SUBMODULAR MUTUAL INFORMATION FOR TARGETED DATA SUBSET SELECTION

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, Rishabh Iyer

THE UNIVERSITY OF TEXAS AT DALLAS

UNIVERSITY of WASHINGTON

INTRODUCTION

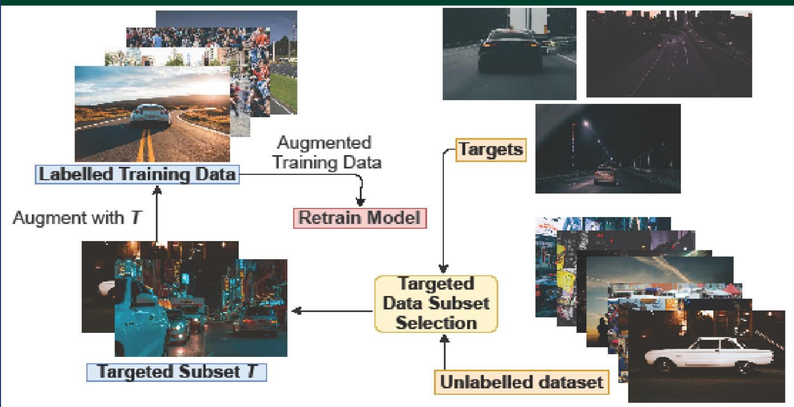
- With the rapid growth of data, it is becoming increasingly difficult to train or improve deep learning models with the right subset of data.
- This problem can be solved for an additional labeling cost by adding a subset of unlabeled data points similar to an auxiliary set to the training data.
- We do so by using Submodular Mutual Information (SMI) functions which can be used to target a certain slice of data that is critical for deployment on which we desire better performance.
- We empirically demonstrate the performance of targeted data subset selection (TSS) for improving the performance on an image classification task for imbalanced datasets.

PROBLEM FORMULATION

Goal: To select a "targeted" data subset for improving data imbalance or accuracy of the task DNN.

- The Submodular Mutual Information (SMI) is defined as $I_F(A;Q) = F(A) + F(Q) - F(A \cup Q)$. This measures the similarity between A and Q, where Q is the query/target set.
- Q is from an auxiliary set V' different from the ground set V.
- For TSS, V is the source set of data instances and the target is a subset of data points (validation set or the specific set of examples of interest).
- Define $f : 2^{V \cup V'} \rightarrow \Re$.
- Although f is defined on $V \cup V'$, discrete optimization is only defined on $A \subseteq V$.
- To find an optimal subset we maximize $g_Q(A) = I_f(A;Q)$

Targeted Data Subset Selection

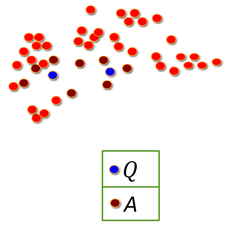


Motivating example for targeted data subset selection (TSS)

- The night images (target)are under-represented in training data.
- TSS mines for night images from a large unlabeled dataset and augments the training data to improve the performance of the final model

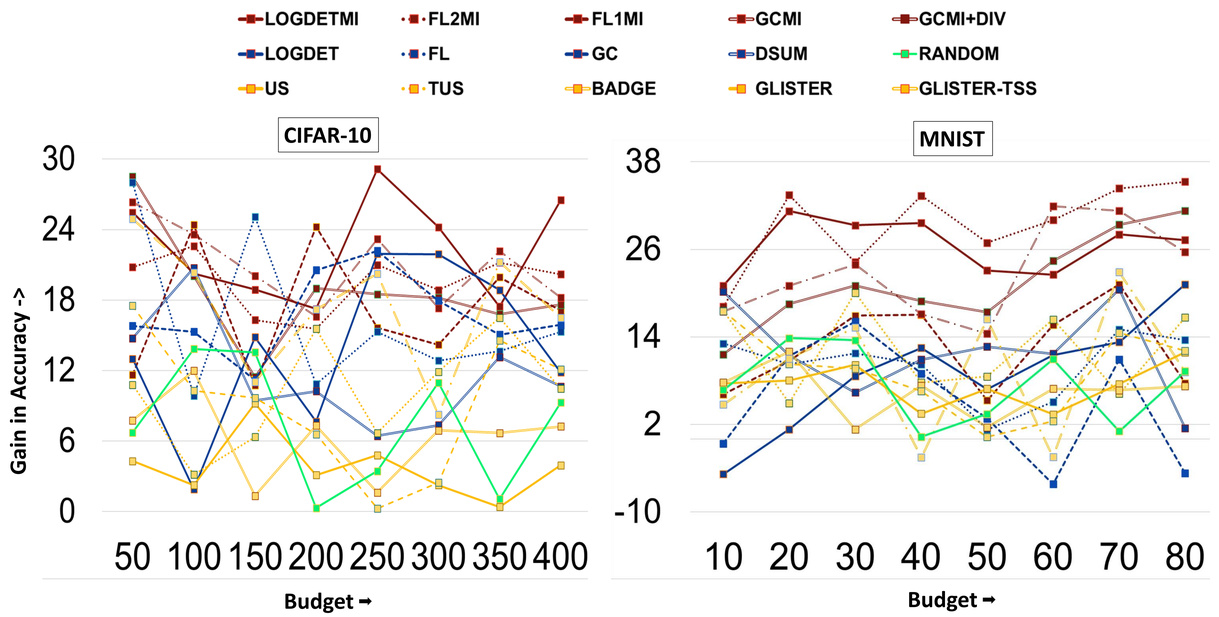
Submodular Mutual Information

$I_F(A;Q) = F(A) + F(Q) - F(A \cup Q)$



Name	$I_F(A;Q)$
Graph Cut MI	$2\sum_{i \in A}\sum_{j \in Q} s_{ij}$
Facility Location MI – V1	$\sum_{i \in V} \min(\max_{j \in A} s_{ij}, \eta \max_{j \in Q} s_{ij})$
Facility Location MI – V2	$\sum_{i \in Q} \max_{j \in A} s_{ij} + \eta \sum_{i \in A} \max_{j \in Q} s_{ij}$
Log Determinant MI	$logdet(S_A) - logdet(S_A - \eta^2 S_{A,Q} S_{Q^{-1}} S_{A,Q}^T)$

RESULTS



SMI functions achieve ≈20-30% gain in accuracy on the targeted classes on re-training with the targeted subset.

TSS Algorithm

Given: Initial Labeled set of Examples: E, large unlabeled dataset: U, A target subset/slice where we want to improve accuracy: T, Loss function L for learning

1. Train model with loss L on labeled set E and obtain parameters θ_E.
2. Compute the gradients $\{\nabla_{\theta_E} L(x_i, y_i), i \in U\}$ and $\{\nabla_{\theta_E} L(x_i, y_i), i \in T\}$.
3. Using the gradients, compute the similarity kernels and define the submodular function f and diversity function g.
4. $\hat{A} \leftarrow \max_{A \subseteq U, |A| \leq K} I_f(A;T) + \gamma g(A)$
5. Obtain the labels of elements in A^*: $L(\hat{A})$
6. Train a model on the combined labeled set $E \cup L(\hat{A})$

CONCLUSIONS

- We demonstrate the effectiveness of SMI functions for improving a model's performance by augmenting the training data with samples that match a target distribution (targeted data subset selection).
- Through experiments on CIFAR-10 and MNIST datasets, we empirically verify the superiority of SMI functions over existing methods.
- Using SMI functions, we observe ≈20-30% gain over the model's performance before re-training with added targeted subset; ≈12% more than other methods.

PAPER

Get the paper for more technical details and results:

https://arxiv.org/abs/2103.00128

SCAN ME