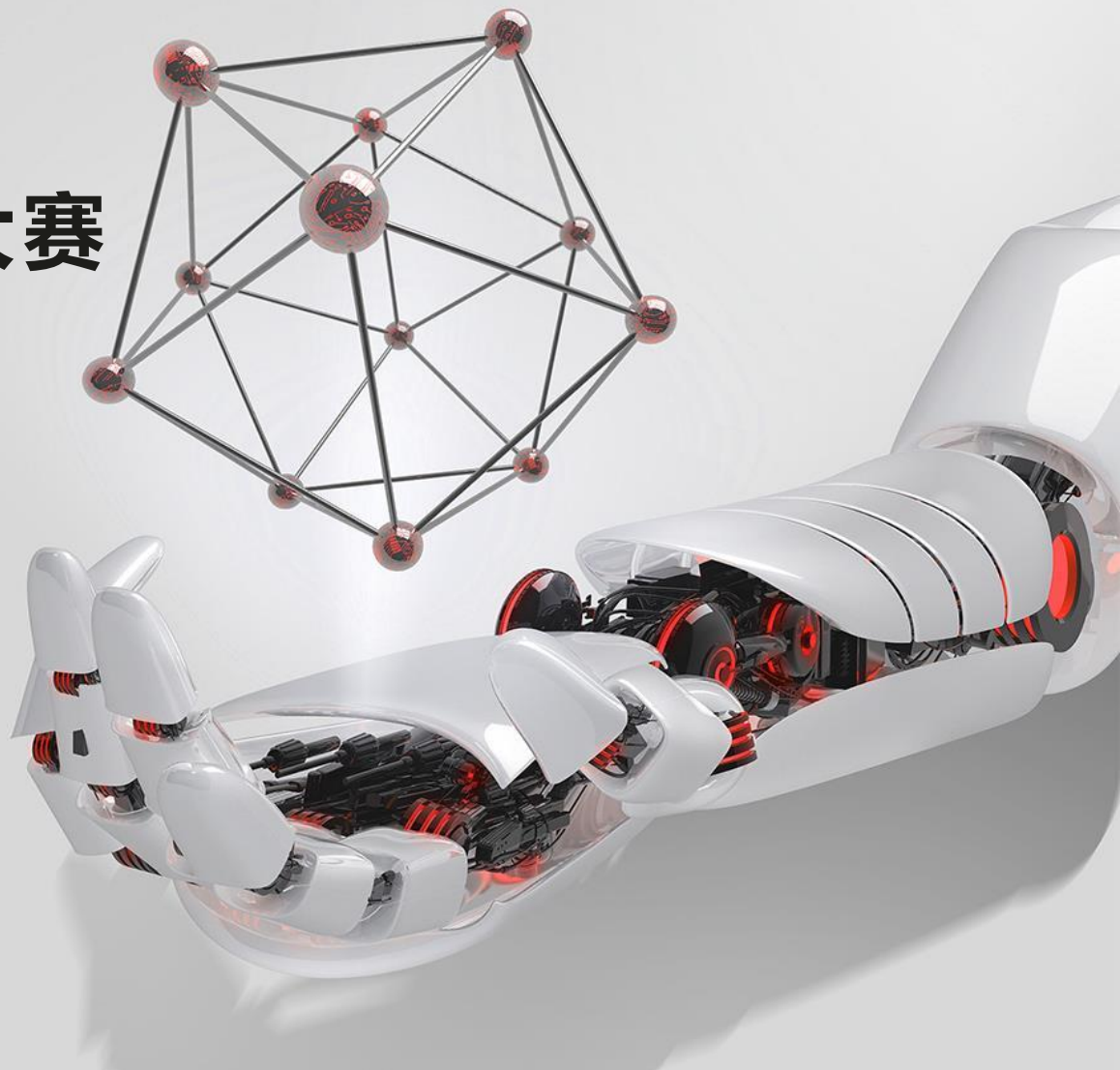


全国大学生嵌入式暨智能互联大赛 海思赛道赋能课件

NNIE精度损失问题解决方法

主讲人：王振坤





**每一位开发者
都是海思要汇聚的星星之火**

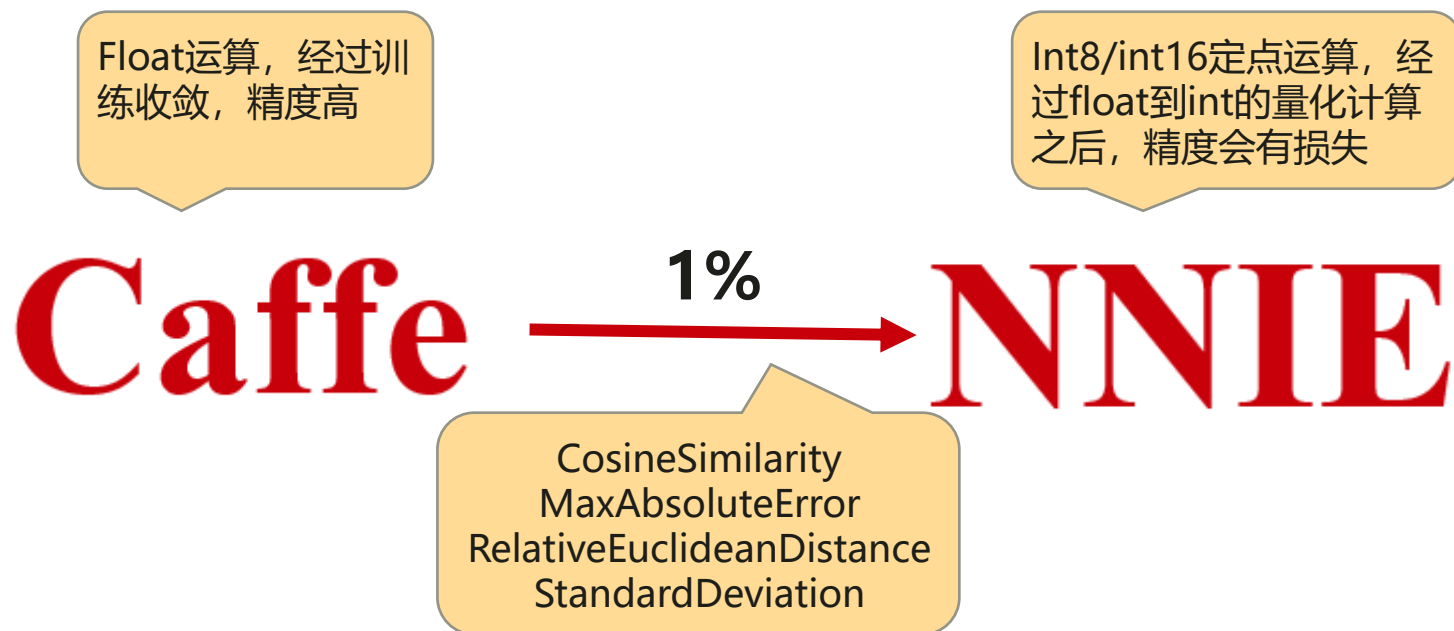
目录

- ① NNIE精度损失简介
- ② 常见情况与解决方法

NNIE精度损失简介

精度损失是指板端feature map和caffe中间结果的余弦相似度有差异，这种差异是经过量化之后产生的。Caffe或其他框架使用float运算，NNIE使用int8或int16运算，所以会有精度损失。海思量化算法的设计目标是精度损失控制在1%以内，大多数网络的精度损失要小于0.5%。

精度损失一般通过比较向量余弦相似度来确定，其他的参考值还有绝对误差、标准差以及欧式距离。板端输出feature map的过程比较繁琐，所以相似度一般比较仿真和caffe每层的相似度。



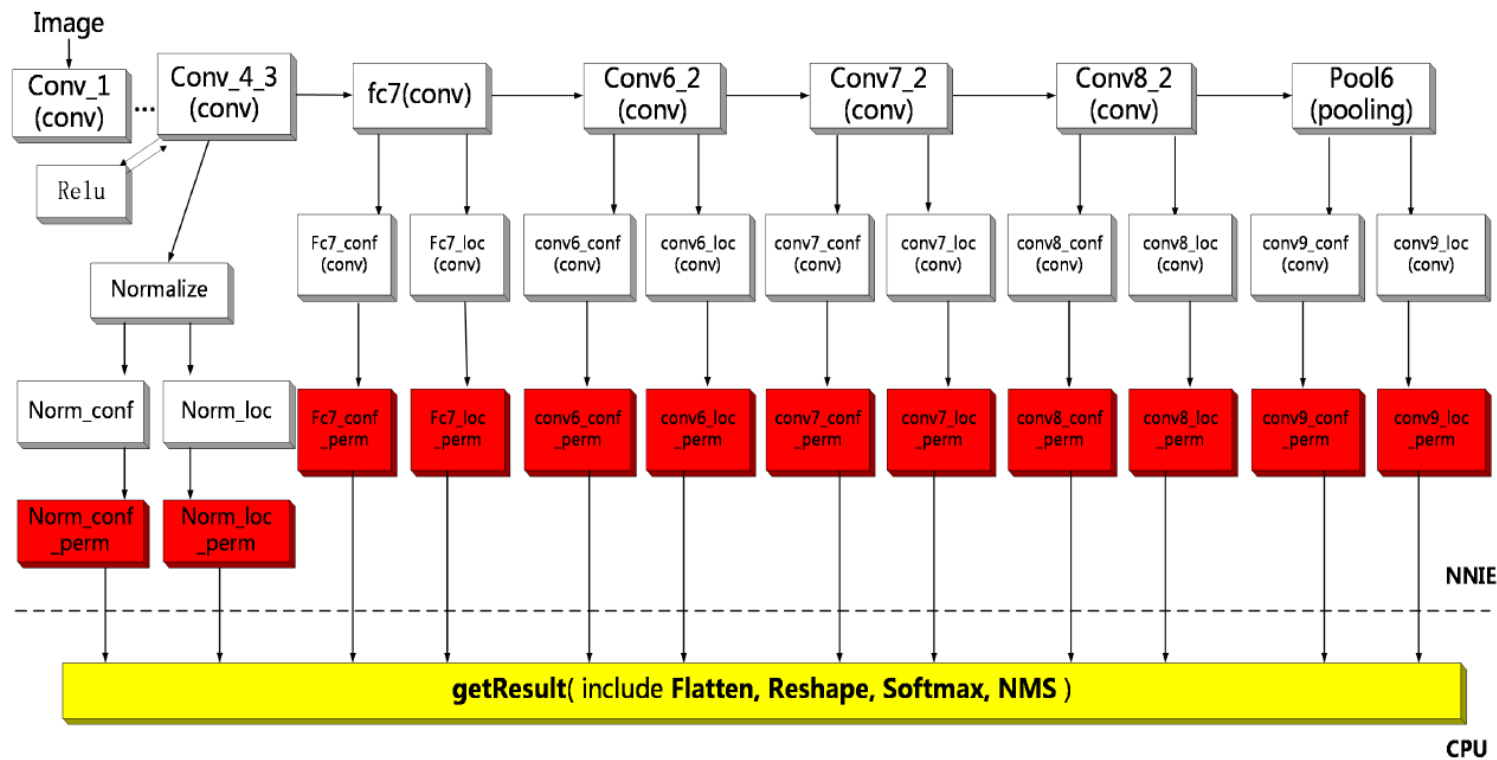
目录

- ① NNIE精度损失简介
- ② 常见情况与解决方法

NNIE精度损失常见情况

常见的精度损失情况有以下几种：

1. Data层的数据相似度对比达不到0.99；
2. Data层是0.99+，后面逐层下降，最后一层小于0.95；
3. 最后一层相似度达到0.99，中间某些层0.90以下；
4. 所有层相似度都达到0.99+，绝对误差也很小，那最终结果的误差可能是后处理的问题。



常见精度损失的解决方法

1. Data层的数据相似度对比达不到0.99

原因分析：

Data层相似度的问题说明输入不一致，或者预处理的配置有差异；

解决方法：

需要检查均值[mean_file]、缩放[data_scale]、预处理方式
[norm_type]是否和caffe一致，比如mxnet和darknet网络训练时
默认输入是RGB顺序，所以在编译时，cfg里的[RGB_order]也要配置为RGB

常见精度损失的解决方法

2. Data层是0.99+, 后面逐层下降, 最后一层小于0.95

原因分析:

这种情况下问题看起来是由于量化工具的误差引起的, 由于从float类型转换成int8或int16类型, 必然会有精度上的损失;

解决方法:

case1: 如果cfg里配置compile_mode为int8的话, 可尝试修改nnie_mapper的配置项, 将[compile_mode] 0改为1, 即把8bit低精度改成16bit高精度, 在重新编译仿真比较相似度;

case2: 如果相似度有明显提升, 所有层都达到0.99, 说明是量化误差导致, 但使用16bit相比8bit性能会下降一倍, 如果不能接受, 请尝试改为[compile_mode] 2, 即自定义高精度, 从首层开始, 逐层layer name加后缀"_hp", 直到精度和性能平衡

常见精度损失的解决方法

3. 最后一层相似度达到0.99，中间某些层0.90以下

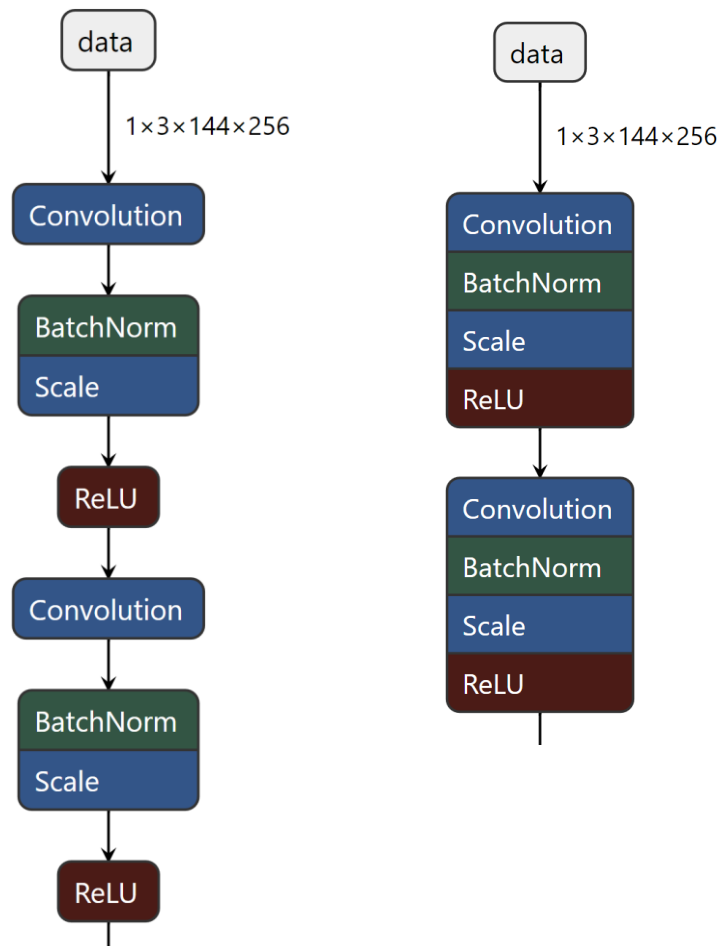
原因分析：

这种情况很可能是比较的网络层未匹配，nnie_mapper会优化网络结构，以适应硬件执行，所以相似度比较时有可能层和caffe的不匹配；

解决方法：

case 1: 是否inplace层。某些层nnie不支持inplace，会拆开，以非inplace的方式处理。如 concat + relu, fc + scale, bn + scale 等，请修改prototxt，改为非inplace方式，再跑仿真和caffe比较相似度；

case 2: 是否nnie_mapper修改了网络，请查看cnn_net_tree.dot 和原来的prototxt比较，看是否修改了网络结构，如upsample, ROI Pooling, PSROI Pooling等层会加permute 做转换，所以要和permute 的结果比，或直接看后面层的相似度



常见精度损失的解决方法

4. 所有层相似度都达到0.99+，绝对误差也很小，可能是后处理的问题

原因分析：

由于这类网络的最后几层可能NNIE不支持处理，所以需要在CPU上处理，但是这些网络层实现很容易出错

解决方法：

假设 caffe 的结果经过 caffe 的后处理画框或分类，则把仿真的结果也使用 caffe 的后处理，看是否正常画框或分类；

case 1: 如果正常，则说明是板端后处理问题，请比较板端和caffe的后处理代码；

case 2: 如果不正常，而数据的相似度0.99且绝对误差很小，则说明caffe的后处理代码对数据很敏感，请检查caffe的后处理代码

总结





筑就智能时代基石

Copyright©2021 Shanghai HiSilicon Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Shanghai HiSilicon may change the information at any time without notice.