Tina NPU 混合量化说明

版本号: 1.0

发布日期: 2021.07.21

版本历史

版本号	日期	制/修订人	内容描述
1.0	2021.07.21	PDC	NPU 开发混合量化说明

目　录

插　图

1 前言

1.1 读者对象

本文档（本指南）主要适用于以下人员：

- 技术支持工程师
- 软件开发工程师
- AI 应用案客户

1.2 约定

1.2.1 符号约定

本文中可能出现的符号如下：

警告

警告

技巧

1. 技巧
2. 小常识

说明

说明

2 正文

2.1 NPU 开发简介

- 支持 int8/uint8/int16 量化精度，运算性能可达 1TOPS.
- 相较于 GPU 作为 AI 运算单元的大型芯片方案，功耗不到 GPU 所需要的 1%.
- 可直接导入 Caffe, TensorFlow, Onnx, TFLite，Keras, Darknet, pyTorch 等模型格式.
- 提供 AI 开发工具：支持模型快速转换、支持开发板端侧转换 API、支持 TensorFlow, TF Lite, Caffe, ONNX, Darknet, pyTorch 等模型.
- 提供 AI 应用开发接口：提供 NPU 跨平台 API.

2.2 开发流程

NPU 开发完整的流程如下图所示：

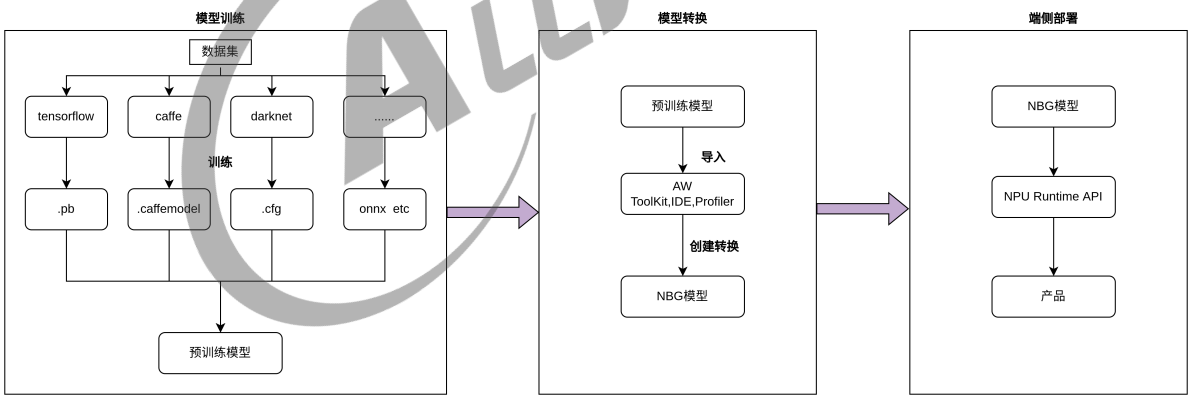

图 2-1: npu_1.png

本篇以 yolov5s 模型为例，来说明混合量化的具体步骤.

2.3 浮点部署

浮点部署的目的是获取 golden 数据，目的是可以和后面混合量化得到的数据比较相似度，来衡量混合量化的效果。

```
pegasus.py import onnx --model yolov5s.onnx --output-data yolov5s.data --output-model
    yolov5s.json

pegasus.py generate inputmeta --model yolov5s.json --input-meta-output yolov5s-inputmeta.
    yml

pegasus.py generate postprocess-file --model yolov5s.json --postprocess-file-output yolov5s
    -postprocess-file.yml

pegasus.py inference --model yolov5s.json --model-data yolov5s.data --batch-size 1 --dtype
    float32  --device CPU --with-input-meta yolov5s-inputmeta.yml --postprocess-file
    yolov5s-postprocess-file.yml

pegasus.py export ovxlib --model yolov5s.json --model-data yolov5s.data --dtype float32 --
    batch-size 1 --save-fused-graph --target-ide-project 'linux64' --with-input-meta
    yolov5s-inputmeta.yml --postprocess-file yolov5s-postprocess-file.yml --output-path
    ovxlib/yolov5s/yolov5sprj --pack-nbg-unify --optimize "VIP9000PICO_PID0XEE" --viv-sdk $
    {VIV_SDK}
```

要注意在第三步完成之后，需要将 input yml 文件的 mean 和 scale 参数修改为符合网络实际的训练时的参数，对于 yolov5s 来讲，scale 需要修改为 0.0039.

图 2-2: scale

结束后，最终得到了输出层的 golden tensor:

```
iter_0_attach_Concat_Concat_303_out0_0_out0_1_25200_85.tensor
```

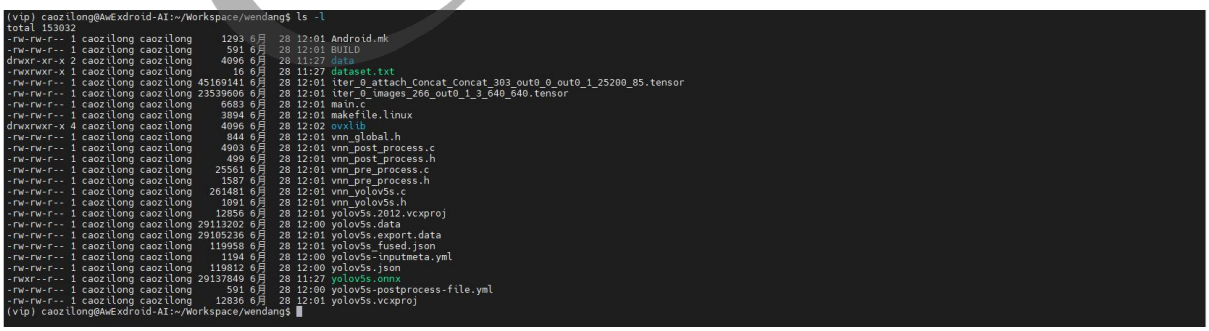

图 2-3: tensor

2.4 混合量化部署

前两步操作相同：

```
pegasus.py import onnx --model yolov5s.onnx --output-data yolov5s.data --output-model
    yolov5s.json

pegasus.py generate inputmeta --model yolov5s.json --input-meta-output yolov5s-inputmeta.
    yml

pegasus.py generate postprocess-file --model yolov5s.json --postprocess-file-output yolov5s
    -postprocess-file.yml
```

之后修改归一化系数，均值，方差 (scale).

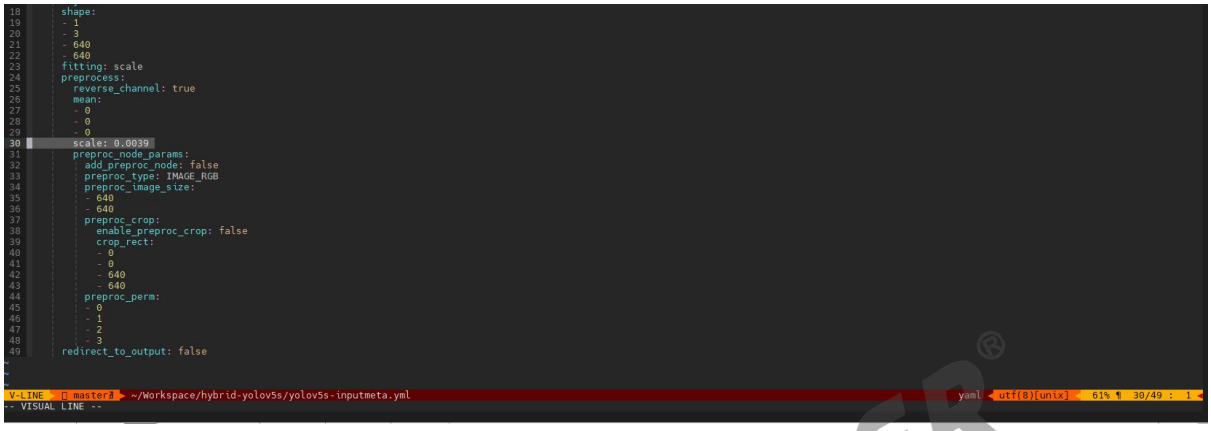

图 2-4: normallize

2.4.1 PCQ+int8 量化

```
pegasus.py quantize --model yolov5s.json --model-data yolov5s.data --batch-size 1 --device
    CPU --with-input-meta yolov5s-inputmeta.yml --rebuild --model-quantize yolov5s.quantize
     --quantizer perchannel_symmetric_affine --qtype int8
```

此步骤中得到量化表文件 yolov5s.quantize，

```
pegasus.py quantize --model yolov5s.json --model-data yolov5s.data --device CPU --with-
    input-meta yolov5s-inputmeta.yml --hybrid --model-quantize yolov5s.quantize --quantizer
     perchannel_symmetric_affine --qtype int8
```

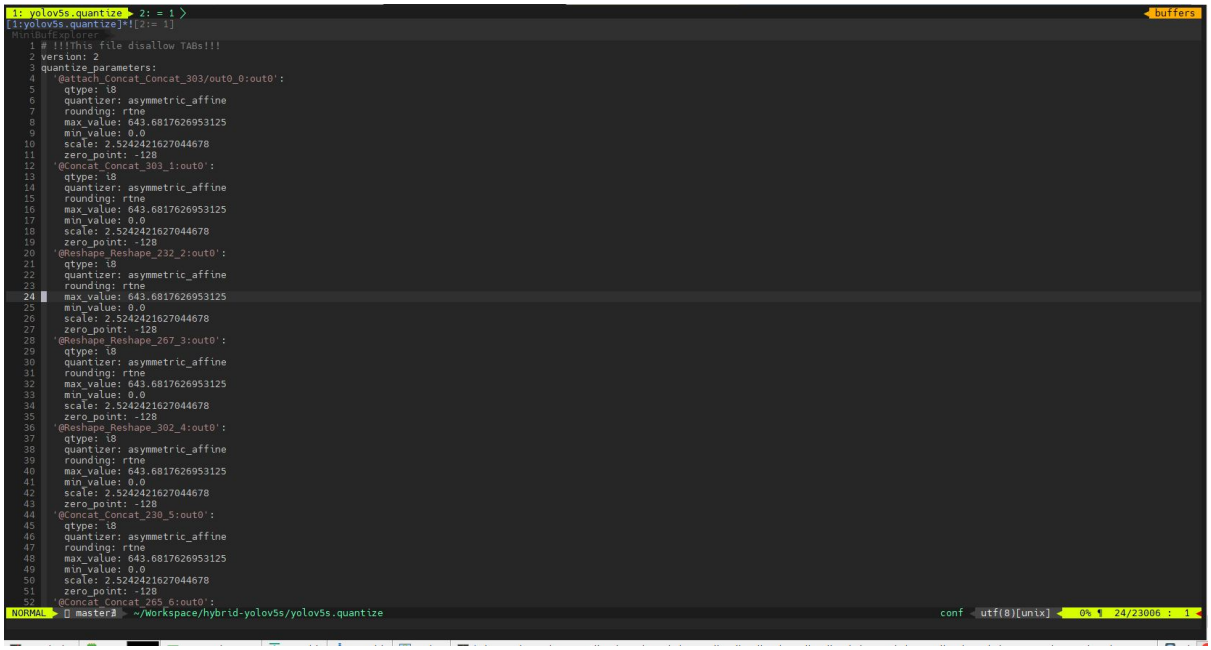

图 2-5: quantilize

YOLOV5S 精度问题的主要原因是，后处理部分也加入到网络中执行了，后处理不太适合量化，量化精度损失很大，如下图所示：

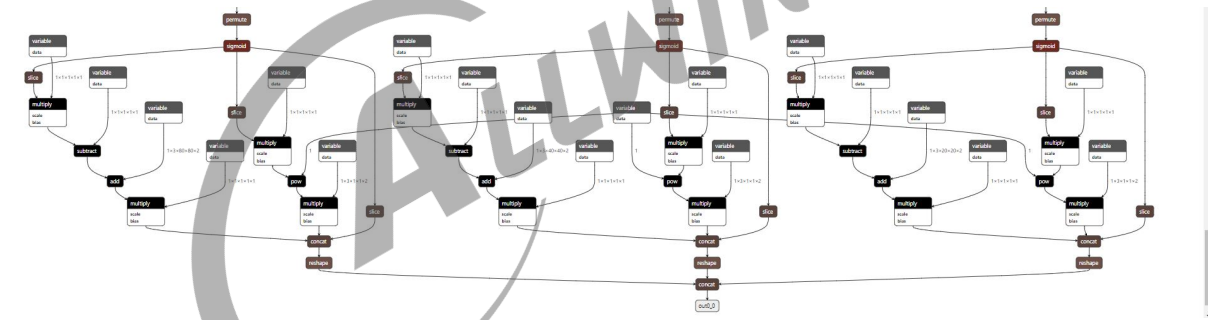

图 2-6: output

permute 下面的层都属于后处理的部分，这部分量化精度损失特别大，需要进行混合量化。

2.4.2 混合量化

修改默认的 yolov5s.quantilize 文件，将 permute 下面需要混合量化的层加入进来，进行 int16 量化。

```
customized_quantize_layers:
Sigmoid_Sigmoid_202_21: dynamic_fixed_point-i16
Initializer_342_62: dynamic_fixed_point-i16
Slice_Slice_207_61: dynamic_fixed_point-i16
Initializer_344_48: dynamic_fixed_point-i16
Mul_Mul_209_47: dynamic_fixed_point-i16
```

```
Sub_Sub_211_32: dynamic_fixed_point-i16
Initializer_346_33: dynamic_fixed_point-i16
Add_Add_213_17: dynamic_fixed_point-i16
Mul_Mul_215_8: dynamic_fixed_point-i16
Initializer_348_18: dynamic_fixed_point-i16
Slice_Slice_220_49: dynamic_fixed_point-i16
Initializer_355_50: dynamic_fixed_point-i16
Mul_Mul_222_34: dynamic_fixed_point-i16
Initializer_460_20: dynamic_fixed_point-i16
Pow_Pow_223_19: dynamic_fixed_point-i16
Mul_Mul_224_9: dynamic_fixed_point-i16
Concat_Concat_230_5: dynamic_fixed_point-i16
Slice_Slice_229_10: dynamic_fixed_point-i16
Reshape_Reshape_232_2: dynamic_fixed_point-i16
Sigmoid_Sigmoid_237_26: dynamic_fixed_point-i16
Initializer_385_66: dynamic_fixed_point-i16
Slice_Slice_242_65: dynamic_fixed_point-i16
Mul_Mul_244_52: dynamic_fixed_point-i16
Initializer_387_53: dynamic_fixed_point-i16
Sub_Sub_246_37: dynamic_fixed_point-i16
Add_Add_248_22: dynamic_fixed_point-i16
Initializer_389_38: dynamic_fixed_point-i16
Mul_Mul_250_11: dynamic_fixed_point-i16
Initializer_391_23: dynamic_fixed_point-i16
Initializer_461_35: dynamic_fixed_point-i16
Initializer_398_55: dynamic_fixed_point-i16
Slice_Slice_255_54: dynamic_fixed_point-i16
Mul_Mul_257_39: dynamic_fixed_point-i16
Pow_Pow_258_24: dynamic_fixed_point-i16
Mul_Mul_259_12: dynamic_fixed_point-i16
Initializer_464_25: dynamic_fixed_point-i16
Slice_Slice_264_13: dynamic_fixed_point-i16
Concat_Concat_265_6: dynamic_fixed_point-i16
Reshape_Reshape_267_3: dynamic_fixed_point-i16
Concat_Concat_303_1: dynamic_fixed_point-i16
attach_Concat_Concat_303/out0_0: dynamic_fixed_point-i16
Sigmoid_Sigmoid_272_31: dynamic_fixed_point-i16
Initializer_428_70: dynamic_fixed_point-i16
Slice_Slice_277_69: dynamic_fixed_point-i16
Mul_Mul_279_57: dynamic_fixed_point-i16
Initializer_430_58: dynamic_fixed_point-i16
Sub_Sub_281_42: dynamic_fixed_point-i16
Initializer_432_43: dynamic_fixed_point-i16
Add_Add_283_27: dynamic_fixed_point-i16
Mul_Mul_285_14: dynamic_fixed_point-i16
Initializer_434_28: dynamic_fixed_point-i16
Initializer_441_45: dynamic_fixed_point-i16
Slice_Slice_290_59: dynamic_fixed_point-i16
Mul_Mul_292_44: dynamic_fixed_point-i16
Initializer_468_30: dynamic_fixed_point-i16
Mul_Mul_294_15: dynamic_fixed_point-i16
Slice_Slice_299_16: dynamic_fixed_point-i16
Concat_Concat_300_7: dynamic_fixed_point-i16
Reshape_Reshape_302_4: dynamic_fixed_point-i16
Pow_Pow_293_29: dynamic_fixed_point-i16
```

```
23025 customized_quantize_layers:
23026   Sigmoid_Sigmoid_202_21: dynamic_fixed_point-i16
23027   Initializer_342_62: dynamic_fixed_point-i16
23028   Slice_Slice_207_61: dynamic_fixed_point-i16
23029   Initializer_344_48: dynamic_fixed_point-i16
23030   Mul_Mul_209_47: dynamic_fixed_point-i16
23031   Sub_Sub_211_32: dynamic_fixed_point-i16
23032   Initializer_346_33: dynamic_fixed_point-i16
23033   Add_Add_213_17: dynamic_fixed_point-i16
23034   Mul_Mul_215_8: dynamic_fixed_point-i16
23035   Initializer_348_18: dynamic_fixed_point-i16
23036   Slice_Slice_220_49: dynamic_fixed_point-i16
23037   Initializer_355_50: dynamic_fixed_point-i16
23038   Mul_Mul_222_34: dynamic_fixed_point-i16
23039   Initializer_460_20: dynamic_fixed_point-i16
23040   Pow_Pow_223_19: dynamic_fixed_point-i16
23041   Mul_Mul_224_9: dynamic_fixed_point-i16
23042   Concat_Concat_230_5: dynamic_fixed_point-i16
23043   Slice_Slice_229_10: dynamic_fixed_point-i16
23044   Reshape_Reshape_232_2: dynamic_fixed_point-i16
23045   Sigmoid_Sigmoid_237_26: dynamic_fixed_point-i16
23046   Initializer_385_66: dynamic_fixed_point-i16
23047   Slice_Slice_242_65: dynamic_fixed_point-i16
23048   Mul_Mul_244_52: dynamic_fixed_point-i16
23049   Initializer_387_53: dynamic_fixed_point-i16
23050   Sub_Sub_246_37: dynamic_fixed_point-i16
23051   Add_Add_248_22: dynamic_fixed_point-i16
23052   Initializer_389_38: dynamic_fixed_point-i16
23053   Mul_Mul_250_11: dynamic_fixed_point-i16
23054   Initializer_391_23: dynamic_fixed_point-i16
23055   Initializer_461_35: dynamic_fixed_point-i16
23056   Initializer_398_55: dynamic_fixed_point-i16
23057   Slice_Slice_255_54: dynamic_fixed_point-i16
23058   Mul_Mul_257_39: dynamic_fixed_point-i16
23059   Pow_Pow_258_24: dynamic_fixed_point-i16
23060   Mul_Mul_259_12: dynamic_fixed_point-i16
23061   Initializer_464_25: dynamic_fixed_point-i16
23062   Slice_Slice_264_13: dynamic_fixed_point-i16
23063   Concat_Concat_265_6: dynamic_fixed_point-i16
23064   Reshape_Reshape_267_3: dynamic_fixed_point-i16
23065   Concat_Concat_303_1: dynamic_fixed_point-i16
23066   attach_Concat_Concat_303/out0_0: dynamic_fixed_point-i16
23067   Sigmoid_Sigmoid_272_31: dynamic_fixed_point-i16
23068   Initializer_428_70: dynamic_fixed_point-i16
23069   Slice_Slice_277_69: dynamic_fixed_point-i16
23070   Mul_Mul_279_57: dynamic_fixed_point-i16
23071   Initializer_430_58: dynamic_fixed_point-i16
23072   Sub_Sub_281_42: dynamic_fixed_point-i16
23073   Initializer_432_43: dynamic_fixed_point-i16
```

图 2-7: mix

```
23073   Initializer_432_43: dynamic_fixed_point-i16
23074   Add_Add_283_27: dynamic_fixed_point-i16
23075   Mul_Mul_285_14: dynamic_fixed_point-i16
23076   Initializer_434_28: dynamic_fixed_point-i16
23077   Initializer_441_45: dynamic_fixed_point-i16
23078   Slice_Slice_290_59: dynamic_fixed_point-i16
23079   Mul_Mul_292_44: dynamic_fixed_point-i16
23080   Initializer_468_30: dynamic_fixed_point-i16
23081   Mul_Mul_294_15: dynamic_fixed_point-i16
23082   Slice_Slice_299_16: dynamic_fixed_point-i16
23083   Concat_Concat_300_7: dynamic_fixed_point-i16
23084   Reshape_Reshape_302_4: dynamic_fixed_point-i16
23085   Pow_Pow_293_29: dynamic_fixed_point-i16
NORMAL  master  ~/Workspace/yolov5s-custom/yolov5s.quantize                    conf  utf8[unix]  100% ¶23085/23085 :  3
```

图 2-8: mix

2.4.3 执行混合量化

```
pegasus.py quantize --model yolov5s.json --model-data yolov5s.data --device CPU --with-
    input-meta yolov5s-inputmeta.yml --hybrid --model-quantize yolov5s.quantize --quantizer
     perchannel_symmetric_affine --qtype int8
```

```
D Quantize tensor(@Resize_Resize_118_174:out0) with tensor(@Sigmoid_Sigmoid_115_120_Mul_Mul_116_105:out0)
D Quantize tensor(@Resize_Resize_140_108:out0) with tensor(@Sigmoid_Sigmoid_137_115_Mul_Mul_138_101:out0)
D Quantize tensor(@Reshape_Reshape_200_51_acuity_mark_perm_268:out0) with tensor(@Conv_Conv_198_64:out0)
D Quantize tensor(@Reshape_Reshape_200_51:out0) with tensor(@Reshape_Reshape_200_51_acuity_mark_perm_268:out0)
D Quantize tensor(@Slice_Slice_229_10:out0) with tensor(@Sigmoid_Sigmoid_202_21:out0)
D Quantize tensor(@Slice_Slice_207_61:out0) with tensor(@Sigmoid_Sigmoid_202_21:out0)
D Quantize tensor(@Reshape_Reshape_232_2:out0) with tensor(@Concat_Concat_230_5:out0)
D Quantize tensor(@Reshape_Reshape_235_56_acuity_mark_perm_269:out0) with tensor(@Conv_Conv_233_68:out0)
D Quantize tensor(@Reshape_Reshape_235_56:out0) with tensor(@Reshape_Reshape_235_56_acuity_mark_perm_269:out0)
D Quantize tensor(@Slice_Slice_264_13:out0) with tensor(@Sigmoid_Sigmoid_237_26:out0)
D Quantize tensor(@Slice_Slice_242_65:out0) with tensor(@Sigmoid_Sigmoid_237_26:out0)
D Quantize tensor(@Reshape_Reshape_267_3:out0) with tensor(@Concat_Concat_265_6:out0)
D Quantize tensor(@Reshape_Reshape_270_60_acuity_mark_perm_267:out0) with tensor(@Conv_Conv_268_63:out0)
D Quantize tensor(@Reshape_Reshape_270_60:out0) with tensor(@Reshape_Reshape_270_60_acuity_mark_perm_267:out0)
D Quantize tensor(@Slice_Slice_299_16:out0) with tensor(@Sigmoid_Sigmoid_272_31:out0)
D Quantize tensor(@Slice_Slice_277_69:out0) with tensor(@Sigmoid_Sigmoid_272_31:out0)
D Quantize tensor(@attach_Concat_Concat_303/out0_0:out0) with tensor(@Concat_Concat_303_1:out0)
I End quantization...
I Dump net to yolov5s.quantize.json
I Dump net quantize tensor table to yolov5s.quantize
I Save net to yolov5s.data
I ----------------Error(0),Warning(0)----------------
(vip) caozilong@AwExdroid-AI:~/Workspace/hybrid-yolov5s$
(vip) caozilong@AwExdroid-AI:~/Workspace/hybrid-yolov5s$
```

图 2-9: hybrid

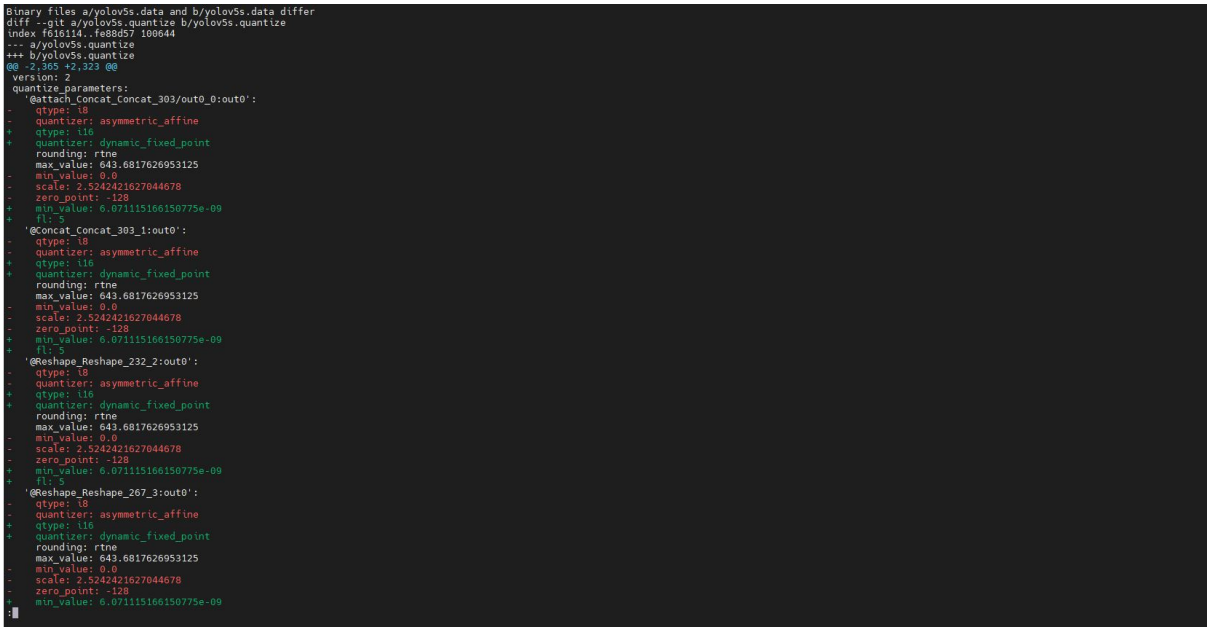

图 2-10: diff 变化

执行后，可以看到量化层输出的变化。

2.5 推理

```
pegasus.py inference --model yolov5s.json --model-data yolov5s.data --batch-size 1 --dtype
    quantized --model-quantize yolov5s.quantize  --device CPU --with-input-meta yolov5s-
    inputmeta.yml --postprocess-file yolov5s-postprocess-file.yml
```

2.6 模型导出

```
pegasus.py export ovxlib --model yolov5s.quantize.json --model-data yolov5s.data --dtype
    quantized --model-quantize yolov5s.quantize --batch-size 1 --save-fused-graph --target-
    ide-project 'linux64' --with-input-meta yolov5s-inputmeta.yml --postprocess-file
    yolov5s-postprocess-file.yml --output-path ovxlib/yolov5s/yolov5sprj --pack-nbg-unify
    --optimize "VIP9000PICO_PID0XEE" --viv-sdk ${VIV_SDK}
```

2.7 相似度对比

将前面生成的 golden tensor 和此时生成的输出 tensor 对比余弦相似度：

```
python /home/caozilong/VeriSilicon/acuity-toolkit-whl-6.6.1/bin/tools/
    compute_tensor_similarity.py ./iter_0_attach_Concat_Concat_303_out0_0_out0_1_25200_85.
    tensor ../wendang/iter_0_attach_Concat_Concat_303_out0_0_out0_1_25200_85.tensor
```

```
(vip) caozilong@AwExdroid-AI:~/Workspace/hybrid-yolov5s$ python /home/caozilong/VeriSilicon/acuity-toolkit-whl-6.6.1/bin/tools/compute_tensor_similarity.py ./iter_0_attach_Concat_Concat_303_out0_0_out0_1_25
200_85.tensor ../wendang/iter_0_attach_Concat_Concat_303_out0_0_out0_1_25200_85.tensor
2022-06-28 12:51:20.381330: W tensorflow/stream_executor/platform/default/dso_loader.cc:59] Could not load dynamic library 'libcudart.so.10.1'; dlerror: libcudart.so.10.1: cannot open shared object file: No
 such file or directory
2022-06-28 12:51:20.381379: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have a GPU set up on your machine.
2022-06-28 12:51:45.635406: W tensorflow/stream_executor/platform/default/dso_loader.cc:59] Could not load dynamic library 'libcuda.so.1'; dlerror: libcuda.so.1: cannot open shared object file: No such file
 or directory
2022-06-28 12:51:45.635483: W tensorflow/stream_executor/cuda/cuda_driver.cc:312] failed call to cuInit: UNKNOWN ERROR (303)
2022-06-28 12:51:45.635532: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel driver does not appear to be running on this host (AwExdroid-AI): /proc/driver/nvidia/version does not exist
2022-06-28 12:51:45.665360: I tensorflow/core/platform/profile_utils/cpu_utils.cc:104] CPU Frequency: 2399990000 Hz
2022-06-28 12:51:45.667511: I tensorflow/compiler/xla/service/service.cc:168] XLA service 0x55989650ad70 initialized for platform Host (this does not guarantee that XLA will be used). Devices:
2022-06-28 12:51:45.667562: I tensorflow/compiler/xla/service/service.cc:176]   StreamExecutor device (0): Host, Default Version
WARNING:tensorflow:From /home/caozilong/anaconda3/envs/vip/lib/python3.6/site-packages/tensorflow/python/util/dispatch.py:201: calling cosine_distance (from tensorflow.python.ops.losses.losses_impl) with di
m is deprecated and will be removed in a future version.
Instructions for updating:
dim is deprecated, use axis instead
euclidean_distance 1127.4191
cos_similarity 0.999912
(vip) caozilong@AwExdroid-AI:~/Workspace/hybrid-yolov5s$
(vip) caozilong@AwExdroid-AI:~/Workspace/hybrid-yolov5s$
(vip) caozilong@AwExdroid-AI:~/Workspace/hybrid-yolov5s$
(vip) caozilong@AwExdroid-AI:~/Workspace/hybrid-yolov5s$
```

图 2-11：diff 变化

可以看到余弦相似度还是非常高的，达到了 0.999912，混合量化部署步骤到此结束。

3 结束

著作权声明

本文档及内容受著作权法保护，其著作权由珠海全志科技股份有限公司（“全志”）拥有并保留一切权利。

本文档是全志的原创作品和版权财产，未经全志书面许可，任何单位和个人不得擅自摘抄、复制、修改、发表或传播本文档内容的部分或全部，且不得以任何形式传播。

商标声明

（不完全列举）均为珠海全志科技股份有限公司的商标或者注册商标。在本文档描述的产品中出现的其它商标，产品名称，和服务名称，均由其各自所有人拥有。

免责声明

您购买的产品、服务或特性应受您与珠海全志科技股份有限公司（“全志”）之间签署的商业合同和条款的约束。本文档中描述的全部或部分产品、服务或特性可能不在您所购买或使用的范围内。使用前请认真阅读合同条款和相关说明，并严格遵循本文档的使用说明。您将自行承担任何不当使用行为（包括但不限于如超压，超频，超温使用）造成的不利后果，全志概不负责。

本文档作为使用指导仅供参考。由于产品版本升级或其他原因，本文档内容有可能修改，如有变更，恕不另行通知。全志尽全力在本文档中提供准确的信息，但并不确保内容完全没有错误，因使用本文档而发生损害（包括但不限于间接的、偶然的、特殊的损失）或发生侵犯第三方权利事件，全志概不负责。本文档中的所有陈述、信息和建议并不构成任何明示或暗示的保证或承诺。

本文档未以明示或暗示或其他方式授予全志的任何专利或知识产权。在您实施方案或使用产品的过程中，可能需要获得第三方的权利许可。请您自行向第三方权利人获取相关的许可。全志不承担也不代为支付任何关于获取第三方许可的许可费或版税（专利税）。全志不对您所使用的第三方许可技术做出任何保证、赔偿或承担其他义务。