

(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
2024 年 12 月 19 日 (19.12.2024)



(10) 国际公布号
WO 2024/255130 A1

(51) 国际专利分类号:
G06N 3/08 (2023.01) H04L 41/16 (2022.01)

(21) 国际申请号: PCT/CN2023/134991

(22) 国际申请日: 2023 年 11 月 29 日 (29.11.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310713616.2 2023年6月16日 (16.06.2023) CN

(71) 申请人: 南京邮电大学 (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) [CN/CN]; 中国江苏省南京市栖霞区文苑路9号, Jiangsu 210023 (CN)。

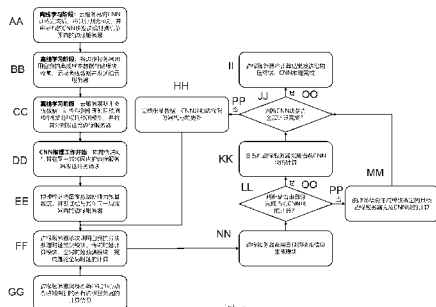
(72) 发明人: 郭永安(GUO, Yongan); 中国江苏省南京市栖霞区文苑路9号, Jiangsu 210023 (CN)。周金粮(ZHOU, Jinliang); 中国江苏省南京市栖霞区文苑路9号, Jiangsu 210023 (CN)。奚城科(XI, Chengke); 中国江苏省南京市栖霞区文苑路9号, Jiangsu 210023 (CN)。王宇翱(WANG, Yuao); 中国江苏省南京市栖霞区文苑路9号, Jiangsu 210023 (CN)。

(74) 代理人: 南京经纬专利商标代理有限公司 (NANJING JINGWEI PATENT & TRADEMARK AGENCY CO., LTD); 中国江苏省南京市鼓楼区中山路179号12楼B座, Jiangsu 210005 (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: GLOBAL DELAY OPTIMIZATION-ORIENTED CLOUD-EDGE-END COLLABORATIVE CNN INFERENCE METHOD AND SYSTEM

(54) 发明名称: 面向全局时延优化的云边端协同CNN推理方法及系统



AA Offline learning phase: after a cloud server completes the training of a CNN, the cloud server divides the CNN into n blocks and sends all the CNN blocks to edge servers within the communication range of the cloud server

BB Offline learning phase: each edge server calls its own offline sample data storage module, collects and records offline data, and sends same to the cloud server

CC Offline learning phase: the cloud server performs training on the basis of the offline data to obtain an inference delay prediction model and a delay loss prediction model, and separately sends same to the edge servers

DD CNN inference work begins: a physical terminal sends a task request to the edge server in the same local area network as the physical terminal

EE The physical terminal processes image data into tensor data, and sends same to the edge server in the same local area network as the physical terminal

FF The edge server sequentially calls its own block inference delay prediction module, transmission delay calculation module and global delay prediction module to complete the calculation of theoretical global delay

GG Calculate information of all the edge servers dynamically sensed by a situational awareness center of the edge server

HH Complete the update of the tensor data, the CNN blocks and an edge server code

II The edge server sends the calculation result to the physical terminal, and CNN inference is completed

JJ Determine whether the calculation of all the CNN blocks has been completed

KK The current edge server completes the calculation of the current CNN block

LL Determine whether to complete the calculation of the current CNN block by itself

MM Another edge server selected by the decision-making information generation module completes the calculation of the CNN block

NN The edge server calls its own decision-making information generation module

OO Yes

PP No

(57) Abstract: Disclosed in the present invention is a global delay optimization-oriented cloud-edge-end collaborative CNN inference method. A block inference delay prediction model is constructed and trained, theoretical data transmission delay is calculated, a delay loss prediction model is constructed and trained, and a global delay prediction model is constructed; the global delay prediction model is used to calculate the theoretical global delay of each edge server, and the minimization of global delay is taken as an optimization objective, so as to determine the selection of an edge server participating in collaborative inference and the allocation of inference task volumes. The present invention also provides a global delay optimization-oriented cloud-edge-end collaborative CNN inference system. Compared with conventional collaborative inference research work which focuses only on delay optimization of current inference tasks in terms of delay indicator settings, the method takes possible delay influence of processing of current inference tasks on other inference tasks processed in parallel into consideration for decision making. In addition, the present invention further provides a CNN segmentation optimization method, which can effectively reduce the decision-making delay in collaborative inference.

(57) 摘要: 本发明公开了面向全局时延优化的云边端协同CNN推理方法, 通过构建训练分块推理时延预测模型、计算理论数据传输时延、构建并训练时延损耗预测模型、构建全局时延预测模型; 利用全局时延预测模型计算各个边缘服务器的理论全局时延, 以全局时延最小化为优化目标, 决定参与协同推理的边缘服务器的选择和推理任务量的分配。本发明同时提出面向全局时延优化的云边端协同CNN推理系统。相较于传统的协同推理研究工作在时延指标制定方面仅关注当前推理任务的时延优化, 本方法将处理当前推理任务对其他并行处理的推理任务可能产生的时延影响也纳入了决策考量范围, 同时, 本发明还提供一种CNN分割优化方法, 能够有效节省协同推理中的决策时延。

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, 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
(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))。

面向全局时延优化的云边端协同 CNN 推理方法及系统

技术领域

[0001] 本发明属于云边端协同计算领域，具体涉及面向全局时延优化的云边端协同 CNN 推理方法及系统。

背景技术

[0002] CNN (Convolutional Neural Networks, 卷积神经网络) 的最新进展推动了越来越多的网络边缘侧智能应用，例如智能家居、智能工厂和智能城市。要在资源受限的物理设备上部署计算密集型 CNN，传统方法依赖于将推理工作卸载到远程云或在本地端设备上优化计算。然而，云辅助方法受到不可靠和延迟显著的广域网的影响，本地计算方法受限于端设备有限的计算能力。

[0003] 为了满足低时延和高准确率的 CNN 推理需求，新兴的云边端协同推理计算范式成为研究重点。然而，目前的协同推理研究工作在时延指标制定方面仅关注当前推理任务的时延优化，未考虑对其他并行处理的推理任务可能产生的时延影响。

发明内容

[0004] 本发明所要解决的技术问题在于：针对现有的云边端协同推理计算中未考虑对其他并行处理的推理任务可能产生的时延影响问题，提供了面向全局时延优化的云边端协同 CNN 推理方法及系统，更好地将云服务器和边缘计算范式结合起来，充分挖掘边缘服务器的计算潜力，同时本发明提出的 CNN 分割优化方法，在 CNN 块与块之间做协同推理决策，能够有效节省协同推理中的决策时延。

[0005] 为解决以上技术问题，本发明提供如下技术方案：面向全局时延优化的云边端协同 CNN 推理方法，包括如下步骤：

S1、基于已训练好的并且已划分为 N 层 CNN 模型，对该 N 层 CNN 模型划分为 n 块： $CNN_1, CNN_2, \dots, CNN_j, \dots, CNN_n$ ， $1 \leq j \leq n$ ， $n > 1$ ；

S2、离线学习阶段：以已知的边缘服务器 e_I 自身浮点计算能力 FS_I 、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的操作层层数为输入，以对应的边缘服务器 e_I 计算待处理的 CNN 块产生的实际分块推理时延为输出，构建并训练分块推理时延预测模型；所述操作层包括卷积层、全连接层和池化层；

S3、以已知的边缘服务器自身的浮点计算能力、待处理的 CNN 块的计算复杂度、正在处理的 CNN 块的平均计算复杂度、正在处理的 CNN 块包含的操作层层数为输入，以对应的边缘服务器计算待处理的 CNN 块对正在处理的 CNN 块产生的实际时延损耗为输出，构建并训练时延损耗预测模型；

S4、基于物理终端将初始图像数据处理为分辨率相同、数据量大小相等的张量数据，并将张量数据发送给与其在同一局域网内的边缘服务器；边缘服务器采集其与局域网内其他边缘服务器之间网络带宽，计算理论数据传输时延；

S5、将实际的边缘服务器自身浮点计算能力、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的层数输入至步骤 S2 得到的分块推理时延预测模型，计算得到边缘服务器计算待处理 CNN 块产生的理论分块推理时延；

将实际的边缘服务器自身的浮点计算能力、待处理的 CNN 块的计算复杂度、正在处理的 CNN 块的平均计算复杂度、正在处理的 CNN 块包含的操作层层数输入至步骤 S3 得到的时延损耗预测模型中，计算得到边缘服务器计算待处理的 CNN 块对正在处理的 CNN 块产生的理论时延损耗；

S6、以 S4 得到的理论数据传输时延、步骤 S5 得到的边缘服务器计算待处理 CNN 块产生的理论分块推理时延、边缘服务器计算 CNN 块对正在处理的 CNN 块产生的理论时延损耗为输入、以边缘服务器理论全局时延为输出，构建全局时延预测模型；

S7、利用全局时延预测模型计算各个边缘服务器的理论全局时延；

S8、判断当前边缘服务器的理论全局时延是否小于等于其他边缘服务器的理论全局时延，是则由当前边缘服务器计算待处理的 CNN 块，然后执行步骤 S9；否则将待处理的 CNN 块传输至其他边缘服务器计算，由选定的其他边缘服务器完成待处理 CNN 块的计算，然后执行步骤 S9，完成对下一个 CNN 块的决策；

S9、判断 CNN 块是否全部计算完成，是则边缘服务器将 CNN 块的计算结果发送给发出任务请求的物理终端，否则返回执行步骤 S7，完成对下一个 CNN 块的决策。

[0006] 进一步地，前述的步骤 S2 中，分块推理时延预测模型如下式：

$$e_I_tmd_j_p = f_1(FS_I, e_I_Fs_{CNN_pt}, Fs_{CNN_avg}, m),$$

式中， FS_I 是边缘服务器 e_I 自身的浮点计算能力， $e_I_Fs_{CNN_pt}$ 是自身正在处理的

CNN 块 $e_I_CNN_j_pt$ 的计算复杂度， Fs_{CNN_avg} 是待处理的 CNN 块 CNN_j 的

平均计算复杂度、 m 是 CNN_j 包含的操作层层数， $e_I_tmd_j_p$ 是边缘服务器 e_I 计算 CNN_j 产生的理论分块推理时延输出。

[0007] 进一步地，前述的步骤 S3 中，时延损耗预测模型如下式：

$$e_I_t_{loss}_p = f_2(FS_I, Fs_{CNN_j}, e_I_Fs_{CNN_j}_avg_pt, m_pt),$$

式中， FS_I 是边缘服务器的浮点计算能力， Fs_{CNN_j} 是待处理的 CNN 块 CNN_j 的计算复杂度， $e_I_Fs_{CNN_j}_avg_pt$ 是正在处理的 CNN 块 $e_I_CNN_j_pt$ 的平均计算复杂度， m_pt 是正在处理的 CNN 块包含的操作层层数， $e_I_t_{loss}_p$ 是边缘服务器 e_I 计算 CNN_j 对正在处理的 $e_I_CNN_j_pt$ 产生的理论时延损耗。

[0008] 进一步地，前述的步骤 S4 中，按如下公式计算理论数据传输时延：

$$t_{II} = Td_{q-1} / bw_{II},$$

式中， Td_{q-1} 是张量数据， bw_{II} 是边缘服务器之间的网络带宽。

[0009] 进一步地，前述的步骤 S6 中，所述全局时延预测模型如下式：

$$\begin{cases} e_I_T_{CNN} = e_I_tmd_j_p + t_{II} + e_I_t_{loss}_p \\ e_I_t_{loss}_p = f_2(FS_I, Fs_{CNN}, e_I_Fs_{CNN}_avg_pt, m_pt) \end{cases},$$

$$\begin{cases} e_I_T_{CNN} = e_I_tmd_j_p + e_I_t_{loss}_p \\ e_I_t_{loss}_p = f_2(FS_I, Fs_{CNN}, e_I_Fs_{CNN}_avg_pt, m_pt) \end{cases}.$$

[0010] 进一步地，前述的步骤 S1 中， CNN_j 中含有至少两个连续的张量数据，对于任一含有 m 个张量数据的 CNN_j 而言，它含有的张量数据和操作层分别为：

$Td_{j-1}, Td_{j-2}, \dots, Td_{j-m}, Ly_{j-1}, Ly_{j-2}, \dots, Ly_{j-m}, m > 1$ ，其中 Td_{j-1} 的计算复杂度大于 CNN_j 中其他任意一个张量数据的计算复杂度；对于任意两个相邻的 CNN 块，即 CNN_p 和 CNN_q ： $CNN_1, \dots, CNN_p, CNN_q, \dots, CNN_n, 1 \leq p < q \leq n, q - p = 1$ ，

CNN_p 中的最后一个张量数据和 CNN_q 中的第一个张量数据相同。

[0011] 进一步地，前述的步骤 S4 中，基于物理终端将初始图像数据处理为分辨率相同、数据量大小相等的张量数据，并将张量数据发送给与其在同一局域网内的边缘服务器，具体为：初始图像数据被处理为数据量大小相同、分辨率大小相等的图像特征数据，图像特征数据作为张量数据 Td_{1-1} 输入 CNN_1 ， Td_{1-1} 经过 CNN_1 中的操作层计算后，生成张量数据 Td_{2-1} 并传入 CNN_2 ，即：在 CNN_j 中，张量数据 Td_{j-1} 经过操作层 $Ly_{j-1}, Ly_{j-2}, \dots, Ly_{j-m}$ 的计算后，生成张量数据 Td_{j-m} 并传入下一个 CNN 块，CNN 块 CNN_n 中最后一个张量数据的生成标志着 CNN 推理的结束。

[0012] 进一步地，前述的面向全局时延优化的云边端协同 CNN 推理方法，CNN 块 CNN_j 中所有操作层的计算复杂度共计为： $Fs_{CNN_j} = \sum_{i=1}^{M_j} Fs_{Ly_i}$ ， Fs_{CNN_j} 记为 CNN_j 的计算复杂度， CNN_j 的平均计算复杂度记为： $Fs_{CNN_j-avg} = \frac{Fs_{CNN_j}}{m}$ ；

计算 CNN_j 包含的所有操作层所产生的分块推理时延为 tmd_j ， Fs_{Ly_i} 为任一操作层 Ly_i 的计算复杂度。

[0013] 进一步地，前述的面向全局时延优化的云边端协同 CNN 推理方法，CNN 各个操作层 Ly_i 的计算复杂度为：

$$Fs_{Ly_i-CONV} = 2 * H * W * (C_{in} * K^2 + 1) * C_{out}$$

$$Fs_{Ly_i-FC} = (2 * I - 1) * Q$$

式中，操作层 Ly_i 为卷积层，其计算复杂度为 Fs_{Ly_i-CONV} ；操作层 Ly_i 为全连接层，其计算复杂度为 Fs_{Ly_i-FC} ； H 、 W 、 C_{in} 依次代表输入操作层 Ly_i 的张量数据 Td_i 的高度、宽度、通道数， K 代表输入操作层 Ly_i 的卷积核的边长， C_{out} 代表输出操作层 Ly_i 的张量数据 Td_{i+1} 的通道数， I 、 Q 分别代表输入操作层 Ly_i 的张量数据 Td_i 的维

数、输出操作层 Ly_i 的张量数据 Td_{i+1} 的维数。

[0014] 本发明另一方面提出面向全局时延优化的云边端协同 CNN 推理系统，包括：云服务器 C ，云服务器 C 的通信范围内至少设有两个边缘服务器，所述边缘服务器部署在 WIFI 接入点或基站上，各边缘服务器的通信范围内至少设有一个物理终端；针对云服务器 C 通信范围内的任意一个边缘服务器 e_I ，边缘服务器 e_I 通信范围内与其物理距离小于预设距离 X 的其他 O 个边缘服务器记为： $e_1, e_2, \dots, e_J, \dots, e_O$ ， $1 \leq J \leq O$ ， $O > 1$ ，且这 O 个边缘服务器也在云服务器 C 的通信范围内，此 O 个边缘服务器同边缘服务器 e_I 一起组成边缘集群；

所述云服务器包括：卷积神经网络、模型训练中心、通信模块；

所述边缘服务器包括：态势感知中心、卷积神经网络、策略生成中心、通信模块；

所述物理终端包括：通信模块；

所述模型训练中心用于训练卷积神经网络、推理时延预测模型和时延损耗预测模型；

所述卷积神经网络用于已经训练完备的、服务于智能应用程序的、被分割为 n 块 CNN 块 CNN_j 的 n 层 CNN；

所述通信模块用于云服务器、边缘服务器和物理终端之间数据发送、接收；

所述态势感知中心包括工作负载感知模块和网络遥测模块；所述工作负载感知模块用于采集边缘服务器自身的浮点计算能力 FS_I 、自身正在负责推理的 CNN 块 $e_I_CNN_j_pt$ 的计算复杂度 $e_I_Fs_{CNN_j_pt}$ 和平均计算复杂度 $e_I_Fs_{CNN_j_avg_pt}$ 、其他 O 个边缘服务器 e_J 的浮点计算能力 FS_J 、其他 O 个边缘服务器 e_J 正在负责推理的 CNN 块 $e_J_CNN_j_pt$ 的计算复杂度 $e_I_Fs_{CNN_j_pt}$ 和平均计算复杂度 $e_I_Fs_{CNN_j_avg_pt}$ ；

所述网络遥测模块用于采集边缘服务器 e_I 和边缘服务器 e_J 间的网络带宽 bw_{IJ} ；

所述策略生成中心包括分块推理时延预测模块、传输时延计算模块、全局时延预测模块、离线样本数据存储模块和决策信息生成模块；

所述分块推理时延预测模块用于以已知的边缘服务器 e_I 自身浮点计算能力 FS_I 、正在处理

的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的层数为输入，以对应的边缘服务器 e_I 计算待处理的 CNN 块产生的实际分块推理时延为输出，构建并训练分块推理时延预测模型；

所述传输时延计算模块用于基于物理终端将图像数据处理为分辨率相同、数据量大小相等的张量数据，并将张量数据发送给与其在同一局域网内的边缘服务器；边缘服务器采集其与局域网内其他边缘服务器之间网络带宽，计算理论数据传输时延；

所述全局时延预测模块用于以理论数据传输时延、边缘服务器计算待处理 CNN 块产生的理论分块推理时延、边缘服务器计算 CNN 块对正在处理的 CNN 块产生的理论时延损耗为输入、以边缘服务器理论全局时延为输出，构建全局时延预测模型；

所述离线样本存储模块用于采集、存储边缘服务器 e_I 在自身浮点计算能力为 FS_I 、自身正在处理的 CNN 块 $e_I_CNN_j_pt$ 的计算复杂度为 $e_I_Fs_{CNN_j_pt}$ 时，计算层数为 m 层且平均计算复杂度为 $Fs_{CNN_j_avg}$ 的 CNN 块 CNN_j 产生的实际分块推理时延 $e_I_tmd_j$ ；采集、存储边缘服务器 e_I 在自身浮点计算能力为 FS_I 、自身正在处理的操作层层数为 m_pt 且平均计算复杂度为 $e_I_Fs_{CNN_j_avg_pt}$ 的 CNN 块 $e_I_CNN_j_pt$ 时，处理计算复杂度为 Fs_{CNN_j} 的 CNN 块 CNN_j 对正在计算的 $e_I_CNN_j_pt$ 产生的实际时延损耗 t_{loss} ；

所述决策信息生成模块用于判断当前边缘服务器的理论全局时延是否小于等于其他边缘服务器的理论全局时延，是则由当前边缘服务器计算待处理的 CNN 块，然后判断 CNN 块是否全部计算完成，是则边缘服务器将 CNN 块的计算结果发送给发出任务请求的物理终端；否则将待处理的 CNN 块传输至其他边缘服务器计算，直到 CNN 块全部计算完成。

[0015] 相较于现有技术，本发明的有益效果如下：区别于以端设备和云计算中心为主导的 CNN 协同推理系统，本发明通过将云服务器和边缘计算范式结合起来，其中云服务器负责模型的训练，在边缘集群内部完成 CNN 的协同推理工作，充分挖掘边缘服务器的计算潜力。

[0016] 本发明通过对 CNN 协同推理过程中的分块推理时延、数据传输时延和时延损耗进行理论建模，并以全局时延最小化为优化目标，决定参与协同推理的边缘服务器的选择和推理

任务量的分配。

[0017] 相较于传统的协同推理研究工作在时延指标制定方面仅关注当前推理任务的时延优化，本发明将处理当前推理任务对其他并行处理的推理任务可能产生的时延影响也纳入了决策考量范围。同时，本发明还提供一种 CNN 分割优化方法，在 CNN 块与块之间做协同推理决策，能够有效节省协同推理中的决策时延。以边缘集群为主导的 CNN 协同推理工作，将计算工作放在离物理终端更近的边缘侧，可有效保证计算数据的安全性和降低网络带宽的占用率。

附图说明

[0018] 图 1 为本发明的 CNN 分割优化原理图。

[0019] 图 2 为本发明的技术原理图。

[0020] 图 3 为本发明策略生成中心的模块组成示意图。

[0021] 图 4 为本发明的工作流程图。

具体实施方式

[0022] 为了更了解本发明的技术内容，特举具体实施例并配合所附图式说明如下。

[0023] 在本发明中参照附图来描述本发明的各方面，附图中示出了许多说明性实施例。本发明的实施例不局限于附图所述。应当理解，本发明通过上面介绍的多种构思和实施例，以及下面详细描述的构思和实施方式中的任意一种来实现，这是因为本发明所公开的构思和实施例并不限于任何实施方式。另外，本发明公开的一些方面可以单独使用，或者与本发明公开的其他方面的任何适当组合来使用。

[0024] 结合图 1，CNN 为多层结构，对于一个 N 层 CNN 而言，它的操作层有 N 个：

$Ly_1, Ly_2, \dots, Ly_i, \dots, Ly_N$ ， $1 \leq i \leq N+1$ ， $N > 1$ ，它的张量数据有 $N+1$ 个：

$Td_1, Td_2, \dots, Td_i, \dots, Td_{N+1}$ ，其中张量数据 Td_i 经过操作层 Ly_i 计算后产生张量数据

Td_{i+1} 。CNN 推理产生的时延消耗主要由计算各个操作层 Ly_i 产生的分层计算时延 Tmd_i

组成。操作层 Ly_i 可分为三类：卷积层、全连接层和池化层，其中卷积层的计算复杂度最高，

全连接层次之，池化层运算最简单，可以忽略。CNN 各个操作层 Ly_i 的计算复杂度用下式

来表示：

$$Fs_{Ly_i-CONV} = 2 * H * W * (C_m * K^2 + 1) * C_{out};$$

$$Fs_{Ly_i-FC} = (2 * I - 1) * Q$$

[0025] 式中，操作层 Ly_i 为卷积层，其计算复杂度为 Fs_{Ly_i-CONV} ；操作层 Ly_i 为全连接层，其计算复杂度为 Fs_{Ly_i-FC} ； H 、 W 、 C_m 依次代表输入操作层 Ly_i 的张量数据 Td_i 的高度、宽度、通道数， K 代表输入操作层 Ly_i 的卷积核的边长， C_{out} 代表输出操作层 Ly_i 的张量数据 Td_{i+1} 的通道数， I 、 Q 分别代表输入操作层 Ly_i 的张量数据 Td_i 的维数、输出操作层 Ly_i 的张量数据 Td_{i+1} 的维数。

[0026] 对于一个 N 层 CNN，本发明对其进行分割优化操作并分为 n 块：

$CNN_1, CNN_2, \dots, CNN_j, \dots, CNN_n$ ， $1 \leq j \leq n$ ， $n > 1$ 。任一 CNN 块 CNN_j 的组成特征如下：

每个 CNN_j 中含有至少两个连续的张量数据；

对于任一含有 m 个张量数据的 CNN_j 而言，它含有的张量数据和操作层分别为：

$Td_{j-1}, Td_{j-2}, \dots, Td_{j-m}$ ， $Ly_{j-1}, Ly_{j-2}, \dots, Ly_{j-m}$ ， $m > 1$ ，其中 Td_{j-1} 的计算复杂度大于 CNN_j 中其他任意一个张量数据的计算复杂度；

对于任意两个相邻的 CNN 块： CNN_p 和 CNN_q （ $CNN_1, \dots, CNN_p, CNN_q, \dots, CNN_n$ ， $1 \leq p < q \leq n$ ， $q - p = 1$ ），则 CNN_p 中的最后一个张量数据和 CNN_q 中的第一个张量数据是相同的；

在具体的 CNN 推理过程中，初始图像数据被处理为数据量大小相同、分辨率大小相等的图像特征数据，图像特征数据作为张量数据 Td_{1-1} 输入 CNN_1 ， Td_{1-1} 经过 CNN_1 中的一系列操作层计算后，生成张量数据 Td_{2-1} 并传入 CNN_2 ，即：在 CNN_j 中，张量数据 Td_{j-1} 经

过操作层 $Ly_{j-1}, Ly_{j-2}, \dots, Ly_{j-m}$ 的计算后, 生成张量数据 Td_{j-m} 并传入下一个 CNN 块, CNN 块 CNN_n 中最后一个张量数据的生成标志着 CNN 推理的结束。

[0027] CNN 块 CNN_j 中所有操作层的计算复杂度共计为 $FS_{CNN_j} = \sum_{i=1}^m FS_{Ly_i}$, FS_{CNN_j}

记为 CNN_j 的计算复杂度, CNN_j 的平均计算复杂度记为:

$$FS_{CNN_j-avg} = FS_{CNN_j} / m;$$

计算 CNN_j 包含的所有操作层所产生的分块推理时延记为: tmd_j , FS_{Ly_i} 为任一操作层 Ly_i 的计算复杂度。

[0028] 参考图 4, 本发明提出的 1.面向全局时延优化的云边端协同 CNN 推理方法, 包括如下步骤:

S1、基于已训练好的并且已划分为 N 层 CNN 模型, 对该 N 层 CNN 模型划分为 n 块:

$$CNN_1, CNN_2, \dots, CNN_j, \dots, CNN_n, \quad 1 \leq j \leq n, \quad n > 1;$$

S2、离线学习阶段: 以已知的边缘服务器 e_I 自身浮点计算能力 FS_I 、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的操作层层数为输入, 以对应的边缘服务器 e_I 计算待处理的 CNN 块产生的实际分块推理时延为输出, 构建并训练分块推理时延预测模型; 所述操作层包括卷积层、全连接层和池化层;

所述分块推理时延预测模型如下式:

$$e_I - tmd_j - p = f_1(FS_I, e_I - FS_{CNN_j} - pt, FS_{CNN_j} - avg, m),$$

[0029] 式中, FS_I 是缘服务器 e_I 自身的浮点计算能力, $e_I - FS_{CNN_j} - pt$ 是自身正在处理的 CNN 块 $e_I - CNN_j - pt$ 的计算复杂度, $FS_{CNN_j} - avg$ 是待处理的 CNN 块 CNN_j 的平均计算复杂度、 m 是 CNN_j 包含的操作层层数, $e_I - tmd_j - p$ 是边缘服务器 e_I 计算 CNN_j 产生的理论分块推理时延输出。

[0030] S3、以已知的边缘服务器自身的浮点计算能力、待处理的 CNN 块的计算复杂度、正在处理的 CNN 块的平均计算复杂度、正在处理的 CNN 块包含的操作层层数为输入，以对应的边缘服务器计算待处理的 CNN 块对正在处理的 CNN 块产生的实际时延损耗为输出，构建并训练时延损耗预测模型；所述时延损耗预测模型如下式：

$$e_I - t_{loss} - p = f_2(FS_I, Fs_{CNN_j}, e_I - Fs_{CNN_j} - avg - pt, m - pt),$$

式中， FS_I 是边缘服务器的浮点计算能力， Fs_{CNN_j} 是待处理的 CNN 块 CNN_j 的计算复杂度， $e_I - Fs_{CNN_j} - avg - pt$ 是正在处理的 CNN 块 $e_I - CNN_j - pt$ 的平均计算复杂度， $m - pt$ 是正在处理的 CNN 块包含的操作层层数， $e_I - t_{loss} - p$ 是边缘服务器 e_I 计算 CNN_j 对正在处理的 $e_I - CNN_j - pt$ 产生的理论时延损耗。

[0031] S4、基于物理终端将初始图像数据处理为分辨率相同、数据量大小相等的张量数据，并将张量数据发送给与其在同一局域网内的边缘服务器；边缘服务器采集其与局域网内其他边缘服务器之间网络带宽，按如下公式计算理论数据传输时延：

$$t_{IT} = \frac{Td_{q-1}}{bw_{IJ}},$$

式中， Td_{q-1} 是张量数据， bw_{IJ} 是边缘服务器之间的网络带宽。

[0032] S5、将实际的边缘服务器自身浮点计算能力、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的层数输入至步骤 S2 得到的分块推理时延预测模型，计算得到边缘服务器计算待处理 CNN 块产生的理论分块推理时延；将实际的边缘服务器自身的浮点计算能力、待处理的 CNN 块的计算复杂度、正在处理的 CNN 块的平均计算复杂度、正在处理的 CNN 块包含的操作层层数输入至步骤 S3 得到的时延损耗预测模型中，计算得到边缘服务器计算待处理的 CNN 块对正在处理的 CNN 块产生的理论时延损耗；

S6、以 S4 得到的理论数据传输时延、步骤 S5 得到的边缘服务器计算待处理 CNN 块产生的理论分块推理时延、边缘服务器计算 CNN 块对正在处理的 CNN 块产生的理论时延损耗为输入、以边缘服务器理论全局时延为输出，构建全局时延预测模型；所述全局时延预测模型如下式：

$$\begin{cases} e_j - T_{CNN} = e_j - tmd_j - p + t_{ij} + e_i - t_{loss} - p \\ e_j - t_{loss} - p = f_2(FS_j, FS_{CNN}, e_j - FS_{CNN} - avg_pt, m_pt) \end{cases}$$

$$\begin{cases} e_i - T_{CNN} = e_i - tmd_i - p + e_i - t_{loss} - p \\ e_i - t_{loss} - p = f_2(FS_i, FS_{CNN}, e_i - FS_{CNN} - avg_pt, m_pt) \end{cases}$$

[0033] S7、利用全局时延预测模型计算各个边缘服务器的理论全局时延；

S8、判断当前边缘服务器的理论全局时延是否小于等于其他边缘服务器的理论全局时延，是则由当前边缘服务器计算待处理的 CNN 块，然后执行步骤 S9；否则将待处理的 CNN 块传输至其他边缘服务器计算，由选定的其他边缘服务器完成待处理 CNN 块的计算，然后执行步骤 S9，完成对下一个 CNN 块的决策；

S9、判断 CNN 块是否全部计算完成，是则边缘服务器将 CNN 块的计算结果发送给发出任务请求的物理终端，否则返回执行步骤 S7，完成对下一个 CNN 块的决策。

[0034] 本发明另一方面提出面向全局时延优化的云边端协同 CNN 推理系统，包括：云服务器 C，云服务器 C 的通信范围内至少设有两个边缘服务器，所述边缘服务器部署在 WIFI 接入点或基站上，各边缘服务器的通信范围内至少设有一个物理终端；针对云服务器 C 通信范围内的任意一个边缘服务器 e_i ，边缘服务器 e_i 通信范围内与其物理距离小于预设距离 x 的其他 o 个边缘服务器记为： $e_1, e_2, \dots, e_j, \dots, e_o$ ， $1 \leq j \leq o$ ， $o > 1$ ，且这 o 个边缘服务器也在云服务器 C 的通信范围内，此 o 个边缘服务器同边缘服务器 e_i 一起组成边缘集群；

所述云服务器包括：卷积神经网络、模型训练中心、通信模块；

所述边缘服务器包括：态势感知中心、卷积神经网络、策略生成中心、通信模块；

所述物理终端包括：通信模块；

所述模型训练中心用于训练卷积神经网络、推理时延预测模型和时延损耗预测模型；

所述卷积神经网络用于已经训练完备的、服务于智能应用程序的、被分割为 n 块 CNN 块 CNN_j 的 N 层 CNN；

所述通信模块用于云服务器、边缘服务器和物理终端之间数据发送、接收；

所述态势感知中心包括工作负载感知模块和网络遥测模块；所述工作负载感知模块用于采集边缘服务器自身的浮点计算能力 FS_i 、自身正在负责推理的 CNN 块 $e_i - CNN_j - pt$ 的计

算复杂度 $e_I - Fs_{CNN_j} - pt$ 和平均计算复杂度 $e_I - Fs_{CNN_j} - avg - pt$ 、其他 O 个边缘服务器 e_J 的浮点计算能力 FS_I 、其他 o 个边缘服务器 e_J 正在负责推理的 CNN 块 $e_J - CNN_j - pt$ 的计算复杂度 $e_J - Fs_{CNN_j} - pt$ 和平均计算复杂度 $e_J - Fs_{CNN_j} - avg - pt$

[0035] 所述网络遥测模块用于采集边缘服务器 e_I 和边缘服务器 e_J 间的网络带宽 bw_{IJ} ；

参考图 3，所述策略生成中心包括分块推理时延预测模块、传输时延计算模块、全局时延预测模块、离线样本数据存储模块和决策信息生成模块；

所述分块推理时延预测模块用于以已知的边缘服务器 e_I 自身浮点计算能力 FS_I 、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的层数为输入，以对应的边缘服务器 e_I 计算待处理的 CNN 块产生的实际分块推理时延为输出，构建并训练分块推理时延预测模型；

所述传输时延计算模块用于基于物理终端将图像数据处理为分辨率相同、数据量大小相等的张量数据，并将张量数据发送给与其在同一局域网内的边缘服务器；边缘服务器采集其与局域网内其他边缘服务器之间网络带宽，计算理论数据传输时延；

所述全局时延预测模块用于以理论数据传输时延、边缘服务器计算待处理 CNN 块产生的理论分块推理时延、边缘服务器计算 CNN 块对正在处理的 CNN 块产生的理论时延损耗为输入、以边缘服务器理论全局时延为输出，构建全局时延预测模型；

所述离线样本存储模块用于采集、存储边缘服务器 e_I 在自身浮点计算能力为 FS_I 、自身正在处理的 CNN 块 $e_I - CNN_j - pt$ 的计算复杂度为 $e_I - Fs_{CNN_j} - pt$ 时，计算层数为 m 层且平均计算复杂度为 $Fs_{CNN_j} - avg$ 的 CNN 块 CNN_j 产生的实际分块推理时延 $e_I - tmd_j$ ；采集、存储边缘服务器 e_I 在自身浮点计算能力为 FS_I 、自身正在处理的操作层层数为 $m - pt$ 且平均计算复杂度为 $e_I - Fs_{CNN_j} - avg - pt$ 的 CNN 块 $e_I - CNN_j - pt$ 时，处理计算复杂度为 Fs_{CNN_j} 的 CNN 块 CNN_j 对正在计算的 $e_I - CNN_j - pt$ 产生的实际时延损耗 t_{loss} ；

所述决策信息生成模块用于判断当前边缘服务器的理论全局时延是否小于等于其他边缘服务器的理论全局时延，是则由当前边缘服务器计算待处理的 CNN 块，然后判断 CNN 块是否全部计算完成，是则边缘服务器将 CNN 块的计算结果发送给发出任务请求的物理终端；否则将待处理的 CNN 块传输至其他边缘服务器计算，直到 CNN 块全部计算完成。

[0036] 虽然本发明已以较佳实施例阐述如上，然其并非用以限定本发明。本发明所属技术领域中具有通常知识者，在不脱离本发明的精神和范围内，当可作各种的更动与润饰。因此，本发明的保护范围当视权利要求书所界定者为准。

1. 面向全局时延优化的云边端协同 CNN 推理方法，其特征在于，包括如下步骤：

S1、基于已训练好的并且已划分为 N 层 CNN 模型，对该 N 层 CNN 模型划分为 n 块： $CNN_1, CNN_2, \dots, CNN_j, \dots, CNN_n$ ， $1 \leq j \leq n$ ， $n > 1$ ；

S2、离线学习阶段：以已知的边缘服务器 e_I 自身浮点计算能力 FS_I 、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的操作层层数为输入，以对应的边缘服务器 e_I 计算待处理的 CNN 块产生的实际分块推理时延为输出，构建并训练分块推理时延预测模型；所述操作层包括卷积层、全连接层和池化层；

S3、以已知的边缘服务器自身的浮点计算能力、待处理的 CNN 块的计算复杂度、正在处理的 CNN 块的平均计算复杂度、正在处理的 CNN 块包含的操作层层数为输入，以对应的边缘服务器计算待处理的 CNN 块对正在处理的 CNN 块产生的实际时延损耗为输出，构建并训练时延损耗预测模型；

S4、基于物理终端将初始图像数据处理为分辨率相同、数据量大小相等的张量数据，并将张量数据发送给与其在同一局域网内的边缘服务器；边缘服务器采集其与局域网内其他边缘服务器之间网络带宽，计算理论数据传输时延；

S5、将实际的边缘服务器自身浮点计算能力、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的层数输入至步骤 S2 得到的分块推理时延预测模型，计算得到边缘服务器计算待处理 CNN 块产生的理论分块推理时延；

将实际的边缘服务器自身的浮点计算能力、待处理的 CNN 块的计算复杂度、正在处理的 CNN 块的平均计算复杂度、正在处理的 CNN 块包含的操作层层数输入至步骤 S3 得到的时延损耗预测模型中，计算得到边缘服务器计算待处理的 CNN 块对正在处理的 CNN 块产生的理论时延损耗；

S6、以 S4 得到的理论数据传输时延、步骤 S5 得到的边缘服务器计算待处理 CNN 块产生的理论分块推理时延、边缘服务器计算 CNN 块对正在处理的 CNN 块产生的理论时延损耗为输入、以边缘服务器理论全局时延为输出，构建全局时延预测模型；

S7、利用全局时延预测模型计算各个边缘服务器的理论全局时延；

S8、判断当前边缘服务器的理论全局时延是否小于等于其他边缘服务器的理论全局时延，是则由当前边缘服务器计算待处理的 CNN 块，然后执行步骤 S9；否则将待处理的 CNN 块传输至其他边缘服务器计算，由选定的其他边缘服务器完成待处理 CNN 块的计算，然后执行步骤 S9，完成对下一个 CNN 块的决策；

S9、判断 CNN 块是否全部计算完成，是则边缘服务器将 CNN 块的计算结果发送给发出任务请求的物理终端，否则返回执行步骤 S7，完成对下一个 CNN 块的决策。

2. 根据权利要求 1 所述的面向全局时延优化的云边端协同 CNN 推理方法，其特征在于，步骤 S2 中，分块推理时延预测模型如下式：

$$e_I_tmd_j_p = f_1(FS_I, e_I_Fs_{CNN_j_pt}, Fs_{CNN_j_avg}, m),$$

式中， FS_I 是边缘服务器 e_I 自身的浮点计算能力， $e_I_Fs_{CNN_j_pt}$ 是自身正在处理的 CNN 块 $e_I_CNN_j_pt$ 的计算复杂度， $Fs_{CNN_j_avg}$ 是待处理的 CNN 块 CNN_j 的平均计算复杂度、 m 是 CNN_j 包含的操作层层数， $e_I_tmd_j_p$ 是边缘服务器 e_I 计算 CNN_j 产生的理论分块推理时延输出。

3. 根据权利要求 2 所述的面向全局时延优化的云边端协同 CNN 推理方法，其特征在于，步骤 S3 中，时延损耗预测模型如下式：

$$e_I_t_{loss_p} = f_2(FS_I, Fs_{CNN_j}, e_I_Fs_{CNN_j_avg_pt}, m_pt),$$

式中， FS_I 是边缘服务器的浮点计算能力， Fs_{CNN_j} 是待处理的 CNN 块 CNN_j 的计算复杂度， $e_I_Fs_{CNN_j_avg_pt}$ 是正在处理的 CNN 块 $e_I_CNN_j_pt$ 的平均计算复杂度， m_pt 是正在处理的 CNN 块包含的操作层层数， $e_I_t_{loss_p}$ 是边缘服务器 e_I 计算 CNN_j 对正在处理的 $e_I_CNN_j_pt$ 产生的理论时延损耗。

4. 根据权利要求 3 所述的面向全局时延优化的云边端协同 CNN 推理方法，其特征在于，步骤 S4 中，按如下公式计算理论数据传输时延：

$$r_{IJ} = Id_{q-1} / bw_{IJ},$$

式中， Id_{q-1} 是张量数据， bw_{IJ} 是边缘服务器之间的网络带宽。

5. 根据权利要求 4 所述的面向全局时延优化的云边端协同 CNN 推理方法，其特征在于，步骤 S6 中，所述全局时延预测模型如下式：

$$\begin{cases} e_j - T_{\text{CNN}_j} = e_j - tmd_j - p + t_{ij} + e_i - t_{\text{oss}} - p \\ e_j - t_{\text{oss}} - p = f_2(FS_j, FS_{\text{CNN}_j}, e_j - FS_{\text{CNN}_j} - \text{avg} - pt, m - pt) \end{cases}$$

$$\begin{cases} e_i - T_{\text{CNN}_i} = e_i - tmd_i - p + e_i - t_{\text{oss}} - p \\ e_i - t_{\text{oss}} - p = f_2(FS_i, FS_{\text{CNN}_i}, e_i - FS_{\text{CNN}_i} - \text{avg} - pt, m - pt) \end{cases}$$

6. 根据权利要求 1 所述的面向全局时延优化的云边端协同 CNN 推理方法, 其特征在于, 步骤 S1 中, CNN_j 中含有至少两个连续的张量数据, 对于任一含有 m 个张量数据的 CNN_j 而言, 它含有的张量数据和操作层分别为: $Td_{j-1}, Td_{j-2}, \dots, Td_{j-m}$, $Ly_{j-1}, Ly_{j-2}, \dots, Ly_{j-m}$, $m > 1$, 其中 Td_{j-1} 的计算复杂度大于 CNN_j 中其他任意一个张量数据的计算复杂度; 对于任意两个相邻的 CNN 块, 即 CNN_p 和 CNN_q : $\text{CNN}_1, \dots, \text{CNN}_p, \text{CNN}_q, \dots, \text{CNN}_n$, $1 \leq p < q \leq n$, $q - p = 1$, CNN_p 中的最后一个张量数据和 CNN_q 中的第一个张量数据相同。

7. 根据权利要求 1 所述的面向全局时延优化的云边端协同 CNN 推理方法, 其特征在于, 所述步骤 S4 中, 基于物理终端将初始图像数据处理为分辨率相同、数据量大小相等的张量数据, 并将张量数据发送给与其在同一局域网内的边缘服务器, 具体为: 初始图像数据被处理为数据量大小相同、分辨率大小相等的图像特征数据, 图像特征数据作为张量数据 Td_{1-1} 输入 CNN_1 , Td_{1-1} 经过 CNN_1 中的操作层计算后, 生成张量数据 Td_{2-1} 并传入 CNN_2 , 即: 在 CNN_j 中, 张量数据 Td_{j-1} 经过操作层 $Ly_{j-1}, Ly_{j-2}, \dots, Ly_{j-m}$ 的计算后, 生成张量数据 Td_{j-m} 并传入下一个 CNN 块, CNN 块 CNN_n 中最后一个张量数据的生成标志着 CNN 推理的结束。

8. 根据权利要求 1 所述的面向全局时延优化的云边端协同 CNN 推理方法, 其特征在于, CNN 块 CNN_j 中所有操作层的计算复杂度共计为: $FS_{\text{CNN}_j} = \sum_{i=1}^m FS_{Ly_{j-i}}$, FS_{CNN_j} 记为 CNN_j 的计算复杂度, CNN_j 的平均计算复杂度记为: $FS_{\text{CNN}_j} - \text{avg} = FS_{\text{CNN}_j} / m$;

计算 CNN_j 包含的所有操作层所产生的分块推理时延为 tmd_j , Fs_{Ly_i} 为任一操作层 Ly_i 的计算复杂度。

9. 根据权利要求 8 所述的面向全局时延优化的云边端协同 CNN 推理方法, 其特征在于, CNN 各个操作层 Ly_i 的计算复杂度为:

$$Fs_{Ly_i} - CONV = 2 * H * W * (C_{in} * K^2 + 1) * C_{out},$$

$$Fs_{Ly_i} - FC = (2 * I - 1) * Q$$

式中, 操作层 Ly_i 为卷积层, 其计算复杂度为 $Fs_{Ly_i} - CONV$; 操作层 Ly_i 为全连接层, 其计算复杂度为 $Fs_{Ly_i} - FC$; H 、 W 、 C_{in} 依次代表输入操作层 Ly_i 的张量数据 Td_i 的高度、宽度、通道数, K 代表输入操作层 Ly_i 的卷积核的边长, C_{out} 代表输出操作层 Ly_i 的张量数据 Td_{i+1} 的通道数, I 、 Q 分别代表输入操作层 Ly_i 的张量数据 Td_i 的维数、输出操作层 Ly_i 的张量数据 Td_{i+1} 的维数。

10. 面向全局时延优化的云边端协同 CNN 推理系统, 其特征在于, 包括: 云服务器 C , 云服务器 C 的通信范围内至少设有两个边缘服务器, 所述边缘服务器部署在 WIFI 接入点或基站上, 各边缘服务器的通信范围内至少设有一个物理终端; 针对云服务器 C 通信范围内的任意一个边缘服务器 e_I , 边缘服务器 e_I 通信范围内与其物理距离小于预设距离 x 的其他 O 个边缘服务器记为: $e_1, e_2, \dots, e_J, \dots, e_O$, $1 \leq J \leq O$, $O > 1$, 且这 O 个边缘服务器也在云服务器 C 的通信范围内, 此 O 个边缘服务器同边缘服务器 e_I 一起组成边缘集群;

所述云服务器包括: 卷积神经网络、模型训练中心、通信模块;

所述边缘服务器包括: 态势感知中心、卷积神经网络、策略生成中心、通信模块;

所述物理终端包括: 通信模块;

所述模型训练中心用于训练卷积神经网络、推理时延预测模型和时延损耗预测模型;

所述卷积神经网络用于已经训练完备的、服务于智能应用程序的、被分割为 n 块 CNN 块

CNN_j 的 n 层 CNN;

所述通信模块用于云服务器、边缘服务器和物理终端之间数据发送、接收;

所述态势感知中心包括工作负载感知模块和网络遥测模块; 所述工作负载感知模块用于采集边缘服务器自身的浮点计算能力 FS_I 、自身正在负责推理的 CNN 块 $e_I_CNN_j_pt$ 的计算复杂度 $e_I_Fs_{CNN_j_pt}$ 和平均计算复杂度 $e_I_Fs_{CNN_j_avg_pt}$ 、其他 o 个边缘服务器 e_J 的浮点计算能力 FS_J 、其他 o 个边缘服务器 e_J 正在负责推理的 CNN 块 $e_J_CNN_j_pt$ 的计算复杂度 $e_I_Fs_{CNN_j_pt}$ 和平均计算复杂度 $e_I_Fs_{CNN_j_avg_pt}$;

所述网络遥测模块用于采集边缘服务器 e_I 和边缘服务器 e_J 间的网络带宽 bw_{IJ} ;

所述策略生成中心包括分块推理时延预测模块、传输时延计算模块、全局时延预测模块、离线样本数据存储模块和决策信息生成模块;

所述分块推理时延预测模块用于以已知的边缘服务器 e_I 自身浮点计算能力 FS_I 、正在处理的 CNN 块的计算复杂度、待处理的 CNN 块的平均计算复杂度、待处理的 CNN 块包含的层数为输入, 以对应的边缘服务器 e_I 计算待处理的 CNN 块产生的实际分块推理时延为输出, 构建并训练分块推理时延预测模型;

所述传输时延计算模块用于基于物理终端将图像数据处理为分辨率相同、数据量大小相等的张量数据, 并将张量数据发送给与其在同一局域网内的边缘服务器; 边缘服务器采集其与局域网内其他边缘服务器之间网络带宽, 计算理论数据传输时延;

所述全局时延预测模块用于以理论数据传输时延、边缘服务器计算待处理 CNN 块产生的理论分块推理时延、边缘服务器计算 CNN 块对正在处理的 CNN 块产生的理论时延损耗为输入、以边缘服务器理论全局时延为输出, 构建全局时延预测模型;

所述离线样本存储模块用于采集、存储边缘服务器 e_I 在自身浮点计算能力为 FS_I 、自身正在处理的 CNN 块 $e_I_CNN_j_pt$ 的计算复杂度为 $e_I_Fs_{CNN_j_pt}$ 时, 计算层数为 m 层且平均计算复杂度为 $Fs_{CNN_j_avg}$ 的 CNN 块 CNN_j 产生的实际分块推理时延

$e_I_tmd_j$; 采集、存储边缘服务器 e_I 在自身浮点计算能力为 FS_I 、自身正在处理的操作层层数为 m_pt 且平均计算复杂度为 $e_I_Fs_{CNN_j_avg_pt}$ 的 CNN 块 $e_I_CNN_j_pt$ 时, 处理计算复杂度为 FS_{CNN_j} 的 CNN 块 CNN_j 对正在计算的 $e_I_CNN_j_pt$ 产生的实际时延损耗 t_{loss} ;

所述决策信息生成模块用于判断当前边缘服务器的理论全局时延是否小于等于其他边缘服务器的理论全局时延, 是则由当前边缘服务器计算待处理的 CNN 块, 然后判断 CNN 块是否全部计算完成, 是则边缘服务器将 CNN 块的计算结果发送给发出任务请求的物理终端; 否则将待处理的 CNN 块传输至其他边缘服务器计算, 直到 CNN 块全部计算完成。

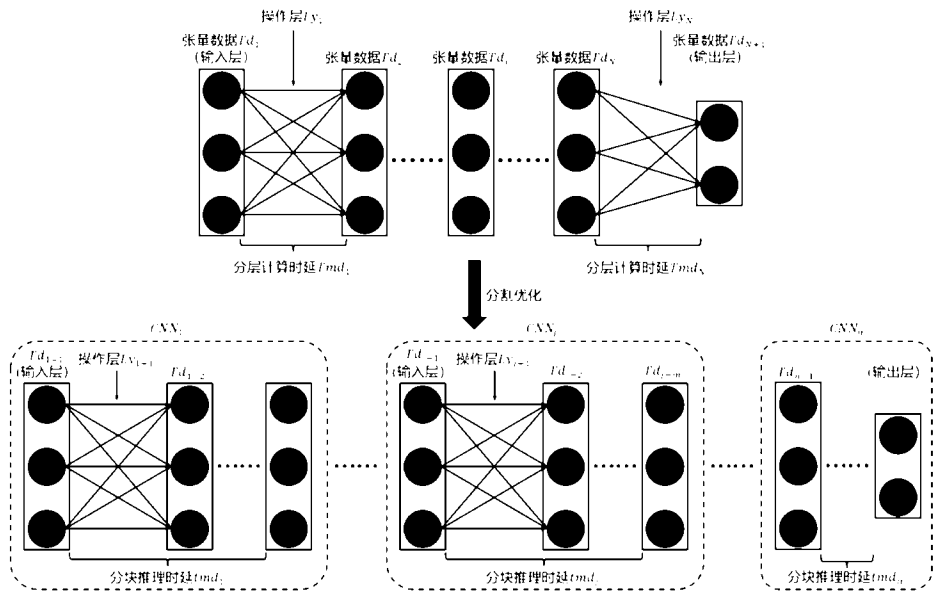


图 1

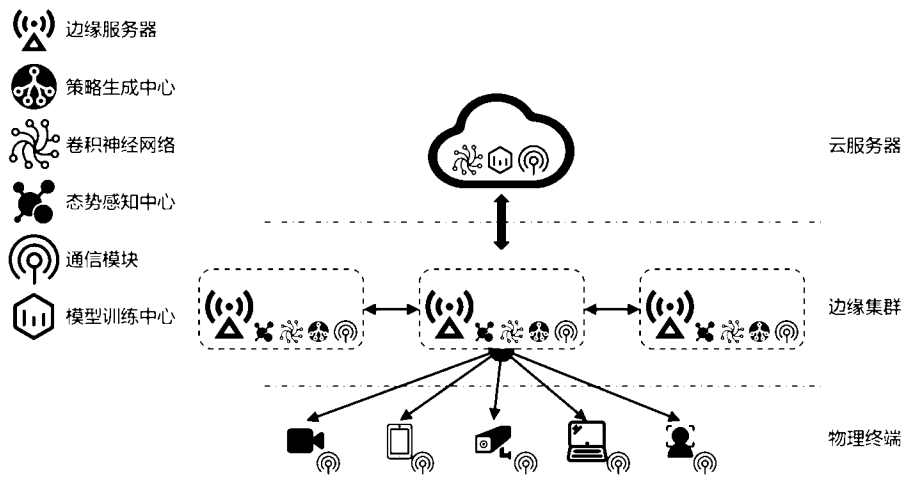


图 2

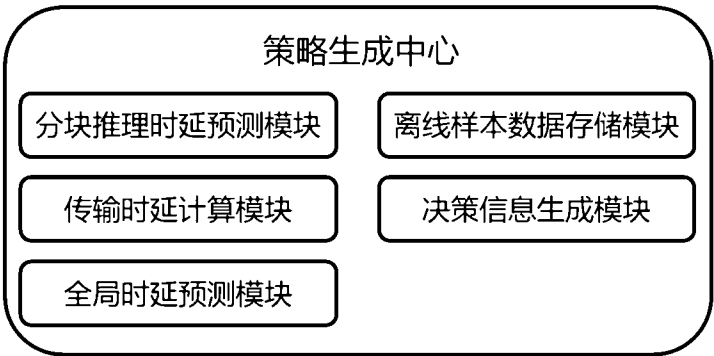


图 3

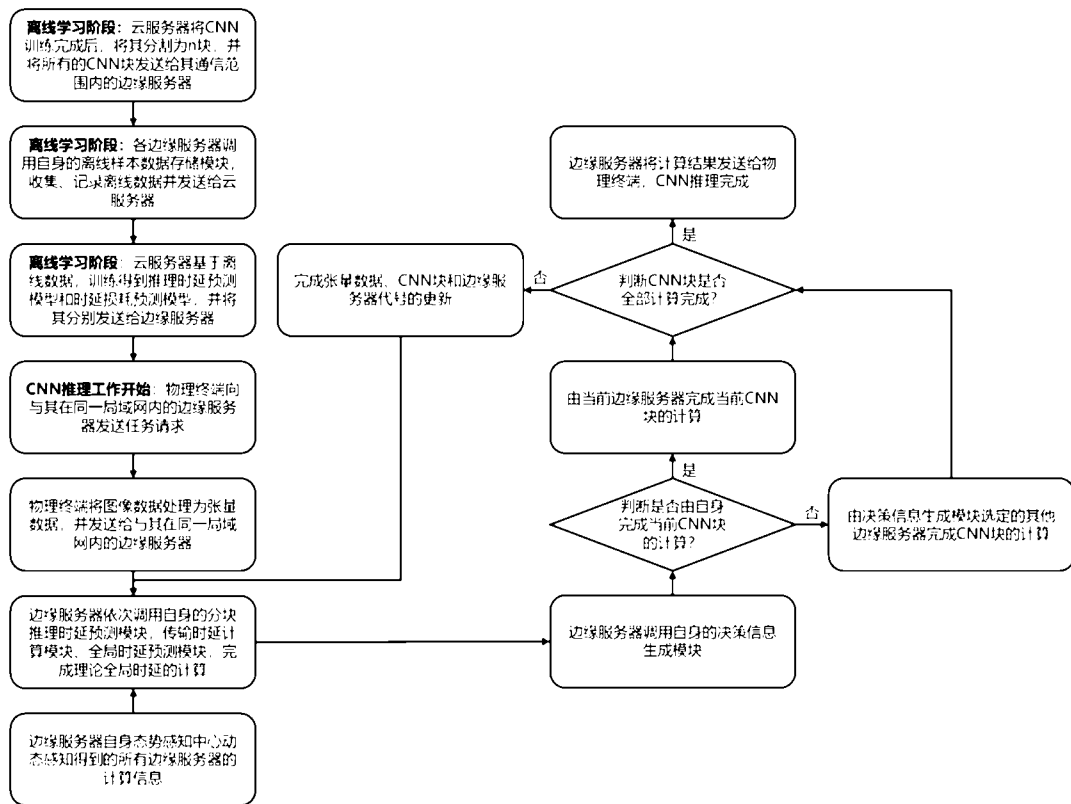


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/134991

A. CLASSIFICATION OF SUBJECT MATTER

G06N3/08(2023.01)i; H04L 41/16(2022.01)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)

IPC: H04L,G06N

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)

CNTXT, ENTXTC, VEN, CNKI, 3GPP: 卷积神经网络, 深度神经网络, 协同, 协作, 联合, 推理, 时延, 预测, 损耗, 模型, 全局, 计算复杂度, 平均, 边缘服务器, 边缘计算, CNN, DNN, collaborat+, cooperat+, federat+, reason, latency, predict+, loss, model, global, computational complexity, averag+, edge server, edge computation,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 116455768 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 18 July 2023 (2023-07-18) entire document	1-10
A	CN 116016538 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 25 April 2023 (2023-04-25) entire document	1-10
A	CN 113067873 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 02 July 2021 (2021-07-02) entire document	1-10
A	WO 2022025303 A1 (LG ELECTRONICS INC.) 03 February 2022 (2022-02-03) 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

“D” document cited by the applicant in the international application

“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

06 March 2024

Date of mailing of the international search report

06 March 2024

Name and mailing address of the ISA/CN

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

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/134991

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	郭永安等 (GUO, Yongan et al.). "基于边端协同的CNN推理加速框架 (A CNN Inference Acceleration Framework based on Edge-End Collaboration)" 南京邮电大学学报 (自然科学版) (<i>Journal of Nanjing University of Posts And Telecommunications (Natural Science)</i>), 25 March 2023 (2023-03-25), entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/134991

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	116455768	A	18 July 2023	None	
CN	116016538	A	25 April 2023	None	
CN	113067873	A	02 July 2021	None	
WO	2022025303	A1	03 February 2022	None	

A. 主题的分类

G06N3/08(2023.01)i; H04L 41/16(2022.01)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)

IPC: H04L,G06N

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))

CNTEXT,ENTXTC,VEN,CNKI,3GPP:卷积神经网络,深度神经网络,协同,协作,联合,推理,时延,预测,损耗,模型,全局,计算复杂度,平均,边缘服务器,边缘计算,CNN,DNN, collaborat+,cooperat+,federat+,reason,latency,predict+,loss,model,global, computational complexity,averag+,edge server,edge computation,

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
PX	CN 116455768 A (南京邮电大学) 2023年7月18日 (2023 - 07 - 18) 全文	1-10
A	CN 116016538 A (南京邮电大学) 2023年4月25日 (2023 - 04 - 25) 全文	1-10
A	CN 113067873 A (北京邮电大学) 2021年7月2日 (2021 - 07 - 02) 全文	1-10
A	WO 2022025303 A1 (LG ELECTRONICS INC) 2022年2月3日 (2022 - 02 - 03) 全文	1-10
A	郭永安等. "基于边端协同的CNN推理加速框架" 南京邮电大学学报(自然科学版), 2023年3月25日 (2023 - 03 - 25), 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

"D" 申请人在国际申请中引证的文件

"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

2024年3月6日

国际检索报告邮寄日期

2024年3月6日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

丁炜

电话号码 (+86) 010-62412239

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/134991

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	116455768	A	2023年7月18日	无	
CN	116016538	A	2023年4月25日	无	
CN	113067873	A	2021年7月2日	无	
WO	2022025303	A1	2022年2月3日	无	