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

Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, Ganesh Ramakrishnan

IEEE 2019 **W**inter Conference on **A**pplications of **C**omputer **V**ision

Motivation

Deep Convolutional Models are very successful for several Computer Vision tasks, *but...*

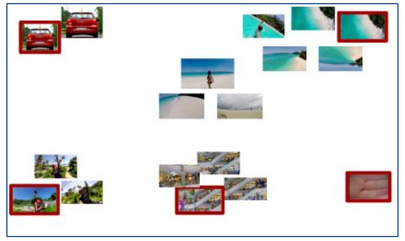
- Increased Model Complexity →
  1) Increased Training Time
  2) Increased Labeling Cost
  3) Increased Experimental Turn around time

- Difficult to get labeled data!

Key Idea

- Using submodular optimization for data subset selection and active learning.
- Submodular fuctions naturally model notions of representation, diversity and coverage which is useful for choosing a good dataset for Deep Learning tasks.
- Facility Location (FL) models representation.
- Minimum Dispersion (DM) models diversity.

Representation functions



Diversity functions

Our Contributions

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

We define a set function $f : 2^V \rightarrow R$ which measures the utility of subset $X \subseteq V$

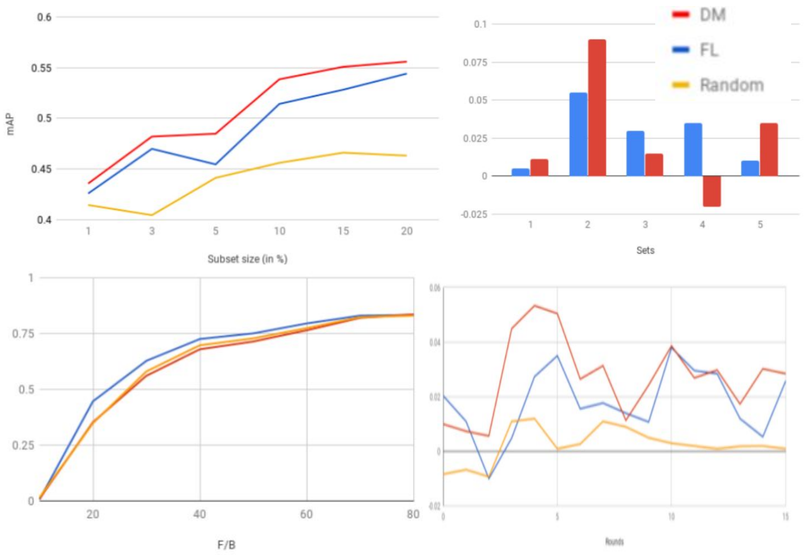
So, $\text{Problem 1: } \max\{f(X) \text{ such that } |X| \leq k\}$

The greedy algorithm obtains an optima with certain approximation guarantees when ***f*** is FL or DM.

Using this, we create a unified framework which performs Supervised and unsupervised data subset selection for

- Quick training and inference
- Quick hyper-parameter tuning
- Labeling from video data
- Diversified Active Learning

Experiments and Results



(TL to BR) - Unsupervised DSS on massive datasets for labeling, Supervised DSS for hyper-parameter tuning, Supervised DSS for KNN Classification, Submodular Active Learning for Gender Classification.

Conclusions

- We demonstrate the utility of subset selection in training models for a variety of Computer Vision tasks.
- Models trained on subsets obtained from certain submodular functions perform better than others.
- Minimum Dispersion works best when there is a higher amount of redundancy in data, while Facility Location works better in other cases.
- Regardless, both out-perform random and uncertainty sampling.