



Denglin HammingTM V2

dINNE Quantization Introduction

DL-DG/SW-035A-01

2025-2-10

Copyright©苏州登临科技有限公司，2019 - 2025，版权所有。

未经苏州登临科技有限公司事先书面同意，不得以任何形式或方式复制或传播本文件的任何部分。

商标和许可



和其它苏州登临科技有限公司的其它登临科技的图标为苏州登临科技有限公司的商标。本手册中提及的所有其他商标均为其各自所有者的财产。

通知

所购买的产品、服务和特性由苏州登临科技有限公司与客户签订的合同规定。本文件中描述的所有或部分产品、服务和特性可能不在采购范围或使用范围内。除非合同中另有规定，本文件中的所有声明、信息和建议均按“原样”提供，无任何明示或暗示的保证或陈述。

本手册中的信息如有更改，恕不另行通知。本文件在编制过程中已尽一切努力确保内容的准确性，本文件中的所有声明、信息和建议不构成任何明示或暗示的保证。

苏州登临科技有限公司

苏州工业园区扬富路11号南岸新地一期商务楼5号1101室，江苏，中国

<http://www.denglin.ai>

Email : support@denglin.ai

更新历史

版本	更新描述
01	第一次发布。

章节目录

章节目录

1. 网络模型量化方法

1.1 介绍

1.2 功能

1.3 量化原理

1.4 量化流程

附录

参考文献

相关文档

1. 网络模型量化方法

1.1 介绍

近年来，定点量化使用更少的比特数 (如8-bit、4-bit等) 表示神经网络的权重和激活已被验证是有效的。定点量化的优点包括低内存带宽、低功耗、低计算资源占用以及低模型存储需求等。在整个量化过程中，我们提供了 (基于TensorFlow平台) 基础的工具包，包括模拟量化节点插入，模型转换等。因实际应用中可能存在不同的量化方式，量化值域统计方法等，用户可基于我们基础的工具包，根据自身的需要，进行修改和扩展。

1.2 功能

- 支持 4-bit (int4/uint4), 8-bit (int8/uint8) 的对称和非对称量化。
- 支持混合精度 (int4/uint4, int8/uint8, float16, float32) 量化。
- 支持 Per-Channel 的方式量化。
- 权值/激活量化值域范围可指定。

1.3 量化原理

量化时使用以下公式近似浮点值：

$$r = S(q - Z) \quad (1)$$

其中：r 为浮点数值，q 为整形数值，S (缩放因子)、Z (零点值) 为量化参数。

根据公式(1)，矩阵乘法可以表示为：

$$S_3(q_3^{(i,k)} - Z_3) = \sum_{j=1}^N S_1(q_1^{(i,j)} - Z_1)S_2(q_2^{(j,k)} - Z_2) \quad (2)$$

可重写为如下等式：

$$q_3^{(i,k)} = Z_3 + M \sum_{j=1}^N (q_1^{(i,j)} - Z_1)(q_2^{(j,k)} - Z_2) \quad (3)$$

其中，

$$M := \frac{S_1 S_2}{S_3} \quad (4)$$

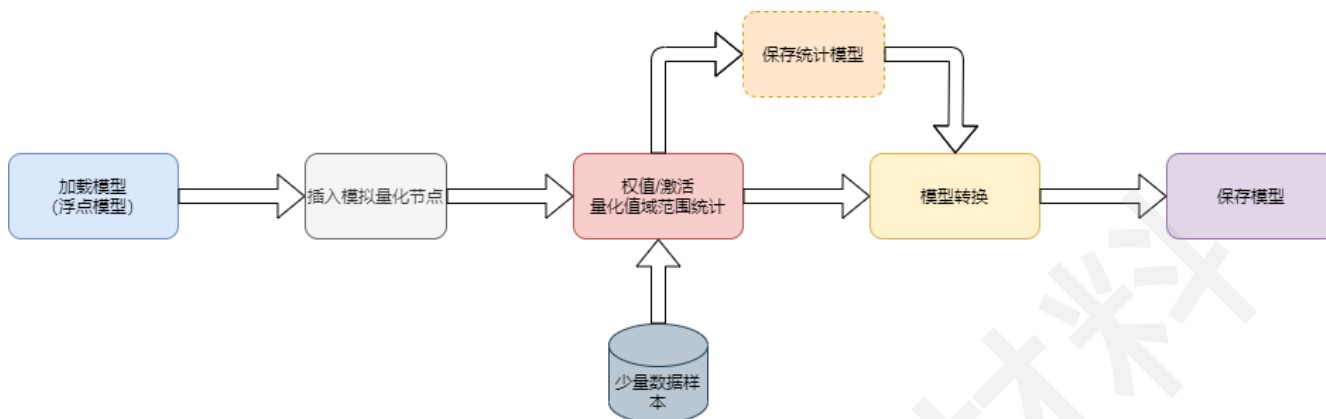
并且计算时M由定点数来表示。

其中的量化参数 (零点值Z和缩放因子S)，可以通过当前浮点 (float32) 数据统计的Min和Max计算得出：

$$S = (r_{max} - r_{min}) / (q_{max} - q_{min})$$

$$Z = q_{min} - (r_{min}) / S$$

1.4 量化流程

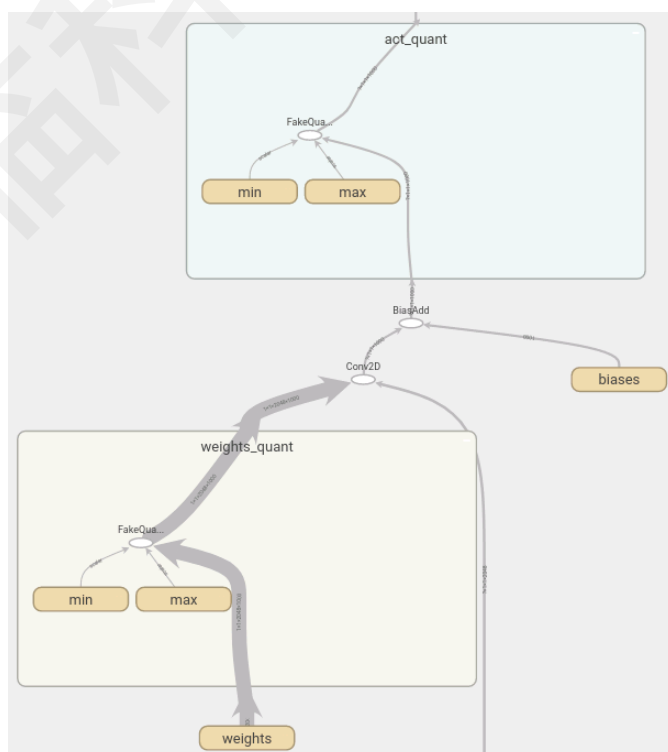


加载模型：加载预先保存的模型文件。模型文件是指通过深度学习框架TensorFlow在完成模型训练并做Frozen后保存的浮点 (float32) 的网络模型，通常是后缀名为 .pb 的文件。

插入模拟量化节点：在模型中需要做量化的节点上插入模拟量化节点。模拟量化节点的作用为：

1. 用浮点 (float32) 模拟权值/激活的量化过程。此过程可以模拟因量化引起的权值/激活的数值变化。
2. 统计权值/激活量化值域范围。通过该值域范围来确定量化参数零点值和缩放因子。
3. 确定需要量化的节点。如果节点上插入了模拟量化节点，模型转换工具默认会将该节点转换为8-bit量化节点；如果节点上未插入模拟量化节点，该节点保持原有精度。

在Conv2D节点上插入模拟量化节点示意图：



权值/激活量化值域范围统计：用少量的数据样本，对插入模拟量化节点后的网络模型做推断过程，以统计权值/激活的量化值域范围。通过该值域范围来确定量化的参数零点值和缩放因子。

保存统计模型：可将统计后的模型进行保存，以便后续使用。其中的量化值域范围统计值也将保存在模型文件中。

模型转换：通过模型转换工具，将统计后的模型转换成推理阶段可加载的模型。该过程会执行量化参数计算、权值量化转换、模拟量化节点消除、量化参数传播等操作。转换后的模型中可能会包含如 [附录](#) 中列举的自定义的算子。

保存模型：将转换后的模型进行保存，以便后续推理引擎使用。

附录

量化后的模型中可能包含的量化相关的自定义的算子如下所示：

算子
DLQuantizedMatMul
DLQuantizedBatchMatMul
DLQuantizedConv2D
DLQuantizedDepthwiseConv2dNative
DLQuantizedConv2dBackpropInput
DLQuantizedConv3D
DLQuantize
DLDequantize

参考文献

【1】Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference

相关文档

- dINNE-Quant-Quantization-Tutorial-with-ResNet-V1-50.pdf 中的 **ResNet-V1-50模型量化教程**
- dINNE-TU-Operator.pdf 中的 **TU (Tensor Unit) 算子介绍**