Turbo Scan: Fast Sequential Nearest Neighbor Search in High Dimensions

Standard Indexed Proximity Search Applications

- Image and video retrieval
- Natural language processing
- Sensor data analysis
- Computer vision

Our Problem: Proximity Searching Without an Index

- Data Archiving
- Historical Data Analysis
- Infrequently Updated Datasets
- Data Visualization

The idea in brief

Figure: Turbo Scan in brief. In each iteration we compute a more precise distance over a smaller fraction of the database.

More details

- Order using a few coordinates only (the slice)
- Split top-ranked first (the split)
- Repeat with a new slice and new split
- Until there are only a few vectors left

Performance analysis

Figure: Search time and recall of the 10k knn queries on the LAION-300K dataset (k=16). Brute force performance is marked with the **black** star.

Performance of computing the all knn graph on the LAION-300K dataset ($k = 16$)

Name	Slice	Δ	Split	α	Recall
BF					1.0
TS	F	8	F	3	0.8692
TS	D	16	F	3	0.9992
TS	F	8	D	12	0.6182
TS	D	16	D	12	0.9728

Take out

- For online applications Turbo Scan is up to 3 times faster than brute force, at high recall rates.
- It has **zero** preprocessing time

Future work

- GPU-based implementation
- Organizing sequential access in secondary memory with no random seeks

Edgar Chavez, Eric S. Tellez

{elchavez,stellez}@cicese.edu.mx

CICESE & INFOTEC/IxM CONAHCyT, México

sisap