凡是在 API 中出现了 LargeScale 的简单说明

在 zAI 的数据集中，我们会看到很多 LargeScale 开头的 API，或则带有 LargeScale 的参数开关

标准的 LargeScale 使用

第一步，TAI_ImageMatrix 通过 LargeScale 方式来读入大数据，比如
LargeScale_LoadFromStream(stream)， LargeScale_LoadFromFile(myfile)

第二步，训练，Train(LargeScale(这里要给 true),RSeri(这是序列光栅化的存储器),matrix,param)

zAI 针对大数据做了特殊 cache 处理，会让常用的数据驻留内存，高频率使用的数据会常驻内存，不常用的都会放到一个临时文件。大数据的训练和普通训练几乎相差无异。

训练大数据前，最好做一次数据抽样做训练测试，摸索出最好的超参数搭配，确保程序没有 bug，一般来说，无效步数，输入长度，这两个东西是关键中的关键，其次是，确保训练后，类似 Metric 的 vector 可以正确转换，类似 RNIC 的 index 可以正确保存。训练一次大数据，少则数小时，多则数天甚至数周，预抽样学习是科学的处理办法。

凡是出现了 LargeScale 字样的东西，它们的共性如下

- 从文件读取的数据，先从文件一小部分读取到内存，然后再降内存转存到序列化文件
- 使用了数据前，会从序列化文件读取数据，然后根据时间释放不用的东西
- 做 Prepare dataset 时，需要更多时间做准备工作，一般是普通训练的 2 倍时间

By.qq600585
2019-4