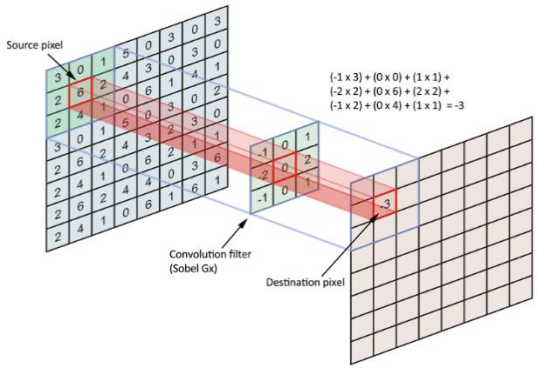
《计算机组成原理》大作业

一、相关知识准备：

1. 学习 MIPS 汇编语言，结合课上所学，熟练掌握相关 MIPS 汇编知识。
2. 通过自己阅读参考文献，理解神经网络推理的过程。
3. 阅读 daBNN 及其源代码，结合其他相关神经网络论文，了解如何实现神经网络卷积层的计算，并学习对其进行优化的思想，包括减少内存访问、重复利用寄存器等操作。
4. 了解二值神经网络与普通卷积神经网络在卷积层计算上的差别，并分析其带来的好处。

二、课程设计内容：

1. 使用 MIPS 汇编和 MIPS 仿真器，设计并实现一个普通整数卷积计算算子，要求有完整的输入输出，输入为 7*7*1 格式的张量，对应的卷积核一个，尺寸为 3*3*1，步长为 1，输出为经过卷积计算后的对应的 5*5*1 张量，要求计算结果正确。参考卷积操作图：



2. 设计并实现一个二值卷积计算算子，要求有完整的输入输出，输入为 7*7*16bit 格式的张量，对应的卷积核一个，尺寸为 3*3*16bit，输出为经过卷积计算后的对应的 5*5*1 的张量，要求计算结果正确。参考文献：论文 1、论文 2
3. 在 2 的基础上，将卷积层的偏置项、BN 层整合到同一层内。参考文献：论文 2 （3,4 二选一）
4. 在 3 的基础上，对二值卷积计算算子进行寄存器复用优化，参考文献：论文 1，论文 1 已经开源基于 ARM 的 BNN 代码，可以参照学习 （3,4 二选一）

三、要求

1. 完成课程设计报告，阐述神经网络前向推理引擎以及优化原理，描述汇编代码实现过程，以及展示实验结果
2. 提交基于 MIPS 的代码，并能正确在 MARs 上运行并输出结果（结果输出在一个文件里）
3. 提交时间：12 月 12 日

四、参考内容：

1. 论文 1：daBNN: A Super Fast Inference Framework for Binary Neural Networks on ARM devices，https://arxiv.org/abs/1908.05858
2. 代码 1：https://github.com/JDAI-CV/dabnn/blob/master/dabnn
3. 论文 2: PhoneBit: Efficient GPU-Accelerated Binary Neural Network Inference Engine for Mobile Phones, https://ieeexplore.ieee.org/document/9116236（二值神经网络参考）