深度学习框架Caffe学习与应用 第13课

本节课内容

- 1. GPU下使用Caffe
- 2. GPU下做深度学习的硬件知识
- 3. 卷积神经网络的基本内存需求
- 4. 减少内存占用技术

1. GPU下使用Caffe

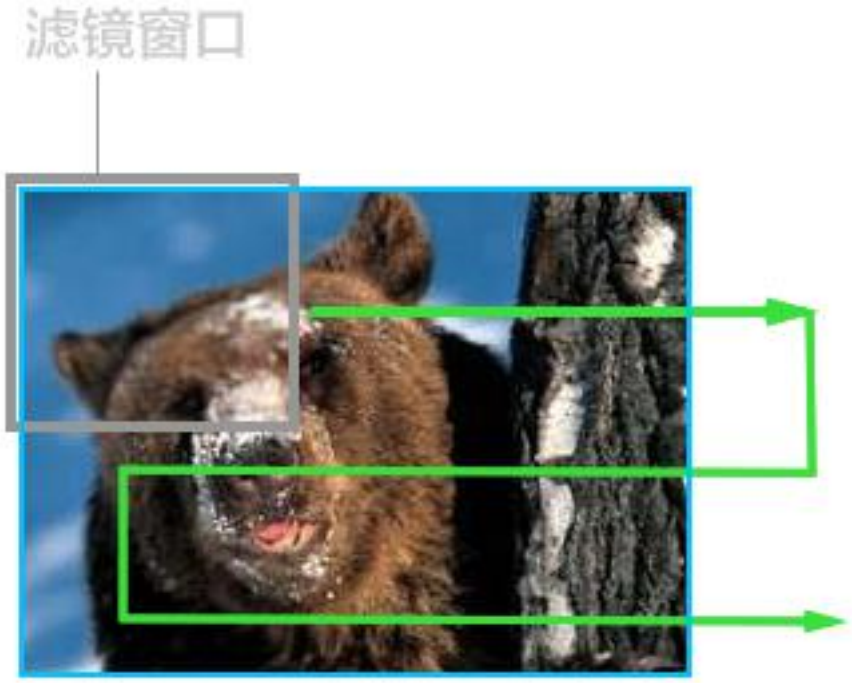
- 1. Caffe对GPU有很好的支持，源码中.cu后缀结尾的文件都是GPU下运行的文件代码。
- 2. 深度学习中主要是用cuDNN做卷积，所以，几乎所有的深度学习框架在GPU下都需要依赖cnDNN(CUDA下的DNN库)
- 3. 安装Caffe前要先安装好CUDA驱动和cnDNN库，安装Caffe时，使用默认安装，不需要设定CPU_ONLY相关的参数。
- 参考资料：（在“资料”里面）
- AWS上选用GPU云主机安装GPU环境教程
- 本地安装GPU环境教程

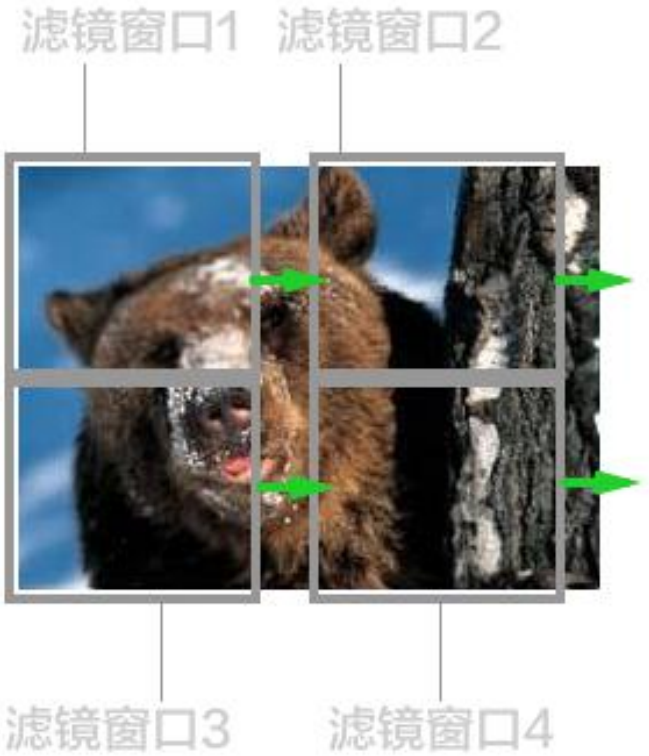
2.GPU下做深度学习的硬件知识

- **NVIDIA的优势：**
- 1. 有CUDA这一标准库（没适合AMD的OpenCL般强大的库）
- 2. N卡的GPU计算和GPGPU社区很大（很多开源CUDA解决方案，AMD很小）
- 3. NVIDIA公司押注深度学习（AMD没有信心）

- **是否需要多个GPU？**
- 在多GPU 上，神经网络难以进行有效的并行化。

为什么GPU比CPU更适合做深度学习？





- **NVIDIA的GPU的类型：**
- 1.主要面向3D游戏应用的GeForce系列，几个高端型号分别是GTX1080、Titan X和GTX980，分别采用最新的Pascal架构和Maxwell架构，因为面向游戏玩家，对双精度计算能力没有需求，单价相比采用相同架构的Tesla系列产品要便宜很多，也经常被用于机器学习。
- 2.面向专业图形工作站应用的Quadro系列，主要是针对CAD、3DMaxs、Maya这一类的设计软件做过驱动层的优化，所以采用相同架构的Quadro售价比GeForce高出许多。
- 3.专用GPU加速计算的Tesla系列，Tesla本是第一代产品的架构名称，后来演变成了这个系列产品的名称了，最新的第五代架构名为Pascal，对应的产品型号就是前面提到的P100。而采用前两代架构Kepler和Maxwell的产品目前也还在销售，分别对应K系列和M系列的产品，目前市面上常见的也就是K40/K80、M4/M40/M60等几个型号。K系列更适合用作HPC科学计算，M系列则更适合机器学习用途。

3.卷积神经网络的基本内存需求

- 1. 激活和误差占大部分
- 主要占用内存的是网络中的激活和误差，将它们的大小加起来，就可确定大概的内存需求（这里说得有点笼统，后面会详细来说）。但确定网络在某状态下的激活和误差的尺寸大小很难，一般而言，前几层网络会占用大量内存，即主要内存需求来自输入数据大小。

- 2.输入维度
- 维度在224×224×3，即224×224像素的3色通道图像。在ImageNet上得到当下最好结果至少需要12GB内存，而若是112×112×3维的数据集上只需4-6GB内存。另一方面，对于输入尺寸为25×75×75×3的视频数据来说，12GB内存可能就达不到很好结果。

- 3.训练样本的规模
- 若只取ImageNet的10%样本作训练，模型会很快过拟合，图像越少所需内存越少。

- 标签数量
- 若只建立2类的模型，相比对于1000类的分类模型来说，2类模型消耗的内存更少。同时，更少的分类意味着彼此区分更少（实质是参数越少），越容易出现过拟合。
- 有多个GPU反而不一定好，可能任务的数据量不大，训练出的模型反而容易过拟合，多个GPU在速度和大显存上并没优势，一个小显存的GPU也可达到好的效果。如果遇到内存不足，可以使用一些简单的减小内存占用的技术。

4. 减少内存占用技术

- 1 更大的stride
- 对卷积核使用更大的stride，达到减少输出数据的目的。消耗大量内存的输入层通常用这种方法。

- 2 使用1×1卷积核
- 96个1×1卷积核可使64×64×256的输入数据降为64×64×96。

- 3 池化
- 2×2的池化层将减少4层的数据量，从而大大减少后续层内存占用。

- 4 减少mini-batch大小
- size为64个样本的mini-batch比128个样本的batch减少一半的内存消耗。缺点就是训练时间会更长，大多数卷积运算对大小为64或者更大的batchsize做了优化。低至32个样本只作为最后策略。

- 5 改变数据类型
- 将数据类型由32位换位16位，可以额减半内存同时不会降低性能。如用在Tesla P100中可带来巨大提速。