

(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
2022 年 6 月 9 日 (09.06.2022)



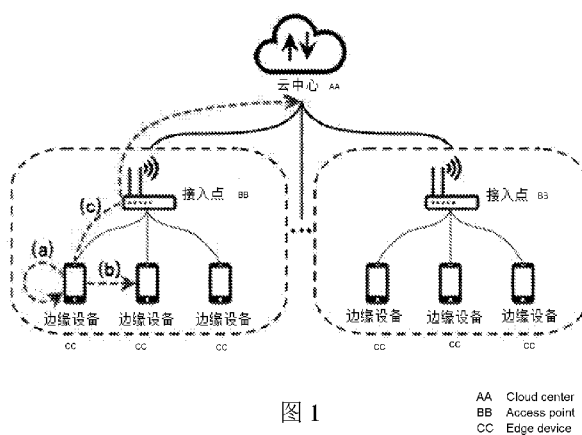
(10) 国际公布号
WO 2022/116325 A1

- (51) 国际专利分类号:
H04W 24/06 (2009.01) **G06K 9/62** (2006.01)
H04L 29/08 (2006.01)
- (21) 国际申请号: PCT/CN2020/140017
- (22) 国际申请日: 2020 年 12 月 28 日 (28.12.2020)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202011400815.0 2020 年 12 月 4 日 (04.12.2020) CN
- (71) 申请人: 中国科学院深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES) [CN/CN]; 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

- (72) 发明人: 代浩(DAI, Hao); 中国广东省深圳市南山区深圳大学城学苑大道1068号, Guangdong 518055 (CN)。 王洋(WANG, Yang); 中国广东省深圳市南山区深圳大学城学苑大道1068号, Guangdong 518055 (CN)。 须成忠(XU, Chengzhong); 中国广东省深圳市南山区深圳大学城学苑大道1068号, Guangdong 518055 (CN)。 叶可江(YE, Kejiang); 中国广东省深圳市南山区深圳大学城学苑大道1068号, Guangdong 518055 (CN)。
- (74) 代理人: 北京中巡通大知识产权代理有限公司 (BEIJING ZHONG XUN TONG DA INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市海淀区杏石口路 50 号中间建筑 1 区 148 号, Beijing 100195 (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,

(54) Title: ENERGY-SAVING METHOD FOR NEURAL NETWORK INFERENCE UNDER EDGE COMPUTING

(54) 发明名称: 一种边缘计算下神经网络推理的节能方法



(57) Abstract: The present invention relates to the technical field of edge computing networks, and in particular to an energy-saving method for neural network inference under edge computing. According to the present invention, energy consumption comparison and calculation are performed, so that an optimal path having the lowest energy consumption is selected, thereby ensuring that on the premise that inferences of all requests randomly accessed on a device are completed, high efficiency, energy saving, and no waste energy consumption are achieved. Thus, the number of inference requests that a device end can serve under the same energy consumption is increased, and the standby time of the device is prolonged.

(57) 摘要: 本发明涉及边缘计算网络技术领域, 特别涉及一种边缘计算下神经网络推理的节能方法; 本发明进行能耗比较计算, 从而选择出最优的能耗最低的路径来, 从而保证完成所有随机在设备上访问的请求的推理的前提下, 进行高效节能, 不浪费能耗, 从而提高了在同样的能耗下设备端能够服务的推理请求数量, 提升设备的待机时长。

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, 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))。

一种边缘计算下神经网络推理的节能方法

技术领域

本发明涉及边缘计算网络技术领域，特别涉及一种边缘计算下神经网络推理的节能方法。

背景技术

近年来，随着智能设备和无处不在的传感器的快速增长，边缘系统产生了大量的数据，如此之大量的数据导致传统的云架构在通信、存储等方面面临着相当大的挑战。此外，许多新型应用(如协同自动驾驶)都有严格的延迟要求，需要在接近用户设备的地方部署计算来提供实时响应，因此研究者提出了移动边缘计算(mobile edge computing, MEC)方式，将人工智能推向网络的边缘，可以有效解决通信和存储问题，充分发挥边缘计算的潜力。

MEC 中使用深度学习的主要缺点是资源受限，这些资源包括计算、存储、带宽等，此外，与传统的云中心不同，MEC 需要考虑的另一个资源是能源消耗，因为大多数终端设备都是电池供电，因此，将整个神经网络(DNN)放在边缘网络上是不可行的，通常的方法是结合云和边缘的优势，协同完成神经网络推理的任务。

目前，在 MEC 中 DNN 的部署方式主流为边缘协同推理和云中心训练，即在云中心训练 DNN 模型，边缘设备进行推理 DNN 模型，这种模式有一些具体的特点:1)云中心会根据接收到的数据不断训练模型并更新参数；2)推理边缘需要与云中心的模型参数一致，以保证每个请求的推理结果正确，请求通常会随机地

继续到达边缘设备，而且推理也需要继续提供服务，然而，因为边缘设备的能源通常有限，保持推理程序长期在线是一种能源浪费。

现有的技术较少在进行推理服务时考虑到边缘端设备的能耗问题，通常的方式是实时保持设备上的模型和云中心的设备进行同步，并且一直保持推理服务的持续运行，这种方式在请求比较稀疏的情况下十分浪费网络带宽和设备能耗，如一个预训练的 YOLO 模型约为 200MB，一个预训练的 VGG 模型约为 500MB，在多个设备需要同步时，每次需要上 GB 的流量来从云中心下载最新的模型到每个设备，造成了不必要的带宽和能耗浪费。

发明内容

本发明主要解决的技术问题是提供一种边缘计算下神经网络推理的节能方法，进行能耗比较计算，从而选择出最优的能耗最低的路径来，从而保证完成所有随机在设备上访问的请求的推理的前提下，进行高效节能，不浪费能耗。

为解决上述技术问题，本发明采用的一个技术方案是：提供一种边缘计算下神经网络推理的节能方法，其中，包括：

在可预知请求序列情况下，将边缘设备从云中心下载模型的能耗与该边缘设备从子网内其它边缘设备转移模型的能耗进行比较，如果边缘设备从云中心下载模型的能耗小于或等于边缘设备从子网内其它设备转移模型的能耗，则边缘设备直接从云中心下载模型；如果边缘设备从云中心下载模型的能耗大于边缘设备从子网内其它设备转移模型的能耗，则边缘设备从子网内其它设备转移模型；其中，子网由与云中心进行通信的接入点和若干个与该接入点通信的边缘设备组成；

在不可预知请求序列情况下，让边缘设备请求发生，维持一段时间，再将边

缘设备从云中心下载模型的能耗与该边缘设备从子网内其它边缘设备转移模型的能耗进行比较。

作为本发明的一种改进，在可预知请求序列情况下，该边缘设备进行请求时，该边缘设备的持续运行的能耗为该边缘设备从上一次进行请求一直持续运行到该边缘设备的这次请求的能耗。

作为本发明的进一步改进，在可预知请求序列情况下，该边缘设备从子网内其它边缘设备转移模型时，比较该边缘设备从子网内每个其它边缘设备的转移模型的所要能耗，选择从所要能耗最少的子网内那个边缘设备进行转移模型。

作为本发明的更进一步改进，在可预知请求序列情况下，该边缘设备从子网内其它边缘设备转移模型时，该边缘设备进行请求时，先将模型转移至子网内持续运行能耗最少的边缘设备上，再将该模型从持续运行能耗最少的边缘设备上转移至发出请求的边缘设备上。

作为本发明的更进一步改进，在可预知请求序列情况下，当上一个请求不来时，该边缘设备直接从云中心下载模型。

作为本发明的更进一步改进，在不可预知请求序列情况下，如果边缘设备从云中心下载模型的能耗小于或等于边缘设备从子网内其它设备转移模型的能耗，则边缘设备直接从云中心下载模型。

作为本发明的更进一步改进，在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗且小于该边缘设备从子网内其它设备转移模型的能耗的两倍时，则该边缘设备从子网内其它设备转移模型。

作为本发明的更进一步改进，在不可预知请求序列情况下，如果该边缘设备

从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗且小于该边缘设备从子网内其它设备转移模型的能耗的两倍时，而且子网内其它设备没有所需的模型，则该边缘设备从云中心下载模型。

作为本发明的更进一步改进，在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗的两倍时，该边缘设备从子网内其它设备转移模型。

作为本发明的更进一步改进，在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗的两倍时，先将模型转移至子网内持续运行能耗最少的边缘设备上持续运行，当有边缘设备发出请求后，再将该模型从持续运行能耗最少的边缘设备上转移至发出请求的边缘设备上。

本发明的有益效果是：与现有技术相比，本发明进行能耗比较计算，从而选择出最优的能耗最低的路径来，从而保证完成所有随机在设备上访问的请求的推理的前提下，进行高效节能，不浪费能耗。

附图说明

图 1 为边缘计算架构的示意图；

图 2 为在可预知请求序列情况下，边缘设备发出请求序列时的时空图；

图 3 为在可预知请求序列情况下，边缘设备进行持续运行的时空图；

图 4 为在可预知请求序列情况下，边缘设备从子网其它边缘设备转移模型的一种实施方式的时空图；

图 5 为在可预知请求序列情况下，边缘设备从子网其它边缘设备转移模型的

另一种实施方式的时空图；

图 6 为在不可预知请求序列情况下，边缘设备只与云中心同步的时空图；

图 7 为在不可预知请求序列情况下，边缘设备从子网其它边缘设备转移模型的时空图；

图 8 为在不可预知请求序列情况下，转移模型至能耗最少的边缘设备上的时空图；

图 9 为云中心主动更新模型的时空图；

图 10 为在不可预知请求序列情况下，实验结果对比图。

具体实施方式

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

已有的云中心结合边缘计算的人工智能服务通常是采用“云中心训练-边缘端推理”的架构来实现，由于训练是在云中心进行的，因此云中心的模型是不断更新的，而推理则是在用户设备上进行的，每个设备会连续接收多个请求，然后根据这些请求的输入数据进行推理，但由于一些实际的原因(例如，隐私问题)，通常不可能将请求路由到其他设备或者云中心进行推理，因此它只能在当前本地设备上请求推理。

目前已有的技术有以下几点缺陷：

1、需要分别实时保持每个设备上的模型和云中心进行同步，因此需要大量的同步数据传输，造成网络带宽和设备能耗的浪费；

2、需要保持设备上的推理服务持续的运行，造成设备产生不必要的能耗；

3、针对随机发生的请求，已有的一些方法并没有理论分析能保证取得较好的在线竞争比。

因此，本发明的目的是研究在这样一个场景下，如何减少不必要的能耗损失，实现高效的模型同步，提供一种边缘计算下神经网络推理的节能方法，包括：

在可预知请求序列情况下，将边缘设备从云中心下载模型的能耗与该边缘设备从子网内其它边缘设备转移模型的能耗进行比较，如果边缘设备从云中心下载模型的能耗小于或等于边缘设备从子网内其它设备转移模型的能耗，则边缘设备直接从云中心下载模型；如果边缘设备从云中心下载模型的能耗大于边缘设备从子网内其它设备转移模型的能耗，则边缘设备从子网内其它设备转移模型；其中，子网由与云中心进行通信的接入点和若干个与该接入点通信的边缘设备组成；

在不可预知请求序列情况下，让边缘设备请求发生，维持一段时间，再将边缘设备从云中心下载模型的能耗与该边缘设备从子网内其它边缘设备转移模型的能耗进行比较。

本发明进行能耗比较计算，从而选择出最优的能耗最低的路径来，从而保证完成所有随机在设备上访问的请求的推理的前提下，进行高效节能，不浪费能耗。

在本发明内，边缘计算架构为三层架构，如图 1 所示，包含了云中心（“Internet”表示），接入点（“access point”表示），以及边缘设备（“device”表示）组成，其中一个接入点和多个边缘设备组成了一个子网，而云中心往往需要通过复杂的网络与设备间联通，可以看出在同一个子网下的设备间通信的代价是远远小于设备与云中心的通信代价的。推理的请求发生在每个设备上，设备需要根据最新的模型来推理请求数据并返回结果，因此不光要保证设备上存在最新

的模型，还需要在请求到达时设备正在运行推理服务。

在本发明内，可以用以下公式来计算所有设备上的能耗：

$$C = CI + R_u \sum_{i=1}^m \beta_i + T_{\max} \sum_{i=1}^m \mu_i ;$$

其中CI是推理所有请求需要的能耗，这部分可以看作是一个常值，一旦请求序列确定了，这个常数也就确定了； R_u 表示模型更新的次数， β_i 表示每台设备与云中心通信消耗的单位能耗，这一部分表示所有设备都跟云中心同步模型造成的能耗； $T_{\max}$ 表示最后一个请求发生的时间， μ_i 表示每台设备持续运行推理服务消耗的单位能耗，这一部分表示所有设备都一直运行推理服务所带来的能耗。

从图 1 可以看出，除了从云中心获取最新的模型 (c) 外，还可以从同一子网下的其它设备上获取模型 (b)，当然也可以通过算法来保证在设备上持续运行的模型是最新的 (a)，因此只需要保持一个子网内的某台设备上的模型持续更新，即可通过设备间的转移来让所有设备上的模型都更新。

在本发明内，定义三种方式来服务每个请求：持续运行 (holding)，从其它设备转移 (transferring) 以及从云中心下载 (pulling)，并分别定义三种方式需要消耗的能耗：

1、持续运行 (holding): $\mu_i * \delta(p(i), i)$ ，其中， μ_i 是边缘设备 s_i 上持续运行服务的单位消耗，而 $p(i)$ 表示该边缘设备上一次请求发生的时间， i 表示当前请求的时间，因此，持续运行 (holding) 的成本即是从上次请求持续运行到当前请求的能耗。

2、从其它设备转移 (transferring): 由于各个边缘设备都处于同一个子网，因此边缘设备之间的传输消耗可以看作同构的，定义为 λ 。

3、从云中心下载 (pulling): 同样由于边缘设备都处于同一个子网, 因此所有边缘设备与云中心的传输消耗也是同样的, 定义为 β 。

因此, 优化边缘计算中设备推理的节能问题, 转化为了如何组合上述三种方式, 来为所有的请求提供服务; 定义不同的设备为 s_i , 请求 r_i 可以表示为一个时空点 (s_i, t_i) , t_i 表示请求发生的时间, 因此请求序列可以转化为一个如图2的时空图表示, 其中, s_0 表示云中心, 即是从 s_0 下来的竖线是通过从云中心下载 (pulling) 满足的请求, 从除 s_0 之外的 $s(i)$ 下来的竖线表示从其它设备转移 (transferring) 满足的请求, 除 s_0 之外的 $s(i)$ 的横线表示通过持续运行 (holding) 来服务推理请求。

对于请求序列预先可知的情况下, 即是一个离线问题, 可以看出目标是在这样的图2中时空图上找出最小的斯坦纳树来连接所有的请求点, 这样的树即是在设备之间最节省能耗的迁移方案。

在本发明内, 在可预知请求序列情况下, 该边缘设备进行请求时, 该边缘设备的持续运行的能耗为该边缘设备从上一次进行请求一直持续运行到该边缘设备的这次请求的能耗。

在本发明内, 分为下面几种情况:

1、在可预知请求序列情况下, 该边缘设备从子网内其它边缘设备转移模型时, 比较该边缘设备从子网内每个其它边缘设备的转移模型的所要能耗, 选择从所要能耗最少的子网内那个边缘设备进行转移模型。

2、在可预知请求序列情况下, 该边缘设备从子网内其它边缘设备转移模型时, 该边缘设备进行请求时, 先将模型转移至子网内持续运行能耗最少的边缘设备上, 再将该模型从持续运行能耗最少的边缘设备上转移至发出请求的边缘设备

上。

3、在可预知请求序列情况下，当上一个请求不来时，该边缘设备直接从云中心下载模型。

具体地讲，在本发明内，考虑通过递推公式来分析，可以将问题分解为子问题来考虑，假设前*i*-1个请求是最优的，用*C*(*i*-1)来表示，而对于第*i*个请求的最优解*C*(*i*)，可以分为以下几种情况：

1) 通过持续运行 (holding) 来服务，则这种情况下的 *i* 是通过在 *s_i* 上一次请求一直持续运行到这次请求的模型来满足的，如图 3 所示，我们记为 *D*(*i*)，可以看出当 *i* 到 *p*(*i*) 之间存在连线后，相应中间的请求将会变为通过从其它设备转移 (transferring) 或者比从其它设备转移 (transferring) 小的持续运行 (holding) 来满足，可以定义一个最小成本 *b*(*i*)：

$$b_i = \min \{ \alpha * \lambda + (1 - \alpha) * \beta, \mu_j * \delta(p(i), i) \},$$

其中， $1 \leq i \leq n$ ；表示请求 *i* 在其它边缘设备有活跃的推理服务时，只需要最小的成本就可将推理服务要发生在所需边缘设备上的请求，用 *B_i* 来表示 *b_i* 的求和，即是在一个区间上所有点都可以用最小的成本来满足，当在子网中存在其它边缘设备活跃的推理程序，因此 *D*(*i*) 的递推式可以表示为：

$$D(i) = \begin{cases} +\infty & -m \leq i \leq 0 \\ \min \left\{ \begin{array}{l} C(p(i)) + \mu_i \delta_{p(i), i} + B_{i-1}^{p(i)} \\ \min_{j \in \pi_{p(i)}(i)} \{ D(j) + \mu_i \delta_{p(i), i} + B_{i-1}^j \} \end{array} \right. & 1 \leq i \leq n \end{cases};$$

2) 通过从其它设备转移 (transferring) 来满足，此时其它边缘设备上应该存在着活跃的推理服务，当前边缘设备通过从其它边缘设备上获取最新的模型来完成推理，如图 4 所示，这时候的能耗成本记为 *T*(*i*)，需要考虑一个转移成本，用以下递推式表示：

$$T(i) = \begin{cases} +\infty & -m \leq i \leq 0 \\ \lambda + \min_{j \in \mathcal{F}(i)} \left\{ \min_{k \in \pi_j(i)} \{C(j) + \mu_i \delta_{j,i} + B_{i-1}^j, D(k) + \mu_i \delta_{j,i} + B_{i-1}^k\} \right\} & i > 0 \end{cases}$$

3) 从其它设备转移 (transferring) 还有一种特殊的形式, 由于边缘计算下所有设备的成本 i 不一致, 所以还有一种解决方案, 即是将最新的模型转移到成本最低的边缘设备上持续运行, 然后需要时再转移到对应的服务器上, 如图 5 所示, 这种情况下, 需要考虑有可能需要两次从其它设备转移 (transferring) 才能到达需要服务的设备 (先转移到最便宜的设备上, 再从该设备转移到目标设备), 用以下递推式表达:

$$E(i) = \begin{cases} +\infty & -m \leq i \leq 0 \\ \min \left\{ \begin{array}{l} E(p(i)) + \mu_{\min} \delta_{p(i),i} + B_{i-1}^{p(i)} \quad s_i = s_{\min} \\ E(p(i)) + 2\lambda + \mu_{\min} \delta_{p(i),i} + B_{i-1}^{p(i)} \quad s_i \neq s_{\min} \\ \min_{j \in \mathcal{F}(i)} \left\{ \begin{array}{l} \min(D(j), T(j)) + 2\lambda + \mu_{\min} \delta_{j,i} + B_{i-1}^j \quad s_j \neq s_{\min}, s_i \neq s_{\min} \\ E(j) + \lambda + \mu_{\min} \delta_{j,i} + B_{i-1}^j \quad s_j \neq s_{\min}, s_i = s_{\min} \\ C(j) + \lambda + \mu_{\min} \delta_{j,i} + B_{i-1}^j \quad s_j = s_{\min} \\ T(i) \quad s_i = s_{\min} \\ E(j) + \mu_{\min} \delta_{j,i} + B_{i-1}^j \quad s_i = s_{\min} \end{array} \right. \end{array} \right\} & i > 0 \end{cases}$$

4) 通过从云中心下载 (pulling) 来满足, 当请求迟迟不来时, 没有必要在任何设备上维护最新的模型, 因此下一个请求需要从云中心下载, 这时的成本为 β 。

因此, 最终最优解的递推式可以表示如下:

$$C(i) = \begin{cases} \beta & i = 0 \\ \min \left\{ \begin{array}{l} T(i) \\ D(i) \\ E(i) \\ C(i-1) + \beta \end{array} \right\} & 1 \leq i \leq n \end{cases}$$

通过上述递推式, 从而可以在已知请求序列的情况下, 得到最优的迁移方式,

来最大程度节省能耗。

在上述中，可预先知道请求序列的情况下可以有最优解，但实际情况中，还有是不可能预知请求什么时候发生的情况，所以在这种情况下，本发明提供了一种在不知道请求什么时候发生的情况下，尽可能的节省每个设备的能耗的节能方法，该方法的主要思想是在每个边缘设备请求发生后，多维持一段运行时间来等待是否还会发生下一个请求，因此每个边缘设备需要维持多长时间的推理程序，具体分为如下：

1、在不可预知请求序列情况下，如果边缘设备从云中心下载模型的能耗小于或等于边缘设备从子网内其它设备转移模型的能耗，则边缘设备直接从云中心下载模型。

2、在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗且小于该边缘设备从子网内其它设备转移模型的能耗的两倍时，则该边缘设备从子网内其它设备转移模型；进一步，如果子网内其它设备没有所需的模型，则该边缘设备从云中心下载模型；

3、在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗的两倍时，该边缘设备从子网内其它设备转移模型；

4、在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗的两倍时，先将模型转移至子网内持续运行能耗最少的边缘设备上持续运行，当有边缘设备发出请求后，再将该模型从持续运行能耗最少的边缘设备上转移至发出请求的边缘设备上。

具体地讲，分三种情况分别讨论：

1) $\beta \leq \lambda$; 这种情况下, 跟云中心同步模型的能耗小于和其它边缘设备同步的能耗, 因此只与云中心进行同步才是最优的选择, 如图 6 所示, 只需要让每个边缘设备持续运行的代价和与云中心通信的代价一样即可; 这种情况下, 很容易证明这个算法的代价与离线最优算法最差的情况下也不超过 2 倍, 即是它的竞争比为 2, 如下公式:

$$c.r.(a) = \frac{ALG}{OPT} = \frac{\mu_i \Delta t}{\mu_i \Delta t} = 1;$$

$$c.r.(b) = \frac{ALG}{OPT} = \frac{\beta + \beta}{\beta} = 2。$$

2) $\lambda < \beta \leq 2\lambda$; 这种情况下, 当有其它活跃的推理服务在运行时, 可以通过转移来获取最新的模型, 而没有时则需要与云中心同步, 如图 7 所示, 如 (a), 当有其它边缘设备在运行时, 当前边缘设备只需要维持成本与设备间转移一样的时间即可, 当没有的情况下 (如 c), 则需要维持到与云中心同步的成本一致的时间, 从而可以看出的竞争比是 2:

$$c.r.(a) = \frac{ALG}{OPT} = \frac{u_i \Delta t}{u_i \Delta t} = 1;$$

$$c.r.(b) = \frac{ALG}{OPT} = \frac{\lambda + \lambda}{\lambda} = 2;$$

$$c.r.(c) = \frac{ALG}{OPT} = \frac{\beta + \beta}{\beta} = 2。$$

3) $\beta > 2\lambda$; 这种情况下, 与云中心同步的通信成本远高于设备间同步, 尽可能的考虑在设备上让最新的模型持续运行, 因此在每个请求结束后, 可以将模型转移到代价最低的设备上持续运行, 直到其它设备上有请求到来, 通过转移来服务, 如图 8 所示, 只有当前设备运行时, 可以持续运行两倍的转移成本时间, 然后将其转移到 s_{min} 上, 运行 $\beta - 2\lambda$ 时间, 可以到达 2.5 的竞争比:

$$c.r. = \frac{2\lambda + \lambda + \beta - 2\lambda + \beta}{\beta} = 2 + \frac{\lambda}{\beta} < 2 + \frac{\lambda}{2\lambda} = 2.5。$$

在本发明内，当云中心需要主动更新模型时，如图 9 所示，只需要选择其中一台边缘设备与云中心进行同步即可，其它边缘设备将立即删除该设备上运行的模型，以保证一致性，之后的请求同样可以通过上述的方法来满足。

如图 10 所示，在本发明内，经过实验数据的对比，在调整边缘设备的单位能耗成本、与云中心同步能耗以及设备间能耗的不同组合的情况下，面对未知的请求序列时可以低于 2.5 倍的能耗比，也就是说，在面对未知的请求序列时，也可以保证以不比预先知道序列最优的算法差 2.5 倍的能耗来服务所有的推理请求。

在本发明内，边缘推理架构往往需要边缘设备实时与云中心保持同步来保证模型的更新，同时还需要设备持续运行推理服务来满足随机发生的请求，这无疑牺牲了大量的设备能耗，本发明则将该问题形式化成了时空图上的节点连接，使用同一子网下的设备间通信来保证模型的更新，比起传统方法，我们可以大大节省设备与云中心通信的次数和带宽，并降低设备的能耗。

除此之外，根据这个新模型，我们可以在减少通信次数的同时，动态关闭和启用推理服务，针对随机产生的请求序列，可以达到最差情况下与最优解至多 2.5 的能耗比，大大减少了能耗损失，延长了边缘设备的使用时间，从而减少了边缘计算中不必要的能耗和带宽，节约如智能制造、智能工厂等的运营成本；从而提高了在同样的能耗下设备端能够服务的推理请求数量，提升设备的待机时长。

本发明具备如下优点：

- 1、在模型更新时，使用边缘设备间的同步来代替边缘设备和云中心同步，减少通信能耗；
- 2、针对随机发生的请求，让模型持续运行的动态时长，来减少不必要的持

续运行能耗；

3、使用迁移来在边缘设备间调度模型来完成请求的推理。

以上所述仅为本发明的实施方式，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

1、一种边缘计算下神经网络推理的节能方法，其特征在于，包括：

在可预知请求序列情况下，将边缘设备从云中心下载模型的能耗与该边缘设备从子网内其它边缘设备转移模型的能耗进行比较，如果边缘设备从云中心下载模型的能耗小于或等于边缘设备从子网内其它设备转移模型的能耗，则边缘设备直接从云中心下载模型；如果边缘设备从云中心下载模型的能耗大于边缘设备从子网内其它设备转移模型的能耗，则边缘设备从子网内其它设备转移模型；其中，子网由与云中心进行通信的接入点和若干个与该接入点通信的边缘设备组成；

在不可预知请求序列情况下，让边缘设备请求发生，维持一段时间，再将边缘设备从云中心下载模型的能耗与该边缘设备从子网内其它边缘设备转移模型的能耗进行比较。

2、根据权利要求 1 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在可预知请求序列情况下，该边缘设备进行请求时，该边缘设备的持续运行的能耗为该边缘设备从上一次进行请求一直持续运行到该边缘设备的这次请求的能耗。

3、根据权利要求 2 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在可预知请求序列情况下，该边缘设备从子网内其它边缘设备转移模型时，比较该边缘设备从子网内每个其它边缘设备的转移模型的所要能耗，选择从所要能耗最少的子网内那个边缘设备进行转移模型。

4、根据权利要求 2 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在可预知请求序列情况下，该边缘设备从子网内其它边缘设备转移模型时，该边缘设备进行请求时，先将模型转移至子网内持续运行能耗最少的边缘设备上，再将该模型从持续运行能耗最少的边缘设备上转移至发出请求的边缘设备

上。

5、根据权利要求 2 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在可预知请求序列情况下，当上一个请求不来时，该边缘设备直接从云中心下载模型。

6、根据权利要求 1 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在不可预知请求序列情况下，如果边缘设备从云中心下载模型的能耗小于或等于边缘设备从子网内其它设备转移模型的能耗，则边缘设备直接从云中心下载模型。

7、根据权利要求 1 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗且小于该边缘设备从子网内其它设备转移模型的能耗的两倍时，则该边缘设备从子网内其它设备转移模型。

8、根据权利要求 7 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗且小于该边缘设备从子网内其它设备转移模型的能耗的两倍时，而且子网内其它设备没有所需的模型，则该边缘设备从云中心下载模型。

9、根据权利要求 1 所述的一种边缘计算下神经网络推理的节能方法，其特征在于，在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗的两倍时，该边缘设备从子网内其它设备转移模型。

10、根据权利要求 1 所述的一种边缘计算下神经网络推理的节能方法，其特

征在于，在不可预知请求序列情况下，如果该边缘设备从云中心下载模型的能耗大于该边缘设备从子网内其它设备转移模型的能耗的两倍时，先将模型转移至子网内持续运行能耗最少的边缘设备上持续运行，当有边缘设备发出请求后，再将该模型从持续运行能耗最少的边缘设备上转移至发出请求的边缘设备上。

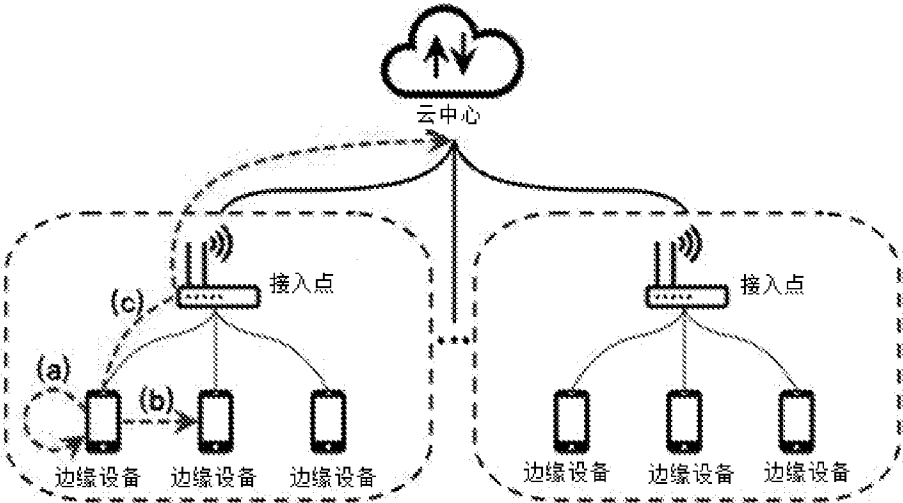


图 1

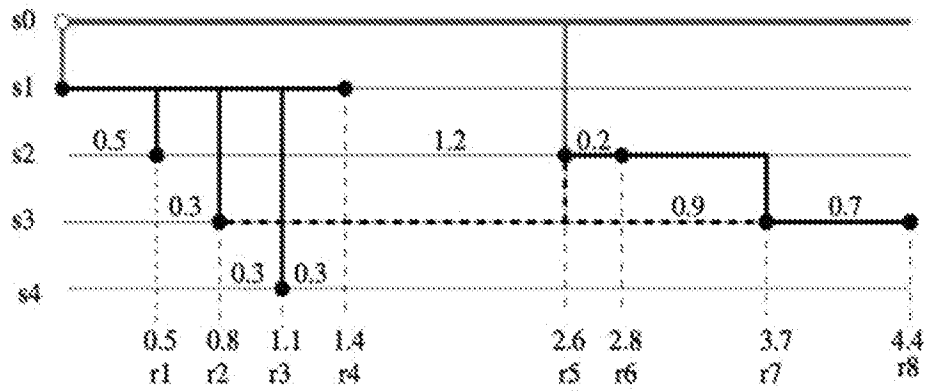


图 2

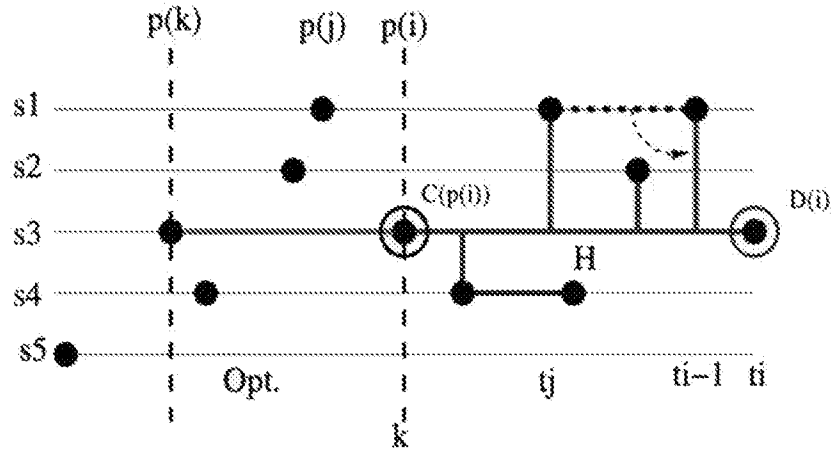


图 3

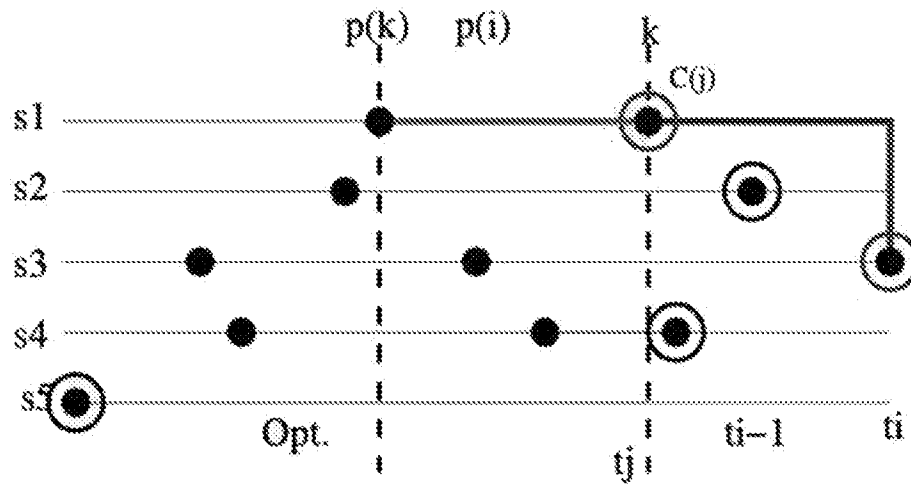


图 4

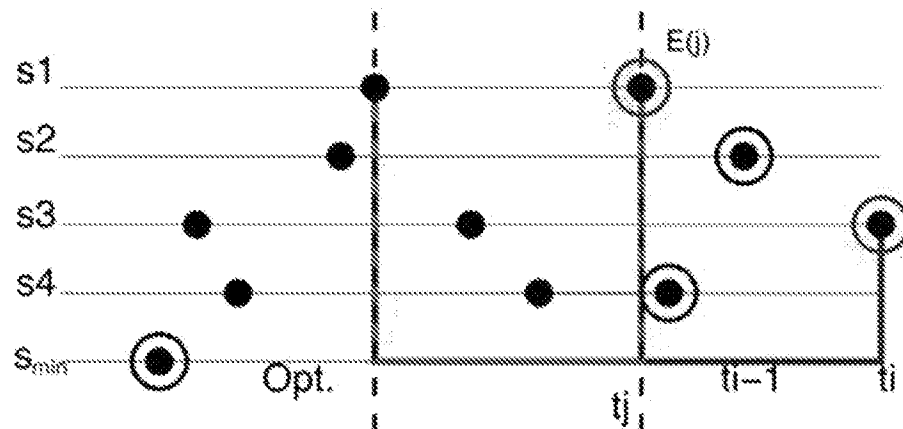


图 5

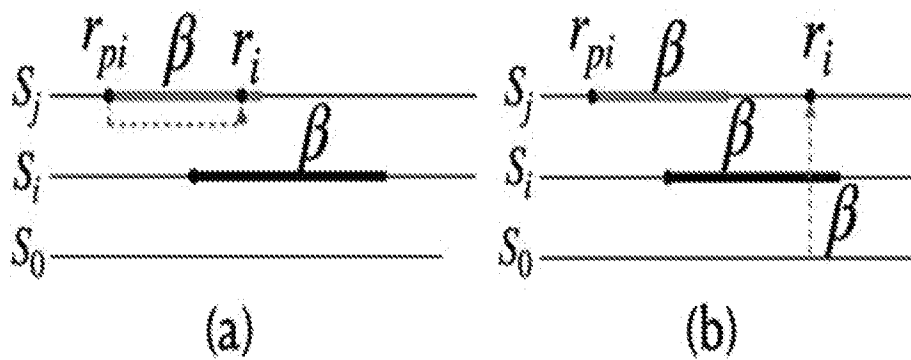


图 6

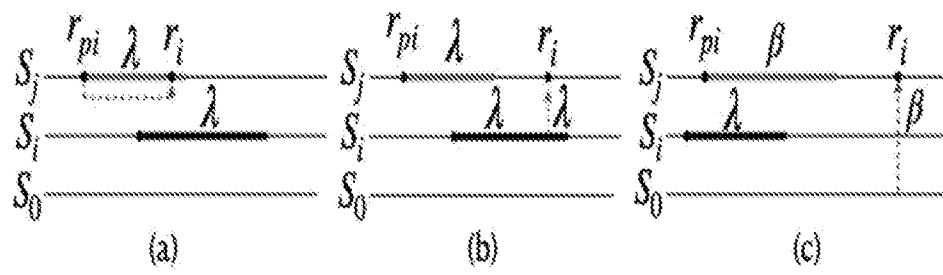


图 7

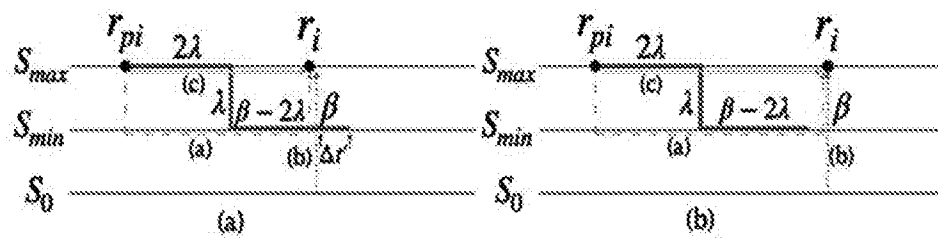


图 8

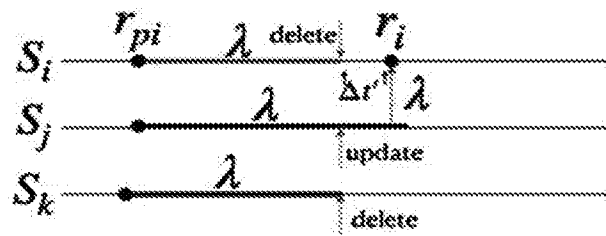


图 9

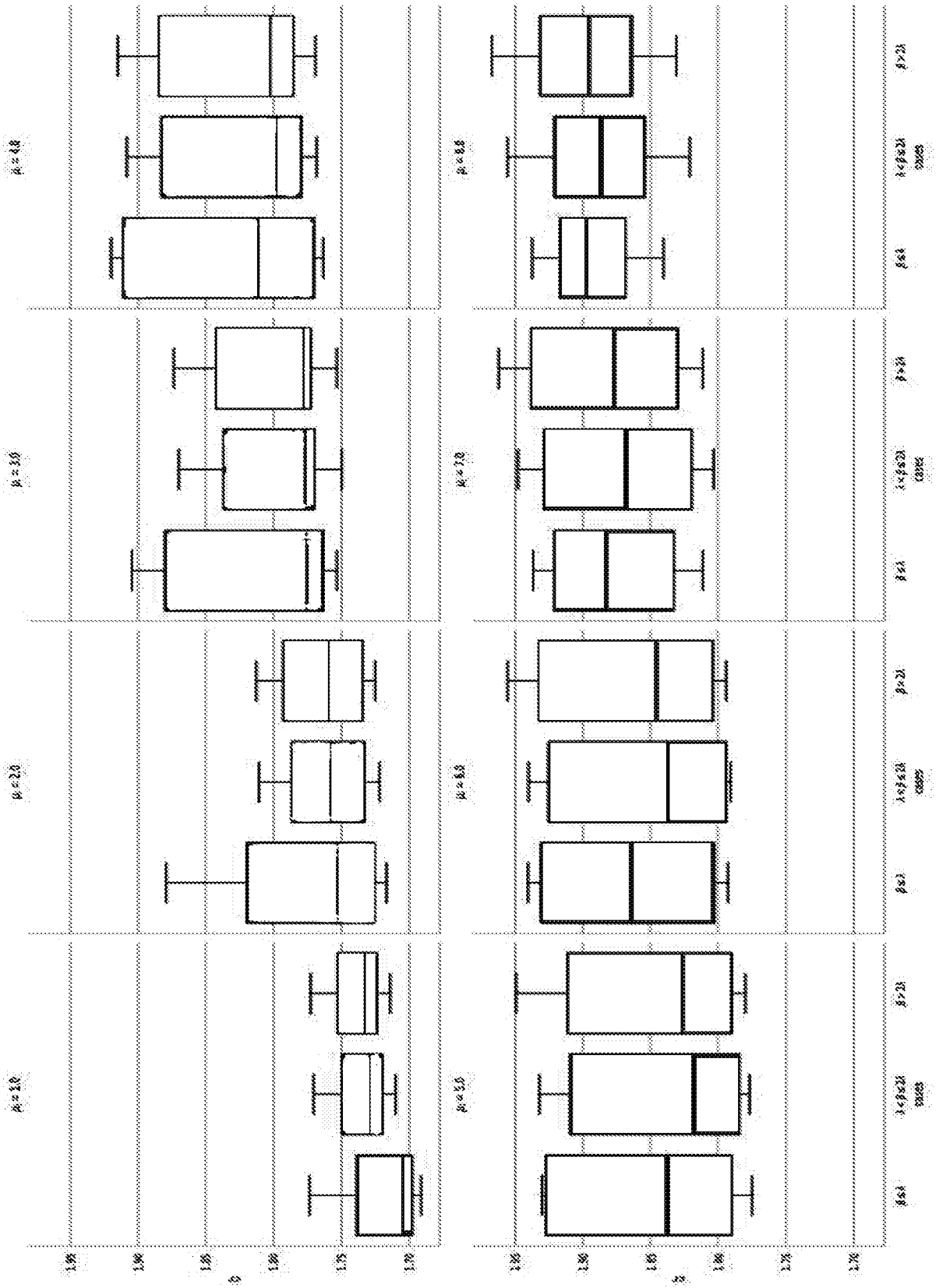


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/140017

A. CLASSIFICATION OF SUBJECT MATTER

H04W 24/06(2009.01)i; H04L 29/08(2006.01)i; G06K 9/62(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04W H04L G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; SIPOABS; DWPI; USTXT; WOTXT; EPTXT; IEEE: 边缘, 计算, 推理, 推断, 推导, 云计算, 神经网络, 能耗, 节能, 损耗, 路径, edge, computing, inference, cloud, network, DNN, energy, consumption, saving, path, link

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 111445026 A (SOUTHEAST UNIVERSITY) 24 July 2020 (2020-07-24) description, paragraphs [0008]-[0117], and figures 1-4	1-10
A	CN 111031102 A (HARBIN INSTITUTE OF TECHNOLOGY) 17 April 2020 (2020-04-17) entire document	1-10
A	CN 107766889 A (JINAN INSPUR HIGH-TECH INVESTMENT DEVELOPMENT CO., LTD.) 06 March 2018 (2018-03-06) entire document	1-10
A	CN 109961097 A (NORTHWEST UNIVERSITY) 02 July 2019 (2019-07-02) entire document	1-10
A	CN 110443298 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 12 November 2019 (2019-11-12) entire document	1-10
A	CN 111459505 A (NANJING UNIVERSITY et al.) 28 July 2020 (2020-07-28) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

04 August 2021

Date of mailing of the international search report

19 August 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/140017

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Zichuan Xu et al. "Energy-Aware Inference Offloading for DNN-Driven Applications in Mobile Edge Clouds" <i>IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS</i> , 10 November 2020 (2020-11-10), ISSN: 1558-2183, pp. 799-814	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/140017

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	111445026	A	24 July 2020	None	
CN	111031102	A	17 April 2020	None	
CN	107766889	A	06 March 2018	None	
CN	109961097	A	02 July 2019	None	
CN	110443298	A	12 November 2019	None	
CN	111459505	A	28 July 2020	None	

国际检索报告

国际申请号

PCT/CN2020/140017

A. 主题的分类 H04W 24/06(2009.01)i; H04L 29/08(2006.01)i; G06K 9/62(2006.01)i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																							
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) H04W H04L G06K 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) CNABS;CNTXT;CNKI;SIPOABS;DWPI;USTXT;WOTXT;EPTXT;IEEE: 边缘, 计算, 推理, 推断, 推导, 云计算, 神经网络, 能耗, 节能, 损耗, 路径, edge, computing, inference, cloud, network, DNN, energy, consumption, saving, path, link																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>CN 111445026 A (东南大学) 2020年 7月 24日 (2020 - 07 - 24) 说明书第[0008]-[0117]段, 图1-4</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 111031102 A (哈尔滨工业大学) 2020年 4月 17日 (2020 - 04 - 17) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 107766889 A (济南浪潮高科技投资发展有限公司) 2018年 3月 6日 (2018 - 03 - 06) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 109961097 A (西北大学) 2019年 7月 2日 (2019 - 07 - 02) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 110443298 A (华中科技大学) 2019年 11月 12日 (2019 - 11 - 12) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 111459505 A (南京大学 等) 2020年 7月 28日 (2020 - 07 - 28) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	A	CN 111445026 A (东南大学) 2020年 7月 24日 (2020 - 07 - 24) 说明书第[0008]-[0117]段, 图1-4	1-10	A	CN 111031102 A (哈尔滨工业大学) 2020年 4月 17日 (2020 - 04 - 17) 全文	1-10	A	CN 107766889 A (济南浪潮高科技投资发展有限公司) 2018年 3月 6日 (2018 - 03 - 06) 全文	1-10	A	CN 109961097 A (西北大学) 2019年 7月 2日 (2019 - 07 - 02) 全文	1-10	A	CN 110443298 A (华中科技大学) 2019年 11月 12日 (2019 - 11 - 12) 全文	1-10	A	CN 111459505 A (南京大学 等) 2020年 7月 28日 (2020 - 07 - 28) 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
A	CN 111445026 A (东南大学) 2020年 7月 24日 (2020 - 07 - 24) 说明书第[0008]-[0117]段, 图1-4	1-10																					
A	CN 111031102 A (哈尔滨工业大学) 2020年 4月 17日 (2020 - 04 - 17) 全文	1-10																					
A	CN 107766889 A (济南浪潮高科技投资发展有限公司) 2018年 3月 6日 (2018 - 03 - 06) 全文	1-10																					
A	CN 109961097 A (西北大学) 2019年 7月 2日 (2019 - 07 - 02) 全文	1-10																					
A	CN 110443298 A (华中科技大学) 2019年 11月 12日 (2019 - 11 - 12) 全文	1-10																					
A	CN 111459505 A (南京大学 等) 2020年 7月 28日 (2020 - 07 - 28) 全文	1-10																					
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																							
国际检索实际完成的日期		国际检索报告邮寄日期																					
2021年 8月 4日		2021年 8月 19日																					
ISA/CN的名称和邮寄地址		授权官员																					
中国国家知识产权局(ISA/CN) 中国 北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		李慧 电话号码 (86-512) 88995679																					

C. 相关文件		
类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	Zichuan Xu等. "Energy-Aware Inference Offloading for DNN-Driven Applications in Mobile Edge Clouds" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 2020年 11月 10日 (2020 - 11 - 10), ISSN: 1558-2183, 第799-814页	1-10

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/140017

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	111445026	A	2020年 7月 24日	无	
CN	111031102	A	2020年 4月 17日	无	
CN	107766889	A	2018年 3月 6日	无	
CN	109961097	A	2019年 7月 2日	无	
CN	110443298	A	2019年 11月 12日	无	
CN	111459505	A	2020年 7月 28日	无	