大纲

TITLE：A Systolic Array Architecture supporting dynamic reshaping with efficient hardware mapping based on nerual network algorithm

1. **Introduction**

伴随着深度学习的发展，深度神经网络如今已经被广泛应用于各种领域。深度神经网络本质上由一系列不同的神经网络层组成，而不同层之间的结构又有所不同，因此不同的网络拓扑结构可以衍生出许多不同的深度神经网络模型，如CNNs,RNNs,MLPs以及Transformers等。针对不同的领域下的各种子任务，一般需要采用合适的深度神经网络模型。

然而，深度神经网络多样的拓扑结构导致了不同的工作负载（workloads），极大地限制了神经网络硬件加速器的工作效率，同时给硬件设计的灵活性带来极大的挑战，因此许多新的硬件加速计算架构被提出。脉动阵列是其中一种优先的架构选择，（脉动阵列的优点....）。然而，脉动阵列天然固定的尺寸通常导致了很低的PE利用率，对于许多深度神经网络层例如LSTM以及depth-wise卷积层，PE的利用率会降低到10%，导致大量的PE资源空闲。因此，目前有许多针对可变尺寸的脉动阵列的研究。

近期研究提出了（。。。。。。。。。）引出脉动阵列尺寸配置的速率慢的问题。然而，目前的研究基本采用强化学习，搜索算法等穷举性的算法，效率低且速度慢。且算法大多部署在软件端，需要额外以耗时进行硬件的配置。

针对上述问题，本文提出了一种可配置脉动阵列架构，可以通过硬件化的MLP神经网络算法来对脉动阵列的尺寸动态配置。主要贡献如下：

1.利用mlp算法，将对脉动阵列尺寸的映射从软件端移至硬件端，充分发挥硬件的天然速度优势，加速脉动阵列尺寸的配置。

2.通过优化设计PE单元，使脉动阵列内部数据可全向流动，设计了支持尺寸可重新配置以及三种dataflow的脉动阵列架构。

3.

**II. 背景和动机**

2.1DNN多样性的工作负载

介绍不同DNN的拓扑结构差异以及不同层的尺寸差异，说明DNN具有多种多样的工作负载，单一的硬件结构面对多样的DNN工作负载时无法充分发挥硬件的效率。

2.2脉动阵列

主要介绍脉动阵列的三种dataflow的计算模式。

2.3相关工作

列举相关的工作gemmini，planaria，SARA，Redas，和本文作对比。

**III. 脉动阵列架构**

3.1PE单元

详细描述PE单元的硬件结构，以及PE单元直接阵列内全向流动的机制。

3.2动态配置尺寸的脉动阵列结构

**IV. 配置和映射**

4.1search space分析

分析针对0-1024尺寸的神经网络，可配置的空间是多少。search space太过于大，穷举很浪费时间，因此选择用神经网络完成配置的映射。

4.2用MLP做分类任务完成映射

主要阐述如何把search space的映射用神经网络MLP做分类任务完成。

4.3数据集构建

利用开源的scale-sim-v2脉动阵列仿真器构建了矩阵乘法尺寸和对应的最佳配置（M-N-K，best\_config)数据集，用于MLP的

4.4 MLP硬件结构

描述MLP的硬件结构实现

**V. 实验性能评估**

5.1实验setup

实验评估方法

5.2性能评估

1.硬件资源消耗

2.映射速度

3.PE利用率

4.计算矩阵乘法的速度

5.3

**VI.总结**