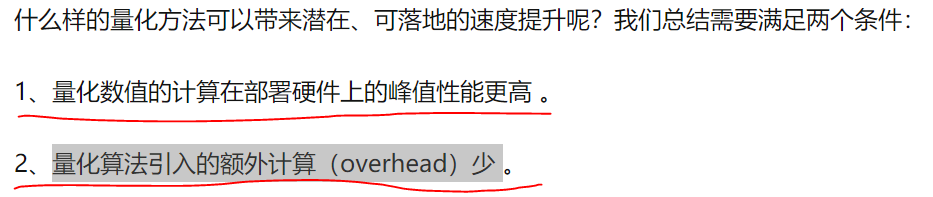
量化论文

[神经网络量化简介 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/64744154)

[模型量化了解一下？ - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/132561405)



Fbgemm

18年年初Google tensorflow-tflite

Ristretto

Google IAO 算法

下面这个连接讲的一般，还是之前那篇论文讲的好一些

[神经网络量化之 Ristretto、增量量化INQ、IAO代码实战分析\_ristretto 量化\_EwenWanW的博客-CSDN博客](https://blog.csdn.net/xiaoxiaowenqiang/article/details/81713131)

[一文搞懂模型量化算法 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/505570612)

A close up of text

Description automatically generated

A screenshot of a phone

Description automatically generated

浮点数、定点数

[计算机中的定点数与浮点数 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/410047526)

## 量化之 Ristretto

## 增量量化INQ

## IAO代码实战分析

Pytorch中的量化

[PyTorch的量化 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/299108528)