(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110074776 B
(45)授权公告日 2020.04.10

(21)申请号 201910362918.3

(22)申请日 2019.04.30

(65)同一申请的已公布的文献号
申请公布号 CN 110074776 A

(43)申请公布日 2019.08.02

(73)专利权人 广东乐之康医疗技术有限公司
地址 510663 广东省广州市广州中新广州知识城九佛建设路333号自编193室

(72)发明人 马振宇

(51)Int.Cl.
A61B 5/0402(2006.01)
A61B 5/053(2006.01)
A61B 5/1455(2006.01)
A61B 7/00(2006.01)
A61B 5/00(2006.01)

(56)对比文件
CN 103263263 A,2013.08.28,全文.
Eric Flamand et al..GAP-8: A RISC-V SoC for AI at the Edge of the IoT.《2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)》.2018,第1-4页.
同亚娜.FPGA功耗模型建立与热点分析.《中国优秀硕士论文学位论文全文数据库 信息科技辑》.2017,第1-64页.
墙威.神经网络在SoC高层次功耗估计中的应用.《2005中国通信集成电路技术与应用研讨会论文集》.2005,第178-184页.
Lijia Wang et al..High-Level Power Estimation Model for SOC with FPGA Prototyping.《2012 Fourth International Conference on Computational Intelligence and Communication Networks》.2012,第491-495页.

审查员 李馥然

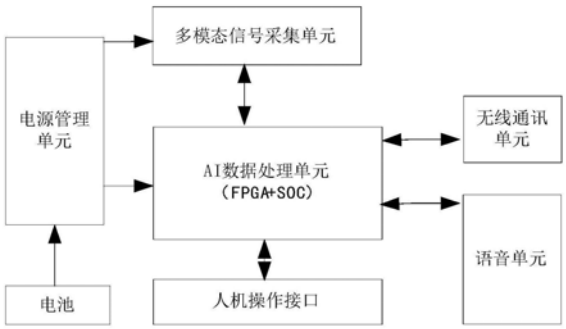
权利要求书3页 说明书12页 附图8页

(54)发明名称

一种人工智能动态心肺监测设备的功耗控制方法及系统

(57)摘要

本发明涉及一种人工智能动态心肺监测设备的功耗控制方法及系统,通过该方法和系统,以计算资源需求和完成时间为控制限制,以功耗作为控制指标,对各层计算资源进行预测计算,作为系统调度的决策依据;可以通过闭环控制对控制决策计算模块进行持续优化,实现动态心肺监测设备系统的能效最大,从而达到用一套最少硬件电路资源的人工智能边缘计算设备,满足各种监测场景而且设备待机时间最长。



CN 110074776 B

1.一种人工智能动态心肺监测设备的功耗控制方法,所述方法包括:

接收多模态数据输入,所述多模态数据输入至少包括心电信号;

根据所述多模态数据输入建立AI推理模型;

根据所述AI推理模型得到各个应用场景AI推理模型的参数需求,所述参数需求至少包括应用场景的计算资源需求和完成时间要求T0;

根据所述AI推理模型的参数进行多模态控制决策计算,得到模态控制参数配置,基于所述模态控制参数配置进行功耗和时间计算,得到模拟功耗Pf和模拟时间Tf;

根据所述模态控制参数配置进行多模态计算单元计算,并进行多模态计算单元的多模态数据融合;所述多模态计算单元计算至少包括心电计算单元计算和心肺音计算单元计算;

对所述多模态计算单元计算进行实际功耗和实际时间监控,得到实际功耗Pr与实际时间Tr;

在所述完成时间要求T0内,以功率最低为控制指标,将所述模拟时间Tf和实际时间Tr反馈到所述多模态控制决策计算进行持续优化,得到最佳完成时间和最佳硬件资源。

2.根据权利要求1所述的功耗控制方法,其特征在于,多模态数据融合后得到本地计算结果输出后,根据不同的需求进行人工智能推理模型参数需求的调整,同时根据不同的应用场景进行闭环参数的更新。

3.根据权利要求1所述的功耗控制方法,其特征在于,得到所述AI推理模型的参数需求包括:

根据所述AI推理模型进行边缘计算和深度学习模型轻量化评估。

4.根据权利要求1所述的功耗控制方法,其特征在于,所述多模态控制决策计算包括:

获取应用场景下的所述计算资源需求和完成时间要求;

读取所述计算资源需求中的系统硬件,对所述系统硬件的资源进行递减式地分级;

按照所述分级的顺序计算在该分级下完成模拟计算的时间,若完成模拟计算的时间低于所述完成时间要求T0,则计算下一级的所述系统硬件的资源下的模拟计算的时间,直至某一级模拟计算的模拟时间高于所述完成时间要求T0;

若最后一级分级下完成模拟计算的时间仍然低于所述完成时间要求,则所述多模态控制决策计算采用所述最后一级的所述系统硬件的资源进行输出。

5.根据权利要求4所述的功耗控制方法,其特征在于,根据所述多模态控制决策计算中计算资源需求中进行模拟计算的硬件资源的模拟功耗与对应的实际功耗的对比以及模拟时间与实际时间的对比,对进行模拟计算的硬件资源的参数进行优化,得到所述最佳硬件资源的参数,并根据该最佳硬件资源的参数得到所述最佳完成时间。

6.根据权利要求5所述的功耗控制方法,其特征在于,所述系统硬件包括AI数据处理单元,所述AI数据处理单元包括SOC芯片和FPGA芯片;

所述SOC芯片包括一个MCU控制单元和一个计算单元,所述计算单元包括具有多个核心的并行RISC-V计算块以及一个硬件卷积计算引擎;

所述FPGA芯片包括多个并行的可编程引擎;

所述分级至少包括递减的全部硬件资源、第二级硬件资源和最少硬件资源;

所述全部硬件资源包括所述AI数据处理单元的全部,在所述全部硬件资源的模拟计算

下,若完成全部数据模拟计算的时间高于或等于所述完成时间要求T0,则判断模型或者硬件参数错误,系统报错;若完成全部数据模拟计算的时间低于所述完成时间要求T0,则转入到第二级硬件资源进行模拟计算;

所述第二级硬件资源包括所述SOC芯片;在所述第二级硬件资源的模拟计算下:

若完成全部数据模拟计算的时间高于或等于所述完成时间要求,则把需要大量硬件资源,同时实时性要求不高的部分数据传送给所述FPGA芯片进行计算,逐步降低所述FPGA芯片的频率,计算出FPGA芯片以最低功耗完成所述部分数据模拟计算的时间T61,使得该模拟计算的时间小于等于所述完成时间要求T0,计算此时所述SOC芯片和所述FPGA芯片的功率之和作为模拟功耗;若完成全部数据模拟计算的时间低于所述完成时间要求,则转入到最少硬件资源进行模拟计算;

所述最少硬件资源包括SOC芯片的硬件卷积计算引擎;在所述最少硬件资源以最大主频进行模拟计算下:

若完成全部数据模拟计算的时间高于或等于所述完成时间要求,则加入所述SOC芯片的具有多个核心的并行RISC-V计算块与所述硬件卷积计算引擎一起进行模拟计算,且根据所述SOC芯片的计算单元的频率-电压曲线找出完成全部数据模拟计算的最低主频并得出所述最低主频相应的电压,使得该模拟计算的时间T51小于等于所述完成时间要求T0,计算所述SOC芯片的计算单元和硬件卷积计算引擎的功耗之和作为模拟功耗;若完成全部数据模拟计算的时间低于所述完成时间要求,则在纯所述硬件卷积计算引擎的计算下,根据所述SOC芯片的硬件卷积计算引擎的频率-电压曲线找出完成全部数据模拟计算的最低主频并得出所述最低主频相应的电压,使得该模拟计算的时间T41小于等于所述完成时间要求,计算此时所述SOC芯片的硬件卷积计算引擎的功耗作为模拟功耗。

7.根据权利要求6所述的功耗控制方法,其特征在于,设置初始时间余量Tm,在所述模拟计算的时间Tx1的基础上加上所述Tm得到模拟时间Tf,即在T61,T51和T41的基础上加上所述Tm作为模拟时间T6,T5和T4,其中,x=4或5或6。

8.根据权利要求7所述的功耗控制方法,其特征在于,所述多模态控制决策计算的持续优化步骤如下:

若所述实际时间Tr小于等于所述完成时间要求T0,则判断所述实际时间Tr是否小于所述模拟时间Tf:

若所述实际时间Tr小于所述模拟时间Tf,则判断所述实际时间是否小于所述模拟计算的时间Tx1:

若小于所述模拟计算的时间Tx1,则将所述实际时间Tr赋给所述模拟计算的时间Tx1,并重新计算所述模拟时间Tf,且当Tr<Tx1min时,所述模拟计算的时间Tx1=Tx1min,其中Tx1min是最少硬件资源所得的模拟计算的时间;

若大于等于所述模拟计算的时间Tx1,则将所述实际时间赋给所述模拟时间,然后用所述模拟时间减去所述初始时间余量Tm,得到中间时间Tz,以中间时间Tz为控制对象,加大对应分级硬件资源的频率,使得模拟计算的时间趋近于所述中间时间Tz;然后以加大后的对应分级硬件资源的参数进行所述多模态计算单元计算,得到更新后的实际时间Tr1;对所述初始时间余量Tm进行调整,得到更新后的时间余量Tm1;将所述更新后的实际时间Tr1赋给所述模拟时间Tf,将所述更新后的时间余量Tm1替换所述初始时间余量Tm,重复该阶段的计

算，直至更新后的实际时间落入【(1-k)*T0,T0】，其中k为误差系数，k的范围为1%-5%。

9.根据权利要求7所述的功耗控制方法，其特征在于，所述多模态控制决策计算的持续优化步骤如下：

若所述实际时间小于等于所述完成时间要求T0，则将所述实际时间作为所述最佳完成时间，将所述实际功耗对应的硬件及参数作为最佳硬件资源；

若实际时间大于所述完成时间要求T0，且所述实际时间大于所述模拟时间，则加大所述初始时间余量Tm直至所述实际时间小于所述模拟时间；

若实际时间大于所述完成时间要求T0，且所述实际时间小于所述模拟时间，则将所述实际时间赋给所述模拟时间，然后用所述模拟时间减去所述初始时间余量Tm，得到中间时间Tz，以中间时间Tz为控制对象，加大对应分级硬件资源的频率，使得模拟计算的时间趋近于所述中间时间Tz；然后以加大后的对应分级硬件资源的参数进行所述多模态计算单元计算，得到更新后的实际时间Tr1；

对所述初始时间余量Tm进行调整，得到更新后的时间余量Tm1；

将所述更新后的实际时间Tr1赋给所述模拟时间，将所述更新后的时间余量Tm1替换所述初始时间余量Tm，重复该阶段的计算，直至更新后的实际时间落入【(1-k)*T0,T0】，其中k为误差系数，k的范围为1%-5%。

10.一种人工智能动态心肺监测系统，所述系统包括多模态信号采集单元、AI数据处理单元、电源管理单元、无线通讯单元、语音单元和人机操作接口，所述AI数据处理单元和电源管理单元实施上述的权利要求1-9中任一项的功耗控制方法。

一种人工智能动态心肺监测设备的功耗控制方法及系统

技术领域

[0001] 本发明涉及心电心音测量领域，具体而言，涉及一种人工智能动态心肺监测设备的功耗控制方法。

背景技术

[0002] 一、穿戴移动式人工智能终端设备目前还处于发展的早期，大多数终端设备目前是采集数据然后通过各类通讯方式把数据发送到服务器后台进行人工智能计算。但是因为通讯稳定带宽、数据安全、本地个性化实时处理的等需求无法通过云计算解决，因此就需要一部分实时性和个性化的功能嵌入到设备终端里进行人工智能推理，因此边缘计算在移动式物联网领域成为一个基础的技术应用。

[0003] 目前能够用于边缘计算的芯片和硬件电路非常少，而且主要关注点是如何提高算力，对功耗控制的最主要的方案依赖人工智能芯片本身的设计，比如选型低功耗MCU、嵌入式低功耗FPGA等。另外一种方式是借用传统电量控制方式，比如进入睡眠模式、关闭不用的外围硬件方式等。

[0004] 总之，目前还没有真正的功耗优化控制方案。

[0005] 二、目前，医院普遍使用的心电、心音设备为12导联动态检测仪(holter)，这种hotler功能都是单一的，在同一时间段，要么采集用户的心电信号，要么采集用户的心音信号，不能综合监测用户的多个信号从而集中判断并发情况。Hotler的功耗比较大，体积也很大，穿戴时一般在医生的指导下进行操作，用户一个满电量情况下使用时间短。对于需要连续采集心电、心音信号的用户来说(心脏的部分病症从检测到异常到非常糟糕的情况只需要10分钟)，应该也必须连续采集心电心音的数据。因此，目前的设备无法满足3-7天的医用连续检测需求。并且，目前的Holter的心电监测判断需要把数据传输到后台服务器计算，这种计算滞后非常明显，不能实现心肺的实时计算和全天候即时预警的需求。

发明内容

[0006] 有鉴于此，本发明提供一种人工智能动态心肺监测仪的功耗控制方法，该方法采用多模态控制思路，根据人工智能边缘计算的需要计算出最优的人工智能硬件资源调配策略并在设备上实施，从而达到控制整机功耗的目的。

[0007] 本发明具体的技术方案如下：

[0008] 一种人工智能动态心电检测仪的功耗控制方法，

[0009] 接收多模态数据输入，所述多模态数据输入至少包括心电信号；

[0010] 根据所述多模态数据输入建立AI推理模型；

[0011] 根据所述AI推理模型得到各个应用场景AI推理模型的参数需求，所述参数需求至少包括应用场景的计算资源需求和完成时间要求TO；

[0012] 根据所述AI推理模型的参数进行多模态控制决策计算，得到模态控制参数配置，基于所述模态控制参数配置进行功耗和时间计算，得到模拟功耗Pf和模拟时间Tf；

[0013] 根据所述模态控制参数配置进行多模态计算单元计算，并进行多模态计算单元的多模态数据融合；所述多模态计算单元计算至少包括心电计算单元计算和心肺音计算单元计算；

[0014] 对所述多模态计算单元计算进行实际功耗和实际时间监控，得到实际功耗Ps与实际时间Tr；

[0015] 在所述完成时间要求TO内，以功率最低为控制指标，将所述模拟时间Tf和实际时间Tr反馈到所述多模态控制决策计算进行持续优化，得到最佳完成时间和最佳硬件资源。

[0016] 进一步地，多模态数据融合后得到本地计算结果输出后，根据不同的需求进行人工智能推理模型参数需求的调整，同时根据不同的应用场景进行闭环参数的更新。

[0017] 进一步地，所述多模态数据输入包括心电信号、心音信号、呼吸音信号、胸阻抗信号、血氧信号中的至少两种。

[0018] 进一步地，得到所述AI推理模型的参数需求包括：

[0019] 根据所述AI推理模型进行边缘计算和深度学习模型轻量化评估。

[0020] 进一步地，所述边缘计算采用轻量化的GRU或者LSTM模型。

[0021] 进一步地，所述多模态控制决策计算包括：

[0022] 获取应用场景下的所述计算资源需求和完成时间要求；

[0023] 读取所述计算资源需求中的系统硬件，对所述系统硬件的资源进行递减式地分级；

[0024] 按照所述分级的顺序计算在该分级下完成模拟计算的时间，若完成模拟计算的时间低于所述完成时间要求TO，则计算下一级的所述系统硬件的资源下的模拟计算的时间，直至某一级模拟计算的模拟时间高于所述完成时间要求TO；

[0025] 若最后一级分级下完成模拟计算的时间仍然低于所述完成时间需求，则所述多模态控制决策计算采用所述最后一级的所述系统硬件的资源进行输出。

[0026] 进一步地，根据所述多模态控制决策计算中计算资源需求中进行模拟计算的硬件资源的模拟功耗与对应的实际功耗的对比以及模拟时间与实际时间的对比，对进行模拟计算的硬件资源的参数进行优化，得到所述最佳硬件资源的参数，并根据该最佳硬件资源的参数得到所述最佳完成时间。

[0027] 进一步地，所述系统硬件包括AI数据处理单元，所述AI数据处理单元包括SOC芯片和FPGA芯片；

[0028] 所述SOC芯片包括一个MCU控制单元和一个计算单元，所述计算单元包括具有多个核心的并行RISC-V计算块以及一个硬件卷积计算引擎(简称HWCE)；

[0029] 所述FPGA芯片包括多个并行的可编程引擎；

[0030] 所述分级至少包括递减的全部硬件资源、第二级硬件资源和最少硬件资源；

[0031] 所述全部硬件资源包括所述AI数据处理单元的全部，在所述全部硬件资源的模拟计算下，若完成全部数据模拟计算的时间高于或等于所述完成时间要求TO，则判断模型或者硬件参数错误，系统报错；若完成全部数据模拟计算的时间低于所述完成时间要求TO，则转入到第二级硬件资源进行模拟计算；

[0032] 所述第二级硬件资源包括所述SOC芯片；在所述第二级硬件资源的模拟计算下：

[0033] 若完成全部数据模拟计算的时间高于或等于所述完成时间要求，则把需要大量硬

件资源,同时实时性要求不高的部分数据(如心音数据)传送给所述FPGA芯片进行计算,逐步降低所述FPGA芯片的频率,计算出FPGA芯片以最低功耗完成所述部分数据模拟计算的时间T61,使得该模拟计算的时间小于等于所述完成时间要求T0(实施例中需要说明SOC部分的时间一定是小于FPGA的时间的),计算此时所述SOC芯片和所述FPGA芯片的功率之和作为模拟功耗;若完成全部数据模拟计算的时间低于所述完成时间要求,则转入到最少硬件资源进行模拟计算;

[0034] 所述最少硬件资源包括SOC芯片的硬件卷积计算引擎;在所述最少硬件资源以最大主频进行模拟计算下:

[0035] 若完成全部数据模拟计算的时间高于或等于所述完成时间要求,则加入所述SOC芯片的具有多个核心的并行RISC-V计算块与所述硬件卷积计算引擎一起进行模拟计算,且根据所述SOC芯片的计算单元的频率-电压曲线(固有的)找出完成全部数据模拟计算的最低主频并得出所述最低主频相应的电压,使得该模拟计算的时间T51小于等于所述完成时间要求T0,计算所述SOC芯片的计算单元和硬件卷积计算引擎的功耗之和作为模拟功耗;若完成全部数据模拟计算的时间低于所述完成时间要求,则在纯所述硬件卷积计算引擎的计算下,根据且根据所述SOC芯片的硬件卷积计算引擎的频率-电压曲线(固有的)找出完成全部数据模拟计算的最低主频并得出所述最低主频相应的电压,使得该模拟计算的时间T41小于等于所述完成时间要求,计算此时所述SOC芯片的硬件卷积计算引擎的功耗作为模拟功耗。

[0036] 进一步地,设置初始时间余量Tm,在所述模拟计算的时间Tx1的基础上加上所述Tm得到模拟时间Tf,即在T61,T51和T41的基础上加上所述Tm作为模拟时间T6,T5和T4,其中,x=4或5或6。

[0037] 进一步地,所述多模态控制决策计算的持续优化步骤如下:

[0038] 若所述实际时间Tr小于等于所述完成时间要求T0,则判断所述实际时间Tr是否小于所述模拟时间Tf:

[0039] 若所述实际时间Tr小于所述模拟时间Tf,则判断所述实际时间是否小于所述模拟计算的时间Tx1:

[0040] 若小于所述模拟计算的时间Tx1,则将所述实际时间Tr赋给所述模拟计算的时间Tx1,并重新计算所述模拟时间Tf,且当Tr<Tx1min时,所述模拟计算的时间Tx1=Tx1min,其中Tx1min是最少硬件资源所得的模拟计算的时间;

[0041] 若大于等于所述模拟计算的时间Tx1,则将所述实际时间赋给所述模拟时间,然后用所述模拟时间减去所述初始时间余量Tm,得到中间时间Tz,以中间时间Tz为控制对象,加大对应分级硬件资源的频率,使得模拟计算的时间趋近于所述中间时间Tz;然后以加大后的对应分级硬件资源的参数进行所述多模态计算单元计算,得到更新后的实际时间Tr1;对所述初始时间余量Tm进行调整,得到更新后的时间余量Tm1;将所述更新后的实际时间Tr1赋给所述模拟时间Tf,将所述更新后的时间余量Tm1替换所述初始时间余量Tm,重复该阶段的计算,直至更新后的实际时间落入【(1-k)*T0,T0】,其中k为误差系数,k的范围为1%-5%。

[0042] 进一步地,所述多模态控制决策计算的持续优化步骤如下:

[0043] 若所述实际时间Tr小于等于所述完成时间要求T0,则将所述实际时间作为所述最佳完成时间,将所述实际功耗对应的硬件及参数作为最佳硬件资源;

[0044] 若实际时间大于所述完成时间要求T0，且所述实际时间大于所述模拟时间，则加大所述初始时间余量Tm直至所述实际时间小于所述模拟时间；

[0045] 若实际时间大于所述完成时间要求T0，且所述实际时间小于所述模拟时间，则将所述实际时间赋给所述模拟时间，然后用所述模拟时间减去所述初始时间余量Tm，得到中间时间Tz，以中间时间Tz为控制对象，加大对应分级硬件资源的频率，使得模拟计算的时间趋近于所述中间时间Tz(趋近于可以理解为±5%的Tz时间段内)；然后以加大后的对应分级硬件资源的参数进行所述多模态计算单元计算，得到更新后的实际时间Tr1；

[0046] 对所述初始时间余量Tm进行调整，得到更新后的时间余量Tm1；

[0047] 将所述更新后的实际时间Tr1赋给所述模拟时间，将所述更新后的时间余量Tm1替换所述初始时间余量Tm，重复该阶段的计算，直至更新后的实际时间落入【(1-k)*T0，T0】，其中k为误差系数，k的范围为1%-5%。

[0048] 一种人工智能动态心电检测系统，所述系统包括多模态信号采集单元、AI数据处理单元、电源管理单元、无线通讯单元、语音单元和人机操作接口，所述AI数据处理单元和电源管理单元实施上述的方法。

[0049] 一种人工智能动态心电监测系统，所述系统包括应用层、系统管理层和数据处理层；

[0050] 所述应用层包括人机交互单元；

[0051] 所述系统管理层包括电源管理单元、主控制单元；

[0052] 所述数据处理层包括多模态数据采集单元、数据预处理、SOC低功耗实时AI计算、FPGA大数据非实时计算、计算结果融合单元和无线通讯单元；

[0053] 所述系统管理层和数据处理层实施上述的方法。

[0054] 通过上述的技术方案，以计算资源需求和完成时间为控制限制，以功耗作为控制指标，对各层计算资源进行预测计算，作为系统调度的决策依据；可以通过闭环控制对控制决策计算模块进行持续优化，实现动态心肺监测仪系统的能效最大，从而达到用一套最少硬件电路资源的人工智能边缘计算设备，满足各种监测场景而且设备待机时间最长。

附图说明

[0055] 图1为本发明的人工智能动态心电检测系统图。

[0056] 图2为本发明的多模态信号采集单元的穿戴图，包括右下角集成有AI数据处理单元的控制盒。

[0057] 图3为本发明的部分模态信号采集的示意图。

[0058] 图4为本发明的软件系统框架图。

[0059] 图5为本发明的不同应用场景的AI推理模型图。

[0060] 图6为本发明对AI推理模型进行边缘计算的流程图。

[0061] 图7为对图5和图6的处理结果进行深度学习模型轻量化评估得到的边缘计算模型参数。

[0062] 图8为SOC芯片的芯片架构示意图。

[0063] 图9为FPGA芯片的芯片架构示意图。

[0064] 图10为本发明的穿戴式设备的计算资源与功耗控制层级图。

[0065] 图11为SOC芯片的不用应用场景需求采用不同计算资源功耗评估示例。

[0066] 图12为SOC芯片不同计算资源对应的运算速度和功耗。

[0067] 图13为同一计算资源采用不同的主频及电压对功耗的影响。

[0068] 图14为本发明功耗控制方法的流程图。

[0069] 图15为本发明的多模态控制决策计算流程图。

[0070] 图16为本发明基于时间的功耗控制反馈图。

具体实施方式

[0071] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0072] 在本发明实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。在本发明实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义，“多种”一般包含至少两种，但是不排除包含至少一种的情况。

[0073] 应当理解，本文中使用的术语“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0074] 应当理解，尽管在本发明实施例中可能采用术语第一、第二、第三等来描述……，但这些……不应限于这些术语。这些术语仅用来将……区分开。例如，在不脱离本发明实施例范围的情况下，第一……也可以被称为第二……，类似地，第二……也可以被称为第一……。

[0075] 取决于语境，如在此所使用的词语“如果”、“若”可以被解释成为“在……时”或“当……时”或“响应于确定”或“响应于检测”。类似地，取决于语境，短语“如果确定”或“如果检测(陈述的条件或事件)”可以被解释成为“当确定时”或“响应于确定”或“当检测(陈述的条件或事件)时”或“响应于检测(陈述的条件或事件)”。

[0076] 还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的商品或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种商品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的商品或者系统中还存在另外的相同要素。

[0077] 另外，下述各方法实施例中的步骤时序仅为一种举例，而非严格限定。

[0078] 如附图1所示，本发明的硬件系统组成主要包括多模态信号采集单元、AI数据处理单元、电源管理单元、电池、人机操作接口、无线通信单元、语音单元。

[0079] a) 多模态信号采集模块

[0080] 多模态信号采集模块集成了十二导联心电信号采集，胸阻抗信号采集，同时可以采集心、肺音信号和血氧信号。

[0081] 十二导联心电信号和胸阻抗信号采集电路前端是柔性传感器电极，后端采用低功

耗专用于ECG采集的ADC(Analog-to-Digital Converter)芯片,该芯片集成了滤波模块,具有精度高、功耗低、集成度高等特点,同时具备胸阻抗信号采集。

[0082] 心、肺音采用MEMS(Micro-Electro-Mechanical System)传感器技术,经过特殊的拾音腔体和放大滤波电路处理,高保真采集患者肺部呼吸的声音信号,经过ADC模拟转换成数字信号,数据由数据模块进行分析处理,通过人工智能算法推理实现对肺部异常的检测、诊断、预警和干预。

[0083] 血氧信号采集采用光电传感器,直接采集人体皮肤的血氧浓度信息,同时可以采集心率信号,信号经过低通滤波器和模数转换模块,变成数字信号给数据处理模块分析。

[0084] 如附图2所示, 图为发明优选地一个多模态信号采集单元的设备,具体为可穿戴移动式心电监测设备。该设备可以同时采集12导联心电信号、心音信号、呼吸音信号、胸阻抗信号。当然也可以集成光电传感器采集血氧信号。该设备体积小、重量轻,便携、穿戴舒适,低功耗,具有本地人工智能计算功能,实现实时监测、预警。图2中左下角的控制盒中集成有AI数据处理单元。设备为采集数据的硬件,并非本申请保护的核心,申请人会在其他申请中涉及。

[0085] 多模态信号采集单元数据处理如附图3所示。多模态数据采集之后,进行多模态生物识别。多模态生物识别是指整合或融合两种及两种以上生物识别特征,利用多重生物识别技术的独特优势,并结合数据融合技术,使得认证和识别过程更加精准、安全。与传统的单一生物识别方式的主要区别在于,多模态生物识别技术可通过独立的或多种采集方式合而为一的采集器,采集不同的生物特征,并通过分析、判断多种生物识别方式的特征值进行识别和认证。因此,相比于单一的生物识别,多模态生物识别技术具有独特的优势:首先,极高的可靠性和安全性。多模态生物识别一般整合或融合了两种以及两种以上的生物识别方式,相比于单一的生物识别技术更加安全、可靠,多重认证过程保证了认证结果的准确性;其次,个性化定制,满足用户不同需求。用户可根据应用环境选择不同的生物识别技术,可以有效克服单一生物识别技术的固有缺陷,极大降低了生物识别对环境的依赖程度,更加贴心的满足用户不同的业务需求;

[0086] b)数据处理和主控制单元

[0087] 数据处理单元采用当今最新工艺高性能FPGA(Field－Programmable Gate Array)和低功耗SOC(System-on-a-Chip)芯片相结合的架构,扩展DRAM(Dynamic Random Access Memory);FPGA内部具有数万个逻辑单元和乘加器,通过选择合适的架构,设计成并行计算的人工智能数据处理单元及硬件加速器,单芯片处理能效率最高可以达到1TOPS/W以上,计算能力强大,可快速对心肺信号处理和人工智能算法推理,实时对心肺功能的异常预警和干预。另外,采用具备神经网络计算功能的超低功耗SOC芯片作为主控制器形成主控制单元,对信号进行预处理,SOC功耗低,待机功耗可以达到1mW以下,可以实时采集和处理心电数据并缓存以及对系统IO管理,配合FPGA运行心肺分析人工智能算法,从而使得整机功耗更低。

[0088] c)人机交互单元

[0089] 人机操作接口采用小尺寸TFT(Thin Film Transistor)触摸屏,可以实时查看心电数据和设置运行参数。

[0090] 语音模块集成了麦克风与扬声器,可以实时语音提醒患者以及进行双向通话,以

便出现心肺信号异常预警时进行主动干预。

[0091] d)无线通信单元

[0092] 集成了蓝牙模块,同时可以扩展WIFI、NB-IOT、3G/4G通讯功能,可以通过软件系统配置设定选择最佳通信方式。

[0093] e)电池和电源管理单元

[0094] 采用了大容量锂电池,保证终端可以持续工作48小时以上。电源管理模块对电池充放电管理,并精确检测剩余电量,保证仪器按照预定功能安全稳定工作。电源管理单元采用普通的电源芯片即可。

[0095] 整个可穿戴式的设备的软件系统如附图4所示,包括应用层、系统管理层和数据处理成。功能模块主要为主控制单元和运算单元两个,硬件挨踢分别是SOC控制单元、SOC计算单元及FPGA芯片。

[0096] a)主控制单元

[0097] 主控制单元运行在SOC控制芯片上,基于一个实时的轻量级嵌入式操作系统(如FREERTOS)进行多模态任务控制和计算数据调度,实现系统的功耗和并行运算的最优控制。

[0098] b)运算单元

[0099] 运算单元有两部分,一部分运行在低功耗SOC的神经网络模块里,主要做实时性要求比较高的12导联ECG计算;另外一部分运行在FPGA高性能人工智能运算模块里,主要运行心、肺音和其他实时性要求不高但数据量比较大的信号的计算。通过这样的分布式计算调度,能够在保证多模态数据实时性的同时使功耗最低。

[0100] 如附图5所示,接受至少包括心电数据、呼吸音和心音数据的信号后,建立AI推理模型,如对于心率衰竭的应用场景,需要心音信号、胸阻抗信号、12导联中的3-6个导联心电信号;如对于心房颤动的应用场景,需要3导联的心电数据即可;如对于猝死的应用场景,需要12导联心电信号、心音信号;如对于心梗或者冠心病,需要12导联心电信号等等。对于不同的应用场景,模型可以在多个模态输入的数据中心进行针对性选择,得到该应用场景下的准确模型。

[0101] 整个设备的人工智能边缘计算模型流程如附图5所示,运行在设备上的边缘计算模型如附图6所示:

[0102] 这个系统可以运行任何符合需求的人工智能模型,因为本设备主要是针对心肺信号,因此本实施例采用的是轻量化的GRU或者LSTM模型(两个模型的性能类似,都是RNN的变种)。

[0103] 图7 表示对本发明的心音监测部分的计算模型在轻量化后对系统的计算资源的需求,这个是后续进行计算资源优化的输入基准。因为在本设备的多模态数据输入中,心音是计算量最大的部分,但是对实时性要求不高,需用计算能力强的芯片进行计算。而多导联心电信号的特性是实时性要求高,需要连续计算,但是波形不复杂,采样频率低,比较适合低功耗、计算能力一般的多路计算资源进行计算。

[0104] 本发明基于Greenwave的人工智能SOC芯片以及Xilinx的FPGA芯片构成。图8是Greenwave GAP芯片架构图示意图。图9是Xilinx下属的深鉴科技的人工智能FPGA(基于Ku115芯片)架构示意图。

[0105] 如图8的架构所示,SOC芯片具有一个MCU控制单元,厂家称作Fabric controller

(简称FC)。还有一个计算单元,厂家称作cluster,包括8个核的并行RISC-V计算单元,可以用于人工智能计算加速;以及一个硬件卷积计算引擎,厂家称作HardWare Convolution Engine(简称 HWCE)。这个芯片的特点是具有超低低功耗控制能力,同时具有人工智能计算加速模块能够用于常用的计算,是本实施例的控制重点。关于基于Greenwave的人工智能SOC芯片,可参考"2018 IEEE 29th International Conference on Application-specific Systems,Architectures and Processors(ASAP),10-12 JULY 2018,DOI:10.1109/ASAP.2018.8445101"。

[0106] 如图9的架构所示,FPGA芯片针对人工智能计算,尤其是LSTM算法做了针对性的架构优化,通过专门的sparse LSTM(稀疏LSTM)模块提高计算带宽,有效提升了运行效率。同时,和SOC的架构类似,FPGA设计了多个并行的HW/SW可编程引擎(简称PE),每个PE计算单元由几千上万个MAC dsp核心构成(本实施例有4096个工作在500MHz的构成MAC沟通,理论峰值计算能力4Tflops)。FPGA的特点是计算能力强,但是功耗比较大。所以在进行功耗控制的时候能不启动就尽量不启动。本实施例中FPGA根据应用模型的需要,主要用于心音信号的计算。关于FPGA的架构,可以参考http://www.deephi.com/technology。

[0107] 图10表示的是本实施例的穿戴式设备的计算资源与功耗控制层级。因为本设备的输入是多模态的数据输入源,但是对于心肺监测的应用而言,并不需要每次都启动所有的计算资源进行计算。比如心房颤动,只需要12导联的心电波形即可,并不需要心音检测;而心力衰竭需要心音检测,但是6个导联的心电波形即可。因此,我们可以根据应用端的模型的需求,调配计算资源的使用,从而达到优化系统功耗的目的。

[0108] 系统的计算资源就包括芯片级别的SOC和FPGA。再往下一个层级就是两个芯片内部的各个计算模块,他们可以分别单独控制。同时芯片的计算主频和电压也是影响功耗的重要因素,所以也是本方法的控制对象。

[0109] 图11表示的是不同的应用模型需求采用不同计算资源功耗评估示例。这个表格说明了不同的应用其实投入的计算资源可以是不一样的,同时他们的功耗也是不一样的。这里需要注意的是计算速度和功耗本身并不是完全对应关系。我们可以看到CIFAR-10投入所有的计算资源后速度提升了10.9倍,但是功耗只降低了4.9倍。所以有时候在速度要求没有那么高的应用场合,并不需要投入所有计算资源,反而能获得更好的能效。

[0110] 图12表示的是不同计算资源对应的运算速度和功耗。这个表格基于同一个应用模型,投入不同的计算资源,产生的能效比是不一样的。特别明显的是第一行和第二行,投入同样的8核cluster,主频从175Mhz降低到15.4Mhz,是降低了11.4倍,但是功耗是从70mW降低到3.7mW,功耗降低了18.9倍。再次验证了在速度满足的情况下,可以通过降低主频实现功耗最低。

[0111] 图13表示的是同一个计算资源采用不同的主频及电压对功耗的影响。这个表格同样说明了降低运行频率和电压对功耗降低是有积极作用的。

[0112] 结合上述的对图5-13的阐述,本发明提出了人工智能边缘计算设备的多模态功耗控制的方法。基本思路是根据系统运行的需求或者状态采用不同调度策略的控制方式,即根据系统的实际运行状态,不同运行条件下调度最合适的资源,实现系统计算及时性和功耗消耗结合后的能效最高。

[0113] 具体如图14所示,这个方法是可以闭环控制方法,首先获取应用模型的计算资源

需求和完成时间要求，然后通过多模态控制决策计算，得到最优的计算资源调度。进入到系统后实际运行，同时通过监测系统的实际功耗及计算结果评估，反馈到控制决策计算模块后持续优化。

[0114] 步骤一，接收多模态数据输入，所述多模态数据输入可以所述多模态数据输入包括心电信号、心音信号、呼吸音信号、胸阻抗信号、血氧信号中的至少两种。

[0115] 步骤二，根据所述多模态数据输入建立AI推理模型。AI推理模型应用到的应用场景和对象在针对附图5的说明中已经详细阐述。

[0116] 步骤三，根据所述AI推理模型得到各个应用场景AI推理模型的参数需求，所述参数需求至少包括应用场景的计算资源需求和完成时间要求TO。本发明以功耗为控制对象，通过控制时间要求，达到满足时间需求下的最低功耗。

[0117] 步骤四，根据所述AI推理模型的参数进行多模态控制决策计算，得到模态控制参数配置，基于所述模态控制参数配置进行功耗和时间计算，得到模拟功耗和模拟时间。这部分是本发明的重点控制步骤，通过经验值、预设时间余量，在完成时间要求的基础上，不断优化计算。

[0118] 步骤五，根据所述模态控制参数配置进行多模态计算单元计算，并进行多模态计算单元的多模态数据融合；所述多模态计算单元计算至少包括心电计算单元计算和心肺音计算单元计算；

[0119] 步骤六，对所述多模态计算单元计算进行实际功耗和实际时间监控，得到实际功耗与实际时间；

[0120] 步骤七。在所述完成时间要求TO内，以功率最低为控制指标，将所述模拟时间和实际时间反馈到所述多模态控制决策计算进行持续优化，得到最佳完成时间和最佳硬件资源。

[0121] 在图14中的反馈中实际存在时间反馈(图14中的反馈中未画出，实际存在)，功耗是目标，时间是手段。

[0122] 步骤八，多模态数据融合后得到本地计算结果输出后，根据不同的需求进行人工智能推理模型参数需求的调整，同时根据不同的应用场景进行闭环参数的更新。

[0123] 得到AI推理模型的参数需求包括：根据所述AI推理模型进行边缘计算和深度学习模型轻量化评估。所述边缘计算采用轻量化的GRU或者LSTM模型。

[0124] 如附图15所示，所述多模态控制决策计算包括：

[0125] 获取应用场景下的所述计算资源需求和完成时间要求；

[0126] 读取所述计算资源需求中的系统硬件，对所述系统硬件的资源进行递减式地分级；

[0127] 按照所述分级的顺序计算在该分级下完成模拟计算的时间，若完成模拟计算的时间低于所述完成时间要求TO，则计算下一级的所述系统硬件的资源下的模拟计算的时间，直至某一级模拟计算的模拟时间高于所述完成时间要求TO；

[0128] 若最后一级分级下完成模拟计算的时间仍然低于所述完成时间需求，则所述多模态控制决策计算采用所述最后一级的所述系统硬件的资源进行输出。

[0129] 根据所述多模态控制决策计算中计算资源需求中进行模拟计算的硬件资源的模拟功耗与对应的实际功耗的对比以及模拟时间与实际时间的对比，对进行模拟计算的硬件

资源的参数进行优化，得到所述最佳硬件资源的参数，并根据该最佳硬件资源的参数得到所述最佳完成时间。

[0130] 针对上述多模态控制决策计算的一个实施例如下：

[0131] 系统硬件包括AI数据处理单元，所述AI数据处理单元包括图8所示的SOC芯片和图9所示的FPGA芯片。

[0132] 所述分级至少包括递减的全部硬件资源、第二级硬件资源和最少硬件资源；

[0133] 所述全部硬件资源包括所述AI数据处理单元的全部，在所述全部硬件资源的模拟计算下，若完成全部数据模拟计算的时间高于或等于所述完成时间要求T0，则判断模型或者硬件参数错误，系统报错；若完成全部数据模拟计算的时间低于所述完成时间要求T0，则转入到第二级硬件资源进行模拟计算；

[0134] 所述第二级硬件资源包括所述SOC芯片；在所述第二级硬件资源的模拟计算下：

[0135] 若完成全部数据模拟计算的时间高于或等于所述完成时间要求，则把需要大量硬件资源，同时实时性要求不高的部分数据(如心音数据)传送给所述FPGA芯片进行计算，逐步降低所述FPGA芯片的频率，计算出FPGA芯片以最低功耗完成所述部分数据模拟计算的时间T61，使得该模拟计算的时间小于等于所述完成时间要求T0，计算此时所述SOC芯片和所述FPGA芯片的功率之和作为模拟功耗；若完成全部数据模拟计算的时间低于所述完成时间要求，则转入到最少硬件资源进行模拟计算。

[0136] 由于SOC进行的是实时计算，因此，实际上，在FPGA完成计算的时间内，SOC必然已经完成了再SOC芯片中的实时计算。

[0137] 所述最少硬件资源包括SOC芯片的硬件卷积计算引擎；在所述最少硬件资源以最大主频进行模拟计算下：

[0138] 若完成全部数据模拟计算的时间高于或等于所述完成时间要求，则加入所述SOC芯片的具有多个核心的并行RISC-V计算块与所述硬件卷积计算引擎一起进行模拟计算，且根据所述SOC芯片的计算单元的频率-电压曲线(固有的)找出完成全部数据模拟计算的最低主频并得出所述最低主频相应的电压，使得该模拟计算的时间T51小于等于所述完成时间要求T0，计算所述SOC芯片的计算单元和硬件卷积计算引擎的功耗之和作为模拟功耗；若完成全部数据模拟计算的时间低于所述完成时间要求，则在纯所述硬件卷积计算引擎的计算下，根据且根据所述SOC芯片的硬件卷积计算引擎的频率-电压曲线(固有的)找出完成全部数据模拟计算的最低主频并得出所述最低主频相应的电压，使得该模拟计算的时间T41小于等于所述完成时间要求，计算此时所述SOC芯片的硬件卷积计算引擎的功耗作为模拟功耗。

[0139] 类似，在采用多个核心进行计算时，我们默认硬件卷积计算引擎已经完成了计算。

[0140] 在计算的过程中，根据系统的特性，本发明创造性的设置了初始时间余量Tm，在所述T61，T51和T41的基础上加上所述Tm作为模拟时间T6，T5和T4。模拟时间为在多模态功耗控制算法流程中硬件资源计算模拟计算的时间加上上述的初始时间余量。针对该初始时间余量，本发明进一步地：

[0141] 如附图16所示，在一个实施例中，所述多模态控制决策计算的持续优化步骤如下：

[0142] 若所述实际时间Tr小于等于所述完成时间要求T0，则判断所述实际时间Tr是否小于所述模拟时间Tf：

[0143] 若所述实际时间Tr小于所述模拟时间Tf，则判断所述实际时间是否小于所述模拟计算的时间Tx1：

[0144] 若小于所述模拟计算的时间Tx1，则将所述实际时间Tr赋给所述模拟计算的时间Tx1，并重新计算所述模拟时间Tf，且当Tr<Tx1min时，所述模拟计算的时间Tx1=Tx1min，其中Tx1min是最少硬件资源所得的模拟计算的时间；

[0145] 若大于等于所述模拟计算的时间Tx1，则将所述实际时间赋给所述模拟时间，然后用所述模拟时间减去所述初始时间余量Tm，得到中间时间Tz，以中间时间Tz为控制对象，加大对应分级硬件资源的频率，使得模拟计算的时间趋近于所述中间时间Tz(趋近于可以理解为±5%的Tz时间段内)；然后以加大后的对应分级硬件资源的参数进行所述多模态计算单元计算，得到更新后的实际时间Tr1；对所述初始时间余量Tm进行调整，得到更新后的时间余量Tm1(为了尽快求得最优解，一般采用比例调整，如Tm1/Tm=(Tr1- T0)/(Tr- T0))；将所述更新后的实际时间Tr1赋给所述模拟时间Tf，将所述更新后的时间余量Tm1替换所述初始时间余量Tm，重复该阶段的计算，直至更新后的实际时间落入【(1-k)*T0，T0】，其中k为误差系数，k的范围为1%-5%。

[0146] 若所述实际时间Tr大于等于所述模拟时间Tf，则将所述实际时间赋给所述模拟计算的时间Tx1，并重新计算所述模拟时间Tf；

[0147] 若所述实际时间Tr大于所述完成时间要求T0，则将所述实际时间赋给所述模拟计算的时间Tx1，并重新计算所述模拟时间Tf。

[0148] 在另一个实施例中，所述多模态控制决策计算的持续优化步骤如下：

[0149] 若所述实际时间Tr小于等于所述完成时间要求T0，则将所述实际时间作为所述最佳完成时间，将所述实际功耗对应的硬件及参数作为最佳硬件资源；

[0150] 若实际时间Tr大于所述完成时间要求T0，且所述实际时间大于所述模拟时间，则加大所述初始时间余量Tm直至所述实际时间小于所述模拟时间；

[0151] 若实际时间Tr大于所述完成时间要求T0，且所述实际时间小于所述模拟时间，则将所述实际时间赋给所述模拟时间，然后用所述模拟时间减去所述初始时间余量Tm，得到中间时间Tz，以中间时间Tz为控制对象，加大对应分级硬件资源的频率，使得模拟计算的时间趋近于所述中间时间Tz(趋近于可以理解为±5%的Tz时间段内)；然后以加大后的对应分级硬件资源的参数进行所述多模态计算单元计算，得到更新后的实际时间Tr1；

[0152] 对所述初始时间余量Tm进行调整，得到更新后的时间余量Tm1，

[0153] 将所述更新后的实际时间Tr1赋给所述模拟时间，将所述更新后的时间余量Tm1替换所述初始时间余量Tm，重复该阶段的计算，直至更新后的实际时间落入【(1-k)*T0，T0】，其中k为误差系数，k的范围为1%-5%。

[0154] 例如，完成时间要求T0=100ms，在所述第二级硬件资源的模拟计算下，计算得到的T51为100ms，我们设置的初始时间余量一般可以往大的设置，因为是第一次计算，我们容许实际运算时具有足够的时间，因此，我们将该余量Tm设置为50ms。此时，模拟时间为T51+Tm=150ms。(由于硬件的功耗节省、时间节省并非是线性的，如附图11和12都可以看出，因此，无法直接作出调整，需要多次逼近)

[0155] 在得到T51的主频的参数下，第二级硬件资源进行实际计算的时间如果小于100ms，由于上述的T51已经是在最低功耗下(频率和电压)下得到的，因此，这种情况下实际

运行的状态已经是最优的状态。

[0156] 在得到T51的主频的参数下，第二级硬件资源进行实际计算的时间如果大于100ms，如是130ms，由于130ms大于100ms，则需要进行优化，利用130减去初始设置的余量50ms，得到中间时间Tz=80ms，显然，低于T51的100ms，要想在更快的时间内完成，必须加大频率(如附图13所示，电压随着频率变化，因此，我们实质上只需要调准频率即可)，使得在该频率下运行的硬件完成计算的时间趋近于80ms，由于无线逼近需要多次尝试，因此，我们一般设置一个容许范围，如±5%。确定了该频率，然后以该频率进行实际的计算，得到更新后的实际时间，例如实际时间为105，还是高于完成时间要求T0。

[0157] 此时，对所述初始时间余量Tm进行调整，得到更新后的时间余量Tm1，为了尽快求得最优解，一般采用比例调整，如Tm1/Tm=(Tr1- T0)/(Tr- T0)。

[0158] 将所述更新后的实际时间Tg赋给所述模拟时间，将所述更新后的时间余量Tg替换所述初始时间余量Tm，重复该阶段的计算，直至更新后的实际时间落入【(1-k)*T0，T0】。

[0159] 本发明的硬件系统如下：

[0160] 一种人工智能动态心电检测系统，所述系统包括多模态信号采集单元、AI数据处理单元、电源管理单元、无线通讯单元、语音单元和人机操作接口，所述AI数据处理单元和电源管理单元实施上述的方法。

[0161] 本发明的软件系统如下：

[0162] 一种人工智能动态心电监测系统，所述系统包括应用层、系统管理层和数据处理层；

[0163] 所述应用层包括人机交互单元；

[0164] 所述系统管理层包括电源管理单元、主控制单元；

[0165] 所述数据处理层包括多模态数据采集单元、数据预处理、SOC低功耗实时AI计算、FPGA大数据非实时计算、计算结果融合单元和无线通讯单元；

[0166] 所述系统管理层和数据处理层实施上述的方法。

[0167] 通过上述的技术方案，以计算资源需求和完成时间为控制限制，以功耗作为控制指标，对各层计算资源进行预测计算，作为系统调度的决策依据；可以通过闭环控制对控制决策计算模块进行持续优化，实现动态心肺监测仪系统的能效最大，从而达到用一套最小体积的人工智能边缘计算设备，满足各种监测场景而且设备待机时间最长。

[0168] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

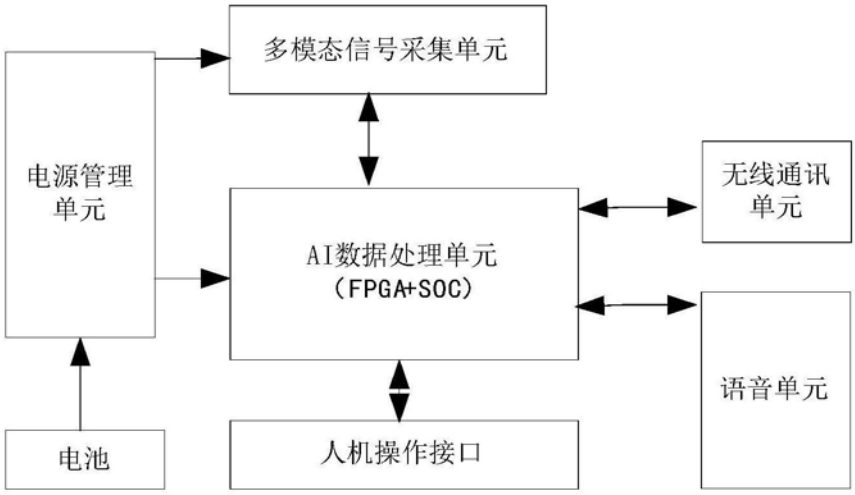

多模态信号采集单元
电源管理单元
AI数据处理单元（FPGA+SOC）
无线通讯单元
语音单元
电池
人机操作接口

图1

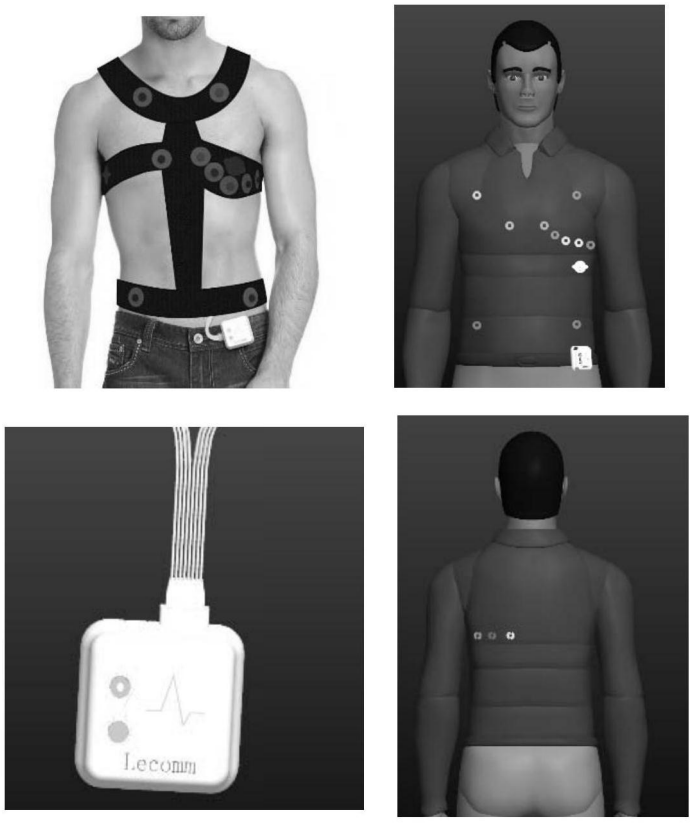

Lecomm

图2

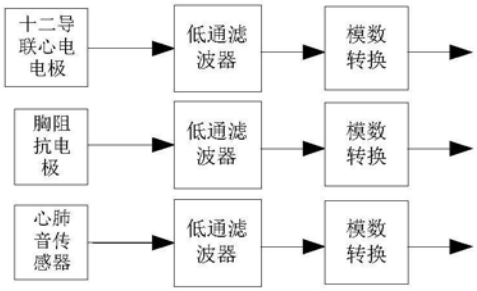

十二导联心电电极
胸阻抗电极
心肺音传感器
低通滤波器
低通滤波器
低通滤波器
模数转换
模数转换
模数转换

图3

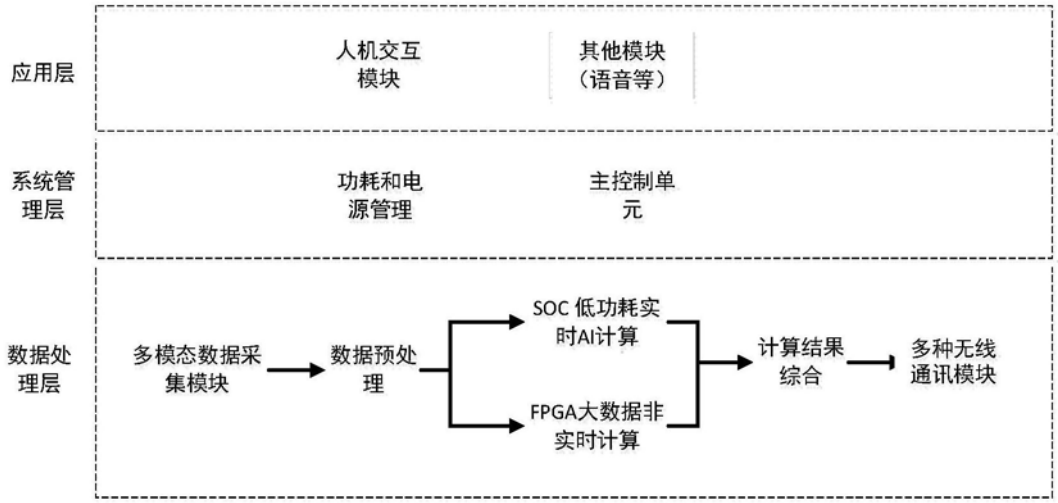

应用层
人机交互模块
其他模块（语音等）
系统管理层
功耗和电源管理
主控制单元
数据处理层
多模态数据采集模块
数据预处理
SOC 低功耗实时AI计算
FPGA大数据非实时计算
计算结果综合
多种无线通讯模块

图4

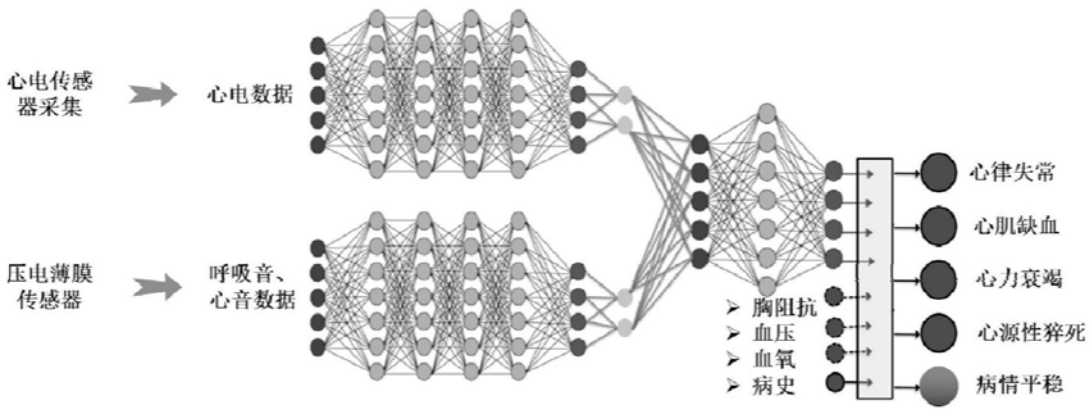

心电传感器采集
心电数据
压电薄膜传感器
呼吸音、心音数据
胸阻抗
血压
血氧
病史
心律失常
心肌缺血
心力衰竭
心源性猝死
病情平稳

图5

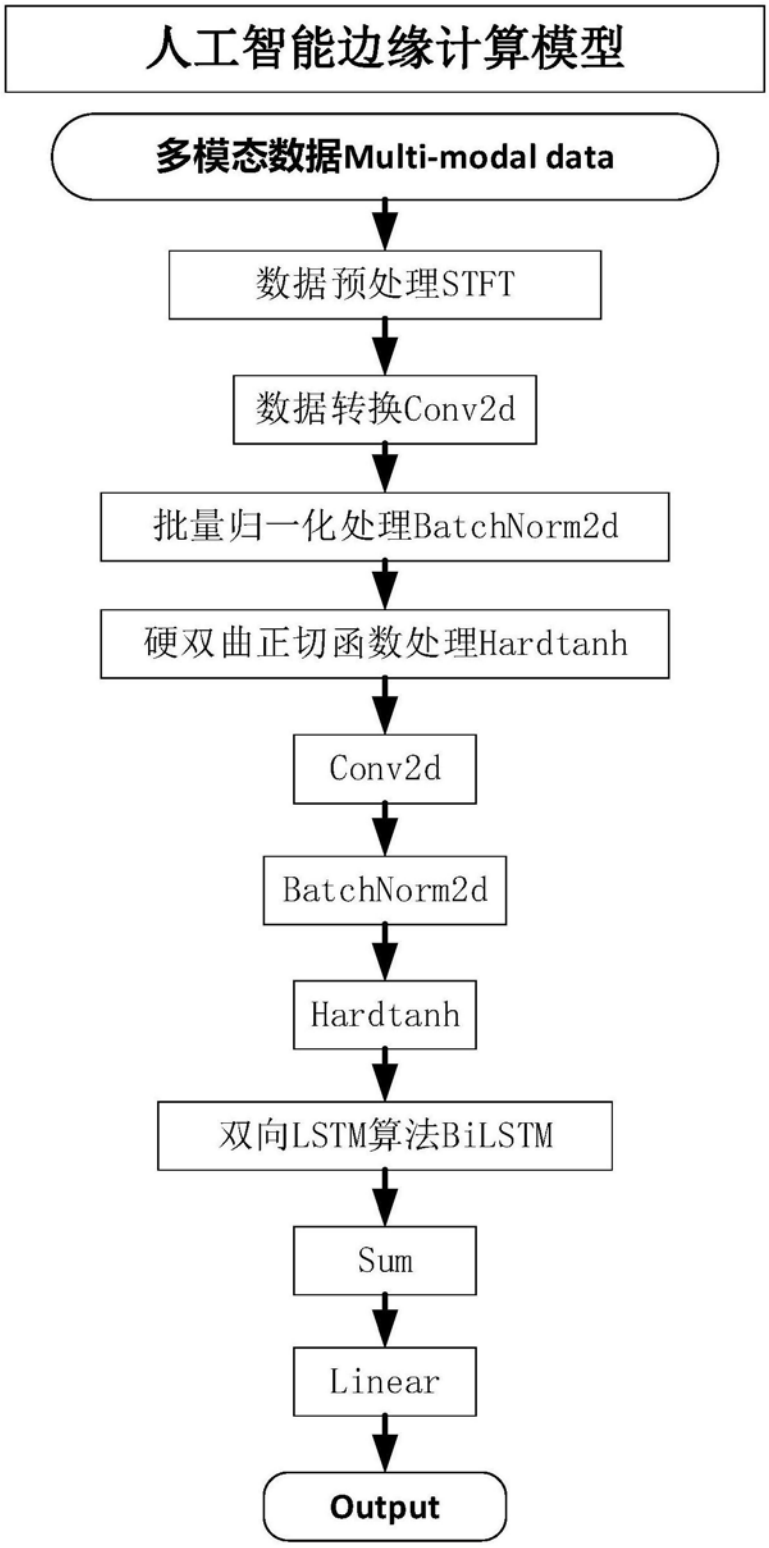

人工智能边缘计算模型
多模态数据Multi-modal data
数据预处理STFT
数据转换Conv2d
批量归一化处理BatchNorm2d
硬双曲正切函数处理Hardtanh
Conv2d
BatchNorm2d
Hardtanh
双向LSTM算法BiLSTM
Sum
Linear
Output

图6

深度学习模型轻量化评估		
输入	音频数据	8Khz/s 采样
输出	诊断类别	10 类
资源	参数个数	357152 float < 1.5MB
	计算量	34915562 FLOPS
	运行内存	4677650 float < 18MB

图7

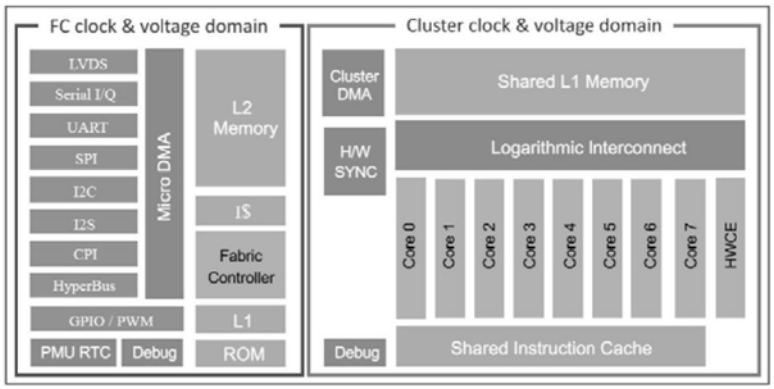


图8

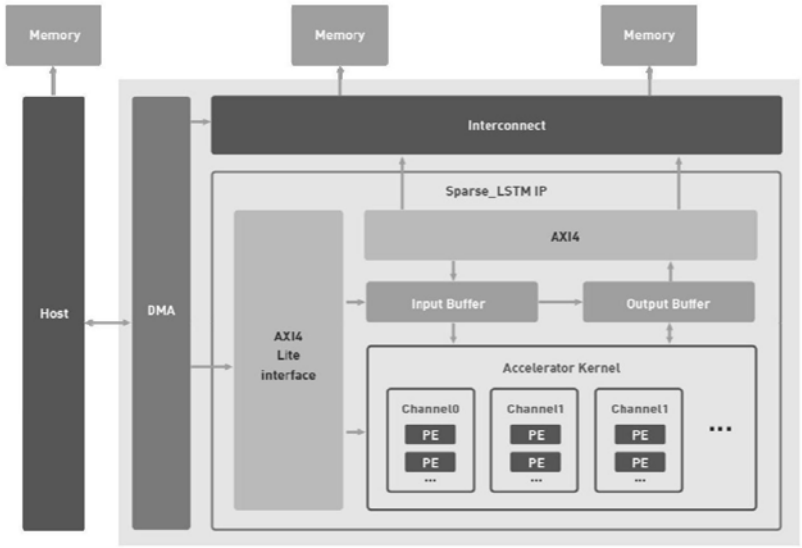


图9

计算资源		
芯片层级	芯片内硬件模块层级	芯片内模块配置层级
FPGA	PE（多核）	
MCU	HWCE	主频和电压
	cluster（多核）	主频和电压

图10

Topology	Cores					Speed-up vs. 1-core	Energy gain vs. 1-core
	1	2	4	8	+ HWCE		
CIFAR10	711 K	415 K	254 K	178 K	65 K	10.9×	4.9×
MNIST	7.6 M	4 M	2.4 M	15 M	0.8 M	9.3×	4.3×
TextReco	98 M	51 M	28 M	17 M	8.3 M	11.7×	5.3×

图11

Greenwave GAP SOC 运行 CIFAR-10 测试结果				
计算资源	主频 Clock	运行 Time	计算 Cycles	功耗 Active Power
8 core cluster	15.4Mhz	99.1ms	1500000	3.7mW
8 core cluster	175Mhz	8.7ms	1500000	70mW
8 cluster with HCE	4.7Mhz	99.1ms	460000	0.8mW

图12

Power Mode	VDD [V]	Nom. Freq. [MHz]	Power @ Nom. Freq.
Deep Sleep	0.8	0.32	3.6 µW
Retentive Deep Sleep	0.8	0.32	30 µW
FC ON	1.0	150	27 mW
FC ON, Cluster ON	1.0	150, 90	37 mW
FC ON	1.2	250	39 mW
FC ON, Cluster ON	1.2	250, 170	75 mW

图13

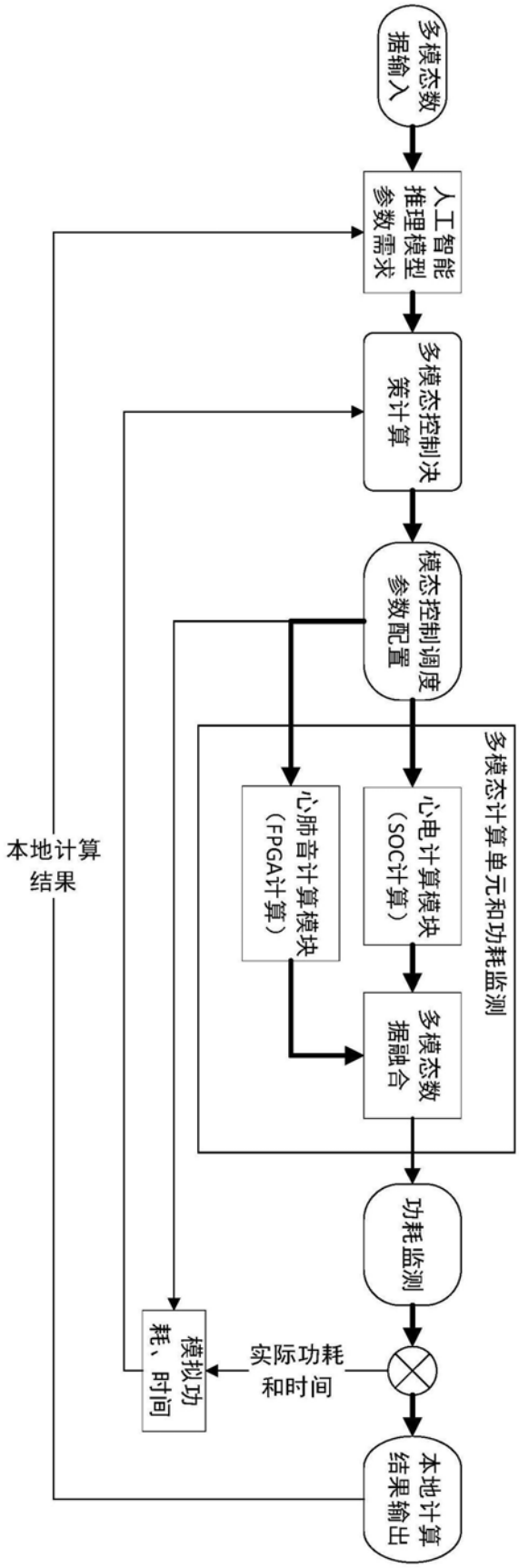

多模态数据输入
人工智能推理模型参数需求
多模态控制决策计算
模态控制调度参数配置
多模态计算单元和功耗监测
心电计算模块（SOC计算）
心肺音计算模块（FPGA计算）
多模态数据融合
功耗监测
实际功耗和时间
模拟功耗、时间
本地计算结果
本地计算结果输出

图14

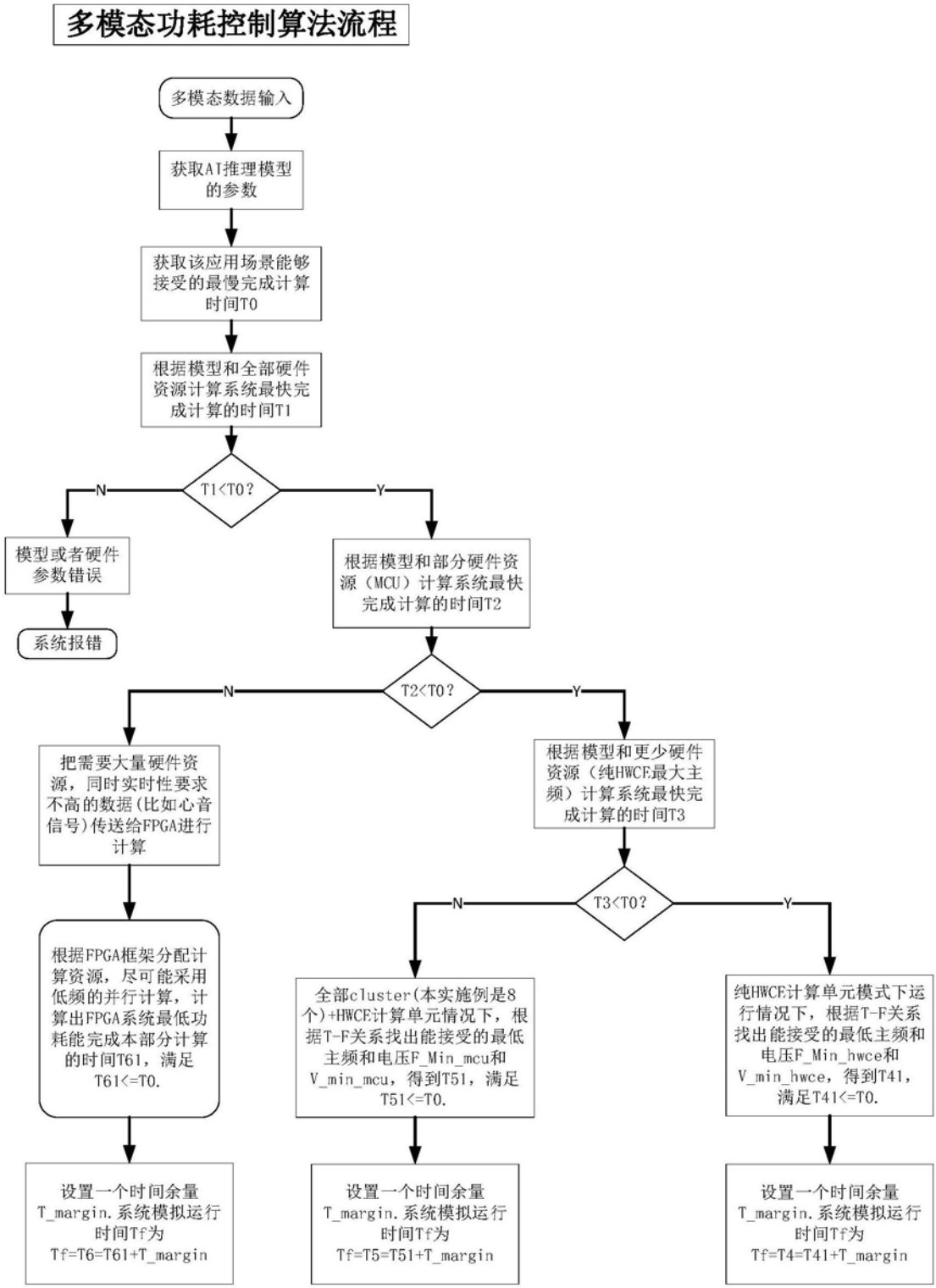

多模态功耗控制算法流程
多模态数据输入
获取AI推理模型的参数
获取该应用场景能够接受的最慢完成计算时间T0
根据模型和全部硬件资源计算系统最快完成计算的时间T1
T1<T0?
N
Y
模型或者硬件参数错误
系统报错
根据模型和部分硬件资源（MCU）计算系统最快完成计算的时间T2
T2<T0?
N
Y
把需要大量硬件资源，同时实时性要求不高的数据(比如心音信号)传送给FPGA进行计算
根据模型和更少硬件资源（纯HWCE最大主频）计算系统最快完成计算的时间T3
T3<T0?
N
Y
根据FPGA框架分配计算资源，尽可能采用低频的并行计算，计算出FPGA系统最低功耗能完成本部分计算的时间T61，满足T61<=T0.
全部cluster(本实施例是8个)+HWCE计算单元情况下，根据T-F关系找出能接受的最低主频和电压F_Min_mcu和V_min_mcu，得到T51，满足T51<=T0.
纯HWCE计算单元模式下运行情况下，根据T-F关系找出能接受的最低主频和电压F_Min_hwce和V_min_hwce，得到T41，满足T41<=T0.
设置一个时间余量T_margin.系统模拟运行时间Tf为Tf=T6=T61+T_margin
设置一个时间余量T_margin.系统模拟运行时间Tf为Tf=T5=T51+T_margin
设置一个时间余量T_margin.系统模拟运行时间Tf为Tf=T4=T41+T_margin

图15

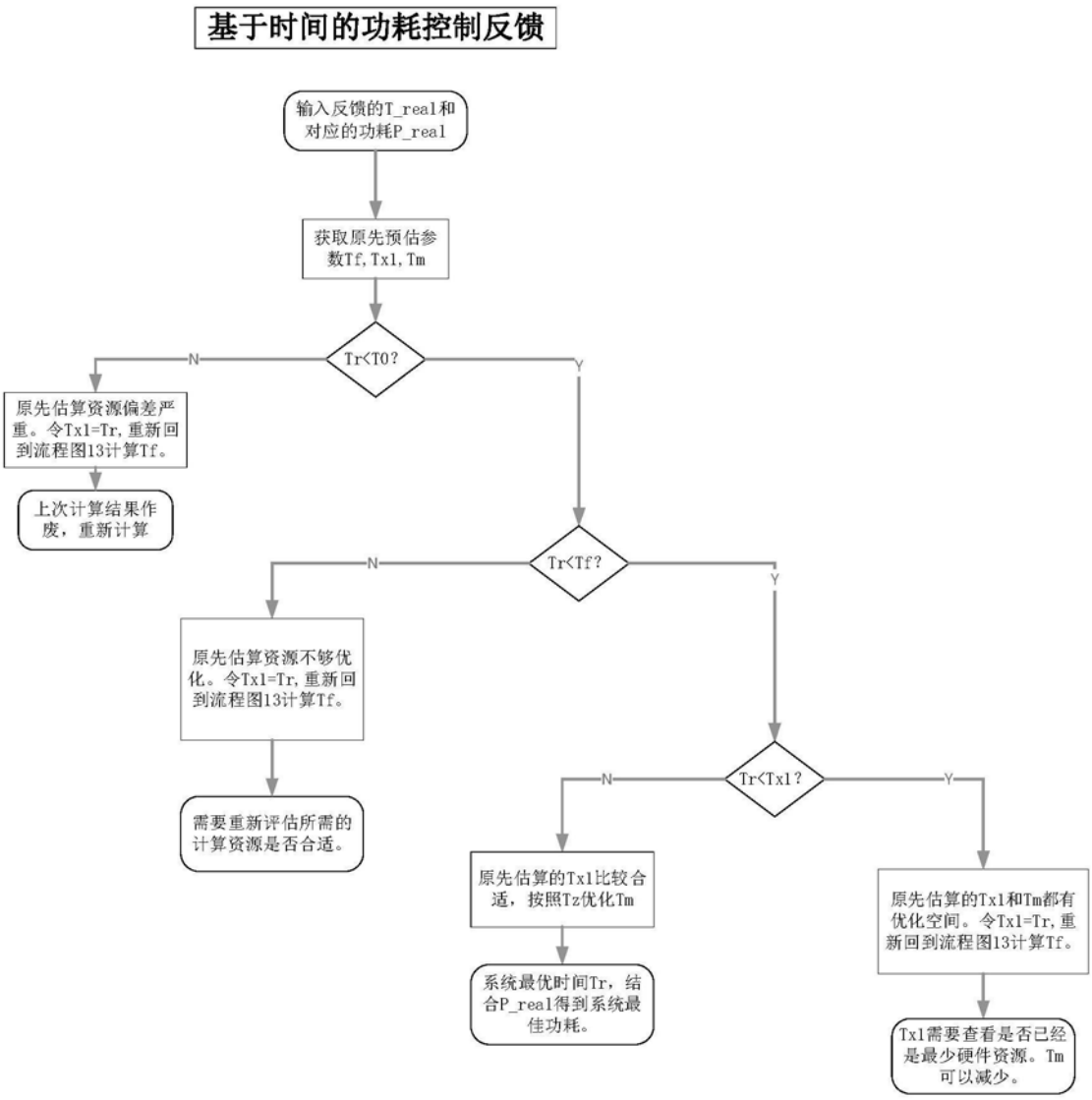

基于时间的功耗控制反馈
输入反馈的T_real和对应的功耗P_real
获取原先预估参数Tf,Tx1,Tm
Tr<T0?
N
Y
原先估算资源偏差严重。令Tx1=Tr,重新回到流程图13计算Tf。
上次计算结果作废，重新计算
Tr<Tf?
N
Y
原先估算资源不够优化。令Tx1=Tr,重新回到流程图13计算Tf。
需要重新评估所需的计算资源是否合适。
Tr<Tx1?
N
Y
原先估算的Tx1比较合适，按照Tz优化Tm
系统最优时间Tr，结合P_real得到系统最佳功耗。
原先估算的Tx1和Tm都有优化空间。令Tx1=Tr,重新回到流程图13计算Tf。
Tx1需要查看是否已经是最少硬件资源。Tm可以减少。

图16

专利名称(译)	一种人工智能动态心肺监测设备的功耗控制方法及系统		
公开(公告)号	CN110074776B	公开(公告)日	2020-04-10
申请号	CN201910362918.3	申请日	2019-04-30
[标]发明人	马振宇		
发明人	马振宇		
IPC分类号	A61B5/0402 A61B5/053 A61B5/1455 A61B7/00 A61B5/00		
CPC分类号	A61B5/0006 A61B5/0402 A61B5/053 A61B5/1455 A61B5/7225 A61B7/003		
其他公开文献	CN110074776A		
外部链接	Espacenet SIPO		

摘要(译)

本发明涉及一种人工智能动态心肺监测设备的功耗控制方法及系统，通过该方法和系统，以计算资源需求和完成时间为控制限制，以功耗作为控制指标，对各层计算资源进行预测计算，作为系统调度的决策依据；可以通过闭环控制对控制决策计算模块进行持续优化，实现动态心肺监测设备系统的能效最大，从而达到用一套最少硬件电路资源的人工智能边缘计算设备，满足各种监测场景而且设备待机时间最长。