高性能GPU芯片 正迎来开发热潮

图形处理单元(GPU)是一种专用电子电路,旨在快速操作和更改内存,以加速在帧缓冲区中创建图像,以输出到显示设备,是现代计算设备的重要组件。

近年来,由于虚拟现实、人工智能和高分辨率游戏等图形密集型应用和技术的兴起,对高性能GPU的需求显著增加。高性能GPU在各个应用领域有着广泛的应用,已是国家经济的支柱技术。

问题背景:为什么要自主开发高性能GPU?

1)广阔的市场需求

GPU在数据中心的应用蕴藏巨大潜力。在数据中心,GPU被广泛应用于人工智能的训练、推理、高性能计算(HPC)等领域,GPU订单火热,主要得益于全球算力需求的激增。

我国是全球电子信息制造和消费大国,对GPU的需求巨大,但我国暂无可替代的产品问世。GPU在中国市场供应确实紧缺,GPU国产替代紧迫性和重要性进一步提升。

2)寡头高度垄断

GPU是一个高技术含量的赛道,是一项系统工程,包含硬件架构、算法、软件生态等多个组成,缺一不可。

而全球GPU行业市场主要由NVIDIA和AMD两家主导,尤其在AI大模型训练芯片市场中,NVIDIA凭借V100、A100、H100、H200等系列产品占据了超过90%的市场份额。

3)芯片封锁

高性能GPU芯片在人工智能、超级计算和数据中心等方面发挥着至关重要的作用。然而,国外对这一领域非常敏感,并采取限制措施来遏制中国行业的发展。

开发高性能GPU的一系列困难和挑战

1)架构设计

GPU的架构包括计算单元、内存子系统、互连和专用硬件组件(如:纹理单元)。高效且能够处理复杂的图形和运算处理任务的架构极其重要。

2)软件栈开发

GPU编程需要编写高效的算法、优化并且执行代码以及利用专门的库和框架。

3)性能优化

最大化性能同时最小化延迟,该瓶颈对于GPU开发至关重要。这涉及优化架构、平衡工作负载以及实施高效的数据处理技术。

4)测试

GPU是个大系统芯片。除了模块级测试、芯片级测试,还需要整机和多机互联测试。

5)功耗管理

管理功耗和散热对于高性能GPU来说至关重要,以确保稳定性、可靠性和效率。必须设计多个电源域,并由软件根据需要自动管理。

6)制造工艺

使用先进半导体制造工艺对于GPU最终性能至关重要。开发高性能GPU需考虑使用最优的生产工艺。

综上所述,开发高性能GPU是一个充满挑战的大系统过程,需要硬件设计、软件优化和先进工艺技术等方面的专业知识,以及多年经验的积累。

我国GPU开发进展

相比于NVIDIA、AMD等国际巨头,国产GPU还处于起步阶段。乐观的是,在积极推动中,一批国内GPU企业崭露头角。初期以购买IP授权的模式为主,目前有不少企业摒弃了以往购买IP授权的模式,选择自主研发。

登临公司的GP+系列产品开创了新一代AI通用处理器/加速器的先河,成功填补了国内高性能AI计算领域技术和产品方面的空白。

华为的昇腾AI芯片是我国自主研发的一款高性能AI芯片,专为人工智能应用而设计。据官方数据显示,该芯片在人工智能推理性能方面表现出色,能够在各种复杂的人工智能计算任务中提供高效能的处理能力。但它不是GPU,只是AI加速芯片。

高性能GPU芯片可应用领域极其广泛:游戏设备、消费电子、云端AI服务器、自动驾驶、边缘计算、智慧安防、加密货币、医疗影像设备等等。推动自主可控GPU国产化,实现进口替代,是发展数字经济的基础性关键核心。

来源:中国科学技术协会《2024重大科学问题、工程技术难题和产业技术问题》、科普中国

AI搜索会取代传统搜索引擎吗?

如果说,近两年科技圈哪项技术热度最高,那一定非AI大模型莫属。作为一种基于深度学习的人工智能技术,AI大模型通过在大规模数据集上进行训练来捕捉数据中的复杂关系和特征,从而在各类任务实现出色的表现。

搜索引擎发展面临困境?

随着AI大模型的飞速发展,搜索引擎已经成为落地应用的重要场景之一。所谓搜索引擎,就是根据用户提供的关键词从互联网上采集信息,在对信息进行组织和处理后,将检索的相关信息展示给用户的系统。可以说,在互联网已经渗透到社会生活各个角落的当下,搜索引擎的重要性日益凸显,已然成为获取信息、解决问题不可或缺的工具。

然而,随着技术的进步和用户需求的演变,搜索引擎的局限性日益明显。由于搜索引擎依赖于基于关键字的算法和索引,通过将查询与相关内容对齐,构成了信息检索的基石。虽然这种方法在许多情况下都是有效的,但它仅限于对查询的字面解释,经常忽略了人类语言的微妙和复杂性。

因此,在使用搜索引擎进行信息检索时,用户往往需要输入若干关键词,浏览大量网页,并从中提取出有用的信息点,以获取答案。然而,有时即便经过这样的过程,也可能无法找到所需的信息。

AI搜索带来全新体验

AI大模型的出现,将重点从单纯的基于关键字的搜索转移到对用户意图的更复杂的理解。与搜索引擎不同的是,AI搜索破译了查询的潜在意图和上下文,提供了更加个性化和精确的结果。

这种上下文洞察力扩展到复杂的查询,使AI驱动的搜索引擎能够合并来自不同领域的信息,例如营养研究和医学期刊,以回答有关地中海饮食的问题。此外,AI搜索通过用户交互不断发展,不断增强个性化搜索体验。

目前,我们已经可以看到人工智能搜索的实际应用。例如,谷歌在2024年I/O开发者大会上宣布了一项名为“AI Overviews(AI概览)”的新搜索体验功能,使用户能够通过提问、聊天的方式进行搜索。百度创始人、董事长兼首席执行官李彦宏在百度2024年第一季度财报电话会上表示,目前百度搜索上已有11%的搜索结果由AI生成,让搜索能更准确、更有组织、更直接地回答用户问题。李彦宏表示,百度搜索的AI重构工作仍处于早期阶段,整体来看,搜索最有可能成为AI时代的Killer App。

AI搜索将引发哪些挑战?

虽然AI搜索的发展前景看似光明,但也面临着一系列的挑战。具体来看:

第一,成本增加。AI搜索会带来计算成本的增加,包括芯片,维护和电力成本。根据瑞银(UBS)的数据,将AI技术整合到搜索中会产生巨大的能源和排放影响,ChatGPT每天约有1300万用户。微软的必应每天处理5亿次搜索,谷歌则处理85亿次。

Alphabet董事长John Hennessy曾表示,与传统的关键字搜索相比,像Bard这样的大模式可能会使搜索成本增加10倍。随着产品的微调,这笔费用显然会“迅速”下降。不过,分析人士仍然认为,这项技术最终可能会蚕食谷歌的利润,即使附带广告。

据摩根士丹利(Morgan Stanley)估计,谷歌搜索查询的成本约为0.5美分。但如果使用人工智能,成本将会飙升。据估计,如果按照类ChatGPT人工智能能用50字的答案处理其收到的半数请求,谷歌的费用到2024年可能会增加60亿美元。

第二,偏见问题。AI模型在训练过程中会吸收大量网络数据,这些数据的质量参差不齐,可能包含恶作剧、误导性内容或低质量信息,从而影响搜索结果的准确性。另外,训练数据中可能存在的偏见会导致AI搜索结果出现不公平或歧视性的情况,如算法歧视问题。

第三,安全与隐私。AI搜索在处理用户查询时需要收集和分析大量个人数据,这引发了隐私保护的担忧。如何在提供精准搜索服务的同时保护用户隐私是一个亟待解决的问题。

由此可见,为了应对这些挑战,需要加强技术研发和应用过程中的隐私保护和道德监管,以确保AI搜索技术的健康、可持续发展。

写在最后:总的来说,通过融入AI大模型这一创新技术,将为更直观、更高效、更个性化的搜索体验奠定基础。搜索的未来已经到来,在AI的赋能下,我们将迎来一个更加互联互通、洞察敏锐、响应迅速的信息世界。

供稿单位:重庆天极网络有限公司

审核专家:李志高

真人场景还是AI生成?
识别“文生视频”的火眼金睛来了

准确率高达93.7%

如今,通过制作逼真的视频内容,AI视频生成工具正在改变设计、营销、娱乐和教育等行业。尤其是Sora、Gen-3等文生视频模型,只需要输入几行prompt文字,便可以生成逼真、连续、高质量的视频大片。

这一技术在为世界各地创作者带来无数可能性的同时,也为普通大众带来了诸多危害和风险,尤其是在传播虚假信息、宣传、诈骗和网络钓鱼等方面。

因此,如何准确识别AI生成的视频,已成为每一个人都需要关心的问题。

日前,哥伦比亚大学杨俊锋(Junfeng Yang)教授团队便开发了一种名为DIVID(DIffusion-generated VIdeo Detector)的文生视频检测工具,对于由SORA、Gen-2和Pika等模型生成的视频,检测准确率达到了93.7%。

相关研究论文(包含开源代码和数据集)已于上月在西雅图举行的计算机视觉与模式识别会议(CVPR)上展示。

DIVID是如何炼成的?

现有的Deepfake检测器在识别GAN生成的样本方面表现出色,但在检测扩散模型生成的视频方面鲁棒性不足。

在这项工作中,研究团队通过DIVID这一新工具来检测由AI生成的视频。据介绍,DIVID基于该团队今年早些时候发布的成果——Raidar,其通过分析文本本身来检测由AI生成的文本,而无需访问大语言模型(LLM)的内部运作。

Raidar使用LLM来重述或修改给定文本,然后测量系统对该文本的编辑次数。编辑次数越多,意味着文本更可能是由人类撰写;编辑次数越少,意味着文本更可能是机器生成的。

他们使用相同的概念开发了DIVID。DIVID通过重构视频并将新重构的视频与原始视频进行对比来工作。它使用DIRE值来检测扩散生成的视频,因为该方法基于这样一个假设:由扩散模型生成的重构图像应彼此非常相似,因为它们是从扩散过程分布中采样的。如果存在显著的变化,原始视频可能是人类生成的,如果没有,则可能是AI生成的。

DIVID的检测流程分为两个步骤。在步骤1中,给定一系列视频帧,研究团队首先使用扩散模型生成每个帧的重建版本。然后通过重建帧和其对应的输入帧计算DIRE值;在步骤2中,基于DIRE值序列和原始RGB帧训练CNN+LSTM检测器。

该框架基于这样一个理念:AI生成工具根据大数据集的统计分布创建内容,导致视频帧中的像素强度分布、纹理模式和噪声特征等“统计均值”内容,以及帧间不自然变化的微小不一致性或更可能出现在扩散生成视频中的异常模式。

相比之下,人类创作的视频表现出个性化,偏离统计常态。DIVID在其基准数据集中对Stable Vision Diffusion、Sora、Pika和Gen-2生成的视频实现了高达93.7%的检测准确率。

来源:学术头条