values

上传

COVID-19 infection

Muhammad Adnan S

1

2

3

4

5

6

7

原文

100%

译文

只看译文

导出翻译结果

Key Lab of Fabrication Technology for Integrated Circuits, Chinese Academy of Sciences, Beijing 100029, China; (Laboratory of Microelectronics Devices and Integrated Technology, Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China; *University of Chinese Academy of Sciences, Beijing 100049, China; *School of Microelectronics, Southern University of Science and Technology, Shenzhen 518055, China Received 5 April 2023/Revised 29 July 2023/Accepted 2 September 2023/Published online 24 September 2023

Abstract Artificial intelligence (AI) has experienced substantial advancements recently, notably with the advent of large-scale language models (LLMs) employing mixture-of-experts (MoE) techniques, exhibiting human-like cognitive skills. As a promising hardware solution for edge MoE implementations, the computing-in-memory (CIM) architecture allocates memory and computing within a single device, significantly reducing the data movement and the associated energy consumption. However, due to diverse edge application scenarios and constraints, determining the optimal network structures for MoE, such as the expert's location, quantity, and dimension in CIM systems remains elusive. To this end, we introduce a software-hardware co-designed neural architecture search (NAS) framework, CIM-based MoE NAS (CMN), focusing on identifying a high-performing MoE structure under specific hardware constraints. The results of the NYUD-v2 dataset segmentation on the BRAM (SRAM) CIM system reveal that CMN can discover optimized MoE configurations under energy, latency, and performance constraints, achieving 29.67% (43.10%) energy savings, 175.44x (109.89x) speedup, and 12.24x smaller model size compared to the baseline MoE-enabled Visual Transformer, respectively. This co-design opens up an avenue toward high-performance MoE deployments in edge CIM systems.

Keywords mixture-of-experts, computing-in-memory, neural architecture search, relative random-access memory, static random-access memory

1 Introduction

The field of Artificial Intelligence (AI) has witnessed significant advancements in recent years, primarily attributed to the emergence of large-scale language models (LLMs), such as the state-of-the-art GPT-4V [1, 2], exhibiting human-like cognitive capacities [3, 4]. To further improve LLM performance while adhering to limited computational constraints, researchers have introduced the sparsely activated Mixture-of-Expert (MoE) model [5, 6], which comprises multiple experts, formed by fully connected networks (FFN), focusing on specific input spaces and a routing network for effective control (Figure 1). This model facilitates the seamless horizontal extension of LLMs and scales the model capabilities with sub-linearly increased computing complexity [7]. Moreover, the router selectively activates a few relevant experts during inference, considerably lowering MoE's computation cost compared to dense models. These advantages make MoE the preeminent pre-trained model, rendering it feasible for training LLMs with parameters exceeding the trillion scale.

*Corresponding author (email: fty120@hk.hk, shangdashan@ime.ac.cn)

†Shihao Han, Sihao Lin, Shucheng Du, and Mingqi Li have equal contributions to this work.

中国标准9990077:
3中国科学数据集成制造技术重点实验室, 北京100049; 4中国科学数据集成制造技术重点实验室, 北京100029; 5
中国科学数据集成制造技术重点实验室, 深圳518055

2024年4月3日收到/2024年7月29日修订/2024年9月2日接受/2024年9月24日在线发表

摘要近年来, 人工智能(AI)取得了长足的发展, 特别是采用混合专家(MoE)技术的大规模语言模型(LLMs)的出现, 展现了类似人类的认知能力。作为边缘MoE实现的一种有前景的硬件解决方案, 计算内存(CIM)架构将内存和计算配置在单个设备中, 显著减少了数据移动和相关的能量消耗。然而, 由于不同的边缘应用场景和约束条件, 确定MoE的最佳网络结构, 如专家在CIM系统上的位置、数量和维度仍然难以实现。为此, 我们介绍了一种软硬件协同设计的神经网络搜索(NAS)框架, NAS (基于CIM)的MoE NAS (CIM-based MoE NAS, CMN), 重点研究在硬件约束下的性能MoE结构。在BRAM (SRAM) CIM系统上对NYUD-v2数据集进行分割的结果表明, CMN可以在能量、延迟和性能约束下发现优化的MoE配置, 与基准MoE-enabled Visual Transformer相比, 分别实现了29.67% (43.10%)的能量节省, 175.44x (109.89x)的加速和12.24x的模型尺寸。这种协同设计为在边缘CIM系统中实现高性能的MoE部署开辟了一条途径。

关键词mixture-of-experts, 存算一体, 神经网络搜索, 静态随机存储器, 静态随机存储器

1引言

近年来, 人工智能(AI)领域取得了重大进展, 主要归因于大规模语言模型(LLMs)的出现。如最新的GPT-4V [1, 2], 表现出类似人类的认知能力[3, 4]。为了在有限的计算约束下进一步提升LLM的性能, 研究人员引入了稀疏激活的专家混合(MoE)模型[5, 6]。该模型由多个专家组成, 由全连接网络(FFN)组成, 专注于特定的输入空间和用于有效控制的路由网络(图1)。该模型有利于LLMs的无缝横向扩展, 并以次线性增加的复杂度扩展模型能力[7]。此外, 在推理过程中, 路由器选择性地激活一些相关专家, 与密集模型相比, 大大降低了MoE的计算成本。这些优点使MoE成为优秀的预训练模型, 使其在训练参数超过万亿规模的LLMs时具有可行性。

*对应作者(邮箱: fty120@hk.hk, shangdashan@ime.ac.cn)

†史浩, 林浩, 杜书成, 和 李鸣齐 对本工作有同等贡献。

3.论文格式检查

配置学校单独学位论文模板, 学生上传论文进行检查, 可检查 100 余格式项目, 导出详细版报告; 单个账号限查 5 次。

4.导师在线指导

支持导师在线指导与审阅学生开题报告, 记录指导过程。

5.AI 功能

包含 AI 专题探究学习: 启发式对话、专题矩阵分析、文献阅读报告; AI 探究式阅读: 渐进式阅读、矩阵式阅读;

AI 创作：自动文献综述、学术规范智能引导、多场景写作模板，自动生成大纲、全文，支持缩写、扩写、续写、改写；学术规范问答、自由问答；AI 问答、AI 选题分析、AI 润色等，接入 DeepSeek 深度思考模式后，针对专业领域知识或跨学科问题，结合知网海量可溯源的高质量知识数据，可以快速给出深度解析和可信可靠的答案，理解更加精准全面，解答更有专业深度。

五、其他说明

导师及学生端使用手册详见附件，研究生学院会组织相关培训，请各位师生关注并积极参加。

联系电话：87657669

附件：1.研究生数智化学习平台使用手册（学生端）
2.研究生数智化学习平台使用手册（导师端）

研究生学院

2025 年 6 月 11 日