Tina Linux NPU AI 模型混合量化指导

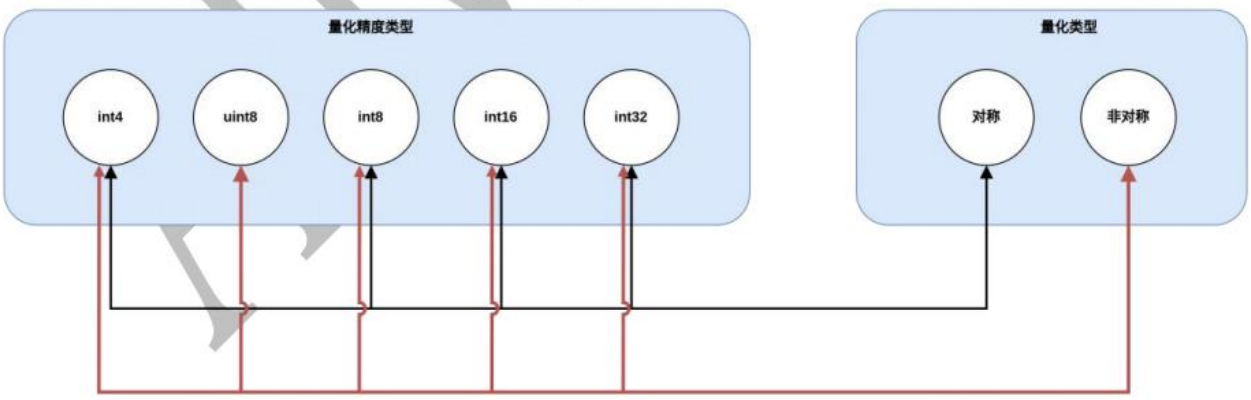
版本号	日期	制/修订人	内容描述
1.0	2022.09.14	AWA1911	初版，模型混合量化操作

1、环境设置

export VIV_SDK=/home/gonghao/VeriSilicon/**VivanteIDE5.7.0**/cmdtools
export ACUITY_PATH=/home/gonghao/project/Verisilicon_Tool_Acuity_Toolkit/acuity/**acuity-toolkit-binary-6.6.6**/bin
export pegasus=/home/gonghao/project/Verisilicon_Tool_Acuity_Toolkit/acuity/**acuity-toolkit-binary-6.6.6**/bin/pegasus

2、混合量化概念

人脸比对网络量化效果不好，由此引出的混合量化概念，究竟什么是混合量化，我们图示表示：



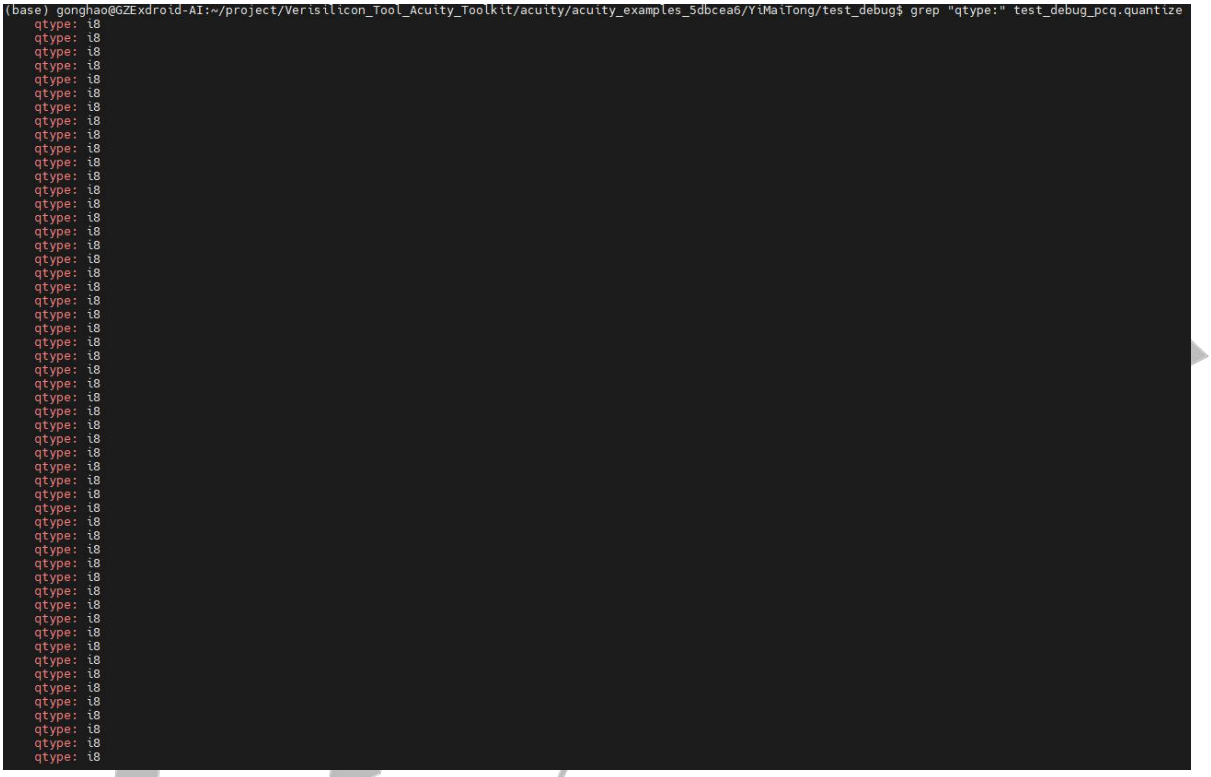
约束：量化的精度和类型组合并不是任意的。对称量化只能和 INT4，INT8，INT16，INT32 等以 0 为中心点的类型搭配，非对称量化可以和任何类型搭配。
混合量化：混合量化是另外一个概念，准确的说，混合量化是指混合精度量化，是指在满足以上约束的前提下，不同精度类型的量化模式同时使用，比如对称 INT16 类型和非对称 INT8 类型，或者非对称 INT16 类型和对称 INT8 类型等等。
使用场景：如果想用 int8 量化，然后发现有的层量化后，模型性能下降，就可以尝试用 int16 去量化这一层。

由于 perchannel_summetric_affine int8 量化的方式（pcq 就是 int8 的 perchannel 量化方式）仍然是 8bit 量化，精度不变，所以，uint8+int8 的量化方式不算是混合量化。

参数配置：混合量化是 acuity tools 具备的功能，准确地说，它是通过$pegasus **quantilize --hybrid** 参数触发生效的。

3、量化效果对比

单使用 pcq 量化



使用 pcq +int16 混合量化

```
(base) gonghao@GZExdroid-AI:~/project/Verisilicon_Tool_Acuity_Toolkit/acuity/acuity-toolkit-binary-6.6.6/test_debug$ grep -rn "qtype: "  test_debug_pcq.quantize
5:    qtype: i8
13:    qtype: i8
21:    qtype: i8
29:    qtype: i16
36:    qtype: i8
44:    qtype: i8
52:    qtype: i8
60:    qtype: i16
67:    qtype: i16
74:    qtype: i8
82:    qtype: i8
90:    qtype: i8
98:    qtype: i16
105:    qtype: i8
113:    qtype: i16
120:    qtype: i8
128:    qtype: i8
136:    qtype: i16
143:    qtype: i8
151:    qtype: i8
159:    qtype: i8
167:    qtype: i8
175:    qtype: i16
182:    qtype: i8
190:    qtype: i8
198:    qtype: i8
206:    qtype: i8
214:    qtype: i8
222:    qtype: i8
230:    qtype: i8
238:    qtype: i8
246:    qtype: i16
253:    qtype: i16
260:    qtype: i8
268:    qtype: i8
276:    qtype: i8
284:    qtype: i8
292:    qtype: i16
299:    qtype: i8
307:    qtype: i8
315:    qtype: i8
323:    qtype: i8
331:    qtype: i8
339:    qtype: i8
347:    qtype: i8
355:    qtype: i8
363:    qtype: i8
371:    qtype: i8
379:    qtype: i8
387:    qtype: i8
395:    qtype: i8
403:    qtype: i8
411:    qtype: i8
419:    qtype: i8
427:    qtype: i8
435:    qtype: i16
```

从量化效果来看，pcq 量化效果不佳，int16 量化效果很好，但模型运行起来估计很慢，计算量大。

```
(base) gonghao@GZExdroid-AI:~/project/Verisilicon_Tool_Acuity_Toolkit/acuity/acuity_examples_5dbcea6/YiMaiTong/test_debug/inf$ python compute_tensor_similarity.py test_debug_non-quantized/iter_0_
iter_0_attach_BatchNormalization_BatchNormalization_172_out0_0_out0_1_512.tensor  iter_0_input.1_173_out0_1_3_112_112.tensor
(base) gonghao@GZExdroid-AI:~/project/Verisilicon_Tool_Acuity_Toolkit/acuity/acuity_examples_5dbcea6/YiMaiTong/test_debug/inf$ python compute_tensor_similarity.py test_debug_non-quantized/iter_0_attac
h_BatchNormalization_BatchNormalization_172_out0_0_out0_1_512.tensor test_debug_pcq/iter_0_attach_BatchNormalization_BatchNormalization_172_out0_0_out0_1_512.tensor
2022-08-30 19:18:02.835340: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.so.11.0: cannot open shared object fi
le: No such file or directory
2022-08-30 19:18:02.835384: 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.
2022-08-30 19:18:05.079849: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'libcuda.so.1'; dlerror: libcuda.so.1: cannot open shared object file: No suc
h file or directory
2022-08-30 19:18:05.079895: W tensorflow/stream_executor/cuda/cuda_driver.cc:269] failed call to cuInit: UNKNOWN ERROR (303)
2022-08-30 19:18:05.079924: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel driver does not appear to be running on this host (GZExdroid-AI): /proc/driver/nvidia/version does not exi
st
WARNING:tensorflow:From /home/gonghao/anaconda3/lib/python3.8/site-packages/tensorflow/python/util/dispatch.py:1096: calling cosine_distance (from tensorflow.python.ops.losses.losses_impl) with dim is
 deprecated and will be removed in a future version.
Instructions for updating:
dim is deprecated, use axis instead
euclidean_distance 11.732007
cos_similarity 0.894119
(base) gonghao@GZExdroid-AI:~/project/Verisilicon_Tool_Acuity_Toolkit/acuity/acuity_examples_5dbcea6/YiMaiTong/test_debug/inf$ python compute_tensor_similarity.py test_debug_non-quantized/iter_0_attac
h_BatchNormalization_BatchNormalization_172_out0_0_out0_1_512.tensor test_debug_int16/iter_0_attach_BatchNormalization_BatchNormalization_172_out0_0_out0_1_512.tensor
2022-08-30 19:19:17.054781: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.so.11.0: cannot open shared object fi
le: No such file or directory
2022-08-30 19:19:17.054823: 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.
2022-08-30 19:19:19.236740: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'libcuda.so.1'; dlerror: libcuda.so.1: cannot open shared object file: No suc
h file or directory
2022-08-30 19:19:19.236780: W tensorflow/stream_executor/cuda/cuda_driver.cc:269] failed call to cuInit: UNKNOWN ERROR (303)
2022-08-30 19:19:19.236809: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel driver does not appear to be running on this host (GZExdroid-AI): /proc/driver/nvidia/version does not exi
st
WARNING:tensorflow:From /home/gonghao/anaconda3/lib/python3.8/site-packages/tensorflow/python/util/dispatch.py:1096: calling cosine_distance (from tensorflow.python.ops.losses.losses_impl) with dim is
 deprecated and will be removed in a future version.
Instructions for updating:
dim is deprecated, use axis instead
euclidean_distance 0.200659
cos_similarity 0.999964
(base) gonghao@GZExdroid-AI:~/project/Verisilicon_Tool_Acuity_Toolkit/acuity/acuity_examples_5dbcea6/YiMaiTong/test_debug/inf$
```

Uint8 量化效果最不好。

```
compute_tensor_similarity.py  test_debug_int16  test_debug_non-quantized  test_debug_pcq  test_debug_uint8
(base) gonghao@GZExdroid-AI:~/project/Verisilicon_Tool_Acuity_Toolkit/acuity/acuity_examples_5dbcea6/YiMaiTong/test_debug/inf$ python compute_tensor_similarity.py test_debug_non-quantized/iter_0_attac
h_BatchNormalization_BatchNormalization_172_out0_0_out0_1_512.tensor test_debug_uint8/iter_0_attach_BatchNormalization_BatchNormalization_172_out0_0_out0_1_512.tensor
2022-08-30 20:23:23.022772: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.so.11.0: cannot open shared object fi
le: No such file or directory
2022-08-30 20:23:23.022812: 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.
2022-08-30 20:23:26.060022: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'libcuda.so.1'; dlerror: libcuda.so.1: cannot open shared object file: No suc
h file or directory
2022-08-30 20:23:26.060068: W tensorflow/stream_executor/cuda/cuda_driver.cc:269] failed call to cuInit: UNKNOWN ERROR (303)
2022-08-30 20:23:26.060098: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel driver does not appear to be running on this host (GZExdroid-AI): /proc/driver/nvidia/version does not exi
st
WARNING:tensorflow:From /home/gonghao/anaconda3/lib/python3.8/site-packages/tensorflow/python/util/dispatch.py:1096: calling cosine_distance (from tensorflow.python.ops.losses.losses_impl) with dim is
 deprecated and will be removed in a future version.
Instructions for updating:
dim is deprecated, use axis instead
euclidean_distance 21.133018
cos_similarity 0.715735
(base) gonghao@GZExdroid-AI:~/project/Verisilicon_Tool_Acuity_Toolkit/acuity/acuity_examples_5dbcea6/YiMaiTong/test_debug/inf$
```

Uint8 量化：

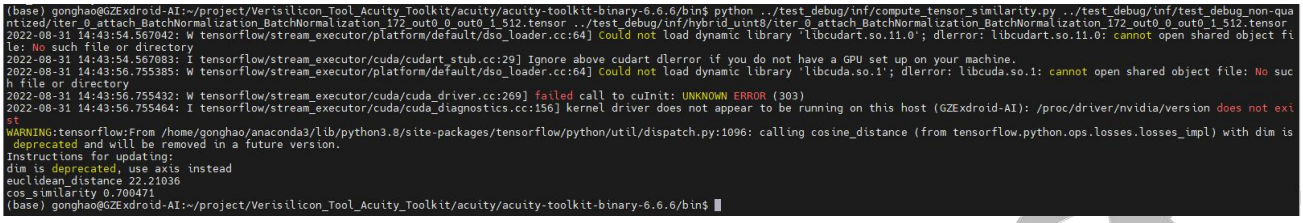
./pegasus quantize --model ../test_debug/test_debug.json --model-data ../test_debug/test_debug.data --batch-size 1 --device CPU --with-input-meta ../test_debug/test_debug_inputmeta.yml --rebuild --iteration 160 --algorithm kl_divergence --model-quantize ../test_debug/test_debug_uint8.quantize --quantizer asymmetric_affine --qtype uint8

再重新执行 uint8+int16 混合量化：

./pegasus **quantize** --model ../test_debug/test_debug.json --model-data ../test_debug/test_debug.data --device CPU --with-input-meta ../test_debug/test_debug_inputmeta.yml --iteration 692 **--hybrid** --model-quantize ../test_debug/test_debug_uint8.quantize --quantizer asymmetric_affine --qtype uint8

混合前项运算

./pegasus inference --model ../test_debug/test_debug_uint8.quantize.json --model-data ../test_debug/test_debug.data --dtype quantized --model-quantize ../test_debug/test_debug_uint8.quantize --device CPU --output-dir ../test_debug/inf/hybrid_uint8 --with-input-meta ../test_debug/test_debug_inputmeta.yml



Int8 PCQ 量化：

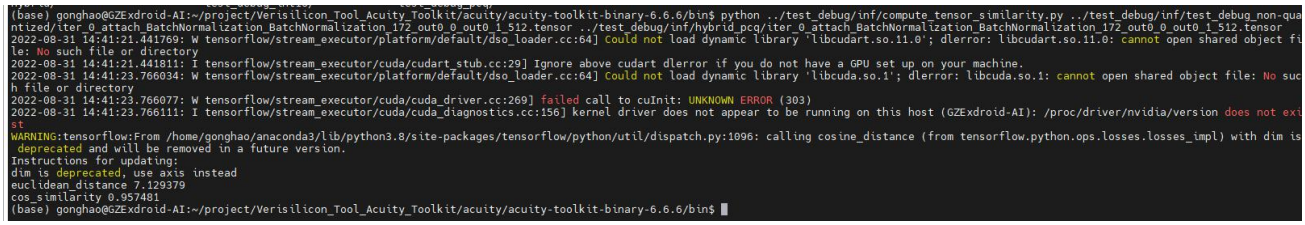
./pegasus quantize --model ../test_debug/test_debug.json --model-data ../test_debug/test_debug.data --batch-size 1 --device CPU --with-input-meta ../test_debug/test_debug_inputmeta.yml --rebuild --iteration 692 --algorithm kl_divergence --model-quantize ../test_debug/test_debug_pcq.quantize --quantizer perchannel_symmetric_affine --qtype pcq

再重新执行 pcq 混合量化：

./pegasus **quantize** --model ../test_debug/test_debug.json --model-data ../test_debug/test_debug.data --device CPU --with-input-meta ../test_debug/test_debug_inputmeta.yml --iteration 692 **--hybrid** --model-quantize ../test_debug/test_debug_pcq.quantize --quantizer asymmetric_affine --qtype int8

混合前项运算

./pegasus inference --model ../test_debug/test_debug_pcq.quantize.json --model-data ../test_debug/test_debug.data --dtype quantized --model-quantize ../test_debug/test_debug_pcq.quantize --device CPU --output-dir ../test_debug/inf/hybrid_pcq --with-input-meta ../test_debug/test_debug_inputmeta.yml



从效果来看使用混合量化中的 pcq +int16 量化的效果比较好

混合量化转换 nb

./pegasus export ovxlib --model ../test_debug/test_debug_pcq.quantize.json

--model-data ../test_debug/test_debug.data --dtype quantized
--model-quantize ../test_debug/test_debug_pcq.quantize --batch-size 1 --save-fused-graph --target-ide-project 'linux64' --with-input-meta ../test_debug/test_debug_inputmeta.yml
--output-path ../test_debug/ovxilb/test_debug/test_debugprj --pack-nbg-unify
--postprocess-file ../test_debug/test_debug_postprocess_file.yml --optimize "VIP9000PICO_PID0XEE" --viv-sdk ${VIV_SDK}