北京郵電大學
Beijing University of Posts and Telecommunications

2016110107-ZQ

硕士研究生学位论文阶段报告

学　　号：2016110107

姓　　名：张大旺

学　　院：信息与通信工程学院

专业(领域)：信息与通信工程

研究方向：多媒体与网络大数据

导师姓名：别志松

北京邮电大学

2018 年 10 月 25 日

论文题目	基于 TTA 的大型卷积神经网络处理器架构设计		
论文类型	应用研究	选题来源	其他
开题日期	2018-01-11	是否开题题目	否
论文开始日期	2018-01-11	报告日期	2018-10-31
报告地点	教三-818	报告时间	上午 10:30-11:00

研究内容简介

1.1 选题背景

基于神经网络的人工智能近年取得了突破性进展，正在深刻改变人类的生产和生活方式，是世界各国争相发展的战略制高点。

卷积神经网络(Convolutional Neural Network, CNN)，是神经网络的一种。由于卷积神经网络具有权值共享以及局部连接的特性，使得卷积神经网络的模型复杂度与参数数量大幅度降低。该优点在网络的输入是多维图像时表现得更明显，使图像可以直接作为网络的输入，避免了传统识别算法中复杂的特征提取和数据重建过程。近年来，卷积神经网络发展迅速，在图像处理以及自然语言处理领域都有着广泛的应用。

卷积神经网络作为实现人工智能任务的有效算法之一，已经在各种应用场景获得广泛的应用。从云端到移动端，不同应用场景也对神经网络的计算能力提出了不同的需求。卷积神经网络的广泛应用离不开核心计算芯片。目前的主流通用计算平台包括通用处理器 CPU 以及图形处理器 GPU，但是由于 CPU 的计算规模过小，以及 GPU 的功耗过高，所以发展神经网络的专用处理器的需求日益强烈。

目前许多人工智能的产品都是采用通用处理器或者专用集成电路（ASIC, Application Specific Integrated Circuit）的实现方式，前者虽然灵活性较好，但是在对实时性或者功耗要求较高的场合并不适合，后者对于某一种算法或者网络，这种实现方式在功耗和性能上可以做到最佳，但是现在的产品通常是多个算法集中在一个设备上，使得这种实现方式的设计成本和功耗猛增，设计周期变长，灵活性很差。

而采用专用指令集处理器（ASIP, Application Specific Instruction Set Processor）实现方式，由于专用指令集处理器是针对某一算法或领域进行裁剪和优化，以满足性能、面积、功耗等约束的处理器，所以在功能与性能之间取得了一个平衡点。

传输触发结构体系(TTA, Transport Triggered Architecture)的核心思想是利用数据传输来触发相应功能单元的具体操作。TTA 结构将寄存器单元也作为一种特殊的基本单元，它有效地减少了寄存器堆的设计压力，成为一种非常适合于专用处理器领域的处理器架构。此外，TTA 架构有着功能单元的灵活性以及可扩展性强等一系列优点，作为神经网络的处理器设计架构，也是非常有意义的。

1.2 研究内容

对于神经网络的应用而言，其实现方式目前主要有三种。第一种，采用通用处理器，如 CPU, GPU, DSP 等，通过软件编程的方式实现。这种方式具有很高的灵活性以及较短的上市时间，但由于通用处理器的设计是面向通用，具有高性能以及高灵活性，但是在一些对实时性以及功耗要求比较高的场合，这种实现方式并不合适。第二种，采用专用集成电路（ASIC, Application Specific Integrated Circuit）的实现方式，对于某一种网络或者算法，生成其固定的物理版图。对于某一种网络或者算法，这种实现方式可以在功耗以及性能上达到最佳，但是如果想使用多个网络或者算法，只能将这些网络或者算法的物理版图独立的集中在一个设备上，使得这种实现方式的设计成本和功耗猛增，设计周期变长，灵活性很差。第三种，采用专用指令集处理器（ASIP, Application Specific Instruction Set Processor）实现，由于专用指令集处理器是针对于某一算法或领域进行裁剪和优化以满足性能、面积、功耗等约束的处理器。因此，它既具有 ASIC 的高性能又具有通用处理器的灵活性，同时还能够有效地缩短设计周期，降低设计风险。

随着 ASIP 技术的发展，其设计流程也产生了很多变化，但是大体上可以分为 5 个步骤：应

用需求分析、体系结构选择、指令集设计、代码综合和硬件综合，其中体系结构选择在整个 ASIP 设计过程中至关重要，将直接影响到系统的性能。目前，主流的体系结构有如下几种。

第一种是复杂指令集结构体系（CISC, Complex Instruction Set Computer），CISC 结构采用微码状态机进行设计，一条汇编指令通常包含若干条微码指令，因此，CISC 结构的一条汇编指令可以执行复杂的功能，具有很高的执行效率，但这也使得 CISC 结构的硬件设计变得十分复杂。

第二种是精简指令集结构体系（RISC, Reduced Instruction Set Computer），在 20 世纪七八十年代，研究人员通过对大量应用程序进行分析发现，CISC 指令集中只有 20%的指令使用频率最大，约占运行时间的 80%，针对这种情况，人们研究出了 RISC 结构。RISC 结构指令集只包含那些使用频率最大的指令，其他指令则通过这些指令编程实现，RISC 结构的显著特点就是硬件结构简单，开发周期短。

第三种是超标量结构体系（Superscalar），RISC 结构虽然硬件实现简单，但是运行效率不高，为了增加运行效率，必须开发指令级并行性，Superscalar 结构系统应运而生。Superscalar 结构体系可以同时执行多条指令，采用硬件的方式检测同时执行的指令间的相关性，以保证程序正确无误地运行。由于采用硬件的方式检测指令间的相关性，显著加大了硬件开销，因此只有在代码兼容性问题成为首要考虑因素时，才会选择这种结构体系进行设计。

第四种是超长指令字结构体系（VLIW, Very Long Instruction Word），VLIW 结构是通过编译器调度，将数据不相关的若干条指令打包成一条长指令执行，从而实现指令级并行性，显然，这种方式的硬件开销相比 Superscalar 结构大大降低，但是，具有指令调度功能的编译器设计成密度也成为一个不可忽略的问题。

第五种是传输触发结构体系（TTA, Transport Triggered Architecture），TTA 结构由 Corporaal 等人提出，其核心思想是利用数据传输来触发相应功能单元的具体操作来触发相应功能单元的具体操作。TTA 架构可以看成 VLIW 的一个超集，我们把 VLIW 看成 SIMO(单指令多操作)类型的体系结构，那么 TTA 则是 SIMT(单指令多传输)类型的体系结构。TTA 相比于 VLIW，将寄存器单元作为一个功能单元（function units），解决了 VLIW 读写寄存器带宽的瓶颈问题，同时采用触发结构，解决了 VLIW 的功能单元之间互联过于复杂的问题。

由于神经网络具有内存密集（memory intensive）的特性，因此采用 VLIW 并不合适，而采用 TTA 架构则可以缓解这一问题。因此本课题最终选择采用 TTA 架构来进行神经网络处理器的设计。

基于 TTA 结构的专用处理器设计主要体现在以下三方面的设计：

1) 指令集的设计。
2) 功能单元的设计。
3) 数据交换网络的设计。

由于 TTA 架构的指令格式统一，只有一种 MOVE 格式，因此难点以及重点在后两个方面，即如何设计针对于神经网络的专用功能单元以及如何设计数据交换网络。

对于功能单元，需要针对神经网络的架构进行单独设计，满足神经网络运算的通用性以及完备性。功能单元的设计的目的是计算代码中运算最为密集的一些操作，从而提升处理器的性能。因此，程序中包含的主要运算操作及数量决定了 TTA 结构中功能单元的种类及数量。根据操作的类型,设计者可以很快确定需要哪些功能单元。根据某种操作占总操作的百分比,设计者可以很快确定需要使用该类型的功能单元的数量。对于寄存器文件这种特殊的功能单元，需要分析其它功能单元需要的存取带宽，从而确定需要多少通用寄存器。在保证性能的前提下，尽可能的节约硬件开销，由此确定处理器寄存器文件的大小与数量。

通常情况下，数据交换网络会成为整个处理器的关键路径，因此，数据交换网络的设计是整个 TTA 处理器设计的重点，数据交换网络的数据传输速度将直接影响到整个处理器的处理性能。数据交互网络包括总线与接口，主要负责将不同的功能单元，寄存器单元联系起来。然而并不是每个功能单元和寄存器单元的输入输出都要连接到每条总线上。这样不仅会增加面积，而且增加输入输出接口电路上的扇出，降低信号的品质，使性能下降。所以在满足性能需求的前提下，可以减少输入输出接口的数目以及连接的总线数量。这也使得数据交换网络的设计的难度大大增加。

而对于本课题而言，由于先前学者所提出的结构对于大型卷积神经网络无法适用，因此需要

额外提出一种可适用于大型卷积神经网络的架构。在传统的实现架构中，对于中间结果以及参数的存储，也主要分为两种做法。一种做法是实现如 LeNet 之类的小型网络，将中间运算结果以及网络参数都放在片内存储上，以减少片外的访存带宽。而这种做法在随着卷积神经网络的模型复杂度越来越高的情况下，由于 FPGA 片内存储资源的限制，并不能满足加速大型网络的需求。另一种做法是将参数放入片外存储，设置输入缓存与输出缓存，计算前将输入特征图传输至输入缓存，计算完毕后将结果从输出缓存写入片外存储，从而降低片外访存。但在大型卷积运算中，有时片内资源并不足以存放中间结果。

在经过调研以后，对比多个 TTA 架构的开发工具集，最终选择了 TTA 协同设计环境(TTA-based Co-Design Environment, TCE)作为本课题的开发工具。TCE 是是芬兰的坦佩雷科技大学（Tampere University of Technology）研发的一个面向 TTA 处理器的架构设计的工具集。TCE 提供了半自动的处理器设计流程，支持设计空间探测。TTA 协同设计环境设计、执行和验证为一体，提供了编译器和指令集仿真器等一些软件工具，为设计过程中的设计空间探测提供了极大的便利。

本课题将使用 TCE 工具集，面向 TTA 处理器架构，设计出一套神经网络处理器的功能单元与数据交换网络结构，旨在满足灵活性的同时，在性能上也达到一定要求。具体开发步骤如下：

a) 使用高级语言写出串行运行代码。

b) 使用 front-end 编译器编译出串行的 MOVE 指令代码。使用一个冗余度比较大的系统架构，由仿真器仿真，得到性能文件。

c) 分析源代码，设计或更改功能单元；分析性能文件，更改系统架构或互联网络。

d) 使用功能单元代替源代码中的操作，使用 back-end 编译，得到性能文件。

e) 重复步骤 3 与 4，直到找到满足需求的功能单元与系统架构的设计。

1.3 关键技术

1.3.1. 功能单元的设计。

功能单元的设计是本课题最基础与最重要的问题之一，功能单元设计的好坏将直接影响到并行度以及数据交换网络的复杂度。如何根据神经网络设计出通用、高效的功能单元，是本课题的难点之一。通用性指的是，对任意规模的网络层，该功能单元都可以通过与其它功能单元互联来进行实现；高效性指的是，功能单元的利用率需要达到一定的值，以免资源面积的浪费。功能单元的设计主要包括输入缓存区的设计、卷积操作功能单元设计、池化操作功能单元设计、激活函数功能单元设计、全连接层功能单元设计以及输出缓存区的设计。

1.3.2. 数据交换网络设计。

数据交换网络提供处理器中各个单元交换数据的通道，它包含两种基本模块，Socket 与总线。除了提供数据交换功能以外，总线还用于传输控制信号，比如源和目标寄存器的 ID，功能单元锁存信号等。Socket 提供了功能单元和寄存器文件与总线的连接，每个 Socket 可以连接到一条或多条总线以及某功能单元的一个或多个寄存器。每个 Socket 与每条总线都相连的方式成为全连接网络，它能简化总线的传输调度，但因为连接点会增加总线负载，延长全局周期时间，增加功耗，因此并不是一种高效的设计方式。尤其在神经网络这种大型网络中，采取全连接型的数据交换网络更不可取，因此如何设计高效的数据交换网络也是难点之一。

1.3.3. 实现大规模卷积神经网络的架构设计

对于传统的卷积神经网络加速架构，其因为片内存储资源的限制，无法实现对大型卷积神经网络的加速。传统加速架构主要有两种。第一种做法是实现如 LeNet 之类的小型网络，由于网络非常小，可以将中间运算结果以及网络参数都放在片内存储上，以减少片外的访存带宽。而这种做法在随着卷积神经网络的模型复杂度越来越高的情况下，很快就不能满足需求。第二种做法是将参数放入片外存储，设置输入缓存与输出缓存，计算前将输入特征图传输至输入缓存，计算完毕后将结果从输出缓存写入片外存储，从而降低片外访存。但在大型卷积运算中，有时片内资源并不足以存放中间结果，这种结构对于如 VGGNet 之类的神经网络也无法适用。因此，提出片内存储资源的使用优化方案以及针对于大型卷积神经网络的架构成为一个专用处理器的必须考虑点。

1.4 论文计划

2017/12-2018/02 对 TTA 架构与卷积神经网络基础知识进行调研，能够对 TTA 架构与神经网络有基本认识。

2018/02-2018/05 熟悉 TCE 工具集，并能使用 TCE 工具集设计一些简单功能的项目。
2018/05-2018/07 专用功能单元的设计，针对各个层能够设计出相应的专用功能单元。
2018/08-2018/11 针对数据交换网络进行设计，进行优化
2018/11-2019/12 对不同网络进行实现可以实现不同网络
2018/12-2019/1 论文撰写与修订完成论文的撰写

1.5 论文进度及目标

1.5.1 论文进度

2017/12-2018/02 已通过许多论文对 TTA 架构以及卷积神经网络有了较深层的了解。

2018/02-2018/05 通过 TCE 工具集官网及其附带的一些示例，对 TCE 工具集的使用有了系统的认识，并使用 TCE 工具集设计了如 CRC 校验等较简单的小项目。

2018/05-2018/07 针对各个层进行了功能单元的设计，其中，对卷积操作的功能单元进行了优化。相比原先的卷积功能单元，在卷积操作步长大于 1 时，新功能单元不会产生计算资源的浪费。

2018/08-2018/11 针对数据交换网络进行了优化。总分为三部分：输入缓存区、计算区、输出缓存区。计算区之间的功能单元互不相连，大大减小了交换网络的复杂度。

1.5.2 论文目标

本课题的研究目标是设计一个基于 TTA 架构的大型卷积神经网络处理器。其共有两个目标。

第一个目标是基于 TTA 架构的卷积神经网络处理器设计，这一目标旨在以专用处理器的形式对卷积神经网络进行加速，从而完成应用灵活性的要求。这一目标包括专用功能单元的设计以及数据交换网络的优化，通过少数的配置信息就可以完成所需要的卷积神经网络的加速，并且在性能上可以满足需求。

第二个目标是大型卷积神经网络处理器的架构设计，其旨在对于任意规模的卷积神经网络，都能够进行加速。其背景是现有的卷积神经网络规模愈加庞大，而 FPGA 资源与成本成正比，若需要加速较大规模的卷积神经网络，所需要的 FPGA 成本也随之增加。这一目标旨在可以通过低成本 FPGA 来加速大规模的卷积神经网络，从而节省成本。

论文进展情况

2.1 工作计划

2017/12-2018/02 对 TTA 架构与卷积神经网络基础知识进行调研，能够对 TTA 架构与神经网络有基本认识。

2018/02-2018/05 熟悉 TCE 工具集，并能使用 TCE 工具集设计一些简单功能的项目。

2018/05-2018/07 专用功能单元的设计，针对各个层能够设计出相应的专用功能单元。

2018/08-2018/11 针对数据交换网络进行设计，进行优化

2018/11-2019/12 对不同网络进行实现可以实现不同网络

2018/12-2019/1 论文撰写与修订完成论文的撰写

2.2 实际进展情况

2017/12-2018/02 已通过许多论文对 TTA 架构以及卷积神经网络有了较深层的了解。

2018/02-2018/05 通过 TCE 工具集官网及其附带的一些示例，对 TCE 工具集的使用有了系统的认识，并使用 TCE 工具集设计了如 CRC 校验等较简单的小项目。

2018/05-2018/07 针对各个层进行了功能单元的设计，其中，对卷积操作的功能单元进行了优化。相比原先的卷积功能单元，在卷积操作步长大于 1 时，新功能单元不会产生计算资源的浪费。

2018/08-2018/11 针对数据交换网络进行了优化。总分为三部分：输入缓存区、计算区、输出缓存区。计算区之间的功能单元互不相连，大大减小了交换网络的复杂度。

工作成果

3.1 目前已完成学位论文工作的内容

目前已完成学位论文工作的内容包括：功能单元的设计与实现、互联网络的设计、大型卷积神经网络的加速方案设计。

功能单元的设计主要包括输入缓存区的设计、卷积操作功能单元设计、池化操作功能单元设计、激活函数功能单元设计、全连接层功能单元设计以及输出缓存区的设计。为了解决在卷积运算步长大于 1 时，传统卷积运算功能单元存在计算资源严重浪费的情况，本课题提出了一种旋转存储的数据存储方式。输入缓存区包括 9 个小存储器，对输入缓存区的操作包括旋转存储、旋转读取、顺序存储、顺序读取。其中旋转存储与旋转读取对应于卷积运算，顺序存储与顺序读取对应于全连接层运算。这样卷积操作与全连接层的功能单元可以进行复用，从而减少计算资源的使用。同时，由于池化操作与卷积操作的相似性，池化操作也可以使用卷积操作的功能单元来进行计算，因此，本设计大大的减少了功能单元的复杂性。最终所需要的功能单元为输入缓存区功能单元、通用计算功能单元、激活函数功能单元、输出缓存区功能单元。

互联网络是连接各功能单元的数据传输网络。在本课题设计中，由于并行加速的原因，通用计算功能单元的数量较多，而通用计算功能单元之间并不需要数据互联。因此互联网络的设计基于将功能单元分为三部分，第一部分为输入缓存区功能单元，第二部分为通用计算功能单元与激活函数功能单元，统称计算功能单元，第三部分为输出缓存区功能单元。其中输入缓存区与计算功能单元之间实现互联，但计算功能单元之间不进行互联，输出缓存区与计算功能单元之间实现

互联。该结构总体为两个二部图，可以大大减少互联网络的复杂度。

大型卷积神经网络的加速方案设计包括两个部分：片内存储优化结构与分块计算技术。其中片内存储优化结构为利用数据的复用性，只存储输入特征图或者输出特征图，从而可以减少一半的存储空间。分块计算技术为将大型卷积操作分解为多个小型卷积操作，从而在片内存储资源有限的情况下，可以通过实现多个小型卷积操作的加速，最终组合形成大型卷积操作的加速。

3.2 取得的阶段性成果

目前所取得的阶段性成果为已发表学术论文《A High Performance Framework for Large-scale 2D Convolution Operation on FPGA》，该论文已在 The 15th International Symposium on Pervasive Systems, Algorithms and Networks 会议进行发表。该论文将被 EI 全文收录。

该学术论文主要包括两个部分，第一部分是卷积运算功能单元的重设计，第二部分是针对大型卷积操作的加速方案。卷积运算功能单元的重设计背景在于，传统的卷积运算功能单元在卷积操作步长大于 1 的情况下，存在非常严重的计算资源浪费。本论文提出了一种旋转存储功能单元，其基本思想为，使用旋转存储的数据存储方式，可以同时读出卷积运算窗口的所有数字，从而避免传统功能单元的数据等待。针对大型卷积操作的加速方案包括片内存储优化结构与分块计算，前者的目的在于减少片内存储资源的使用，后者的目的在于将大型卷积操作分解为多个小型卷积操作，从而在片内资源有限的情况下，可以通过对小型卷及操作的加速来进行大型卷积操作的加速。

3.3 主要创新点

1. 基于 TTA 架构实现了卷积神经网络的可配置方案。相比传统 ASIC 方式，大大提升了灵活性，在实际的应用中可大幅度减少硬件方案的开发周期，从而减少开发成本以及开发时间。

2. 针对于大型卷积神经网络，提出了计算框架与优化方案，使大型卷积神经网络能够在小型 FPGA 上进行加速。此创新点的立足点在于使用低成本的 FPGA 进行大型卷积神经网络的加速，减少硬件成本。

3. 功能单元的重设计。提出了旋转存储卷积计算单元，相比传统 Z 型卷积计算单元，在卷积运算步长大于 1 时，可大幅度减少运算时间。此创新点的出发点在于减少计算时间，从而使卷积神经网络的运算更加满足实时性。

4. 提出了一种通用计算单元，可同时适用于卷积层、池化层、全连接层。减少了计算资源的使用。此创新点的立足点在于在相同资源的情况下，可以进行更多倍数的加速，从而在实时性上能够满足需求。

计划及进度安排

经过将近一年的工作，完成选题，研究方案确立。各设计方案，包括功能单元的设计、TTA 架构的数据互联网络的设计、大型卷积神经网络的加速方案设计都已经确定。其中功能单元的设计已经完成，卷积运算、池化运算、全连接层运算共用一套功能单元，大大减少了资源的消耗。

下一步的工作为在 TCE 工具集上实现数据互联网络、将 TCE 工具集生成的 HDL 源码移植到硬件工程上进行硬件实现以及论文撰写。预计时间分别为：

使用 TCE 工具集实现数据互联网络：15 天

移植到硬件工程：15 天

论文撰写：15 天

问题及整改方案

论文中所遇到的问题： TCE 工具集中没有对片外存储 DDR 的使用设计，而大型卷积神经网络的加速方案需要用到片外存储。

整改方案：设计了单独的功能单元来实现 DDR 的使用，该功能单元简单模拟了 DDR 的使用，操作包括输入读指令、读取数据状态、读取数据、写入数据。在具体的 HDL 语言描述中将使用 DDR 的 IP 核来进行代替。由于 DDR 的 IP 核无法保证数据读取的时延，因此在 DDR 后加上一个 FIFO 来进行数据的缓冲，利用 FIFO 的状态信号，在 FIFO 不空的时候，从 FIFO 中进行数据的读取，可以使数据的读取有固定的 1 个时钟的延迟，从而在 TCE 工具中可以使用。

参考文献

[1] Han S, Kang J, Mao H, et al. ESE: Efficient Speech Recognition Engine with Sparse LSTM on FPGA[J]. 2017.

[2] 刘俊, 谢憬, 王琴. 基于 TTA 技术的专用处理器设计[J]. 微电子学与计算机, 2009, 26(11):161-164.

[3] 李杰. 基于 TTA 架构的 ASIP 设计与应用[D]. 西安电子科技大学, 2014.

[4] 徐争莉. 基于 TTA 的 LTE 符号级处理过程的研究[D]. 北京邮电大学, 2013.

[5] Teittinen J, Hiienkari M, Žliobaitė I, et al. A 5.3 pJ/op approximate TTA VLIW tailored for machine learning[J]. Microelectronics Journal, 2017, 61:106-113.

[6] Li H, Fan X, Jiao L, et al. A high performance FPGA-based accelerator for large-scale convolutional neural networks[C]// International Conference on Field Programmable Logic and Applications. IEEE, 2016:1-9.

[7] 朱学亮, 柴志雷, 钟传杰,等. 基于 FPGA 的图像卷积 IP 核的设计与实现[J]. 微电子学与计算机, 2011, 28(6):188-192.

[8] 方睿, 刘加贺, 薛志辉,等. 卷积神经网络的 FPGA 并行加速方案设计[J]. 计算机工程与应用, 2015, 51(8):32-36.

[9] Yue H, Shen L, Dai K, et al. A TTA-Based ASIP Design Methodology for Embedded Systems[J]. Journal of Computer Research & Development, 2006, 43(4):752-758.

[10] 陆志坚. 基于 FPGA 的卷积神经网络并行结构研究[D]. 哈尔滨工程大学, 2013.

[11] 朱礼波. 基于 TTA 技术的多功能可配置 DSP 处理器设计[D]. 上海交通大学, 2008.

[12] Bouvrie J. Notes on Convolutional Neural Networks[J]. Neural Nets, 2006.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS, 2012, pp. 1097－1105.

[14] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[15] Qiu J, Wang J, Yao S, et al. Going Deeper with Embedded FPGA Platform for Convolutional Neural Network[C]// Acm/sigda International Symposium on Field-Programmable Gate Arrays. ACM, 2016:26-35.

[16] Chen Y, Sun N, Temam O, et al. DaDianNao: A Machine-Learning Supercomputer[C]// Ieee/acm International Symposium on Microarchitecture. IEEE, 2015:609-622.

评审小组

姓 名	职 称	职务	工 作 单 位
别志松	副教授	组长	北京邮电大学
龚萍	副教授	成员	北京邮电大学
李永华	副教授	成员	北京邮电大学
林雪红	副教授	成员	北京邮电大学
王思野	讲师	成员	北京邮电大学

导师评语

张大旺同学自开题以来，针对基于 TTA 的大型神经网络处理器架构设计这一问题，开展了深入的研究工作，已完成学位论文工作的大部分工作，内容包括：功能单元的设计与实现、互联网络的设计、大型卷积神经网络的加速方案设计。 目前所取得的阶段性成果为已发表会议论文《A High Performance Framework for Large-scale 2D Convolution Operation on FPGA》。

主要工作在一些方面取得了创新成果，总体进度符合要求，遗留问题解决方案具体合理，能够按期完成学位论文，可以通过论文阶段检查。

导师：

日期：　　年　　月　　日

阶段报告小组意见：

负责人：

日期：　　年　　月　　日

学院意见：

负责人：

日期：　　年　　月　　日　（签章）