

UGRP Meeting

Aug. 1, 2022, 4:30 pm

Supporter: Jungwoo Kim

Data-Intensive Computing Systems Laboratory(DataLab), DGIST

DeepSketch 재구현

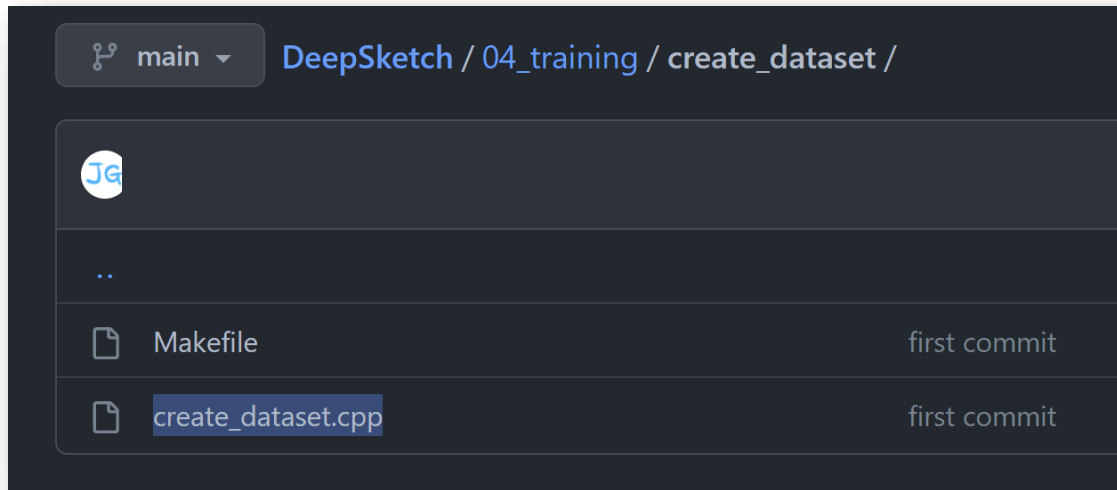
- **deepsketch-fast2022 github** 내에 업로드 되어있는 파일들은 다음과 같다.
 - **NGT** (High-speed approximate nearest neighbor searches library)
 - **bruteforce**
 - Makefile은 존재하지 않는다.
(g++ -O2 -pthread -o bf_xdelta3_all ../lz4.c ../xdelta/xdelta3.c ../xxhash.c bf_xdelta3_all.cpp)
 - **clustering**
 - coarse_xdelta3.cpp
 - fine_xdelta3.cpp
 - 마찬가지로 Makefile이 존재하지 않지만, 어렵지 않다.
 - **deepsketch**
 - libtorch build 작업을 위해 필요한 파일들이 없는데, 이전에 UGRP 학생들에게 libtorch build하는 방법 지도한 적 있다.
 - finesse, multi_cand, results, xdelta3, compress.h, lz4.c, lz4.h

DeepSketch 재구현

- deepsketch-fast2022 github 내에 업로드 되어있는 파일들은 다음과 같다.
 - Training
 - train_baseline.py, train_hashlayer_gh.py, model_converter_gh.py
 - training을 위한 cluster dataset 생성 코드가 없다. (일종의 preprocessing)
- training을 하기 위해서는 아래와 같은 동작을 수행해야하는데, train_baseline.py를 위한 cluster를 생성하는 코드가 존재하지 않기 때문에 실험 수행이 현재의 github code만으로는 불가능하다.
 - python train_baseline.py **cluster_mix_20000**
 - python3 train_hashlayer_gh.py model/model.4096_2048_1e-05.torchsave cluster_mix_20000 128 1 0.001 0.05
 - python3 model_converter_gh.py model/model_hash.128_4096_2048_1_0.002_0.05.torchsave
- DeepSketch private github에서는 아래와 같이 cluster를 생성할 수 있다고 하지만, 해당하는 이름의 실행파일에 대한 소스가 없다.
 - ./coarse mix_random_10 256 > mix_random_10_coarse
 - ./fine mix_random_10 256 < mix_random_10_coarse > mix_random_10_fine
 - **./dataset** **mix_random_10** mix < mix_random_10_fine

DeepSketch 재구현

- 이름이 바뀌어있었다.



- 해당코드 또한 주먹구구식으로 짜여져 있어, 약간의 변경이 필요했다.

```
root@29181dee65dd:/workspace/deepsketch-fast2022/deepsketch/build# ./DeepSketch ../../../../DeepSketch/sf0 ../../../../DeepSketch/04_training/model/model_hash_sf0_128_4096_2048_1_0.001.torchsave.pt 128
Total time: 61999986us
ANN ../../../../DeepSketch/sf0 with model ../../../../DeepSketch/04_training/model/model_hash_sf0_128_4096_2048_1_0.001.torchsave.pt
Final size: 111615521 (41.60%)
```