zAI 的大数据支持说明

概述

当图片数量达到一个“恐怖”的数量级，比如 20 万张以上，普通配置的机器学习电脑就会感到计算吃力：显存开销太大，内存开销太大，拟合计算太慢了。

这时候，zAI 针对大规模图片集提出了完整解决办法：可以节省一笔不菲设备投资，同时也兼顾了大数据训练效率。

大规模训练的工作原理

步数一次都是按一批图片进行 input，当需要 input 时，zAI 才会去申请内存，不需要就用光栅序列技术将图片暂存交换文件中。

当图片规模达到数百 GB 后，大部分都会暂存到序列化文件。

大规模图片度量化

zAI 的图片度量化包含了两个技术体系，分别是 Metric_XXX 和 LMetric_XXX 开头的 API

在 1.19 中，Metric 开头的 API 主要用于人脸，也能支持图片快照

在 1.19 中，LMetric 开头的 API 主要支持图片快照，也能用于人脸，准确度不如 Metric

当大规模图片度量化处理以后，会产生向量，向量多了 Learn 引擎就会很慢

zAI 在 1.19 版本，针对度量化新支持了 KDTree，当度量化规模大了，KDTree 可以作为替代 Learn 引擎的有效解决方案，因为 KDTree 相对 Learn 更省内存，查询更快

zAI 在 1.19 版本训练 Metric+LMetric 都可以指定以快照，或则人脸方式进行

大规模图片分类器

在 zAI 1.19 版本后，提供了 4 种支持大规模训练的图片分类器方案

RNIC：基于残差网络的图片分类器，额定分类 1000

LRNIC：基于残差网络的图片分类器，额定分类 10000

GDCNIC：基于 Google LeNet 的训练加速版分类器，额定分类 10000

GNIC：基于文字识别 NN 部分的图片分类器，额定分类 10000

大规模训练数据集不支持编辑工具

现有的 ZAI_IMGMatrix_Tool 无法支持大规模数据集编辑
针对大规模数据集的生成，编辑，处理，必须通过编程解决
一般来说，你可以预先准备好大数据集，然后使用 TAI_ImageMatrix 提供的 LargeScale-API 体系将数据导入进来，然后训练。在 TAI_ImageMatrix 提供了很多大数据支持方法。

大规模数据集的训练

准备工作：首先，你需要自行对大数据做预处理，并且确定是准化的输入

第一步，使用 TAI_ImageMatrix-LargeScale-API，将数据读取到内存，它只会在内存暂存一个指针，光栅缓冲区都放在序列化文件中。

第二步，调用训练的 API，包括：Metric,LMetric,RNIC,LRNIC,GDCNIC,GNIC，它们都能支持数百 GB 的大规模数据集训练。

第三步，特殊化处理训练输出系统

Metric+LMetric 如果要支持大数据，不能直接使用 Learn，Learn 在海量数据面前非常慢，需要采用 Metric+KDTree 的解决方案，具体细节我已在 Demo 详细描述

图片分类器 RNIC,LRNIC,GDCNIC,GNIC，他们都有额定的最大分类量，如果需求是 1000 万分类：zAI 的分类器不是返回的索引，都是经过了 softmax 处理后的接近度。你可以构建 1000 个模型，每个模型 10000 万分类，分类时遍历 1000 模型采用最佳结果。

针对大数据训练的性能优化

zAI 在 1.19 版本，Metric,LMetric,RNIC,LRNIC,GDCNIC,GNIC 均有大数据训练优化支持
Metric,LMetric,RNIC,LRNIC：属于结构级别优化，优化率得分 60
GDCNIC,GNIC：属于核心算法级别优化，优化率得分 90

针对大数据模型的选择

不同模型，算法，都针对不同的目标问题，
广泛的图片分类：RNIC+LRNIC
对单张图片做分类，如人脸，衣服：Metric+LMetric
以文字，笔迹，图标做分类：GDCNIC+GNIC

By qq600585
2019-4