Learning From Less Data: A Unified Data Subset Selection and Active Learning Framework for Computer Vision

Vishal Kaushal[1], Rishabh Iyer[2], Suraj Kothawade[1], Rohan Mahadev[1] , Khoshrav Doctor[3] , Ganesh Ramakrishnan[1]

Indian Institute of Technology Bombay[1]
Microsoft Corporation[2]
University of Massachusetts Amherst[3]

Motivation

Deep Models are used everywhere today!

Increased Model Complexity →

1) Increased Computing Resources
2) Increased Labeling Cost
3) Increased Turn Around Time

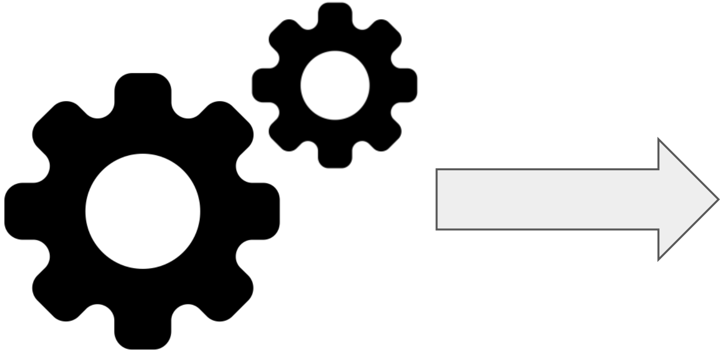
Increased Labelling Cost

Difficult to get Labeled Data!

Increased Turn Around Time

Harder to tune Hyperparameters

Hyperparameter tuning

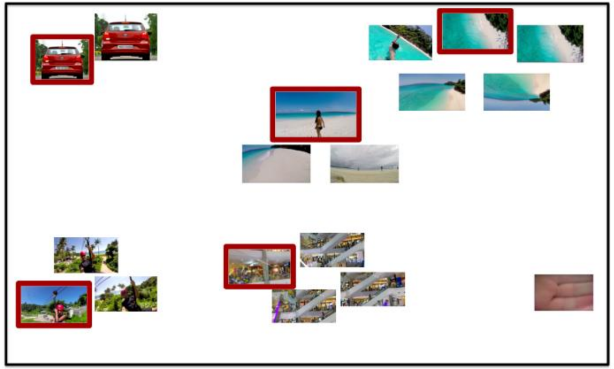


Best hyperparameters

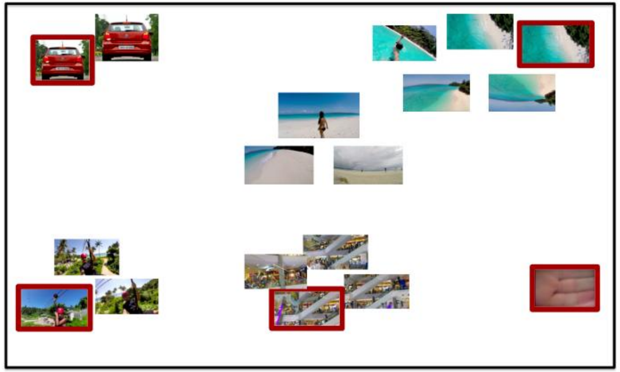
Our Contributions

A unified framework for data subset selection

1. Facility Location (models representation)
2. Minimum Dispersion (models diversity)



Representation Functions



Diversity Functions

Data Subset Selection

Given a ground set $V = \{1,\ 2,\ 3,\ ....\ , n\}$

Define a set function $f : 2^V \rightarrow R$ which measures how good a subset $X \subseteq V$

$$\text{Problem 1: } \max\{f(X) \text{ such that } |X| \leq k\}$$

The greedy algorithm obtains an approximation guarantee of *(1 − 1/e)* for Problem 1 when *f* is the Facility Location function.

Similarly, the greedy algorithm achieves an approximation factor of *1/2* when *f* is the Dispersion function.

Representation Functions

They try to find a representative subset of items, akin to centroids and medoids in clustering

Facility Location: $f(X) = \sum_{i \in V} \max_{j \in X} s_{ij}$

Diversity Functions

They attempt to obtain a diverse set of keypoints

Dispersion Function: $f(X) = \min_{i,j \in X} d_{ij}$

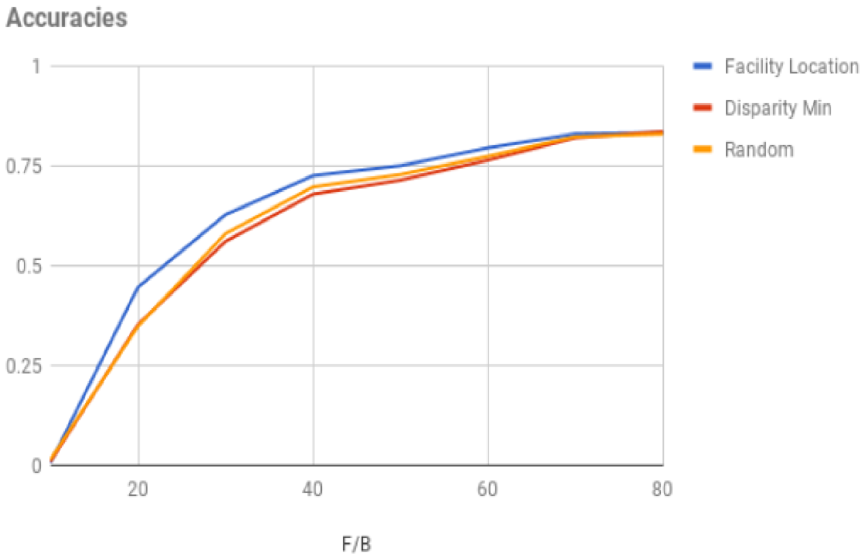
While diversity only looks at the elements in the chosen subset, representativeness also worries about their similarity with the remaining elements in the superset

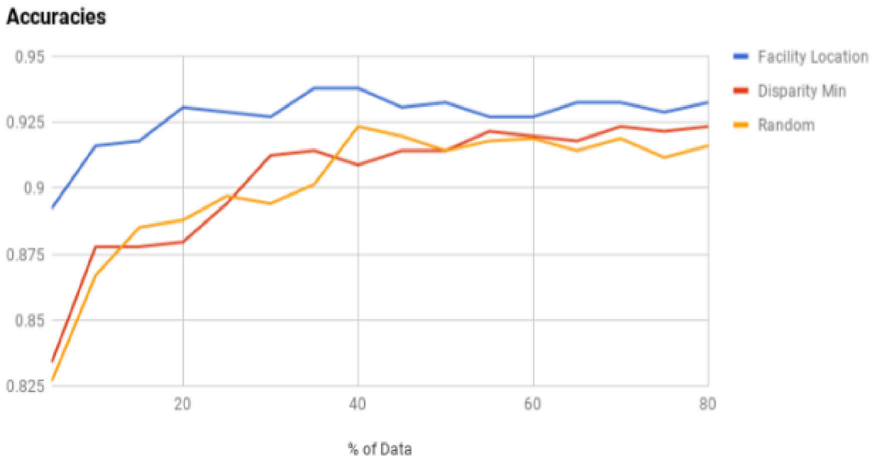
Four Settings / Use Cases

Four concrete use cases of our framework

1. Supervised Data Selection for **Quick Training/Inference**
2. Supervised Data Selection for **Quick Hyper-parameter tuning**
3. Unsupervised Data Selection for **Labeling from Video Data**
4. Diversified **Active Learning**

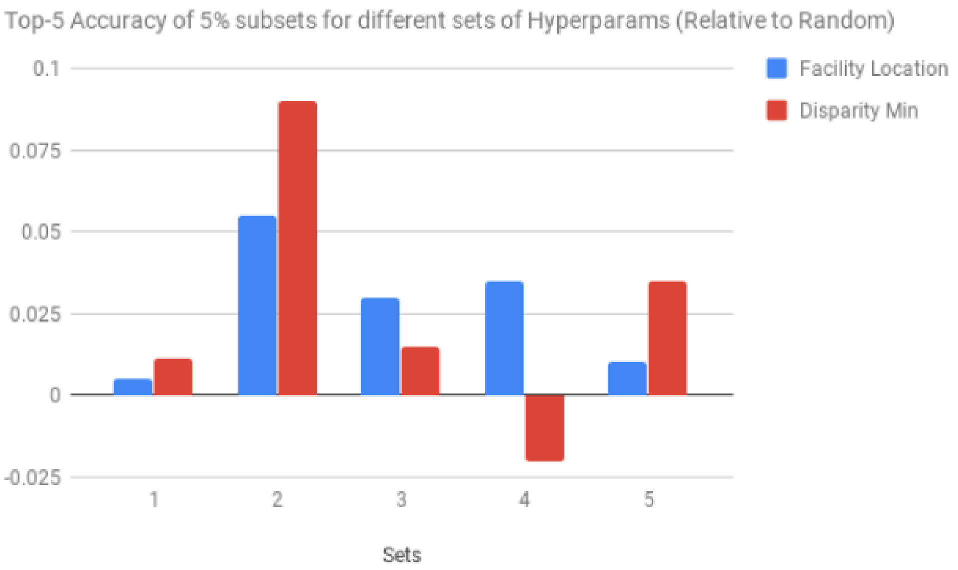
Results: DSS for Quick Training/Inference



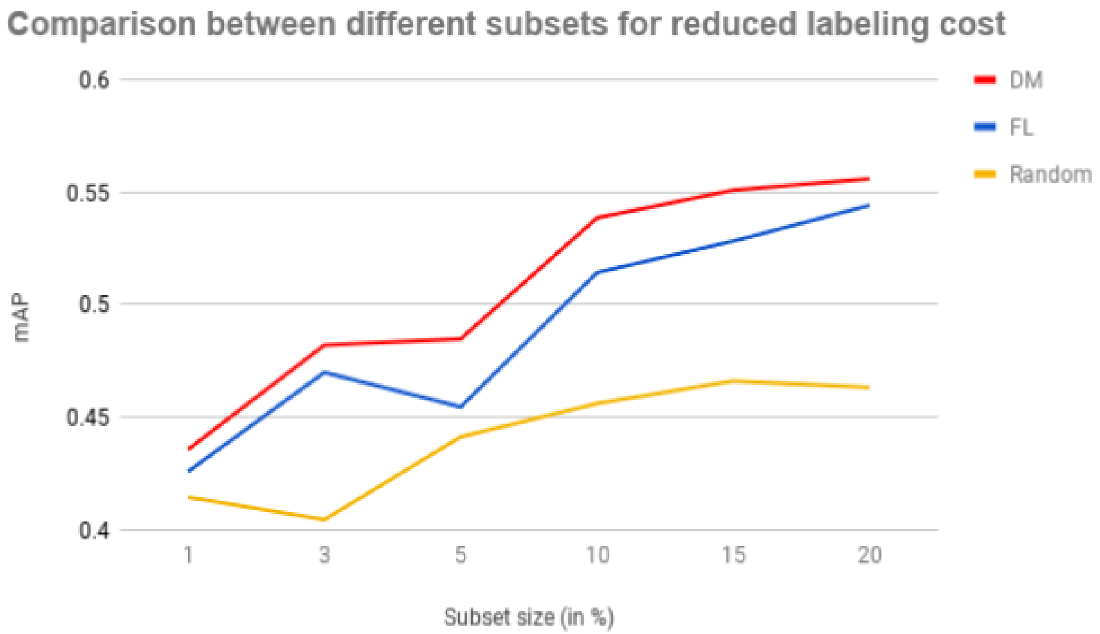


Supervised DSS for Quick Training/Inference (KNN Classification)

Results: DSS for Hyper Parameter Tuning

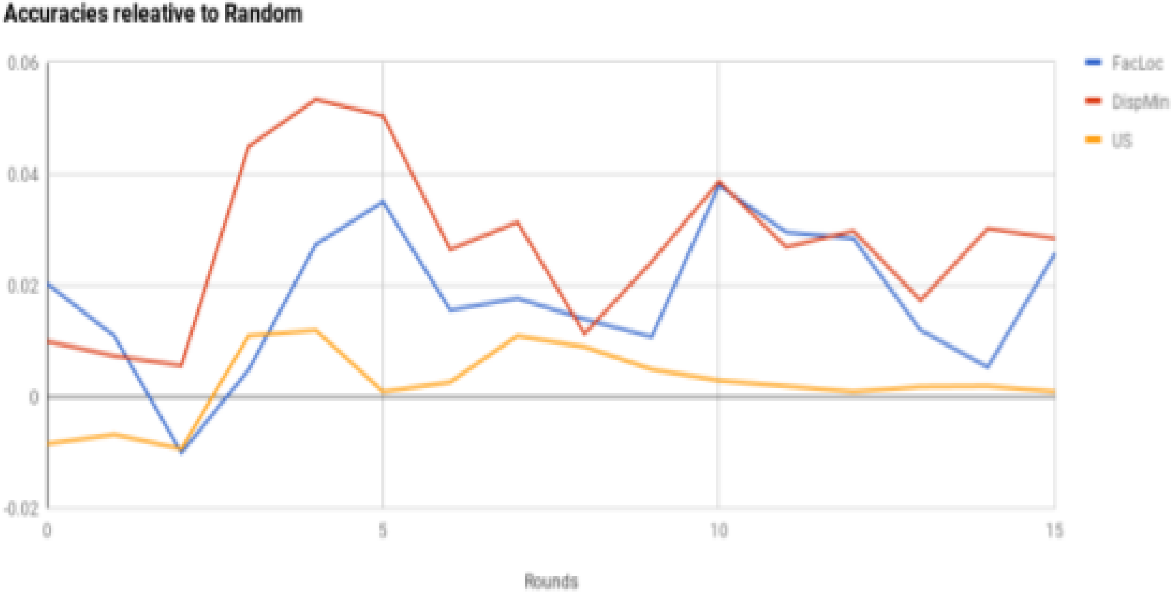


Results: DSS for Labeling Video Data



Data Subset Selection on Massive Datasets for Labeling

Results: Diversified Active Learning



$$\max\{f(X) \text{ such that } |X| \leq B, X \subseteq F\}$$

Submodular Active Learning on the Adience Dataset for Gender Classification