

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号

WO 2024/222552 A1

(43) 国际公布日
2024 年 10 月 31 日 (31.10.2024) WIPO | PCT

- (51) 国际专利分类号:
G06F 8/41 (2018.01) **G06F 9/50** (2006.01)
- (21) 国际申请号: PCT/CN2024/088380
- (22) 国际申请日: 2024 年 4 月 17 日 (17.04.2024)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
2023 10478439.4 2023 年 4 月 25 日 (25.04.2023) CN
- (71) 申请人: 阿里云计算有限公司 (ALIBABA CLOUD COMPUTING LTD.) [CN/CN]; 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区 1-2-A06 室, Zhejiang 310030 (CN)。
- (72) 发明人: 郑祯 (ZHENG, Zhen); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。庄东霖 (ZHUANG, Donglin); 中国浙

江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。宋帅文 (SONG, Shuaiwen); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。邱侠斐 (QIU, Xiafei); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。白俊杰 (BAI, Junjie); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。林伟 (LIN, Wei); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。

(74) 代理人: 广州铸智知识产权代理有限公司 (GUANGZHOU ZHUZHI INTELLECTUAL PROPERTY LIMITED); 中国广东省广州市天河区林和西路 161 号中泰国际广场 B 座 2305 室, Guangdong 510620 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,

(54) Title: MODEL COMPILING METHOD AND APPARATUS, AND DEVICE AND STORAGE MEDIUM

(54) 发明名称: 模型编译方法、装置、设备和存储介质

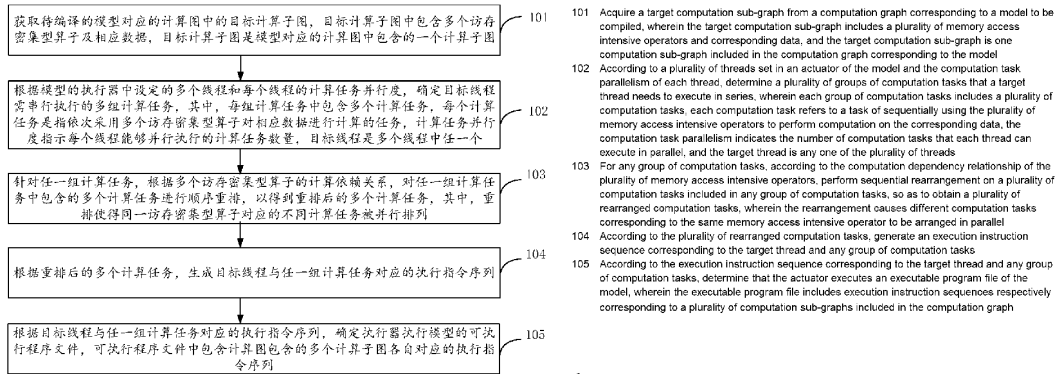


图 1

(57) Abstract: Provided in the present application are a model compiling method and apparatus, and a device and a storage medium. The method comprises: acquiring, from a computation graph corresponding to a model to be compiled, a target computation sub-graph which includes a plurality of memory access intensive operators and corresponding data; according to a plurality of threads set in an actuator of the model and the computation task parallelism of each thread, determining a plurality of groups of computation tasks that a target thread needs to execute in series, wherein each computation task refers to a task of sequentially using the plurality of memory access intensive operators to perform computation on the corresponding data; for any group of computation tasks, according to the computation dependency relationship of the plurality of memory access intensive operators, performing sequential rearrangement on a plurality of computation tasks included in any group of computation tasks, such that different computation tasks corresponding to the same memory access intensive operator are arranged in parallel, so as to generate an execution instruction sequence corresponding to the target thread and any group of computation tasks; and on the basis of the execution instruction sequence corresponding to any group of computation tasks, determining that the actuator executes an executable program file of the model.

[见续页]

BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本申请提供一种模型编译方法、装置、设备及存储介质, 该方法包括: 获取待编译的模型对应的计算图中包含多个访存密集型算子及相应数据的目标计算子图。根据模型的执行器中设定的多个线程和每个线程的计算任务并行度, 确定目标线程需串行执行的多组计算任务, 其中, 每个计算任务是指依次采用多个访存密集型算子对相应数据进行计算的任务。针对任一组计算任务, 根据多个访存密集型算子的计算依赖关系, 对任一组计算任务中包含的多个计算任务进行顺序重排, 使同一访存密集型算子对应的不同计算任务被并行排列, 以生成目标线程与任一组计算任务对应的执行指令序列。基于任一组计算任务对应的执行指令序列, 确定执行器执行模型的可执行程序文件。