哈爾濱工業大學(深圳)

HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

深度学习体系结构（实验）

实验二 YOLO算法量化加速

实验目的

- 了解利用YOLO算法进行目标检测的基本原理
- 了解**浮点数量化**的意义和原理，掌握浮点数的量化方法
- 掌握使用**HLS Directive**优化IP核性能的方法

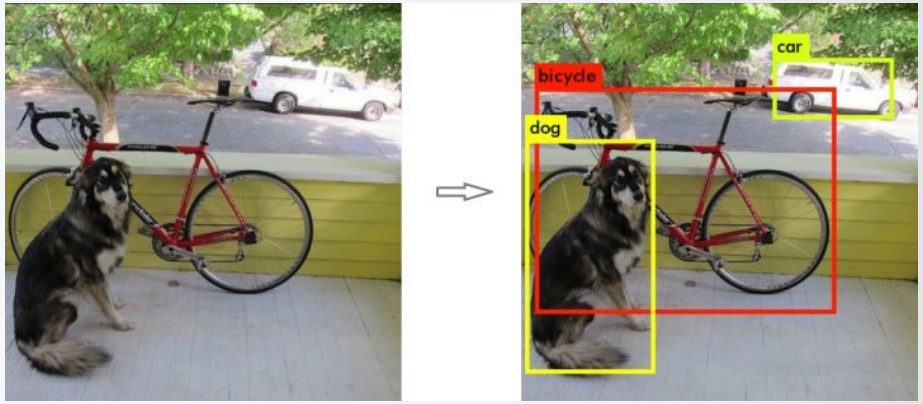
实验内容

- 运行量化前的Tiny YOLOv2，记录识别效果和运行时间
- 量化卷积层和FC层的网络参数，再次运行Tiny YOLOv2，并对比量化前后网络参数的大小变化和识别效果差异
- 使用HLS Directive优化卷积、池化IP核，对比分析优化前后的前向推导性能

实验原理 — YOLO算法简介

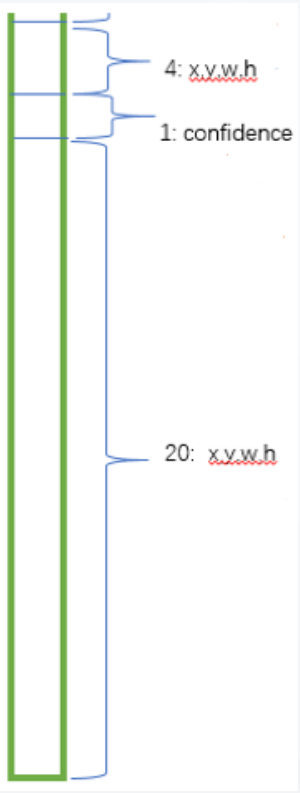
YOLO（You Only Look Once）

- 目标检测问题：在图片中定位和识别目标对象
- 传统算法分3步：
  ① 在图片上**确定**一些**候选区域**
  ② 对每个候选区域进行**特征提取**
  ③ 使用经过训练的**分类器**对候选区域进行分类
- 传统算法分多步走，耗时大，实时性差
- YOLO同时进行目标定位和分类，统一作为回归问题处理



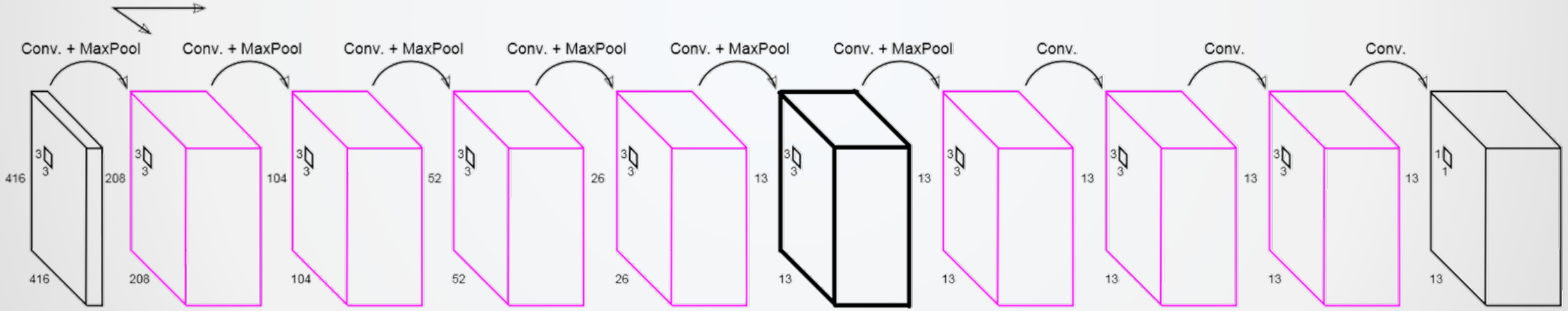
实验原理 — YOLOv2算法流程

- 预 处 理： 将输入图像归一化；以0.5为padding，按比例缩放至416×416
- 网络推导： 将预处理得到的416×416×3图像输入网络推导，得到13×13×5×25的输出张量
- 后 处 理： 使用NMS去除重复检测，将检测到的边框按比例放大到原图中



实验原理 — Tiny YOLOv2网络结构

- Tiny YOLOv2共15层：9个卷积层 + 6个最大池化层
  - 卷积用于提取特征
  - 池化用于缩小特征图、减小数据量



实验原理 — 量化原理与方法

Q1：为什么需要量化?

- 数据量大、设备的存储、带宽资源不足
  - 卷积产生大量的特征图数据
  - 网络层次越深，网络权值越多
- 加快推导速度，提高NN的实时性

Q2：如何量化?

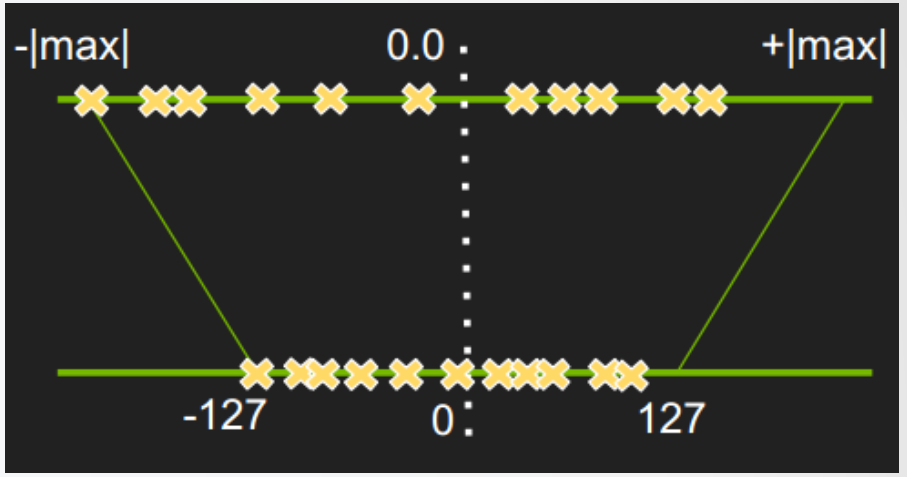
- 使用较低精度的定点数（整数）来表示浮点数
- 线性**对称**量化方法、线性**非对称**量化方法

实验原理 — 线性量化方法（以int8量化为例）

线性**对称**量化

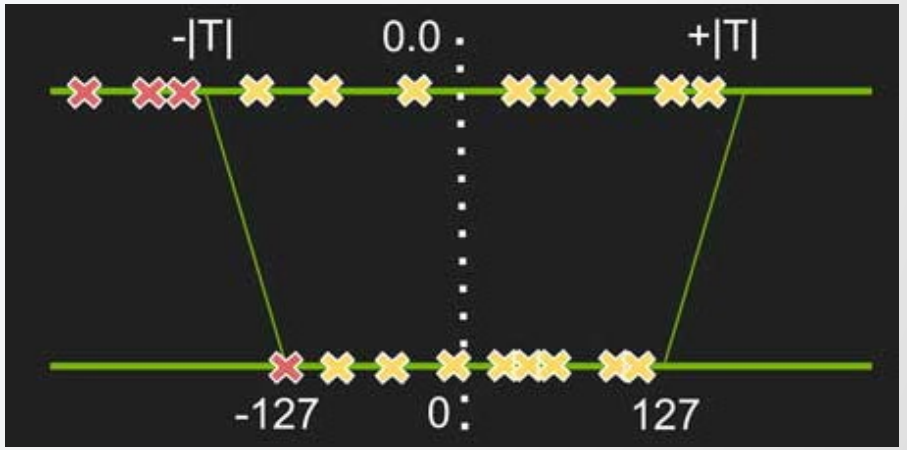
- 存储参数绝对值的最大值MAX
- 将[-MAX, MAX]线性映射到[-127, 127]
- 公式：

$$\begin{cases} MAX = \max_i\{|x_i|\} \\ y_i = \left\lfloor x_i \cdot \frac{127}{MAX} \right\rfloor \end{cases}$$



线性**非对称**量化

- 找到参数所在区间[a, b]的端点a和b
- 将[a, b]线性映射到[0, 255]
- 公式：$y_i = \left\lfloor 256 \cdot \frac{x_i - a}{b - a} \right\rfloor$



实验原理 — HLS优化方法

循环展开：牺牲资源换取速度

- 每次循环时，执行m次循环体
- 语法：#pragma UNROLL factor = <int>

流水线：时间重叠原理

- 语法：#pragma PIPELINE II = <int>

  #pragma PIPELINE

实验步骤

YOLO量化：

1. 运行未量化的Tiny YOLOv2
2. 编写量化算法
3. 修改卷积IP核
4. 更新Block Design，生成并导出Overlay
5. 上板测试

HLS优化：

1. 利用HLS Directive优化卷积、池化IP核
2. 更新Block Design，生成并导出Overlay
3. 上板测试

验收与提交

- 检查量化YOLO的实验现象（**2**分）
- 检查HLS优化的实验现象（**2**分）
- 将源码、运行结果、实验报告打包提交（**4**分）
  - 命名规则：**学号_姓名_DLA实验2.zip**
  - 提交方法：https://hitsz-cslab.gitee.io/dla/ojguide
  - DDL：下周同一上课时间前
- 附加题（+**2**分）：
  - 检查实验现象、回答问题并提交源码、运行结果和报告

哈爾濱工業大學(深圳)

HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

开始实验