

(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

(43) 国际公布日
2024 年 7 月 4 日 (04.07.2024)



(10) 国际公布号
WO 2024/140071 A1

- (51) 国际专利分类号:
G06F 9/50 (2006.01) **G06N 20/00** (2019.01)
- (21) 国际申请号: PCT/CN2023/136621
- (22) 国际申请日: 2023 年 12 月 6 日 (06.12.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202211679737.1 2022 年 12 月 27 日 (27.12.2022) CN
- (71) 申请人: 浙江阿里巴巴机器人有限公司(ZHEJIANG ALIBABA ROBOT CO., LTD.) [CN/CN]; 中国浙江省杭州市滨江区网商路 699 号 1 号楼 3 楼 304 室, Zhejiang 310052 (CN)。
- (72) 发明人: 王文特 (WANG, Went); 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: 北京同钧律师事务所(BEIJING TONGJUN LAW FIRM); 中国北京市海淀区西直门北大街 32 号院 1 号楼 7 层 808 室, Beijing 100082 (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, 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,

(54) Title: DATA PROCESSING SYSTEM AND METHOD FOR MODEL, AND STORAGE MEDIUM

(54) 发明名称: 模型的数据处理系统、方法和存储介质

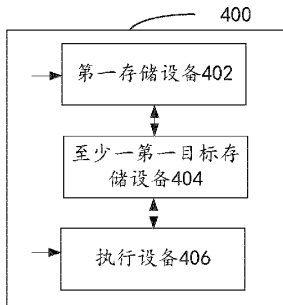


图 4

- 402 First storage device
404 At least one first target storage device
406 Execution device

(57) Abstract: Disclosed in the present application are a data processing system and method for a model, and a storage medium. The system comprises: a first storage device, which is used for determining, on the basis of a predetermined execution order of any one execution unit in a deep learning model, the current target execution unit to be executed; at least one first target storage device, which is used for returning weight data of the target execution unit, which weight data is stored in a first target storage space, to a first storage space sequentially through at least one storage device in a storage device set, and storing the weight data in the first storage space, wherein the storage device set comprises the at least one first target storage device, and the first target storage device is used for sequentially storing weight data of each execution unit in the first target storage space in advance according to the predetermined execution order; and an execution device, which is used for executing the target execution unit on the basis of the weight data of the target execution unit, which weight data is stored in the first storage space. The present application solves the technical problem of a large storage space being occupied by data processing during model execution.

(57) 摘要: 本申请公开了一种模型的数据处理系统、方法和存储介质。其中, 该系统包括: 第一存储设备, 用于基于任意一个执行单元在深度学习模型中的预定执行顺序, 确定当前待执行的目标执行单元; 至少一第一目标存储设备, 用于将第一目标存储空间中存储的目标执行单元的权重数据, 依次经由存储设备集合中至少一存储设备返回存储至第一存储空间, 其中, 存储设备集合包括至少一第一目标存储设备, 第一目标存储设备用于预先按照预定执行顺序, 依次将每个执行单元的权重数据存储至第一目标存储空间; 执行设备, 用于基于第一存储空间中存储的目标执行单元的权重数据, 执行目标执行单元。本申请解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

[见续页]

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)。

根据细则4.17的声明:
— 关于申请人有权要求在先申请的优先权(细则4.17(iii))

本国际公布:
— 包括国际检索报告(条约第21条(3))。