

(12) 按照专利合作条约所公布的国际申请

更正本

(19) 世界知识产权组织
国际局

(43) 国际公布日
2025 年 9 月 11 日 (11.09.2025)



(10) 国际公布号

WO 2025/185304 A8

- (51) 国际专利分类号:
G06N 20/00 (2019.01) G06F 21/60 (2013.01)
G06F 21/62 (2013.01)
- (21) 国际申请号: PCT/CN2024/141657
- (22) 国际申请日: 2024 年 12 月 23 日 (23.12.2024)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202410252065.9 2024年3月5日 (05.03.2024) CN
- (71) 申请人: 中国银联股份有限公司 (CHINA UNIONPAY CO., LTD.) [CN/CN]; 中国上海市浦东新区含笑路36号银联大厦 200135 (CN)。
- (72) 发明人: 胡君一 (HU, Junyi); 中国上海市浦东新区含笑路36号银联大厦 200135 (CN)。 吕伊蒙 (LY, Yimeng); 中国上海市浦东新区含笑路36号银联大厦 200135 (CN)。 周雍恺 (ZHOU, Yongkai); 中国上海市浦东新区含笑路36号银联大厦 200135 (CN)。 王琪 (WANG, Qi); 中国上海市浦东新区含笑路36号银联大厦 200135 (CN)。 谢袁源 (XIE, Yuanyuan); 中国上海市浦东新区含笑路36号银联大厦 200135 (CN)。 高鹏飞 (GAO, Pengfei); 中国上海市浦东新区含笑路36号银联大厦 200135 (CN)。

(54) Title: LARGE-MODEL TRAINING METHOD, APPARATUS AND DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 大模型的训练方法、装置、设备及存储介质

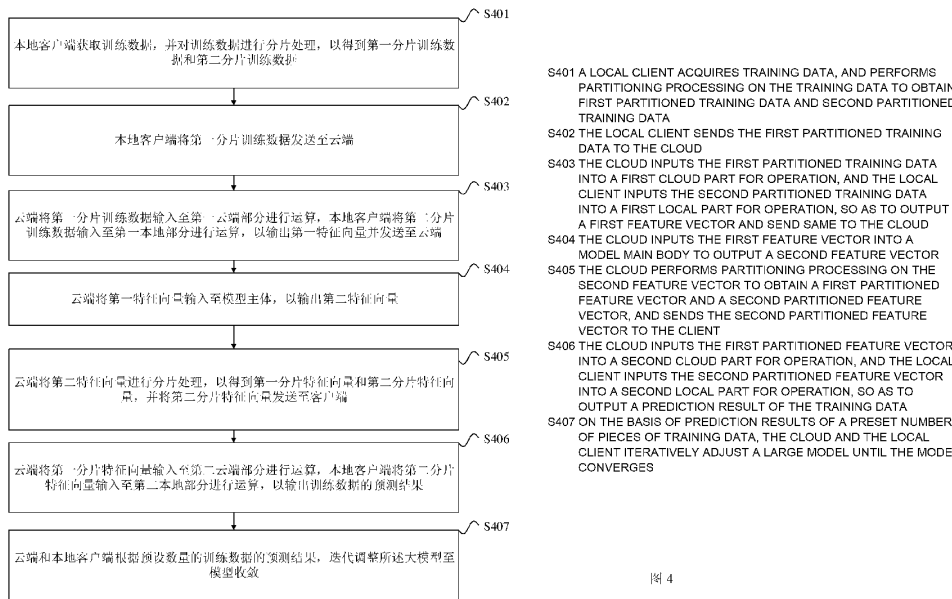


图 4

(57) Abstract: Provided in the present application are a large-model training method, apparatus and device, and a storage medium. A large model comprises an adaptor header, a model main body and an adaptor tail, wherein the adaptor header comprises a first cloud part and a first local part, and the adaptor tail comprises a second cloud part and a second local part; and during the process of training the large model, the first cloud part, the model main body and the second cloud part are deployed on the cloud, and the first local part and the second local part are deployed on a local client. In the method, only the adaptor header and the adaptor tail are distributed to

[续页]

(74) 代理人: 北京同立钧成知识产权代理有限公司(LEADER PATENT & TRADEMARK FIRM); 中国北京市海淀区西直门北大街32号枫蓝国际A座8F-6 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, 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))。

(48) 更正本的公布日:

2025 年 10 月 2 日 (02.10.2025)

(15) 更正内容:

见 2025 年 10 月 2 日 (02.10.2025) 公布的公告

the cloud and the client, encrypted client data is used to perform model training for an adaptor part, the majority of the structure of the model remains on the cloud, and cloud computing power is used to perform training and inference, thereby ensuring the computing power performance; in addition, an adaptor fine-tuning mechanism is effectively used, thereby realizing effective training of the large model while securely and fully utilizing the computing power.

(57) 摘要: 本申请提供一种大模型的训练方法、装置、设备及存储介质, 大模型包括头部适配器、模型主体和尾部适配器; 其中所述头部适配器包括第一云端部分和第一本地部分, 其中所述尾部适配器包括第二云端部分和第二本地部分; 在训练所述大模型的过程中所述第一云端部分、所述模型主体和所述第二云端部分部署于云端, 所述第一本地部分和所述第二本地部分部署于本地客户端, 该方法仅将头尾两个 Adaptor 分散到云端和客户端, 利用加密的客户端数据做 Adaptor 部分的模型训练, 模型大部分结构依然位于云端, 利用 Cloud 算力进行训练与推理, 保证了算力性能, 同时有效利用 Adaptor 微调机制, 实现了在安全、充分利用算力的前提下对大模型的有效训练。