

(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 2022/141924 A1

(43) 国际公布日
2022 年 7 月 7 日 (07.07.2022)

- (51) 国际专利分类号:
G06N 3/04 (2006.01) *G06N 3/063* (2006.01)
- (72) 发明人: 徐磊(XU, Lei); 中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。
- (21) 国际申请号: PCT/CN2021/086229
- (74) 代理人: 北京林达刘知识产权代理事务所 (普通合伙) (LINDA LIU & PARTNERS); 中国北京市东城区北三环东路 36 号北京环球贸易中心 C 座 16 层, Beijing 100013 (CN)。
- (22) 国际申请日: 2021 年 4 月 9 日 (09.04.2021)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202011619783.3 2020 年 12 月 31 日 (31.12.2020) 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, 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,
- (71) 申请人: 上海商汤智能科技有限公司 (SHANGHAI SENSETIME INTELLIGENT TECHNOLOGY CO., LTD.) [CN/CN]; 中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。

(54) Title: NEURAL NETWORK OPERATION METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 神经网络运行方法、装置、电子设备及存储介质

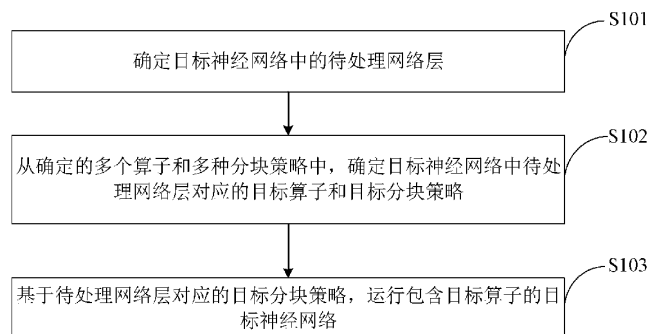


图 1

- S101 Determine a network layer to be processed of a target neural network
- S102 Determine, from determined multiple operators and multiple chunking policies, a target operator and a target chunking policy respectively corresponding to the network layer to be processed of the target neural network
- S103 On the basis of the target chunking policy corresponding to the network layer to be processed, operate the target neural network comprising the target operator

(57) Abstract: A neural network operation method and apparatus, an electronic device, and a storage medium. The method comprises: determining a network layer to be processed of a target neural network (S101); determining, from determined multiple operators and multiple chunking policies, a target operator and a target chunking policy respectively corresponding to the network layer to be processed of the target neural network (S102), each of the multiple operators being used for implementing a function corresponding to the network layer to be processed, and each of the multiple chunking policies matching an operating requirement of a computing device for operating

[见续页]

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))。

the target neural network; and on the basis of the target chunking policy corresponding to the network layer to be processed, operating the target neural network comprising the target operator (S103).

(57) 摘要: 一种神经网络运行方法、装置、电子设备及存储介质, 该方法包括: 确定目标神经网络中的待处理网络层 (S101); 从确定的多个算子和多种分块策略中, 确定所述目标神经网络中待处理网络层分别对应的目标算子和目标分块策略 (S102); 多个算子中的每个算子均用于实现待处理网络层对应的功能, 多种分块策略中的每个分块策略均匹配用于运行目标神经网络的计算设备的运行要求; 基于待处理网络层对应的所述目标分块策略, 运行包含所述目标算子的所述目标神经网络 (S103)。