

(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 2022/206717 A1

(43) 国际公布日
2022 年 10 月 6 日 (06.10.2022)

(51) 国际专利分类号:
G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2022/083526

(22) 国际申请日: 2022 年 3 月 29 日 (29.03.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110350815.2 2021 年 3 月 31 日 (31.03.2021) CN

(71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 郭巍 (GUO, Wei); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 郭慧丰 (GUO, Huifeng); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 高勇 (GAO, Yong); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 唐睿明 (TANG, Ruiming); 中国

广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 刘文志 (LIU, Wenzhi); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 何秀强 (HE, Xiuqiang); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦 B 座 18C2、18D、18E、18E2, Guangdong 518001 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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,

(54) Title: MODEL TRAINING METHOD AND APPARATUS

(54) 发明名称: 一种模型训练方法及装置

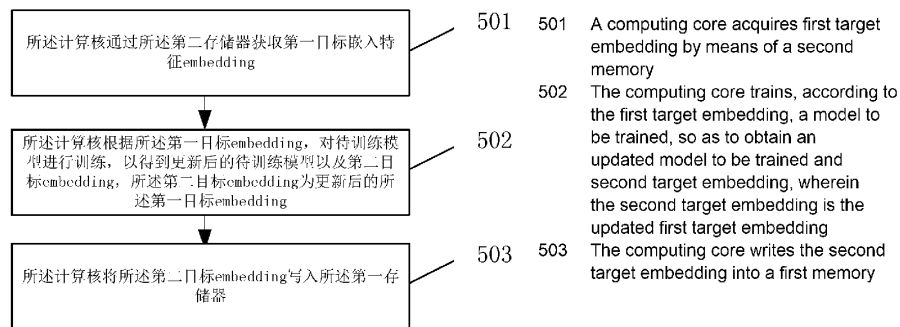


图 5

(57) Abstract: Disclosed in the present application is a model training method, which is applied to the field of artificial intelligence. The method comprises: a computing core in a first processor acquiring embedding for model training, and after the model training is complete, writing updated embedding into a first memory in the first processor instead of transferring same to a second processor. In the present application, after completing the update of the embedding, the first processor stores the updated embedding in the first memory of the first processor. Therefore, on the premise of ensuring that the first processor can acquire the latest updated embedding, the first processor can directly acquire the updated embedding without waiting for the second processor to complete the process of transmitting second target embedding to a GPU, and perform the next round of model training on the basis of the updated embedding. By means of the present application, the speed at which a processor processes training data during model training can be improved.

[见续页]



WO 2022/206717 A1

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, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, 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, 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: 本申请公开了一种模型训练方法, 应用于人工智能领域, 所述方法包括: 第一处理器中的计算核获取用于进行模型训练的embedding, 在完成模型训练后, 将更新后的embedding写入第一处理器中的第一存储器, 而不是传递至第二处理器中。本申请中, 第一处理器在完成embedding的更新后, 将更新后的embedding存储至第一处理器的第一存储器中, 在保证第一处理器可以获取到最新更新的embedding的前提下, 使得第一处理器不需要等待第二处理器将第二目标embedding传递至GPU的过程完成, 就可以直接获取到更新后的embedding, 并基于更新后的embedding进行下一轮模型训练, 本申请可以处理器在进行模型训练时的训练数据的处理速度。