

(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 2020/119268 A1

(43) 国际公布日
2020 年 6 月 18 日 (18.06.2020)

(51) 国际专利分类号:
G06N 20/00 (2019.01)

(21) 国际申请号: PCT/CN2019/112895

(22) 国际申请日: 2019 年 10 月 24 日 (24.10.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811522364.0 2018 年 12 月 13 日 (13.12.2018) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 刘永超 (LIU, Yongchao); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。李思仲 (LI,

Sizhong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。潘国振 (PAN, Guozhen); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。徐建国 (XU, Jianguo); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。黄启印 (HUANG, Qiyin); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 409, Beijing 100085 (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,

(54) Title: MODEL-BASED PREDICTION METHOD AND DEVICE

(54) 发明名称: 一种基于模型进行预测的方法和装置

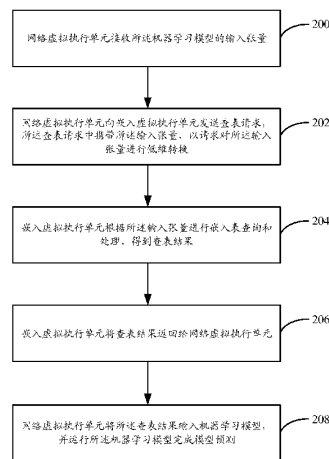


图 2

- 200 A network virtual execution unit receives an input tensor of a machine learning model
202 The network virtual execution unit sends a table lookup request to an embedding virtual execution unit, the table lookup request carrying the input tensor to request low-dimensional transformation of the input tensor
204 The embedding virtual execution unit performs embedding table query and processing according to the input tensor to obtain a table lookup result
206 The embedding virtual execution unit returns the table lookup result to the network virtual execution unit
208 The network virtual execution unit inputs the table lookup result into the machine learning model, and runs the machine learning model to complete model prediction

(57) Abstract: A model-based prediction method and device. The method comprises: a network virtual execution unit receives an input tensor of a machine learning model (200); the network virtual execution unit sends a table lookup request to an embedding virtual execution unit, the table lookup request carrying the input tensor to request low-dimensional transformation of the input tensor (202); the embedding virtual execution unit performs embedding table query and processing according to the input tensor to obtain a table lookup result (204); the embedding virtual execution unit returns the table lookup result to the network virtual execution unit (206); the network virtual execution unit inputs the table lookup result into the machine learning model, and runs the machine learning model to complete model prediction (208).

[见续页]

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, 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, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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: 一种基于模型进行预测的方法和装置, 其中, 所述方法包括: 网络虚拟执行单元接收机器学习模型的输入张量(200); 网络虚拟执行单元向嵌入虚拟执行单元发送查表请求, 所述查表请求中携带所述输入张量, 以请求对所述输入张量进行低维转换(202); 嵌入虚拟执行单元根据所述输入张量进行嵌入表查询和处理, 得到查表结果(204); 嵌入虚拟执行单元将查表结果返回给网络虚拟执行单元(206); 网络虚拟执行单元将所述查表结果输入机器学习模型, 并运行所述机器学习模型完成模型预测(208)。