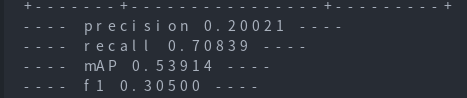
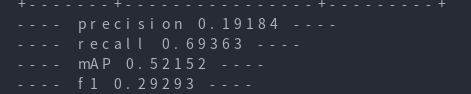
origin-5k

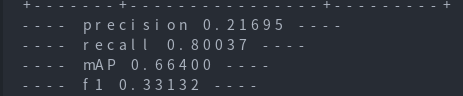


quantisim:

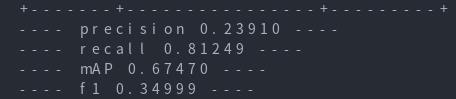
5k



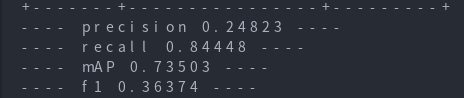
0.6k



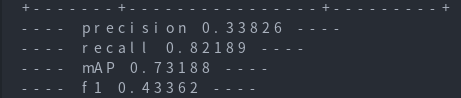
0.2k



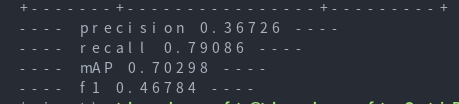
0.1k



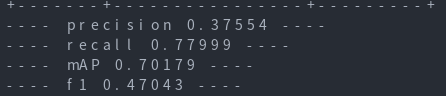
40



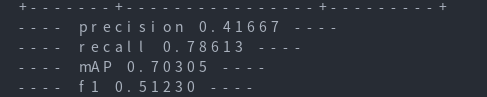
40 : BNF



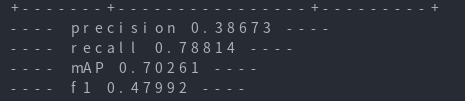
40 : CLE



40: BNF--CLE--BC--QAT



40: BNF--AdaRound-QAT



在aimet的demo中，选取yolov3网络，对其pytorch版本的模型进行了量化，yolov3的darknet53中包含了大量的DBL结构(由Convolution + BatchNormalization + LeakyReLU节点组成)，期待这个结构在aimet的BNF操作下会有一些良好的结果，这是选择aimet进行量化的主要原因。

In the demo of AIMET, yolov3 network was selected and its PyTorch version model was quantified. Darknet53 of Yolov3 includes a large number of DBL structures (consisting of Convolution + BatchNormalization + LeakyReLU nodes) and hopefully this structure will have some good results under AIMET's BNF operation. This is the main reason why yolov3 was chosen for quantification in the AIMET demo.

使用了COCO数据集，但是在量化过程中并没有使用所有的数据，仅仅使用5k，0.6k,0.2k,0.1k,40（在数据集大小为5k的时候，量化感知训练的各项参数相较原模型均略有下降，同时由于没有足够的硬件资源，为降低量化花费的时间，大幅度减少了数据量）大小的自定义数据，评估函数返回了四个主要的模型评估参数：precision（精度），recall（召回率），mAP（平均精度），F1 Score（模型精确率和召回率的一种调和平均），在数据集为40的时候，我获得了一个较好的F1 Score值，所以后续操作的数据集大小都选定为40。

COCO data set was used, but not all data were used in the quantization process, only 5K, 0.6K, 0.2K, 0.1K and 40 were used (When the size of data set was 5K, all parameters of quantization aware training were slightly reduced compared with the original model. Meanwhile, due to insufficient hardware resources, in order to reduce the quantization time,greatly reduce the amount of data) The evaluation function returns four main model evaluation parameters: Precision, recall, mAP and F1 Score (a harmonic mean of model accuracy and recall). When the data set was 40, I obtained a good F1 Score value, so the size of the data set for subsequent operations was selected as 40.

对yolov3单独进行BNF操作的效果略小于单独进行CLE操作的，但均优于原模型评估参数，因此后续采用了BNF--CLE--BC--QAT和BNF--AdaRound--QAT两种组合操作进行量化操作，两种操作都带来了提升，但显然BNF--CLE--BC--QAT的量化结果更加让人满意。

The effects of BNF operation alone on Yolov3 are slightly less than those of CLE operation alone, but both of them are better than the evaluation parameters of the original model. Therefore, two combined operations of BNF--CLE--BC--QAT and BNF--AdaRound--QAT are adopted for quantization operation, both of which bring improvement. But obviously BNF--CLE--BC--QAT quantification results are more satisfactory.

yolov3：A neural network algorithm for target detection,目标检测神经网络算法

darknet53：yolov3`s backbone，Used for feature extraction,用于特征提取

Convolution + BatchNormalization + LeakyReLU：All three are network ops in deep learning,这三个都是深度学习中的网络操作

BNF：BatchNormalization Fold，An interlayer optimization method to reduce intermediate computation，批处理折叠，一种层间优化方式，减少中间计算

CLE：跨层均衡，Cross Layer Equalization

BC ：偏置校正，Bias Correction

QAT ： 量化感知训练，Quantization Aware Training