Architectural Support for Large-Scale Visual Search

A Work In Progress

Carlo C. del Mundo
Vincent Lee
Armin Alaghi
Luis Ceze
Mark Oskin

UNIVERSITY of WASHINGTON
Motivation: Visual Data & Their Applications

Algorithm Across These Apps: k-nearest neighbors (kNN)

Overall Challenges:
- High data movement
- High compute requirements

Idea: Co-design the memory subsystem with search algorithms.
State of the art: Today’s Software Pipeline

Feature Extraction

Multimedia Database

Extract shape, appearance, motion

Index Features

Query Vector

Candidate Nearest Neighbor Vectors

K-nearest neighbors

Distance Calculation

Calculate pairwise metric distance

Query Generation

Extract features

Slow of Index

Trillions x Images
Billions x Videos
Billions x Audio

Reverse Lookup

Recover data

Result Vector

Response

Result Vector

Query Vector


State of the Art: Nearest Neighbor Search on Commodity Systems

Intra-node Challenge:

k-nearest neighbor search is heavily memory bound.

Inter-node Challenge

Each query vector generates k local responses.

To get the global k nearest neighbors we need an all-to-one reduction; this makes the network the bottleneck.
Our Vision: Systems and Architecture Support for Nearest Neighbor Search

Our Vision:
A nearest neighbor content addressable memory (NCAM) to alleviate memory boundedness by moving compute closer to data.

Our Vision:
Systems support for in-network joins and data reductions to reduce global network bandwidth and scalability.
Intra-Node: Rebalancing Bandwidth and Compute

Rebalance Bandwidth
Use binary features\(^2\).

Rebalance Compute
Move compute closer to memory.

Algorithmic Growing Pains\(^1\)
500 GB of communication
150 GFLOP of computation
0.3-0.5 FLOP/byte

Our Vision.
Nearest Neighbor CAM

1. Assume 512 KB feature vectors. \(|N| = 1\) billion. Assume a single query needs to evaluate 10% of \(|N|\) with kd-trees. Distance metric is Euclidean. Note: we assume these numbers are from FLANN.
High-Level Overview:
Nearest Neighbor Content Addressable Memory

Idea: Achieve high internal bandwidth by interposing logic in memory.
Our Vision: A class of memories akin to CAMs for high-throughput distance calculations.
Architectural Overview: Inside the NCAM
Programming Model: Integrating the NCAM in DRAMs

1. NCAM additions are orthogonal to existing DRAM infrastructure.
2. We introduce a memory mapped NCAM configuration and control interface.
3. We introduce a command set which allows for configuration.

<table>
<thead>
<tr>
<th>Command</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>READ</td>
<td>Reads the result out from the priority queues</td>
</tr>
<tr>
<td>WRITE</td>
<td>Writes initial query vector to accelerator memory</td>
</tr>
<tr>
<td>ENABLE</td>
<td>Signals NCAM to process data read from DRAM</td>
</tr>
<tr>
<td>DISABLE</td>
<td>Signals NCAM to ignore any data from the DRAM</td>
</tr>
<tr>
<td>RESET</td>
<td>Reinitializes priority queues in preparation for next execution</td>
</tr>
</tbody>
</table>
Related Works: What came before the NCAM

Software-based solutions

- Libraries

- Algorithms
  - [Datar et al., SCG ‘04] Locality-sensitive Hashing
  - [Y. Weiss et al., NIPS ‘08] Spectral Hashing
  - [Torralba et al., CVPR ‘08] Small codes for Recognition
  - [Gong, Lazebnik, CVPR ‘11] Iterative Quantization

Hardware-based solutions

- Nearest-Neighbors
  - [Roberts, MS Thesis ‘90] PCAM: Proximity Content-Addressable Memory
  - [Garcia et al., CVPRW ‘08] Nearest Neighbors on GPU

- In-memory computing
  - [Patterson et al., IEEE MICRO ‘97] A Case for Intelligent RAM
  - [D.G. Elliott, D&T ‘99] Computational RAM
  - [Guo et al., MICRO ‘11] TCAMs for Data-Intensive Computing
**Current Work: What We’ve Done & Plan To Do**

- **Application Design & Characterization.** Design & Implementation of VIRAL: *Video and Image Retrieval At Large-Scale.*
Current Work: What We’ve Done & Plan To Do

- **Hardware Design.** ASIC, Simulation, Design & Test.

![Diagram of kNN Engine and Priority Shift Register]
**Summary**: Enabling visual applications via co-design

**Our Vision**: Architectural support to enable high dimensional & high cardinality nearest neighbor search.

Our work enables visual applications such as:

- **3D Reconstruction**
- **Content-based Search**
- **Augmented Reality**
- **Activity Recognition**