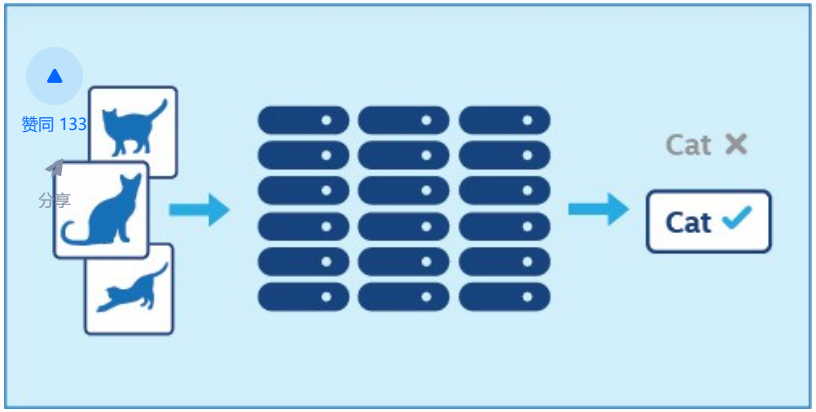



# OpenVINO推理简介

火狐狸
不时地挑战一下不可能



半导体厂商开发的硬件再怎么厉害，也需要软件工具的加持，重复制造轮子不是一个好主意，为了充分挖掘处理器的性能，各个厂家都发布了各种软件框架和工具，比如Intel的OpenVINO，Nvidia的TensorRT等等。

这里重点介绍英特尔发布的针对AI工作负载的一款部署神器--OpenVINO。

OpenVINO是英特尔推出的一款全面的工具套件，用于快速部署应用和解决方案，支持计算机视觉的CNN网络结构超过200余种。

目前OpenVINO已经发布了API 2.0，详情见另一篇介绍OpenVINO推理简介2.0

我们有了各种开源框架，比如tensorflow，pytorch，mxnet，caffe2等，为什么还要推荐OpenVINO来作为部署工具呢?

当模型训练结束后，上线部署时，就会遇到各种问题，比如，模型性能是否满足线上要求，模型如何嵌入到原有工程系统，推理线程的并发路数是否满足，这些问题决定着投入产出比。只有深入且准确的理解深度学习框架，才能更好的完成这些任务，满足上线要求。实际情况是，新的算法模型和所用框架在不停的变化，这个时候恨不得工程师什么框架都熟练掌握，令人失望的是，这种人才目前是稀缺的。

OpenVINO是一个Pipeline工具集，同时可以兼容各种开源框架训练好的模型，拥有算法模型上线部署的各种能力，只要掌握了该工具，你可以轻松的将预训练模型在Intel的CPU上快速部署起来。

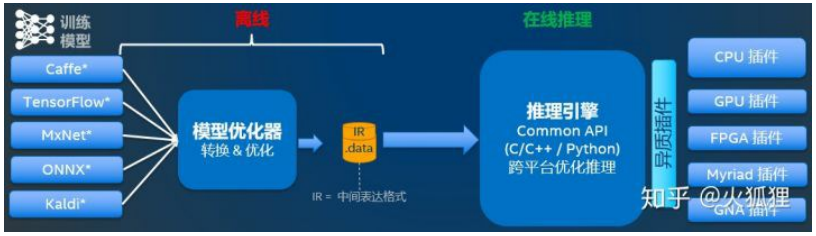
对于AI工作负载来说，OpenVINO提供了深度学习推理套件（DLDT），该套件可以将各种开源框架训练好的模型进行线上部署，除此之外，还包含了图片处理工具包OpenCV，视频处理工具包Media SDK，用于处理图像视频解码，前处理和推理结果后处理等。

在做推理的时候，大多数情况需要前处理和后处理，前处理如通道变换，取均值，归一化，Resize等，后处理是推理后，需要将检测框等特征叠加至原图等，都可以使用OpenVINO工具套件里的API接口完成。

对于算法工程师来说，OpenCV已经非常熟悉，这里重点讲一下深度学习部署套件DLDT。

DLDT分为两部分：

- 模型优化器(Model Optimizer)
- 推理引擎(Inference Engine)




其中，模型优化器是线下模型转换，推理引擎是部署在设备上运行的AI负载。

模型优化器是一个python脚本工具，用于将开源框架训练好的模型转化为推理引擎可以识别的中间表达，其实就是两个文件，xml和bin文件，前者是网络结构的描述，后者是权重文件。模型优化器的作用包括压缩模型和加速，比如，去掉推理无用的操作(Dropout)，层的融合(Conv + BN + Relu)，以及内存优化。

推理引擎是一个支持C\C++和python的一套API接口，需要开发人员自己实现推理过程的开发，开发流程其实非常的简单，核心流程如下：

1. 装载处理器的插件库
2. 读取网络结构和权重
3. 配置输入和输出参数
4. 装载模型
5. 创建推理请求
6. 准备输入Data
7. 推理
8. 结果处理

下面给出一段C++的代码例子

```
// 创建推理core，管理处理器和插件
InferenceEngine::Core core;
// 读取网络结构和权重
CNNNetReader network_reader;
network_reader.ReadNetwork("Model.xml");
network_reader.ReadWeights("Model.bin");
// 配置输入输出参数
auto network = network_reader.getNetwork();
InferenceEngine::InputsDataMap input_info(network.getInputsInfo());
InferenceEngine::OutputsDataMap output_info(network.getOutputsInfo());
/** Iterating over all input info**/
for (auto &item : input_info) {
    auto input_data = item.second;
    input_data->setPrecision(Precision::U8);
    input_data->setLayout(Layout::NCHW);
    input_data->getPreProcess().setResizeAlgorithm(RESIZE_BILINEAR);
    input_data->getPreProcess().setColorFormat(ColorFormat::RGB);
}
/** Iterating over all output info**/
for (auto &item : output_info) {
```

```
    auto output_data = item.second;
    output_data->setPrecision(Precision::FP32);
    output_data->setLayout(Layout::NC);
}
// 装载网络结构到设备
auto executable_network = core.LoadNetwork(network, "CPU");
std::map<std::string, std::string> config = {{ PluginConfigParams::KEY_PERF_COUNT, F
auto executable_network = core.LoadNetwork(network, "CPU", config);
// 创建推理请求
auto infer_request = executable_network.CreateInferRequest();
// 准备输入Data
or (auto & item : inputInfo) {
    auto input_name = item->first;
    /** Getting input blob **/
    auto input = infer_request.GetBlob(input_name);
    /** Fill input tensor with planes. First b channel, then g and r channels **/
    ...
}
// 推理
sync_infer_request->Infer();
// 结果处理
for (auto &item : output_info) {
    auto output_name = item.first;
    auto output = infer_request.GetBlob(output_name);
    {
        auto const memLocker = output->cbuffer(); // use const memory locker
        // output_buffer is valid as long as the lifetime of memLocker
        const float *output_buffer = memLocker.as<const float *>();
        // process result
        ...
    }
}
```

推理过程只需要开发一次，只要模型的输入和输出不变，剩下的就是训练模型和模型优化工作了。

这是一款非常给力的专门做推理的工具，并且有intel在不停的开发和优化新的网络结构，有人维护和开发这件事很重要。

部署上线

另外一篇介绍一种灵活且高效的



关于OpenVINO优化参数配置参考



关于AI设备选型可参考

4 赞同 · 1 评论 文章

对于做AI工程化的初学者可读一下

火狐狸：聊聊算法引擎的工程化问题
13 赞同 · 10 评论 文章

编辑于 2022-03-18 09:19

AI初创 AI技术 持续部署(CD)

## 文章被以下专栏收录

AI架构与优化
人工智能相关的软硬件知识和算法模型优化加速等技术

AI模型部署
主要介绍AI模型的落地与部署，关注AI芯片

## 推荐阅读


可能是最好的能运行在CPU上的深度学习框架：...
天马微云 发表于天马行空

极市直播|周兆靖：如何利用开源OpenVINO™工具集加...
极市平台

OpenVino初体验
OLDPA... 发表于深度学习那...


OpenVINO推理简介2.0
火狐狸

57 条评论 切换为时间排序

写下你的评论...

哈Aa哈 2019-12-25

您好，我的yolov3在转IR时有报错FusedBatchNormV3 (72)，可以分享一下您的思路吗?

1

火狐狸 (作者) 回复 哈Aa哈 2019-12-27

你的OpenVINO是什么版本?

赞

Yeszhuang 2021-11-11

请问火狐狸，如果输入输出是默认的，不需要设置参数，那配置输入输出的部分可以删掉吗?还有，怎么处理多个输入图片🤔

赞

火狐狸 (作者) 回复 Yeszhuang 2021-11-11

如果bs也是确定的，可以默认，否则还是要指定一下，多个图片组合成一个batch。

赞

hildw 2020-04-13

速度大概提升多快啊

赞

火狐狸 (作者) 回复 hildw 2020-07-22

同样精度的情况下，2倍以上速度提升，另外，内存占用也会较少

1

AD哥 回复 火狐狸 (作者) 2021-01-25

是的，就是比mnn还快的

赞

知乎用户 2020-03-15

你好，请问下运行自带的sample程序，报错DLL Load Failed找不到指定的程序是什么原因?

赞

火狐狸 (作者) 回复 知乎用户 2020-03-16

source一下openvino环境，source /opt/intel/openvino/bin/setup cars. sh

赞

知乎用户 回复 火狐狸 (作者) 2020-03-16

用过这个方法，没有解决。还试过配置全部环境变量，也试过更换ie_api.pyd文件。用的2020.01版本。

赞

展开其他 2 条回复

rot.cx 2020-02-23

mark

赞

徐唱 2019-12-16

你好，我用yolov3的xmld但是创建不了可执行网络是什么原因

赞

火狐狸 (作者) 回复 徐唱 2019-12-17

报什么错误?

赞

火狐狸 (作者) 回复 徐唱 2020-01-03

是标准的yolo3吗?

赞

旭旭233 2021-07-22

您好，我想问一下yolov5自训练的模型怎么openvino上部署推理呢？已经转成IR文件了

赞

火狐狸 (作者) 回复 旭旭233 2021-07-23

调用API接口开发引擎，或者用Openvino Model Server部署

赞

火狐狸 (作者) 2021-07-03

你的cpu是i3还是i5？vpu比cpu慢也正常，vpu主要是负载AI，让cpu可以做更多其他事情

赞

一只小飞象 回复 火狐狸 (作者) 2021-07-19

好的，谢谢您，另外我测试异步状态下模型前向时间，一次100ms,一次10ms,时间是这样间歇的，请问这是什么原因呢

赞

火狐狸 (作者) 回复 一只小飞象 2021-07-19

前面几次有个warmup的过程，一直跳还是开始慢一下?

赞

一只小飞象 2021-07-03

您好，我插了加速棒相比于cpu反而变慢了,请问是什么原因呢，只改动了loadnetwork这个函数参数，是需要配套改动其他地方吗

赞

火狐狸 (作者) 2021-07-02

GetBlob是根据input name找到输入节点buffer，然后copy数据进去

赞

知乎用户 2021-07-02

萌新提问，在准备输入数据那里，
auto input = infer_request.GetBlob(input_name);
这里应该用SetBlob而不是GetBlob吧

赞

TShijie 2021-05-05

您好，我是一个小白，我想问一下，怎么在PycharmIDE上使用Openvino进行推理，是直接在设置里面进行导入Openvino的包就行了，还是需要官网下载openvino然后进行相应的环境配置才能用，求一个解决问题的方向

赞

Alva 2021-02-22

你好，我仅有一些cv和qt的基础，目前要做一个前景提取的可执行软件，想问一下写好程序之后如何利用openvino这个“推理框架”。我不太懂openvino中的IR文件与一个c++代码之间的关系

赞

火狐狸 (作者) 回复 Alva 2021-03-18

IR文件是OpenVINO可以读取的模型文件，调用OpenVINO的API去装载，推理就行。IR文件是由其他框架(TF, Pytorch)的模型转换而来。

赞

Alva 回复 Alva 2021-03-18

嗯嗯，谢谢，目前做的差不多了

赞

辉子 2021-01-18

请问自己搭建的CNN网络模型能部署在openvino上面吗？我是把faster rcnn模型加上FPN之后训练的，这样是为了检测小目标

赞

火狐狸 (作者) 回复 辉子 2021-01-18

可以

赞

辉子 回复 火狐狸 (作者) 2021-01-19

openvino是不是不支持merge 这个operator？Supported only when it is fused to the TensorIterator layer

docs.openvinotoolkit.org

...

赞

展开其他 3 条回复

知乎用户 2020-12-24

请问下：推理的话，一般用GPU的还是多吧？用cpu的话，是不是用在比较简单的模型，且对性能要求不太高的场合呢？

赞

火狐狸 (作者) 回复 知乎用户 2020-12-24

适合即可，看具体指标，如吞吐，延时，并发等，不浪费，又能满足场景要求，且做到最低成本，即最高标准

赞

知乎用户 2020-12-23

请问model optimizer部分是开源的吗？能否知道其中使用了什么优化加速的方法？

赞

火狐狸 (作者) 回复 知乎用户 2020-12-23

开源的

赞

火狐狸 (作者) 回复 知乎用户 2021-01-18

开源的

赞

persuelx 2020-11-11

请问我推理一张图片需要87ms，但是加载模型那会比较耗时，这个87是只计算infer的时间，请问这个有问题嘛？而且我的模型只有3.5兆。

赞

火狐狸 (作者) 回复 persuelx 2020-11-11

model只load次，算是warmup吧，统计后面infer时间即可

赞

persuelx 回复 火狐狸 (作者) 2020-11-12

就是这样操作的

赞

展开其他 3 条回复

火狐狸 (作者) 2020-09-29

不行，不支持arm指令

赞

放羊娃王二小 2020-09-29

这个能用在各种Linux anzhuo开发板上吗

赞

火狐狸 (作者) 2020-07-22

现在最新版本是20.4，支持BF16，运行在CPX型号CPU上

赞

okaoka 回复 火狐狸 (作者) 2020-08-25

Failed to initialize Inference Engine backend (device = CPU): Failed to create plugin /opt/intel/openvino_2020.4.287 /deployment_tools/inference_engine/lib/intel64/libMKLDNNPlugin.dylib for device CPU

赞

okaoka 回复 火狐狸 (作者) 2020-08-25

请问下这个是什么原因呀?

赞

查看全部 9 条回复