基于 AMPERE EMAG 实现动态实时目标检测

Contents

人工智能（AI）或机器学习软件程序是可以感知，推理，响应和自动调整的程序。尽管 AI 算法已经存在了很多年，但是最近基于 AI 的应用在不同行业中迅速扩展。 为了提供更有前景的 AI 算法，应用创新者必须能在一个复杂多样的平台上部署其应用程序，该平台应该提供出色的最终用户体验，较低的运营成本并确保节能环保。

每个机器学习程序中都有两个主要任务。 首先，它进行训练，其次，它对在训练过程中获得的数据进行推理。 虽然训练只进行一次，但推理却要连续进行多次，那么推理过程就需要强大的平台来处理。

基于 Ampere 的 Armv8 64 位处理器的服务器平台是为大型公共和私有云环境量身打造的，可以成为如实时对象检测等机器学习应用中承担推理任务的极佳解决方案。 事实证明，Ampere 的云解决方案通过其大量的内核，高速连接，较高的内存吞吐量和成本效益为开发人员提供了明显的优势。 高度集成的专用 Ampere 解决方案为私有云和公共云提供了高性能和很低的总拥有成本（TCO）。

本文将描述如何利用 eMAG™平台来演示 YOLO 算法，这是一种流行的物体检测系统，由单个卷积网络组成，该系统同时预测多个边界框和这些框的类概率。 YOLO 训练完整图像并直接优化检测性能。 与传统的对象检测方法相比，此统一模型具有多个优点。 YOLO 在训练和测试期间会看到整个图像，因此它隐式地编码有关类及其外观的上下文信息。

在此演示中，所有训练和测试均在 eMAG 平台上完成。 本文将总结优化 eMAG 上实时物体检测的最佳方法。

数据和网络架构

YOLO 不使用传统分类器，而是采取了完全不同的方法实现了一套很有竞争力的解决方案。 它只查找图像一次，然后将图像划分为 13 x 13 的单元格，以预测 5 个边界框。 每个单元可容纳 5 个边界框，边界框是包围对象的矩形区域，如下图 2 所示。

YOLO 使用卷积神经网络。 在本演示中，模型被实现为卷积神经网络，并在 PASCAL VOC 检测数据集上进行了评估。 该数据集用于评估图像分类，图像分割和对象检测的算法，该算法总共具有 20 个类和 11,530 个图像。 该网络具有 9 个卷积层，使体系结构能够接受输入并经过卷积层，然后进行最大池化，其中 3 * 3 滤波器用于卷积，2 * 2 滤波器用于最大池化。 没有完全连接的层。 卷积层在 ImageNet 分类中进行了预训练。 该网络的最终输出是预测的 13×13×125 张量。

Layer	kernel	stride	output shape
Input			(416, 416, 3)
Convolution	3x3	1	(416, 416, 16)
MaxPooling	2x2	2	(208, 208, 16)
Convolution	3x3	1	(208, 208, 32)
MaxPooling	2x2	2	(104, 104, 32)
Convolution	3x3	1	(104, 104, 64)
MaxPooling	2x2	2	(52, 52, 64)
Convolution	3x3	1	(52, 52, 128)
MaxPooling	2x2	2	(26, 26, 128)
Convolution	3x3	1	(26, 26, 256)
MaxPooling	2x2	2	(13, 13, 256)
Convolution	3x3	1	(13, 13, 512)
MaxPooling	2x2	1	(13, 13, 512)
Convolution	3x3	1	(13, 13, 1024)
Convolution	3x3	1	(13, 13, 1024)
Convolution	1x1	1	(13, 13, 125)

Figure 1: 该框架具有 9 个卷积层，每个卷积层之后是最大池化层。

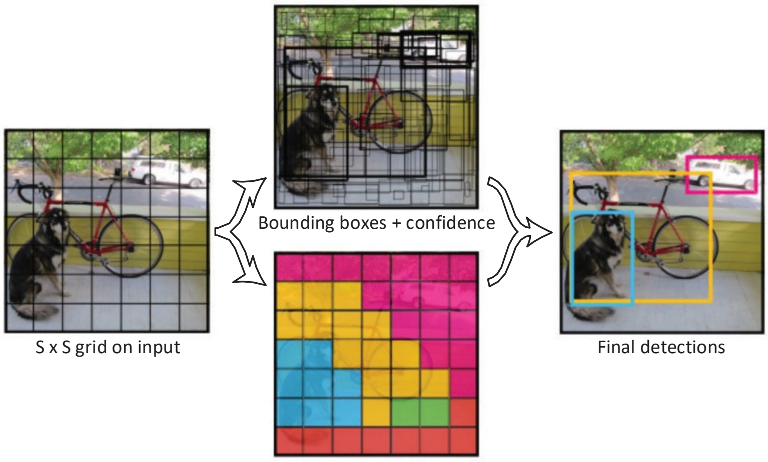

Figure 2: 模型

硬件平台

YOLO 实时对象检测系统采用 Tensorflow 1.0 后端，运行在 Ampere 的 eMAG 平台商，以下是在演示过程中使用的所有硬件的摘要：

架构： Arm®v8 64-bit 服务器

CPU 操作位数： 64-bit

CPU 型号： Ampere eMAG 8180

处理器子系统

- 32 个 Arm v8 64 位 CPU 内核，Turbo 模式下主
- 频高达 3.3 GHz
- 每核 32KB 一级指令缓存、32KB 一级数据缓存
- 每 2 核共享 256KB 二级缓存

内存

- 32MB 全局共享三级缓存
- 8 通道 72 位 DDR4-2667 内存控制器
- 高级 ECC 和 DDR4 RAS 特性
- 支持多达 16 个 DIMM，支持高达 1TB 内存容量

系统资源

- 通用中断控制器 ARM GICv3
- 支持 I/O 虚拟化
- 企业服务器级 RAS
  - 端到端 Data Poisoning 内存错误处理
  - 数据容错与隔离
  - 后台刷洗三级缓存和 DRAM

连接性

- 42 条 PCIe Gen 3 通道，8 个控制器
- 4 个 SATA Gen 3 端口
- 2 个 USB 2.0 端口

技术与功能

- TSMC 16nm FinFET+
- Arm v8.0-A，SBSA 3 级
- EL3，安全内存和安全启动支持
- 高级功耗管理

公号

- TDP: 125 W

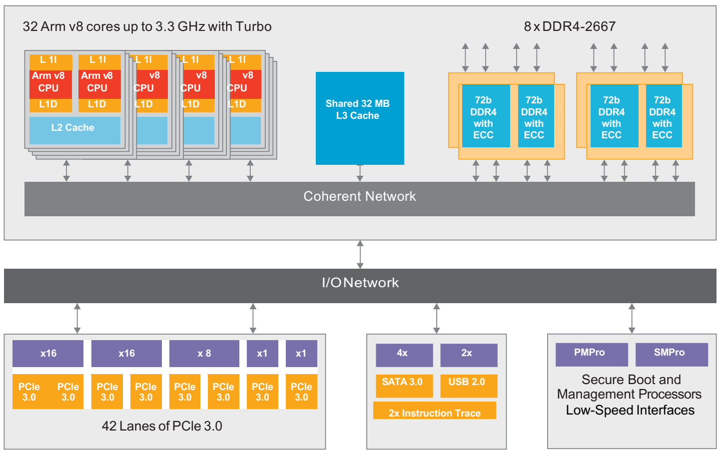

Figure 3: eMAG Block Diagram

软件栈

为了运行 YOLO 演示，安装了以下所用系统：Tensorflow，Numpy，OpenCV3 和 Python3。

首先，第一步是通过提供以下命令来构建适当的 Cython 扩展：$ python3 setup.py build_ ext －inplace。

训练

卷积层已在 ImageNet 1000 类数据集中进行了预训练。在预训练中，使用前 9 个卷积层，然后使用平均池化层。

下一步是转换模型来执行检测。 最后的卷积层具有 1 * 1 内核，该内核的存在是为了将数据减少为 13 * 13 * 125 的形状。这样可以为每个网格单元启用了 125 个通道，这 125 个数字包含边界框和类预测的数据。 每个网格单元预测 5 个边界框和一个边界框，因此共有 125 个通道。

边界框由 25 个数据元素描述，这些数据元素是该类的宽度，高度，置信度得分和概率分布。 YOLO 预测每个网格单元有多个边界框。 在训练期间，理想的情况是只对每个对象使用一个边界框预测变量。

一个预测器被指定为“负责任”的，该预测器基于具有真实的当前 IOU 最高的预测来预测对象。 这导致边界框预测变量之间的专用化。 每个预测器在预测某些大小，宽高比或对象类别方面都有更好的表现，从而改善了总体召回率。 这是一个简单的过程，其中输入图像的大小调整为 416 * 416 像素，并且一次通过了卷积神经网络。 它从另一端出来，为 13 * 13 * 125 张量，描述了网格单元的边界框。 计算边界框的最终分数，并消除小于 30％阈值的分数。

通过在 eMAG 平台上运行训练模型，使训练变得更加容易，因为开发人员能够通过 eMAG 的高性能内核，高速连接性，内存吞吐量和运营商级可靠性在现有数据中心的占地面积中部署定制的深度学习解决方案，同时大幅降低了电力和运营成本。

推理

就像在训练中一样，预测测试图像的推理仅需要进行一次网络评估。 在 PASCAL VOC 上，网络可预测每个图像 98 个边界框以及每个框的类概率。 与基于分类器的方法不同，YOLO 只需要进行一次网络评估，因此测试时间非常快。 网格设计在边界框预测中强制执行空间分集。 通常，很明显，一个对象属于哪个网格单元，并且网络仅为每个对象预测一个框。 但是，一些大对象或多个单元格边界附近的对象可以被多个单元格很好地定位，非最大抑制可用于修复这些多次检测。 尽管对于 R-CNN 或 DPM 而言，对性能并不重要，但非最大抑制会在 mAP 中增加 2-3％。

对于此演示，在 eMAG 平台上运行了以下命令：
为了加载微小的 yolo 重量进行训练，给出了以下命令：python flow--model cfg/tiny-yolo-voc.cfg--load weights/tiny-yolo-voc.weights

为了在 eMAG 平台上执行演示，给出了以下命令：
python flow--model cfg/tiny-yolo-voc.cfg--load weights/tiny-yolo-voc.weights--demo videofile.avi - saveVideo

总结

如上所述，使用 Ampere eMAG 平台运行 YOLO 实时对象检测算法不需要任何其他软件或配置。 该系统开箱即用，用户可以直接下载预构建的 Python 库并利用预训练的模型。

Ampere 的 eMAG 平台将继续开发 AI 培训模型，以提高准确性并支持不断变化的工作负载和用例。 与基于分类器的方法不同，YOLO 在直接与检测性能相对应的损失函数上进行训练，并且整个模型都在一起进行训练。
YOLO 是实时对象检测的最先进技术，它还可以很好地推广到新领域，这使其成为依赖于快速，强大的对象检测的应用程序的理想选择。

在训练了神经网络之后，将其部署为运行推理计算来对新输入进行分类，识别和处理。 由于所有必需的软件和库已经存在，因此简化了 Ampere 的 eMAG 平台用于推理的部署。 开发人员可以利用 eMAG 的灵活性将在线机器学习功能添加到任何定制的训练模型中，来进行实时推理。eMAG 良好的吞吐量和低延迟带来高性能的同时，提供了更好的可扩展性，同时降低了平台的复杂性并削减了成本。 由于在 eMAG 平台上运行的设计和代码的高效性能，Arm®v8 64 位对于运行最新的对象检测算法和传统方法（例如 SIFT，SURF 和 ORB）至关重要。

Ampere 能够利用其高度集成的专用架构来交付此解决方案。事实证明，该架构可为私有云和公共云以较低的总拥有成本（TCO）提供高性能。

参考

https://modelzoo.co/model/yolo-tensorflow
https://www.intel.ai/papers/training-deep-convolutional-neural-networks-with-horovod-on-intel-high-performance-computing-architecture/
https://arxiv.org/pdf/1506.02640.pdf
https://amperecomputing.com/product/
https://www.youtube.com/watch?v=4eIBisqx9_g