Deep Learning Deployment Toolkit

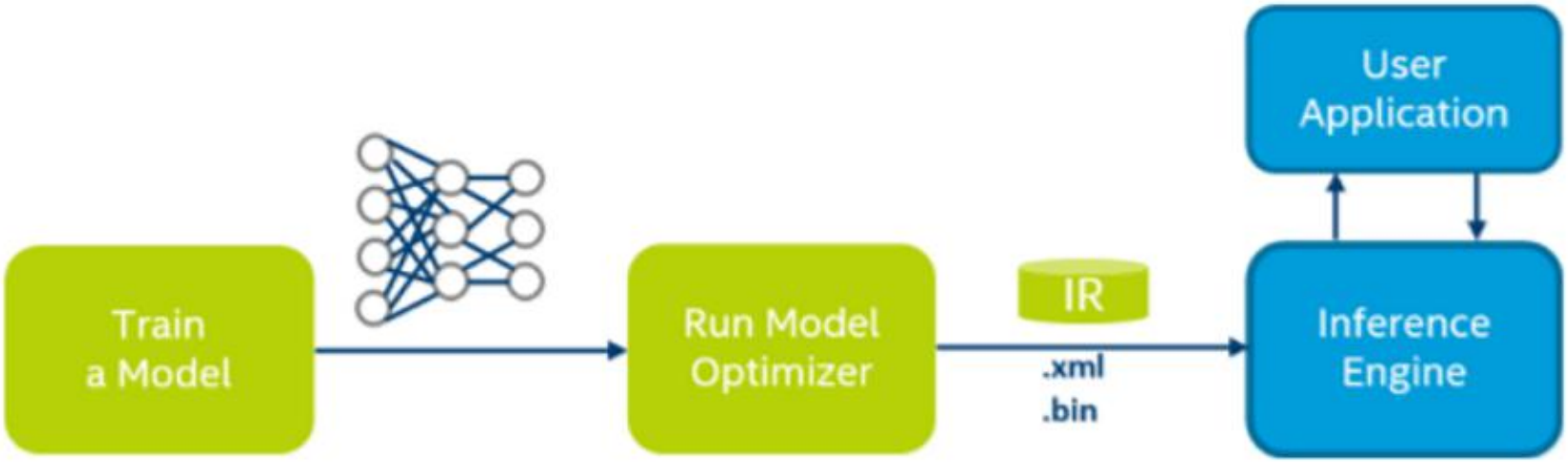
AI图像组 欧阳高询

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

- 简介
- Dldt源码下载地址：https://github.com/opencv/dldt
- Dldt是Inter发布的一套针对Inter CPU/GPU的深度学习开发工具。
- Dldt是为了将已训练的模型快速部署落地，俗称推断(Inference)，对于常见的100种模型能够比较快速支持，对于含有自定义层的模型需要注册之类的操作，相对麻烦一些。
- Dldt主要分为两大模块，Model Optimizer 和 Inference Engine
- Dldt目前支持Caffe/TensorFlow/MxNet/ONNX四种框架，部分支持Kaggle框架。

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Dldt流程图



平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Model Optimizer

- 1.安装好环境：
  TF模型需要安装TF，MxNet模型需要安装MxNet等等

- 2.需要提供包含已训练模型拓扑结构的文件和权重文件
- MxNet最好提供.params和.json文件
- TF最好提供frozen的pb文件，对于.meta文件和checkpoint文件容易出错。
- Caffe需要提供.caffemodel和.prototxt文件

- **注意注意注意，文件不要轻易重命名，比如.params和.json文件要对应起来**

- 一定要清楚模型输入层的input_shape，以及NCHW格式还是NHWC格式等，模型的batchsize不要设定死，因为训练阶段batchsize肯定大于1，而部署的时候batchsize会把强制转成1。

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Model Optimizer

--input_model

--output_dir

--input_shape

--reverse_input_channels

--log_level

--mean_values

--scale_values

--data_type

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Model Optimizer

BatchNorm	BatchNormalization
Crop	Crop
ScaleShift	ScaleShift
Pooling	Pooling
SoftmaxOutput	SoftMax
SoftmaxActivation	SoftMax
null	Ignored, does not appear in IR
Convolution	Convolution
Deconvolution	Deconvolution
Activation(act_type = relu)	ReLU
ReLU	ReLU
LeakyReLU	ReLU (negative_slope = 0.25)
Concat	Concat
elemwise_add	Eltwise(operation = sum)
_Plus	Eltwise(operation = sum)
Flatten	Flatten
Reshape	Reshape
FullyConnected	FullyConnected
UpSampling	Resample
transpose	Permute
LRN	Norm
L2Normalization	Normalize
Dropout	Ignored, does not appear in IR
_copy	Ignored, does not appear in IR
_contrib_MultiBoxPrior	PriorBox

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Model Optimizer

_contrib_MultiBoxDetection	DetectionOutput
broadcast_mul	ScaleShift
sigmoid	sigmoid
Activation (act_type = tanh)	Activation (operation = tanh)
LeakyReLU (act_type = prelu)	PReLU
LeakyReLU (act_type = elu)	Activation (operation = elu)
elemwise_mul	Eltwise (operation = mul)
add_n	Eltwise (operation = sum)
ElementWiseSum	Eltwise (operation = sum)
_mul_scalar	Power
broadcast_add	Eltwise (operation = sum)
slice_axis	Crop
_minus_scalar	Power
Pad	Pad
_contrib_Proposal	Proposal
ROIPooling	ROIPooling
stack	Concat
swapaxis	Permute
zeros	Const
rnn	TensorIterator
rnn_param_concat	Concat
slice_channel	Split
_maximum	Eltwise(operation = max)
_minimum	Power(scale=-1) + Eltwise(operation = max) + Power(scale=-1)
InstanceNorm	scale * (x - mean) / sqrt(variance + epsilon) + B

平安寿险AI团队 平安寿险AI团队

Model Optimizer示例

```
[root@SHB-L0127452 model_backup]# ls -l
total 22000
-rwxrwxrwx 1 root root 22385626 Jan 22 16:23 deploy_ssd_mobilenet_v2-0214.params
-rwxr-x--- 1 root root   138833 Jan 29 13:20 deploy_ssd_mobilenet_v2-symbol.json
[root@SHB-L0127452 model_backup]# python /var/bot-dev/dldt_release/dldt/model-optimizer/mo_mxnet.py  --input_mo
del ./deploy_ssd_mobilenet_v2-0214.params  --input_shape [1,3,540,960]  --mean_values [123.0,117.0,104.0]
```

```
[ SUCCESS ] Generated IR model.
[ SUCCESS ] XML file: /data/ouyanggaoxun/TA-demo_test/model_backup/./deploy_ssd_mobilenet_v2-0214.xml
[ SUCCESS ] BIN file: /data/ouyanggaoxun/TA-demo_test/model_backup/./deploy_ssd_mobilenet_v2-0214.bin
[ SUCCESS ] Total execution time: 4.71 seconds.
[root@SHB-L0127452 model_backup]# ls -l
total 43752
-rw-r----- 1 root root 22157544 Jan 29 13:29 deploy_ssd_mobilenet_v2-0214.bin
-rw-r----- 1 root root    19419 Jan 29 13:29 deploy_ssd_mobilenet_v2-0214.mapping
-rwxrwxrwx 1 root root 22385626 Jan 22 16:23 deploy_ssd_mobilenet_v2-0214.params
-rw-r----- 1 root root    92630 Jan 29 13:29 deploy_ssd_mobilenet_v2-0214.xml
-rwxr-x--- 1 root root   138833 Jan 29 13:20 deploy_ssd_mobilenet_v2-symbol.json
[root@SHB-L0127452 model_backup]#
```

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

xml示例

```
<?xml version="1.0" ?>
<net batch="1" name="deploy_ssd_mobilenet_v2-0214" version="3">
    <layers>
        <layer id="0" name="data" precision="FP32" type="Input">
            <output>
                <port id="0">
                    <dim>1</dim>
                    <dim>3</dim>
                    <dim>540</dim>
                    <dim>960</dim>
                </port>
            </output>
        </layer>
        <layer id="1" name="Add_" precision="FP32" type="ScaleShift">
            <input>
```

注意这里的**version**代表的是版本号，有时候版本会更新，他会校验版本号，我试过偷懒的方式，将升级后**version**改为之前的版本，那样是不行的。

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Inference Engine

•Ubuntu* 16.04 LTS 64-bit or CentOS* 7.4 64-bit
•GCC* 5.4.0 (for Ubuntu* 16.04) or GCC* 4.8.5 (for CentOS* 7.4)
•CMake* version 2.8 or higher

export OMP_NUM_THREADS=1

这个变量决定了使用多少个CPU资源，默认情况下会吃掉所有的CPU资源

要写C++工程，CMakeLists.txt进行编译就好了

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Inference Engine

```
Load(personDetection).into(pluginsForNetworks[FLAGS_d]);
```

```
struct Load {
    BaseDetection& detector;
    explicit Load(BaseDetection& detector) : detector(detector) { }

    void into(InferencePlugin & plg) const {
        if (detector.enabled()) {
            detector.net = plg.LoadNetwork(detector.read(), {});
            detector.plugin = plg;
        }
    }
};
```

```
-rwxr-x--- 1 root root  7931424 Jan 23 18:29 libclDNN64.so
-rwxr-x--- 1 root root  1636905 Jan 23 18:29 libclDNNPlugin.so
-rwxr-x--- 1 root root   726637 Jan 23 18:23 libcpu_extension.so
-rwxr-x--- 1 root root 21455753 Mar 27 13:33 libformat_reader.so
-rwxr-x--- 1 root root   650812 Jan 23 18:22 libHeteroPlugin.so
-rwxr-x--- 1 root root  2545478 Jan 23 18:21 libinference_engine.so
-rwxr-x--- 1 root root 12310463 Jan 23 18:26 libMKLDNNPlugin.so
-rwxr-x--- 1 root root    36801 Jan 23 18:19 libmock_engine.so
```

调用InferenceEngine的API函数LoadNetwork解析.xml和.bin，生成网络结构的对象CNNNetwork

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Inference Engine

```
    virtual void submitRequest() {
        if (!enabled() || !request) return;
        request.StartAsync();
    }
```

```
const float *detections = request.GetBlob(outputName)->buffer().as<float *>();
```

```
-rwxr-x--- 1 root root  7931424 Jan 23 18:29 libclDNN64.so
-rwxr-x--- 1 root root  1636905 Jan 23 18:29 libclDNNPlugin.so
-rwxr-x--- 1 root root   726637 Jan 23 18:23 libcpu_extension.so
-rwxr-x--- 1 root root 21455753 Mar 27 13:33 libformat_reader.so
-rwxr-x--- 1 root root   650812 Jan 23 18:22 libHeteroPlugin.so
-rwxr-x--- 1 root root  2545478 Jan 23 18:21 libinference_engine.so
-rwxr-x--- 1 root root 12310463 Jan 23 18:26 libMKLDNNPlugin.so
-rwxr-x--- 1 root root    36801 Jan 23 18:19 libmock_engine.so
```

调用InferenceEngine的InferRequest对象处理推断请求，并通过GetBlob获得模型输出层outputName的结果，结果为7维的数据，分别为：

1.框的ID号

2.框的label

3.框的置信度

4~7 框的坐标

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Inference Engine

LResNet50E-IR

	1 cpu	2 cpu	3 cpu	4 cpu	5 cpu	6 cpu	7 cpu	8 cpu
	215	109	115	96	147	130	120	118
	232	110	110	96	148	138	118	121
	208	107	108	101	147	150	121	122
	245	117	116	96	142	134	123	123
	249	106	114	96	128	131	133	110
	234	108	81	94	131	134	123	160
	206	112	81	100	144	151	133	125
	206	116	79	95	123	139	129	107
	200	113	114	96	87	138	117	101
	206	112	118	95	81	117	137	120
Ave time(ms):	220.1	111	103.6	96.5	127.8	136.2	125.4	120.7

MobileFaceNet

	1 cpu	2 cpu	3 cpu	4 cpu	5 cpu	6 cpu	7 cpu	8 cpu
	11	5	5	4	4	5	6	3
	9	5	5	5	4	3	5	3
	9	7	5	8	4	7	7	3
	9	11	5	9	3	4	6	3
	10	5	5	7	8	3	12	9
	9	7	6	6	3	4	5	4
	10	5	5	5	3	6	5	9
	9	7	4	5	4	3	5	8
	9	7	4	5	6	3	5	7
	9	7	7	5	4	5	10	3
Ave time(ms):	9.4	6.6	5.1	5.9	4.3	4.3	6.6	5.2

说明：测试环境，I5第8代虚拟机下，输入:640*360视频文件

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

Inference Engine

dldt与MxNet对比

dldt单cpu			mxnet-4 cpu			mxnet-单CPU		
0155_1020.jpg	323.357	16	0155_1020.jpg	16	1399	0155_1020.jpg	16	1883
0155_1050.jpg	321.194	15	0155_1050.jpg	15	1439	0155_1050.jpg	15	1370
0155_1080.jpg	318.219	15	0155_1080.jpg	15	1333	0155_1080.jpg	15	1292
0155_1110.jpg	313.117	16	0155_1110.jpg	16	1556	0155_1110.jpg	16	1354
0155_1140.jpg	316.814	14	0155_1140.jpg	14	1275	0155_1140.jpg	14	1291
0155_1200.jpg	317.68	13	0155_1200.jpg	13	1315	0155_1200.jpg	13	1301
0155_1290.jpg	312.828	15	0155_1290.jpg	15	1196	0155_1290.jpg	15	1370
0155_150.jpg	311.414	21	0155_150.jpg	21	1315	0155_150.jpg	21	1314
0155_180.jpg	322.553	21	0155_180.jpg	21	1350	0155_180.jpg	21	1255
0155_240.jpg	320.798	15	0155_240.jpg	15	1320	0155_240.jpg	15	1261
0155_270.jpg	314.113	14	0155_270.jpg	14	1295	0155_270.jpg	14	1233
0155_30.jpg	320.339	20	0155_30.jpg	20	1263	0155_30.jpg	20	1223
0155_330.jpg	322.097	10	0155_330.jpg	10	1263	0155_330.jpg	10	1300
0155_360.jpg	320.453	11	0155_360.jpg	11	1241	0155_360.jpg	11	1265
0155_420.jpg	320.824	9	0155_420.jpg	9	1209	0155_420.jpg	9	1370
0155_450.jpg	312.799	8	0155_450.jpg	8	1295	0155_450.jpg	8	1249
0155_480.jpg	324.187	7	0155_480.jpg	7	1230	0155_480.jpg	7	1288
0155_570.jpg	322.019	6	0155_570.jpg	6	1276	0155_570.jpg	6	1298
0155_600.jpg	331.871	8	0155_60.jpg	21	1240	0155_60.jpg	21	1337
0155_60.jpg	319.962	21	0155_600.jpg	8	1319	0155_600.jpg	8	1321
0155_630.jpg	319.422	11	0155_630.jpg	11	1240	0155_630.jpg	11	1365
0155_660.jpg	310.6	18	0155_660.jpg	18	1242	0155_660.jpg	18	1312
0155_810.jpg	316.311	16	0155_810.jpg	16	1220	0155_810.jpg	16	1295
0155_840.jpg	317.309	16	0155_840.jpg	16	1394	0155_840.jpg	16	1306
0155_870.jpg	321.425	17	0155_870.jpg	17	1329	0155_870.jpg	17	1334
0647_120.jpg	319.301	55	0647_120.jpg	55	1253	0647_120.jpg	55	1376
0647_180.jpg	310.164	49	0647_180.jpg	49	1191	0647_180.jpg	49	1303
0647_210.jpg	312.263	34	0647_210.jpg	34	1253	0647_210.jpg	34	1240
0647_240.jpg	313.777	46	0647_240.jpg	46	1218	0647_240.jpg	46	1262
0647_270.jpg	317.439	47	0647_270.jpg	47	1313	0647_270.jpg	47	1329
0647_30.jpg	315.402	50	0647_30.jpg	50	1265	0647_30.jpg	50	1244

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

参考资料

https://github.com/opencv/dldt
https://software.intel.com/en-us/articles/OpenVINO-InferEngine
https://software.intel.com/en-us/articles/OpenVINO-ModelOptimizer

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队

THANK YOU !

平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队 平安寿险AI团队