

(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 2019/047795 A1

(43) 国际公布日
2019 年 3 月 14 日 (14.03.2019)

(51) 国际专利分类号:
G06F 21/57 (2013.01)

(21) 国际申请号: PCT/CN2018/103730

(22) 国际申请日: 2018 年 9 月 3 日 (03.09.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710798683.3 2017 年 9 月 7 日 (07.09.2017) 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: 及

(71) 申请人 (仅对 US): 夏巨鹏 (XIA, Jupeng) [CN/CN]; 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)。李才伟 (LI, Caiwei) [US/CN]; 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: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (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, 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,

(54) Title: METHOD AND APPARATUS FOR DETECTING MODEL SECURITY AND ELECTRONIC DEVICE

(54) 发明名称: 一种模型安全检测方法、装置以及电子设备

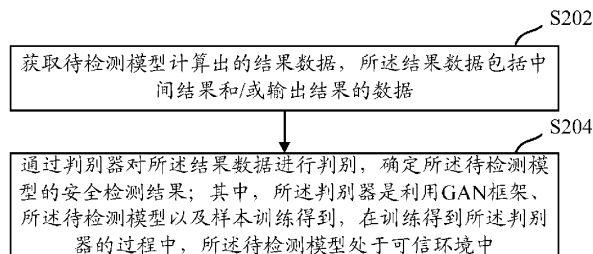


图 2

S202 Acquire result data calculated by a model to be detected, wherein the result data comprises data of an intermediate result and/or an output result

S204 Make a determination for the result data by means of a discriminator so as to determine a security detection result for the model to be detected, wherein the discriminator is obtained by using a generative adversarial nets (GAN) frame, the model to be detected and sample training, and during the process of training to obtain the discriminator, the model to be detected is in a reliable environment

(57) Abstract: Disclosed in the embodiments of the present description are a method and apparatus for detecting model security and an electronic device. The method comprises: using a generative adversarial nets (GAN) frame, a model to be detected and sample training to obtain a discriminator; by means of the discriminator, making a determination for data of an intermediate result and/or output result calculated by the model to be detected for current input data so as to detect whether the model to be detected is currently secure.

(57) 摘要: 本说明书实施例公开了一种模型安全检测方法、装置以及电子设备。所述方法包括: 利用 GAN 框架、待检测模型以及样本训练得到判别器, 通过判别器对待检测模型针对当前输入数据计算出的中间结果和/或输出结果的数据进行判别, 检测待检测模型当前是否安全。

[见续页]



WO 2019/047795 A1

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))。