昵称：侯凯
园龄：8年8个月
粉丝：251
关注：10
+加关注

< 2022年2月 >

日	一	二	三	四	五	六
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	1	2	3	4	5
6	7	8	9	10	11	12

搜索
找找看
谷歌搜索

随笔分类
C++及STL(23)
java开发(5)
python知识与应用(17)
shell编程及小工具(10)
web相关(3)
操作系统(4)
长尾知识(9)
大数据学习(8)
经典算法(12)
深度学习(12)
数据库(18)
图像算法及OpenCV(33)
系统框架(41)

Faiss教程：索引(2)

作者：@houkai

索引的I/O与复制

所有的函数都是深复制，我们不需要关心对象关系。

I/O函数：

- write_index(index, "large.index"): 写索引到文件
- Index * index = read_index("large.index") 读索引

复制函数：

- Index* index2 = clone_index(index): 返回索引的深复制
- Index *index_cpu_to_gpu = index_cpu_to_gpu(resource, dev_no, index): 复制索引到GPU
- Index *index_gpu_to_cpu = index_gpu_to_cpu(index):从GPU到CPU
- index_cpu_to_gpu_multiple: uses an IndexShards or IndexProxy to copy the index to several GPUs.

index_factory

index_factory通过字符串来创建索引，字符串包括三部分：预处理、倒排、编码。
预处理支持：

- PCA：PCA64表示通过PCA降维到64维（PCAMatrix实现）;PCAR64表示PCA后添加一个随机旋转。
- OPQ：OPQ16表示为数据集进行16字节编码进行预处理（OPQMatrix实现），对PQ索引很有效但是训练时也会慢一些。

倒排支持：

- IVF：IVF4096表示使用粗量化器IndexFlatL2将数据分为4096份
- IMI：IMI2x8表示通过Mutil-index使用2x8个bits（MultiIndexQuantizer）建立2^(2*8)份的倒排索引。
- IDMap：如果不使用倒排但需要add_with_ids，可以通过IndexIDMap来添加id

编码支持：

- Flat：存储原始向量，通过IndexFlat或IndexIVFFlat实现
- PQ：PQ16使用16个字节编码向量，通过IndexPQ或IndexIVFPQ实现
- PQ8+16：表示通过8字节来进行PQ，16个字节对第一级别量化的误差再做PQ，通过IndexIVFPQR实现

如：
index = index_factory(128, "OPQ16_64,IMI2x8,PQ8+16"): 处理128维的向量，使用OPQ来预处理数据16是OPQ内部处理的blocks大小，64为OPQ后的输出维度；使用multi-index建立65536（2^16）和倒排列表；编码采用8字节PQ和16字节refine的Re-rank方案。

OPQ是非常有效的，除非原始数据就具有block-wise的结构如SIFT。

自动调参

索引的参数包括两种：bulid-time索引创建时需要设置的、run-time在搜索前可以调整的。针对run-time参数可以进行Auto-tuning。

Key	类名	run-time参数	备注
IVF, *IMI2x*	IndexIVF*	nprobe	控制速度和精度的折中
IMI2x*	IndexIVF	max_codes	平衡倒排列表
PQ*	IndexIVFPQ, IndexPQ	ht	Hamming threshold for polysemous
PQ+	IndexIVFPQR	k_factor	Re-rank时要核实的数据量

AutoTuneCriterion：包含ground-truth，使用搜索结果，评估召回；OperatingPoints：包含（性能，时间，参数集合id），目标是找到最优的operating point——没有其他point可以在更短的时间内达到更好的性能；ParameterSpace：参数空间是指数级的，但是这些参数有一个共同的特性，值越高一般来说速度越慢，性能越好。

faiss/tests/demo_sift1M.cpp中有一个自动调参的示例。自动调参依赖于：评测集合完备且充足，机器环境稳定。

特殊的操作

- 根据索引重建数据，见test_index_composite.py
  支持IndexFlat, IndexIVFFlat (call make_direct_map first), IndexIVFPQ (same), IndexPreTransform (provided the underlying transform supports it)
- 从索引中移除元素，remove_ids方法
  见test_index_composite.py，支持IndexFlat, IndexIVFFlat, IndexIVFPQ, IDMap
- 范围查找，range_search方法
  将返回离查询点一定半径内的向量，在Python中它将返回一个1D元组lims/D/I，针对第i个的查询结果为I[lims[i]:lims[i+1]], D[lims[i]:lims[i+1]]，支持IndexFlat, IndexIVFFlat
- 合并切分索引
  merge_from合并其他索引，copy_subset_to复制当前索引的子集到其他索引，支持IndexIVF

· 百度计算生物研究登Nature子刊！结果超斯坦福MIT，落地制药领域
· 比一粒盐还小的电池问世
· 死前真的会有「跑马灯」，人类首次同步测量大脑濒死状态
· 网传员工猝死，字节跳动内网流出回应：仍在医院抢救中
» 更多新闻...