Convex Optimization for Computer Vision Data Selection

David Jose Florez Rodriguez,[1] Benjamin Jose Martinez,[2] Gabriel Dean Magaña[2]

[1] M.S. Candidate, Electrical Engineering, Stanford University
[2] B.S. Candidate, Computer Science, Stanford University

CS231N, Spring 2022

Introduction

To assist with X-ray image classification we present the ‘Convex Pseudo-Spread Maximizer’ (CPSM) for optimized data selection, this has multiple advantages:

- Decreased need for massive datasets
- Faster computation than alternative data selection methods

Related Work & Motivation

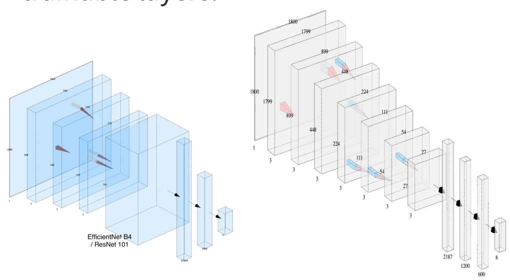
- Demand for radiology and its data
- Data selection : pick valuable data points to boost performance and evaluate individual samples.
- Truncated Monte Carlo Data Shapley (TMC) method is our benchmark

Data

- PadChest , 1000 (1800 x 1800) PA images
- classes:[Normal, COPD, pneumonia, scoliosis, infiltrates, cardiomegaly, pleural effusion, OTHER]

Predictors

We train raw models and sandwich existing ResNet and EfficientNet in trainable layers.



Pseudo-Spread

New metric for evaluating a subset of training data:

$$S = \sum \omega^T D + \frac{\alpha}{N} \sum \omega^T \left((A-\Theta)^T (A-\Theta)\right)$$

$\underline{D}$ is the distance matrix for vectors in dataset $\underline{A}$, an $\underline{M}$ by $\underline{K}$ matrix
Diagonal matrix $\underline{\omega}$ selects datapoints
Matrix $\underline{\Theta}$ is a repeated vector of $\underline{\theta}$, the median of $\underline{A}$.
Weigh $\underline{\alpha}$ regulates the impact of each side.

Left element is a sum of distances of all points with chosen points’ columns present.
Right element is a convex adaptation of variance with respect to $\underline{\omega}$, the problem variable.

Convex Optimization

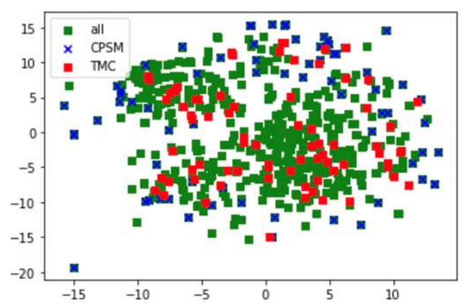
$$\text{Maximize} \quad S,$$

$$\text{such that } 0 \le \omega \le I_M, \text{ and } \sum \omega <= N$$

Where I_M is the size *M* identity matrix and *N* is the number of points to sample from *A*.

Feature Space

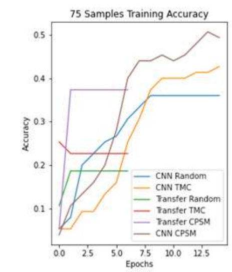
A pre-trained conv model generates size 1200 vectors in its second dense layer. Training images (600) fed through the model produce $\underline{A}$ of size 600 by 1200. CPSM acts on this $\underline{A}$.

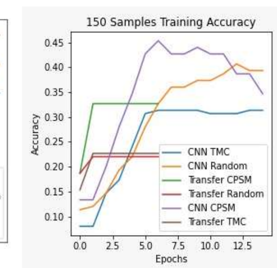


A plot of **t-SNE** representation of $\underline{A}$ and samples selected by CPSM and TMC methods. TMC’s subset is denser than CPSM’s, which includes outliers.

Results

- CPSM and TMC outperform methods without data selection
- CPSM on our CNN architecture increased accuracy by 10% compared to TMC method
- Transfer learning models failed





Conclusion

CPSM outperforms TMC and runs 100x faster, but requires high RAM
Future Work:

- Improve transfer learning
- Investigate density in chosen subsets
- Tune α and *N* hyperparameters

Acknowledgements

Doctors Manuel Acosta and Antonio Bueso for their medical expertise. Professor Stephen Boyd for EE364A

Find the code at:
https://github.com/elsirdavid/Convex_Pseudo_Spread_Maximizer