# 一种基于脉动阵列的二值复数神经网络卷积计算加速装置与方法

## 技术领域

本发明涉及人工智能领域，具体涉及一种基于脉动阵列的二值复数神经网络卷积计算加速装置与方法。

## 背景技术

近年来，随着人工智能的飞速发展，深度神经网络（Deep Neural Networks，DNN）在诸多领域都有着广泛应用，如目标检测、人脸识别、图像视频超分处理、自动驾驶以及语音识别等。DNN算法具有存储密集和计算密集两大特性，因为其中包含了大量的权重数据和特征图数据之间的卷积操作，使得目前的DNN模型无法在边缘侧充分地训练或者推理。

在进行卷积操作时，一般会将一张特征图的卷积操作转换成多个小型的矩阵乘法；如果采用传统的计算系统，即从内存中读取特征图矩阵数据和卷积核权重矩阵数据，完成计算后写回存储器，再继续下一批数据的计算，这种方法会导致数据复用率低，内存访问次数多、耗时久，计算效率低等问题。所以在人工智能芯片领域多采用脉动阵列的计算方式，以权重固定型脉动阵列完成矩阵乘法为例，其中A为输入特征图矩阵，B为卷积核数据矩阵，C为输出特征图矩阵，计算方法为：首先将权重矩阵数据在脉动阵列中自上而下传播，传播完成后将权重数据固定于每个脉动阵列单元中；再将输入特征图矩阵数据按行依次间隔一个周期输入脉动阵列，并自左向右传播；输入数据与权重数据进行乘法计算，乘法结果与上一级计算单元的部分和输出进行累加计算，所得新的部分和结果向下传播给下一级计算单元；最后一级计算单元将最终的部分和结果传入累加器中。

二值复数神经网络有效地结合二值神经网络和复数神经网络，其中每层的输入特征图数据、权重数据和输出特征图数据都使用二值化的复数值来表示，即{1+i,1-i,-1+i,-1-i}中某一个，在进行卷积计算时，原先的全精度乘法计算也相应地映射为单比特的同或运算（xnor）外加一个计“1”（popcount）操作，在数据存储时，使用单比特替代32比特的单精度浮点数或16比特的半精度浮点，该网络具有二值神经网络的计算效率、低硬件开销、高能效比、高鲁棒性、高精度等多种优势。

## 发明内容

为解决深度神经网络高硬件开销、计算效率低、无法在资源受限的边缘侧设备中部署等问题，实现低功耗、低成本、高能效比的边缘智能设备的目的，本发明采用如下的技术方案：

一种基于脉动阵列的二值复数神经网络卷积计算加速装置，装置中包括两个移位寄存器阵列、针对二值复数设计的基本运算单元PE组成的脉动阵列、相邻PE之间的寄存器和一个带累加器的SRAM存储单元。两个移位寄存器阵列分别位于脉动阵列的左侧与下侧，

## 具体实施方法

## 附图

### 附图1

## 参考文献

1. Bansal A, Borgnia E, Chu H M, et al. Cold diffusion: Inverting arbitrary image transforms without noise[J]. arXiv preprint arXiv:2208.09392, 2022.
2. Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[C]// arXiv. arXiv, 2017.