ZAI 计算瓶颈的解决办法(multi gpu)

什么是计算瓶颈？

用于记忆和计算张量和显存开销太大了，超过 GPU 的显存。

训练速度慢算不算瓶颈？

不算，多花时间训练。或则调节超参数，将 batch 给大，并将失效步数给小。也可以对数据样本做优化，诸如使用 scale(0.5)减肥，删掉特征不明显的样本集。

cpu 内存瓶颈：

买内存吧，内存很便宜，就算插到 128G 也只会投资几千块钱。

DNN-OD 有 batch 超参数，它不会发生 GPU 计算瓶颈。
RNIC 和 LRNIC 有 batch 超参数，并且额定分类数量 1000/10000，它不会发生 GPU 计算瓶颈。

在 ZAI 中最容易发生 GPU 计算瓶颈问题的只有 DNN-Metric（深度学习的度量化网络）
以人脸为例

数亿甚至数十亿人的人脸 matric，会在 GPU 中记忆一下拟合后的张量，当数量多了，就会消耗极大的显存，这时候，就发生瓶颈问题。

解决办法 1，前期 diy 开发设备时，考虑好未来场景，做好准备工作，比如 diy 两张 32G 的 v100，以 64G 显存来应对未来的大批量人脸度量化系统。

解决办法 2，租用阿里云 1 个月的 v100*4gpu，一个月大概 30000 多 rmb，直接在云端运行 ZAI 的 metric 训练，待训练完成，本地试用一张大显存卡即可运行识别。