智能计算系统 2022autumn

2.19开卷考试的回忆版本，仅供参考

填空题（35分，一空一分）

注：（number）表示此处填空

注1：此处第7、9题的回忆很模糊，表述并不十分接近原题

1. AI技术的背后除了AI算法，还需要(1), (2), (3)的支撑
2. AI模型的部署方式分为(1), (2), (3)三种
3. AI模型推理应用中，Host内存指的是(1)的内存，Device侧内存指的是 (2)的内存。在AI推理模型的执行过程中，Host和Device之间的数据交流至少需要(3)次
4. Tensorflow、Pytorch等知名AI深度学习框架都采用了基于(1)的计算模型，并且普遍将(2)作为输入/输出参数的描述格式
5. AI框架在数据处理的过程采用batch操作的好处有(1), (2)
6. 面对着深度学习当前(1),(2),(3)的问题，分布式自动训练被广泛应用于各大AI框架中，常见的分布式自动训练模式包括(4), (5), (6)
7. 在模型训练过程中，可以采用(1),(2),(3)的优化技巧提高模型训练的效率
8. 为了隔离不同AI框架的不同架构并沟通机器硬件，我们采用(1)技巧进行优化。在图优化阶段，与硬件无关的优化(2)(3)，与硬件有关的优化有(4)(5).
9. 在计算图执行过程中，为了提高硬件利用率，降低Host和device之间数据搬移的开销，可以采用(1), (2)技巧对模型推理进行优化
10. 卷积神经网络中，卷积算子的计算特性是(1), 访存特性是(2)
11. 以昇腾处理器为例，昇腾处理器的核心是(1)，其中Cube的作用是(2)
12. TVM的核心思想是(1). TVM中调度主要针对(2)问题

问答题

1. 四层卷积神经网络（3*3conv（pad=0，stride=2，输出通道96）、2*2maxpool（pad=0, stride=2，输出通道32）、1*1conv（pad=0,stride=1，输出通道32)、全连接层（64个输出神经元）），当输入32*32*3的张量时，各层的输入、输出、参数量和计算量分别是多少（计算量只考虑乘法计算，且不需要考虑激活函数的计算）（10分）
2. Tensorflow、Pytorch、MindSpore等AI框架有哪两种执行模式，分别说说各自的优缺点和适用场景，如果可以，请设计一个可以混合使用两种执行模式的框架机制（15分）
3.
    1. 请说明AI深度学习框架的作用
    2. 请说出至少三点AI框架的特性
    3. 简述反向自动微分的实现方式
    4. 简述梯度消失现象的原因和解决方案（10分）
4.
    1. 请简述AI Core中Buffer和通用CPU cache的区别
    2. 在AI Core中使用Buffer，对AI Core的I/O性能有什么帮助
    3. 如果输入数据大小大于AI Core的缓冲区大小，请给出对应算子的设计和调度方案（10分）
5. 现在的大模型面临着哪些挑战和问题，对于每种问题有什么解决方案（20分）（本题是开放题，PPT上有问题，但没有解决方案，要自己写）