2024～2025学年度第二学期《智能芯片设计》期末考查题

要求：

（1）课外完成，2025年6月30日17：00前提交给任课教师，逾期不接收。（2）第1-5题，需将答案手写在A4打印纸上（不用抄题）；第6题是报告，需按照模板要求撰写然后打印纸质版提交。

专业班级：_____学号：_____姓名：_____

题号	1	2	3	4	5	6	总分
分数							

1、（10分）对于一个单通道输入特征图（5×5×1，如图(a)），卷积神经网络包含3通道3×3卷积核W（3×3×3，如图(b)），设定卷积步长为1、边界补0、扩展长度为1，经过2×2最大池化。给定一个Transformer的自注意力层查询矩阵Q、键矩阵K、值矩阵V（序列长度为3，长度d=2，如图(c)）。

(1) 神经网络中卷积核和Transformer的作用是什么？

(2) 根据输入特征图和卷积核，通过直接卷积计算输出特征图，请写出计算过程。

(3) 根据输入特征图和卷积核，通过矩阵向量乘计算输出特征图，请写出计算过程。

(4) 观察上述计算得到的输出特征图，请问各个卷积核各提取了何种特征？

(5) 根据给定的Transformer自注意层输入计算自注意力输出，请写出计算过程。

5	5	5	5	5
4	4	4	4	4
0	0	0	0	0
2	2	2	2	2
1	1	1	1	1

(a) 单通道输入特征图

1	0	-1
1	0	-1
1	0	-1

3	3	0
3	0	-3
0	-3	-3

2	2	2
0	0	0
-2	-2	-2

(b) 卷积核

Q		K		V	
1	0	1	1	1	2
0	1	1	0	2	1
1	0	0	1	3	2

(c) Transformer自注意输入

2、（10 分）假设某卷积神经网络中权重数据符合单峰分布，其期望为 1.0，方差为 0.09。

(1) 试给出权重数据的 8 bit 位宽（有符号数）仿射映射量化的公式，使量化数据覆盖 95.45%左右（即 2-sigma 法则）的原权重数据。

(2) 使用上述量化公式，计算权重 d=1.2 的量级整数，并给出二进制形式。

(3) 写出单精度浮点数 0 1000 0001 10110000000000000000000 的十进制数值表示，并给出 8 bit 动态定点数（f=4）的表示形式。

3、（10 分）现有一卷积神经网络，其卷积层部分结构如下： 16 通道 64*64 输入特征图，先通过一常规卷积层，输出 32 通道 62*62 特征图，再输入一深度可分离卷积层，最终输出 16 通道 58*58 特征图， 请回答以下问题：

(1) 求该神经网络卷积层部分的权重个数。

(2) 说明深度卷积与逐点卷积 1）在网络轻量化方面的作用； 2）在信息流动方面的区别； 3）为何要串联二者构成深度可分离卷积？

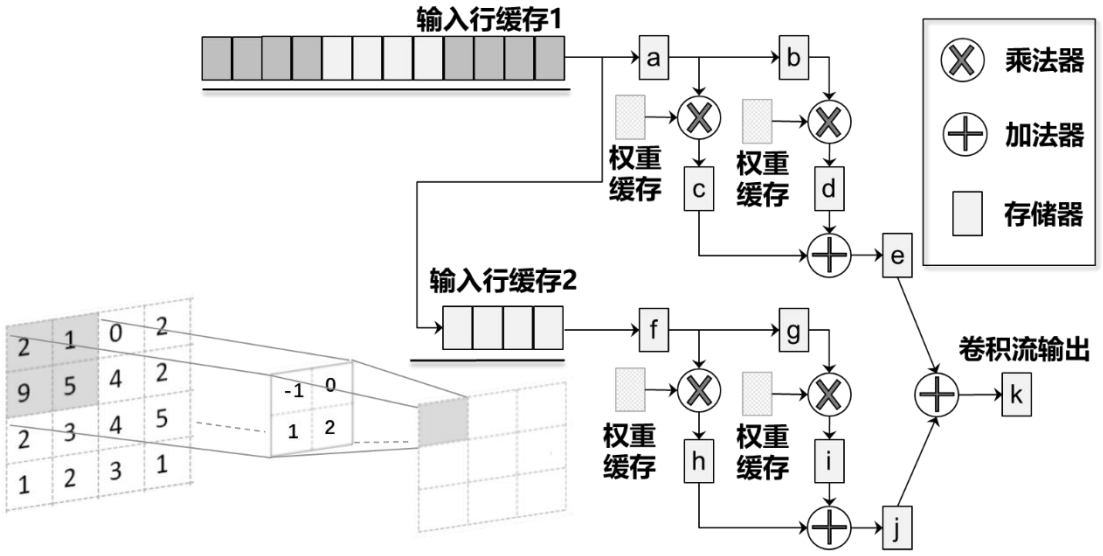
4、（10 分）假设有 M 个尺寸×通道数=K×K×N=2×2×28 的卷积核，输入特征图尺寸×通道数=H×L×N=4×4×28，步长 s 为 1，对应得到的输出特征图尺寸×通道数=R×C×M。下图实现了其中一层通道的权重复用，假设在该架构中所有硬件及操作均只消耗一个时间单位，并且在时刻 0 时输入行缓存 1、输入行缓存 2、权重缓存加载好值。

(1) 若计算核心每次只能加载 Tm 个卷积核及 Tn 层通道计算得到部分输出特征图 Tr 及 Tc，写出权重复用数据加载顺序的伪代码。

(2) 求时刻 0 时输入行缓存 1、输入行缓存 2、权重缓存内每个寄存器所存的数值？可在图中直接写出。

(3) 计算得出时刻 5 时存储器 k 中的值？计算得出在第几个时刻完成下图中完整输入特征图的卷积？

(4) 若 M=12，Tm=3，Tn=4，求输出特征图在计算核心内被完整读取的次数？



5、（10 分）异构 DSA（Domain Specific Architecture）是当前智能芯片的主流架构，请从计算并行形式、内存层次结构、数据精度以及特定领域语言的角度阐述其在智能计算中相对于冯诺依曼架构效率更高、能耗更低的原因，以及软硬件协同的重要性。

6、（50 分）报告：请围绕智能芯片“高效能、低功耗” 这一关键目标，重点从数据复用和高效加速计算的角度来阐述其核心内涵，从数据复用技术、高效并行架构、模型轻量化设计等层面，梳理当前智能芯片的发展脉络，并展望未来的发展方向。报告要求以图文并茂的形式呈现，并体现自己独立思考的内容（按照提供的模板格式编排， 5-6 页纸，标题、内容等自行编排）。