航空科学技术
Aeronautical Science & Technology
May 25 2022 Vol. 33 No.05 89-96

面向航空目标检测的神经网络加速器设计

施立瑞，王帅帅，肖昊
合肥工业大学，安徽 合肥 230000

摘 要：卷积神经网络被广泛应用于航空图像目标检测领域。然而，由于航空图像成像背景环境复杂、目标尺寸小且方向任意，为了提取更高层次的特征信息，神经网络模型的结构复杂度不断提高，使得模型计算复杂度高、计算时间长，从而难以满足航空目标检测的实时性需求。本文提出了一种面向航空目标检测的基于Winograd算法的神经网络加速器，通过Winograd卷积算法可大幅减少卷积计算中的乘法数量，并针对Winograd卷积在神经网络计算中由于时域变换引入额外加法计算的问题，提出了一种深流水的矩阵变换计算结构，通过复用加法计算的中间结果以及调整运算顺序减少输入和输出变换的计算量。同时，针对加速器的现场可编程门阵列（FPGA）实现，提出了一种高效的数据流形式和DSP阵列结构。试验结果表明，本文提出的加速器相比CPU和GPU分别获得了32倍和2.6倍的速度提升。

关键词：目标检测；卷积神经网络；加速器；Winograd算法；FPGA

中图分类号：TP183 **文献标识码：**A DOI：10.19452/j.issn1007-5453.2022.05.013

近年来，航空图像数据分析需求大幅增加，卷积神经网络（convolutional neural network，CNN）由于其优越的性能被广泛应用于航空目标检测[1-5]。然而，由于航空图像成像背景环境复杂、目标尺寸小且方向任意，为了获取更高的预测准确率，神经网络模型层数逐渐加深，结构更加复杂，使得模型计算复杂度高、计算时间长，从而难以满足航空目标检测中实时性的需求。因此，如何根据遥感图像快速检测目标成为一项巨大的挑战。

大量研究从软硬件方面优化提升航空图像目标检测速度。在软件算法优化方面，参考文献[6]和文献[7]通过提出轻量级网络模型减少网络参数和计算复杂度。参考文献[6]提出了一种压缩MobileNet网络模型，在特征映射下采样阶段减少瓶颈架构数量，但在特征映射平台阶段增加更多瓶颈，从而减少网络参数，缩短推理时间并提高预测准确性。参考文献[7]提出了一种轻量级神经网络模型G-YOLOv3，其主要构件为嵌套残差块，每个残差块中都包含了本征特征图与相似特征图，且残差块中添加注意力机制，增强网络的学习能力，从而降低网络所需参数和计算复杂度，提升目标检测速度。在硬件加速方面，参考文献[8]在MPSOC硬件平台对网络计算进行硬件加速，通过丰富的可编程逻辑资源提升计算并行度，缩短网络计算时间。然而，上述加速目标检测的方法均基于传统卷积计算，而在航空目标检测中用来提取特征值的卷积神经网络占据了网络推断阶段70%以上的时间[6,9]，传统卷积由于高计算复杂度成为限制目标检测速度的瓶颈。因此，本文提出了一种面向航空目标检测的基于Winograd算法的神经网络加速器设计，包括：（1）对航空目标检测网络中的基干网络部分进行硬件加速，通过Winograd算法降低了卷积计算的复杂度，与传统卷积相比减少了75%的乘法计算，同时提出了一种深流水结构的矩阵变换方法，优化Winograd算法中的加法计算，与直接Winograd计算相比减少了18.75%的输入变换计算量和至少76.19%的输出变换计算量；（2）在FPGA平台加速卷积神经网络，设计了一种高效的DSP阵列结构和数据流方式，大幅提高DSP

收稿日期：2022-01-06；退修日期：2022-02-11；录用日期：2022-03-19
基金项目：国家自然科学基金(61974039)；航空科学基金(2018ZCP4003)
引用格式：Shi Lirui，Wang Shuaishuai，Xiao Hao. Design of neural network accelerator for aeronautical target detection[J]. Aeronautical Science & Technology，2022，33(05)：89-96. 施立瑞，王帅帅，肖昊. 面向航空目标检测的神经网络加速器设计[J]. 航空科学技术，2022，33(05)：89-96.

的计算效率。本文提出的加速器相比CPU和GPU分别获得了32倍和2.6倍的速度提升。

1 目标检测网络

1.1 网络结构

航空目标检测网络模型结构如图1所示，整体分为基干网络和检测网络两个部分。其中，基干网络主要完成图片特征值提取的功能，检测网络根据基干网络提取的特征对特定目标进行检测。由于航空图像成像背景复杂且目标尺度较小，为了提取更高层次的特征信息，获得更高的准确率，基干网络采用深层卷积神经网络获取有效特征。其中，卷积计算示意图如图2所示，每个通道上的卷积核分别与对应通道上的输入数据相同大小的区域进行乘累加运算，每个通道上的乘累加后的结果相加得到输出数据中的一个值，接着卷积核以一定的步长不断滑动得到所有的输出数据。可见，传统卷积的滑窗计算方式计算复杂度高，因此，本文引入Winograd算法，利用图像像素点之间的结构相似性，将滑窗卷积变为矩阵点乘运算，大幅降低了卷积运算中乘法计算的数量。

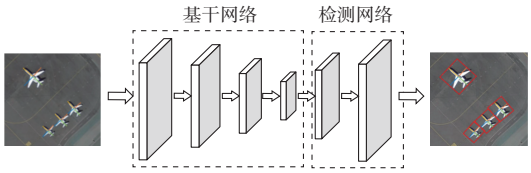

图1 航空目标检测网络模型示意图

Fig.1 Schematic diagram of the aerial target detection network model

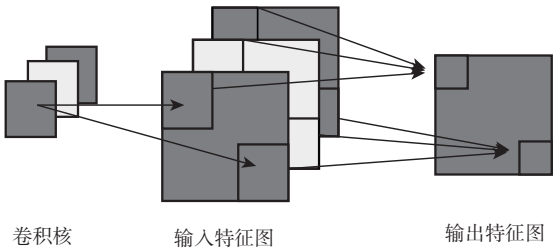

图2 传统卷积计算示意图

Fig.2 Schematic diagram of traditional convolution calculation

1.2 Winograd算法

Winograd算法[10]由数学家Winograd于1980年提出，用于信号处理中的卷积操作。以一维Winograd卷积为例，记输出尺寸为m，卷积核尺寸为r的卷积计算为$\boldsymbol{F}(m,r)$，$\boldsymbol{d}$为输入数据，$\boldsymbol{g}$为卷积核数据，我们通过一组具体的计算阐述Winograd卷积原理。

$$\boldsymbol{d}=\begin{bmatrix}d_0 & d_1 & d_2 & d_3\end{bmatrix}^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{g}=\begin{bmatrix}g_0 & g_1 & g_2\end{bmatrix}^{\mathrm{T}} \tag{2}$$

对于传统卷积，计算过程如下

$$\boldsymbol{F}(2,3)=\begin{bmatrix}d_0 & d_1 & d_2\\ d_1 & d_2 & d_3\end{bmatrix}\begin{bmatrix}g_0\\ g_1\\ g_2\end{bmatrix}=\begin{bmatrix}r_0\\ r_1\end{bmatrix} \tag{3}$$

其中，r_0,r_1为

$$r_0=\left(d_0\cdot g_0\right)+\left(d_1\cdot g_1\right)+\left(d_2\cdot g_2\right) \tag{4}$$

$$r_1=\left(d_1\cdot g_0\right)+\left(d_2\cdot g_1\right)+\left(d_3\cdot g_2\right) \tag{5}$$

对于Winograd卷积，计算过程如下

$$\boldsymbol{F}(2,3)=\begin{bmatrix}d_0 & d_1 & d_2\\ d_1 & d_2 & d_3\end{bmatrix}\begin{bmatrix}g_0\\ g_1\\ g_2\end{bmatrix}=\begin{bmatrix}m_1+m_2+m_3\\ m_2-m_3-m_4\end{bmatrix} \tag{6}$$

其中，

$$m_1=\left(d_0-d_2\right)g_0 \tag{7}$$

$$m_2=\left(d_1+d_2\right)\frac{g_0+g_1+g_2}{2} \tag{8}$$

$$m_3=\left(d_2-d_1\right)\frac{g_0-g_1+g_2}{2} \tag{9}$$

$$m_4=\left(d_1-d_3\right)g_2 \tag{10}$$

由于CNN模型已知，卷积核数据固定不变，故g_0,g_1,g_2相关运算只需计算一次。因此，实际应用中可忽略卷积核的变换，即Winograd卷积共需4次乘法和8次加、减法。相较于传统卷积的6次乘法和4次加法计算，Winograd卷积以增加一倍加法为代价，节省了33%的乘法计算。由于硬件中乘法的实现成本远高于加法，因此Winograd卷积算法较传统卷积算法更友好。

将上述一维Winograd卷积计算表示为矩阵相乘形式，如式(11)所示。其中，$\boldsymbol{A}$、$\boldsymbol{G}$、$\boldsymbol{B}$、$\boldsymbol{Y}$分别为输出特征图转换矩阵、卷积核转换矩阵、输入特征图转换矩阵、输出特征图矩阵。针对二维CNN计算，通过嵌套使用两次一维Winograd卷积得到其矩阵计算表达，如式(15)所示。二维Winograd卷积计算中，矩阵矢量乘法操作计算量为$(m+r-1)^2$，相较于传统卷积的$m^2\times r^2$次乘法有大幅度降低。以本文2.2节采取的6×6分块方法为例，m=4，r=3，Wingrad卷积的乘法计算量为36，传统卷积乘法量为144，Winograd卷积相比传统卷积乘法量减少了75%。

$$\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}\left[\left(\boldsymbol{G}g\right)\odot\left(\boldsymbol{B}^{\mathrm{T}}d\right)\right] \tag{11}$$

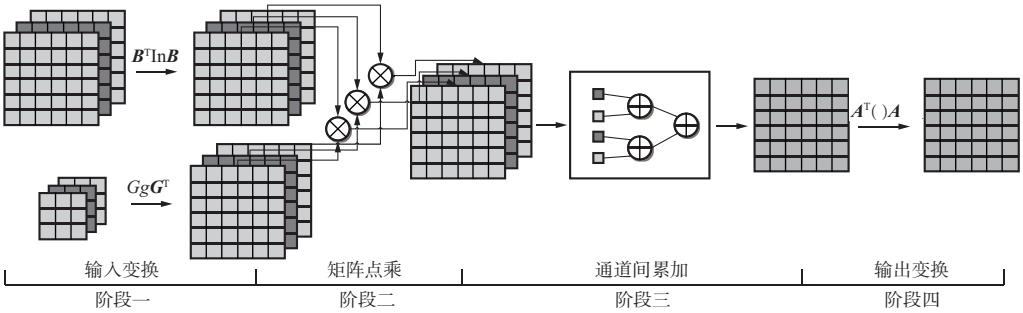

图3 Winograd卷积计算流程示意图

Fig.3 Schematic diagram of Winograd convolution calculation process

$$\boldsymbol{B}^{\mathrm{T}}=\begin{bmatrix}1&0&-1&0\\0&1&1&0\\0&-1&1&0\\0&1&0&-1\end{bmatrix} \tag{12}$$

$$\boldsymbol{G}=\begin{bmatrix}1&0&0\\\frac{1}{2}&\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}\\0&0&1\end{bmatrix} \tag{13}$$

$$\boldsymbol{A}^{\mathrm{T}}=\begin{bmatrix}1&1&1&0\\0&1&-1&-1\end{bmatrix} \tag{14}$$

$$\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}\left[\left(\boldsymbol{GgG}^{\mathrm{T}}\right)\odot\left(\boldsymbol{B}^{\mathrm{T}}d\boldsymbol{B}\right)\right] \tag{15}$$

将Winograd卷积应用到CNN中的计算流程图如图3所示。第一阶段，将输入特征图和卷积核数据从时域转换到Winograd域；第二阶段，变换后的输入特征图和卷积核执行矩阵点乘计算；第三阶段，各通道间的卷积结果执行累加；第四阶段，将Winograd域的卷积结果变换到时域。其中，第三阶段和第四阶段的执行顺序不影响最终卷积结果。

2 CNN加速器设计

2.1 目标检测系统架构

航空目标检测系统架构如图4所示，基于网络部分采用FPGA加速，检测网络在主机完成。其中，加速器主要由输入特征图缓存、权重缓存、输入特征图变换、并行计算单元(PE)、输出特征图缓存组成。由于神经网络参数量巨大而片上存储资源有限，外部存储器中的数据将根据运算阶段分批次加载到片上缓存中。其中，权重变换在片外预处理完成，片上缓存接收到的不再是原始权重数据而是转换后的数据。多个输入特征图转换模块并行处理多通道输入数据，多个PE单元并行处理多个卷积核上的卷积，二者并行度可根据不同硬件平台的资源进行调整。在输出变换中，PE单元循环复用多次，完成所有通道卷积和累加后再执行输出变换，与参考文献[11]相比，输出变换的运算量仅为前者的1/N，N为输入通道数。由于片上存储资源有限，当输出特征图缓存中积累了一定数据后，通过访存控制传输到片外存储器中。当加速器完成特征值提取后，将数据传到主机进行检测网络计算，实现完整目标检测功能。

2.2 输入/输出特征图转换

Winograd算法是针对短卷积的优化算法，而在神经网络计算中，输入特征图尺寸较大，因而需要经过分块处理再逐一进行Winograd卷积计算。不同分块尺寸的选取直接影响输入输出变换系数矩阵的复杂度。同时，分块的数量又影响着流水计算周期。表1对比了不同尺寸分割对计算量的影响，其中，C表示特征图长和宽。对于同一幅输入图，6×6分块方法的块数量是3×3分块方法的1/4，乘法计算量是3×3分块方法的56.25%，加法计算量是3×3分块方法的1.23倍，可见高阶的分块方法带来的计算复杂度和计算量的整体效率更佳。分块尺寸越大，转换矩阵中的常数范围越大且在变换阶段的加法运算量越大，研究表明8×8以上的分块方法在硬件实现时转换精度损失较大，影响卷积神经网络的预测准确率。因此，本设计选取6×6分块方法，此时转换矩阵如式(16)~式(18)所示。输入输出变换矩阵中常数可以由移位操作或移位操作的组合替代，如$5=2^2+1$。由于卷积核变换矩阵中的常数乘法由此方法实现时组合复杂、计算精度难以保证，考虑到神经网络推理阶段卷积核固定不变，本设计将卷积核变换在片外预处理完成。根据式(15)，输入特征图转换可表示为

$$\boldsymbol{U}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{XB}$$

式中：矩阵$\boldsymbol{B}$是常系数转换矩阵，与输入特征图的分块尺寸直接相关；$\boldsymbol{X}$为输入特征图的分块。式(20)是左变换的直接计算公式，右变换与左变换系数相同。现有设计在处理输入特征图转换时存在的问题有：(1)$\boldsymbol{U}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{XB}$计算时需要等待$\boldsymbol{U}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}$全部运算完毕后才执行下一步与$\boldsymbol{B}$的运

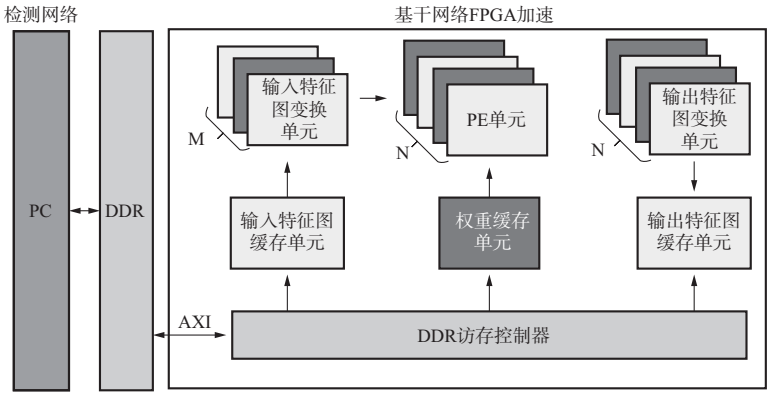

图4 航空目标检测系统架构

Fig.4 Aeronautical target detection system architecture

算，使得加法计算过程的延时可能会大于乘法计算（矩阵点乘单元）的延时，导致加速器性能下降；（2）输入转换矩阵中链式加法较长，直接计算一方面会影响系统的运行频率，另一方面，加法计算的中间结果存在可以复用的部分，直接计算时存在大量冗余计算。

表1 输入特征图（$C\times C$）分割尺寸对比

Table 1 Comparison of different segmentation size of input feature map ($C\times C$)

tile尺寸	分tile数量	乘法计算量	加法计算量
4×4	$(C/2)^2$	$16\times(C/2)^2$	$56\times(C/2)^2$
6×6	$(C/4)^2$	$36\times(C/4)^2$	$276\times(C/4)^2$
注：本文的讨论基于卷积核大小为3×3，步长为1的Winograd卷积			

针对以上问题，本设计提出了一种六级流水结构的输入/输出变换计算结构。图5为输入左变换电路结构，图6为输出左变换电路结构。在输入变换中，设计令$C_{j0}=-x_{2j}+x_{4j}$，$C_{j1}=-4x_{2j}+x_{4j}$，$C_{j2}=x_{1j}-x_{3j}$，此时左变换计算公式转换为式（21）。可见，该方法通过对计算重新拆分组合，使得常数乘法可由移位实现。同时，通过复用C_{j0}、C_{j1}、C_{j2}计算结果，加法计算量由原来的192减少为156，降低了18.75%。

右变换与左变换具有对称性，可由相同计算电路实现三级流水计算，故输入/出特征图变换由六级流水计算实现。各级计算均为两输入加法，相对于直接计算的长链式加法更利于加速器系统时钟频率的提升，同时流水计算的计算效率更高。

$$\boldsymbol{B}^{\mathrm{T}}=\begin{bmatrix}4&0&-5&0&1&0\\0&-4&-4&1&1&0\\0&4&-4&-1&1&0\\0&-2&-1&2&1&0\\0&2&-1&-2&1&0\\0&4&0&-5&0&1\end{bmatrix} \tag{16}$$

$$\boldsymbol{A}^{\mathrm{T}}=\begin{bmatrix}1&1&1&1&1&0\\0&1&-1&2&-2&0\\0&1&1&4&4&0\\0&1&-1&8&-8&1\end{bmatrix} \tag{17}$$

$$\boldsymbol{G}=\begin{bmatrix}\frac{1}{4}&0&0\\-\frac{1}{6}&-\frac{1}{6}&-\frac{1}{6}\\-\frac{1}{6}&\frac{1}{6}&-\frac{1}{6}\\\frac{1}{24}&\frac{1}{12}&\frac{1}{6}\\\frac{1}{24}&-\frac{1}{12}&\frac{1}{6}\\0&0&1\end{bmatrix} \tag{18}$$

$$\boldsymbol{X}=\begin{bmatrix}x_{0j}\\x_{1j}\\x_{2j}\\x_{3j}\\x_{4j}\\x_{5j}\end{bmatrix}\quad j\in[0,5] \tag{19}$$

$$\boldsymbol{B}^{\mathrm{T}}\cdot\boldsymbol{X}=\begin{bmatrix}4x_{0j}-5x_{2j}+x_{4j}\\-4x_{1j}-4x_{2j}+x_{3j}+x_{4j}\\4x_{1j}-4x_{2j}-x_{3j}+x_{4j}\\-2x_{1j}-x_{2j}+2x_{3j}+x_{4j}\\2x_{1j}-x_{2j}-2x_{3j}+x_{4j}\\4x_{1j}-5x_{3j}+x_{5j}\end{bmatrix} \tag{20}$$

$$\boldsymbol{B}^{\mathrm{T}}\cdot\boldsymbol{X}=\begin{bmatrix}4x_{0j}-4x_{2j}+C_{j0}\\-4x_{1j}+x_{3j}+C_{j1}\\4x_{1j}-x_{3j}+C_{j1}\\C_{j0}-2C_{j2}\\C_{j0}+2C_{j2}\\4C_{j0}-x_{3j}+x_{5j}\end{bmatrix} \tag{21}$$

2.3 PE单元

PE单元执行的功能是对转换后的输入特征图和权重进行点乘计算，如何合理地部署DSP资源是PE单元性能的

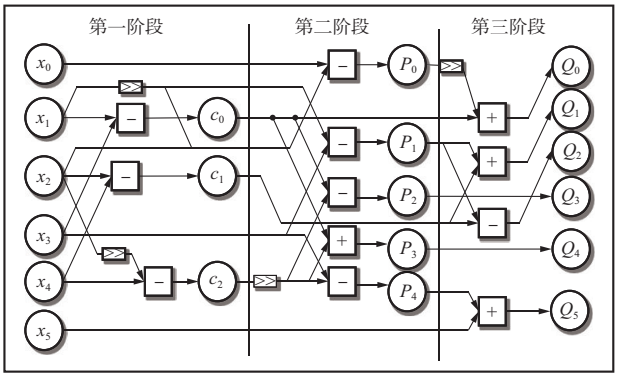

图5 输入左变换电路结构

Fig.5 Left conversion circuit of input feature map

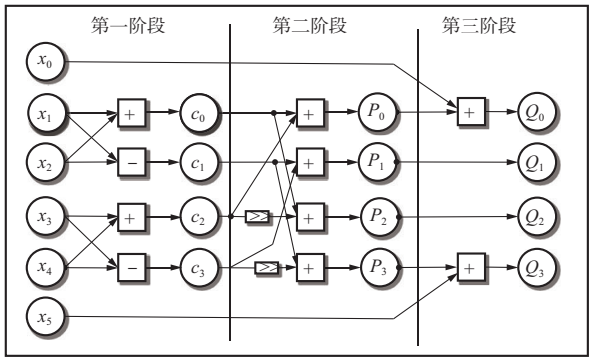

图6 输出左变换电路结构

Fig.6 Left conversion circuit of output feature map

关键。现有设计仅使用DSP中的乘法器处理乘法计算，而DSP中除乘法器外还包含加法器等丰富的计算资源。因此，本文提出了一种高效的DSP计算数据流形式和通道间级联的DSP阵列结构。

DSP内部的计算结构和数据流安排如图7所示，单个DSP中计算两个卷积核和输入特征图的卷积操作。输入特征图固定从A端口输入，两个卷积核的权重数据分别固定从B端口和D端口输入，两个卷积核复用输入特征图，通过控制DSP的配置参数使其乘法器执行$A\times B$和$A\times D$的乒乓操作。同时，加法器乒乓执行$A\times B$和$A\times D$的计算结果与其各自上一通道结果的累加。图8展示了DSP间级联的阵列结构，其中，$W_{a,b}$表示第a个卷积核的第b个通道的权重数据，indata_m为输入特征图第m个通道数据。P_{sum}、p_{sum}分别为上一循环和当前的部分和结果。级联结构主要用来处理通道间累加，级联的首个DSP从C口输入上一通道的结果，其余DSP从PCIN口级联输入上一通道的结果，最后输出两个通道的部分和(p_{sum})结果。PE单元包含多组级联的DSP组，假设有M组DSP组，N个输入通道，Z个卷积核，循环复用计算单元N/M次完成单个卷积核的计算，单次循环的最后一组DSP的输出保存在RAM中参与下一次循环计算，直至完成所有通道的累加。

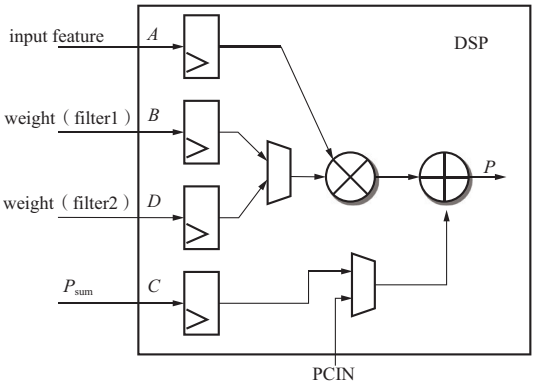

图7 DSP内部计算结构

Fig.7 DSP internal calculation structure

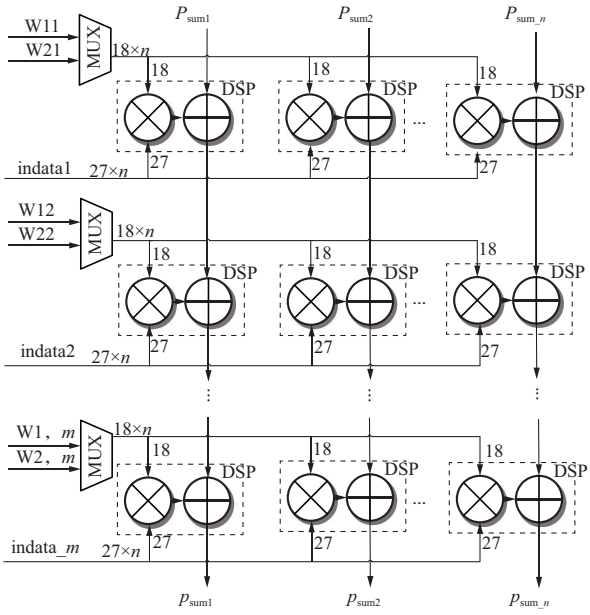

图8 PE中DSP阵列结构

Fig.8 DSP array structure in PE

DSP的数据流安排使得DSP的计算效率在计算过程中提高了4倍；DSP级联的阵列结构与参考文献[11]中单独设置加法单元方法相比，减少了累加计算周期。同时，充分利用了DSP中的计算资源，从而节省了片上LUT资源。

2.4 双缓冲区缓存单元

由于神经网络参数量庞大，减少数据传输时间是加速器性能提升的关键。本文采用图9所示双缓冲区结构对数据进行乒乓读写，隐藏了从片外缓存数据与片上计算之间的延时。

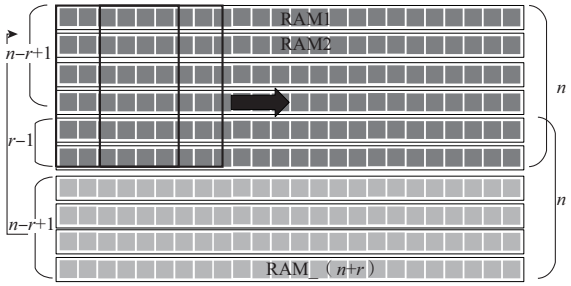

图9 双缓冲区单元示意图

Fig.9 Schematic diagram of double buffer unit

在$\boldsymbol{F}(m^2,r^2)$的二维Winograd卷积中，输入块的大小为$n\times n$，相邻输入块之间重叠r–1个像素点，为了在读取输入图分块时复用重叠数据，在缓存单元中，设计时将输入特征图按行存储在RAM中。首先在n个RAM中分别存储输入特征图的n行数据，然后计算模块开始从缓冲区读取输入特征图块开始计算，同时，缓冲区的$(n-r+1)$个RAM继续接收来自片外的输入特征图数据。

当缓冲区中的前n行数据由计算单元读取完成之后，重新从片外缓存$(n-r+1)$行输入特征图数据，此时计算单元从后n个RAM中读取数据，实现乒乓读写。该方法在复用输入数据的同时使得数据缓存单元和计算单元一直处于工作状态，有效减少了加速器整体运行时间。

3 试验与结果分析

为验证本文提出的基于Winograd算法的神经网络加速器的综合性能，试验选用经典VGG16网络模型作为航空目标检测的基干网络，SSD网络模型作为检测网络，并在Xilinx VCU118硬件平台部署基干网络部分CNN加速器，模型检测结果如图10所示。

将本文设计的加速器部署在FPGA上的推理延时与CPU、GPU上运行结果做对比见表2，其中，CPU采用Intel i5–4590处理器，GPU采用NVIDIA RTX2070super。相比CPU和GPU，本文加速器分别获得了32倍和2.6倍的速度提升。

表3评估了设计的硬件资源使用率、吞吐量、运行时间等性能指标。参考文献[5]加速器采用传统卷积计算，在硬件资源占用率相近的情况下，本文基于Winograd算法的加速器，相比参考文献[12]节省了60.1%的网络推断时间。参考文献[13]和文献[[11]与本文卷积算法相同。其中，参考文献[13]在Winograd卷积计算阶段将输出变换应用在通道累加之前，相比本文增加了71%的冗余计算。因而，在吞吐量相近的情况下，本文的加速器计算延时减少了31.5%。本文相比参考文献[11]增加了不到一倍的DSP资源而计算延时仅为其一半不到。在LUT资源占用上，由于本文优化了变换阶段的计算（主要由LUT实现），与参考文献[9]和文献

图10 航空图像检测结果

Fig.10 Aerial image detection results

表2 不同平台推理延时对比

Table 2 Comparison between of inference delays on different platforms

平台	目标检测系统推断时间/s
CPU	0.74
GPU	0.061
本文（CPU+FPGA）	0.023

表3 加速器性能对比

Table 3 Accelerator performance comparison

性能指标	参考文献[12]	参考文献[13]	参考文献[11]	本文
FPGA	Xilinx VCU118	Xilinx VCU118	Xilinx ZCU102	Xilinx VCU118
网络模型	VGG16	VGG16	VGG16	VGG16
频率/MHz	150	200	200	200
数据精度	NA	16-bit fixed	16-bit fixed	16-bit fixed
DSP	4096	3224	2520	4608
BRAM	1779	1672	1824	1632
LUT	781K	558K	600K	446K
吞吐量/GOPs	2558	3772	2601	3812
运行时间/ms	12.0	7	10.43	4.79
注：参考文献[12]加速器采用传统卷积方法				

[11]相比分别减少了20.1%和25.7%。经上述分析可见，本文加速器加速效果更佳。

4 结束语

本文提出了一种基于航空目标检测的神经网络加速器，通过Winograd卷积算法大幅减少卷积计算中的乘法数量，并针对当前Winograd卷积在CNN计算中由于时域变换引入额外加法计算的问题，提出了一种深流水的矩阵变换计算结构，通过复用加法计算的中间结果以及调整运算顺序减少输入变换和输出变换的计算量。同时，针对加速器的FPGA实现，提出了一种高效的数据流形式和DSP阵列结构。试验结果表明，本文的加速器吞吐量更高、计算延时更短。 AST

参考文献

[1] Yi J，Wu P，Liu B，et al. Oriented object detection in aerial images with box boundary-aware vectors[C]// 2021 IEEE Winter Conference on Applications of Computer Vision（WACV），2021：2149-2158.

[2] Kovbasiuk S，Kanevskyy L，Romanchuk M.Object detection method based on aerial image instance segmentation by unmanned aerial system in the framework of decision making system[C]// 2019 3rd International Conference on Advanced Infor-mation and Communications Technologies（AICT），2019：332-335.

[3] 赵文清,孔子旭,周震东,等.增强小目标特征的航空遥感目标检测[J].中国图象图形学报,2021,26(3):644-653.
Zhao Wenqing, Kong Zixu, Zhou Zhendong, et al. Aerial remote sensing target detection with enhanced small target features[J]. Chinese Journal of Image and Graphics, 2021, 26 (3):644-653. (in Chinese)

[4] 何晓骁,姚呈康.人工智能等新技术在航空训练中的应用研究[J].航空科学技术,2020,31(10):7-11.
He Xiaoxiao, Yao Chengkang. Research on the application of new technologies such as artificial intelligence in aviation training[J]. Aeronautical Science & Technology, 2020,31(10):7-11. (in Chinese)

[5] 邓伟伟,段朝阳.人工智能在导弹控制系统中的应用[J].航空科学技术,2020,31(10):30-35.
Deng Weiwei, Duan Chaoyang. Application of artificial intelligence in missile control system[J]. Aeronautical Science & Technology,2020,31(10):30-35. (in Chinese)

[6] He Y，Pan Z，Li L，et al. Real-time vehicle detection from short-range aerial image with compressed mobileNet[C]//2019 International Conference on Robotics and Automation（ICRA），2019：8339-8345.

[7] 聂子一,张璞.轻量级神经网络与遥感图像目标检测[J].电子制作,2021(18):54-55,62.
Nie Ziyi, Zhang Pu. Lightweight neural network and remote sensing image target detection[J] .Electronic Production, 2021 (18): 54-55, 62. (in Chinese)

[8] 任彬,王宇庆,丛振,等.基于MPSOC的航空图像目标检测系统设计[J].液晶与显示,2021,36(7):1006-1017.
Ren Bin, Wang Yuqing, Cong Zhen, et al. Design of aerial image target detection system based on MPSOC[J]. Liquid Crystal and Display, 2021,36(7):1006-1017. (in Chinese)

[9] Ichim L，Popescu D. Road detection and segmentation from aerial images using a CNN based system[C]//2018 41st International Conference on Telecommunications and Signal Processing（TSP），2018：1-5.

[10] Winograd S. Arithmetic complexity of computations[J]. Society for Industrial & Applied Mathematics Philadelphia，1980，43(2)：625-633.

[11] Liang Y，Lu L，Xiao Q，et al. Evaluating fast algorithms for convolutional neural networks on FPGAs[C]//IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2020，9(4)：857-870.

[12] Yuan T，Liu W，Han J，et al. High performance CNN accelerators based on hardware and algorithm co-optimization [C]//IEEE Transac-tions on Circuits and Systems I：Regular Papers，2021，68(1)：250-263.

[13] Shen J，Huang Y，Wen M，et al. Toward an efficient deep pipelined template-based architecture for accelerating the entire 2D and 3D CNNs on FPGA[C]//IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2020，1442-1455.

Design of Neural Network Accelerator for Aeronautical Target Detection

Shi Lirui, Wang Shuaishuai, Xiao Hao

Hefei University of Technology, Hefei 230000, China

Abstract: Convolutional neural networks are widely used in the field of object detection in aerial images. However, in order to extract higher-level feature information of small size and arbitrary direction target in complex background environment, the structural complexity of the neural network model continues to increase, which makes the model computationally complex and time-consuming. So it is difficult to satisfy real-time requirements in aerial target detection. This paper proposes a neural network accelerator based on the Winograd algorithm for aerial target detection. The Winograd convolution algorithm greatly reduces the number of multiplications in the convolution calculation. However, there are many additional calculations in transforming input data and filter to Winograd domain in current Winograd convolution. To addresses this problem, a deep pipeline calculation structure of matrix transformation is proposed, which reduces the calculation amount of input transformation and output transformation by reusing the intermediate results of addition calculation and adjusting the operation order. At the same time, for the field programmable gate array (FPGA) implementation of the accelerator, an efficient data flow format and DSP array structure are proposed. The experimental results show that the accelerator proposed in this paper achieves a speed improvement of 32 times and 2.6 times compared to CPU and GPU respectively.

Key Words: target detection; convolutional neural network; accelerator; Winograd algorithm; FPGA

Received: 2022-01-06; **Revised:** 2022-02-11; **Accepted:** 2022-03-19

Foundation item: National Natural Science Foundation of China(61974039); Aeronautical Science Foundation of China (2018ZCP4003)