(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
2025 年 9 月 25 日 (25.09.2025)

WIPO | PCT

(10) 国际公布号
WO 2025/195346 A1

(51) 国际专利分类号:
*G06F 9/50* (2006.01)

(21) 国际申请号: PCT/CN2025/083042

(22) 国际申请日: 2025 年 3 月 18 日 (18.03.2025)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202410332551.1 2024年3月21日 (21.03.2024) 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 518129 (CN)。

(72) 发明人:张馨然(**ZHANG, Xinran**); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。赵明宇(**ZHAO, Mingyu**); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。严学强(**YAN, Xueqiang**); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。习燕(**XI, Yan**); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。王俊凡(**WANG, Junfan**); 中国广东省深圳市龙岗区坂田华为总部办公楼 518129 (CN)。

(74) 代理人:北京龙双利达知识产权代理有限公司 (**LONGSUN LEAD IP LTD.**); 中国北京市海淀区北清路81号院二区3号楼8层801-1室 100094 (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, UY, 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

(54) **Title:** COMMUNICATION METHOD, DEVICE AND SYSTEM

(54) 发明名称: 通信方法、装置和系统

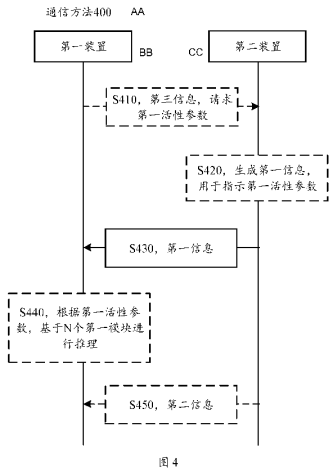



图 4

AA Communication method 400
BB First device
CC Second device
S410 Third information for requesting first activity parameters
S420 Generate first information for indicating the first activity parameters
S430 First information
S440 On the basis of the first activity parameters, perform inference on the basis of N first modules
S450 Second information

(57) **Abstract:** A communication method, device and system, relating to the technical field of communications. The method comprises: receiving first information, the first information being used for indicating first activity parameters, the first activity parameters being activity parameters of N first modules in a first model, and N being a positive integer; and on the basis of the first activity parameters, performing inference on the basis of the N first modules. According to the solution, a first device used for inference obtains the activity parameters in full parameters of the first modules by means of the first information, instead of storing the full parameters of the first module. Compared with the solution of locally storing the full parameters in the first device and reading the activity parameters from the full parameters, the described solution can save storage space of the first device, so that the first device having limited storage space can run a large model having massive parameters, thereby improving the inference performance of the large model.

WO 2025/195346 A1

[见续页]

(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: 一种通信方法、装置和系统，涉及通信技术领域。该方法包括：接收第一信息，该第一信息用于指示第一活性参数，该第一活性参数是第一模型中的N个第一模块的活性参数，N为正整数；根据该第一活性参数，基于该N个第一模块进行推理。通过上述方案，用于推理的第一装置可以不存储第一模块的全量参数，而是通过第一信息获取第一模块的全量参数中的活性参数。与将全量参数存储在第一装置本地，并从全量参数中读取出活性参数的方案相比，上述方案能够节约第一装置的存储空间，使得存储空间有限的第一装置能够运行大参数量的大模型，从而提高了大模型的推理性能。