Balancing Feature Similarity and Label Variability for Optimal Size-Aware One-shot Subset Selection

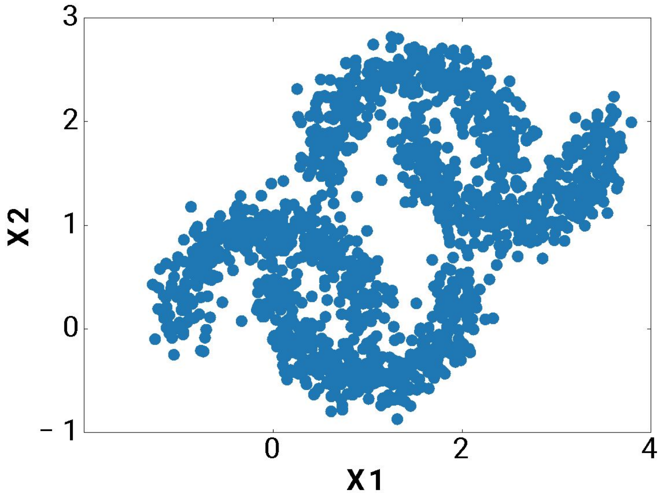
Abhinab Acharya, Dayou Yu, Qi Yu and Xumin Liu
Rochester Institute of Technology

Introduction

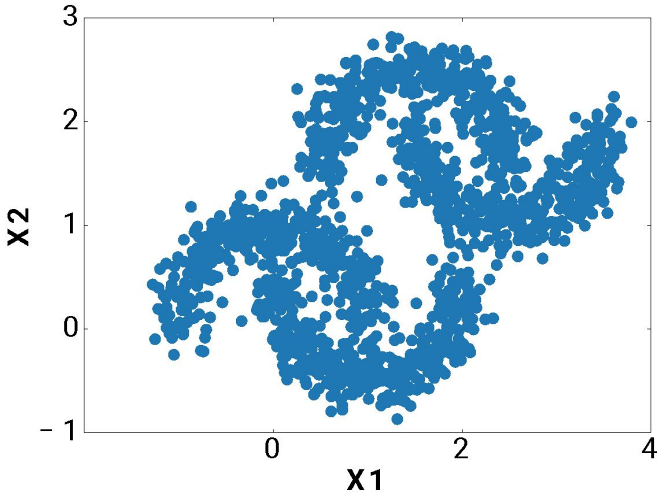
- **Subset selection** finds candidate data points from a large pool, trains model efficiently, and decreases resource consumption.
- **One-shot subset selection** is challenging as subset selection is only performed once and full set data become unavailable after selection.

Introduction

- Existing methods are classified into **diversity-based** and **difficulty-based** subset selection.
- They do not consider the **tradeoff between feature similarity (diversity) and label variability (difficulty)** as they solely rely on the feature or label side.



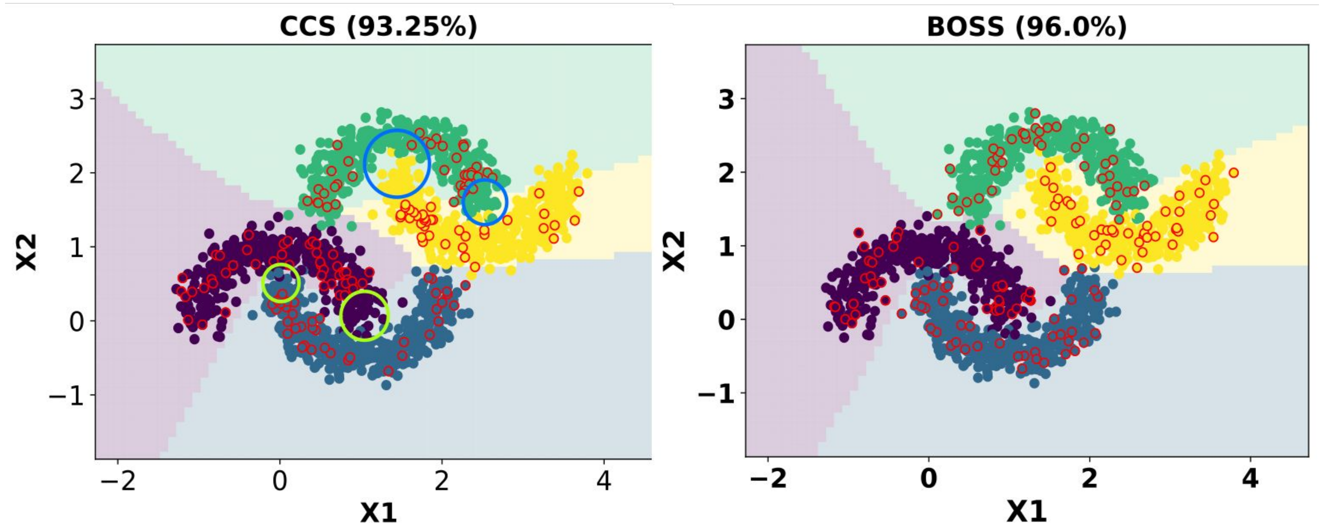
Label Variability Based Selection



Feature Similarity Based Selection

Introduction

- Recent methods (like CCS) lack principled way to balance diversity and difficulty given a subset size. The selection also misses some critical region in the full set.
- We propose to **conduct feature similarity and label variability Balanced One-shot Subset Selection (BOSS)**, aiming to construct an optimal size-aware subset.



Our Contribution

- Our method (BOSS) incorporates the tradeoff between prioritizing feature similarity (diversity) or label variability (difficulty) in relation to the subset size.
- We provide a theoretical insight via a novel core-set loss bound that shows the importance of balancing both diversity and difficulty with respect to the subset size.
- We design a **practical surrogate target** which connects the loss bound to a novel importance function to delicately control the optimal balance of diversity and difficulty.
- We evaluate our method on 4 image classification datasets.

Balanced Core-set Loss Bound

- **Minimization of generalization loss** bounded by the full set loss:

$$\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[l(\boldsymbol{\eta}(\mathbf{x}),\mathbf{y};\boldsymbol{\theta})\right] \leq \left|\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[l(\boldsymbol{\eta}(\mathbf{x}),\mathbf{y};\boldsymbol{\theta})\right] - \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} l(\boldsymbol{\eta}(\mathbf{x}_i),\mathbf{y}_i;\boldsymbol{\theta}_{\mathcal{S}})\right| + \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} l(\boldsymbol{\eta}(\mathbf{x}_i),\mathbf{y}_i;\boldsymbol{\theta}_{\mathcal{S}})$$

Theorem 1 (Balanced Core-set Loss Bound). *Given the full set $\mathcal{V}$ and the subset $\mathcal{S}$, for each $\mathbf{x}_i \in \mathcal{V}$, we can locate a corresponding $\mathbf{x}_j \in \mathcal{S}$, such that $\|\mathbf{x}_j - \mathbf{x}_i\| = \min_{\mathbf{x}_n\in\mathcal{S}} \|\mathbf{x}_n - \mathbf{x}_i\|$ and $l(\boldsymbol{\eta}(\mathbf{x}_j),\mathbf{y}_j) = 0$. Then, we have*

$$\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} l(\boldsymbol{\eta}(\mathbf{x}_i),\mathbf{y}_i,\boldsymbol{\theta}_{\mathcal{S}}) \leq \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \left(\lambda^{\boldsymbol{\eta}}\|\mathbf{x}_i - \mathbf{x}_j\| + \lambda^{y}\|\mathbf{y}_i - \mathbf{y}_j\|\right) + L\sqrt{\frac{\log(1/\gamma)}{2|\mathcal{V}|}}$$

with the probability of $1-\gamma$, where $\lambda^{\boldsymbol{\eta}}$ and λ^{y} are Lipschitz parameters, L is the maximum possible loss and γ is the probability of the Hoeffding's bound not holding true.

Bridging Label Variability and Difficulty Score

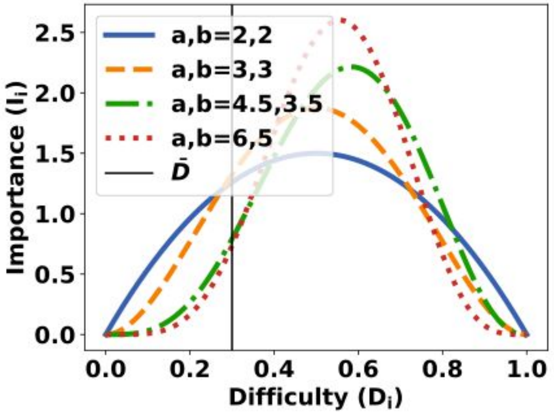
- We make a connection between the label variability and the difficulty score:

Theorem 2 (EL2N lower bounds the label variability). *Assuming a subset sample* $(\mathbf{x}_j, \mathbf{y}_j) \in \mathcal{S}$ *is located in a difficult region* (e.g., *near the decision boundary), where (i) the neighborhood* $\mathcal{N}_j$ *is dense* $(\|\mathbf{x}_j - \mathbf{x}_i\| \leq \delta_x, \forall(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{N}_j$ *for* $|\mathcal{N}_j|$ *closest points) and (ii) the label variability is high* $(p(\|\mathbf{y}_i - \mathbf{y}_j\| > 0) \geq \xi)$*, the EL2N score produced by a smooth model* (e.g., *the initial model* $\boldsymbol{\eta}_0(x; \mathcal{V})$*) will lower bound the label variability in this neighborhood* $\mathcal{N}_j$.

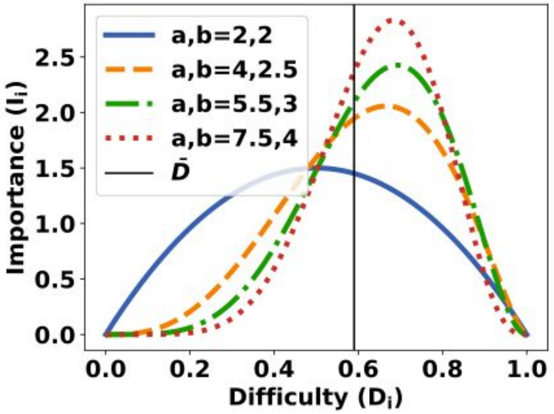
Importance Sampling Function

- We leverage and difficulty score to construct a special beta distribution.
- It helps achieve **fine-grained balance** over **each component.**

$$\mathcal{I}(\mathbf{x}_j, \mathbf{y}_j) = \mathtt{Beta}(D_j | a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} D_j^{a-1} (1 - D_j)^{b-1}$$



CIFAR 10

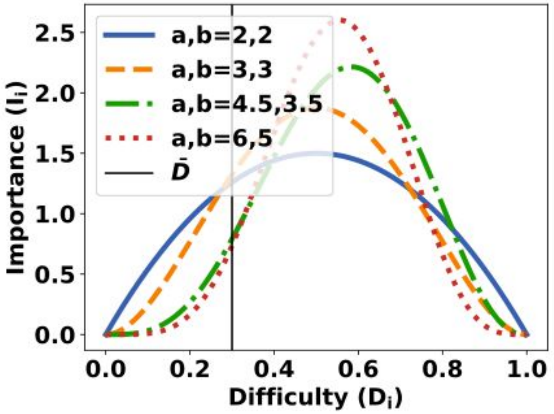


Tiny ImageNet

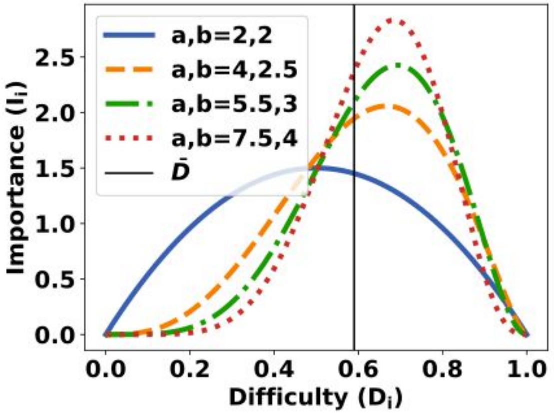
Importance Sampling Function

Proposition 1 (Setting a and b for desired `Mode` and `Variance` for the importance sampling function)**.** *By setting* $a = 1 + \bar{D} + c_a|\mathcal{S}|$ *and* $b = 2 + c_b|\mathcal{S}|$, *where* $c_a > c_b > 0$, *the importance function meets the following three properties:*

- P_1: `Mode` *increases with* $|\mathcal{S}|$ *and* $\bar{D}$*;*
- P_2: `Mode` $> \bar{D}$ *generally holds true;*
- P_3: `Variance` *decreases with* $|\mathcal{S}|$ *and* $\bar{D}$ *under mild conditions* ($c_a < c_b b$)*.*



CIFAR 10



Tiny ImageNet

Balanced Subset Selection Function

- The importance function is combined with a facility location function:

$$F(\mathcal{S}) = \sum_{i \in \mathcal{V}} \max_{j \in \mathcal{S}} \texttt{Sim}(\mathbf{x}_i, \mathbf{x}_j) \mathcal{I}(\mathbf{x}_j, \mathbf{y}_j)$$

- **Optimum subset** is selected using a greedy algorithm that starts with an empty subset and keeps on adding samples to the subset that maximizes the gain:

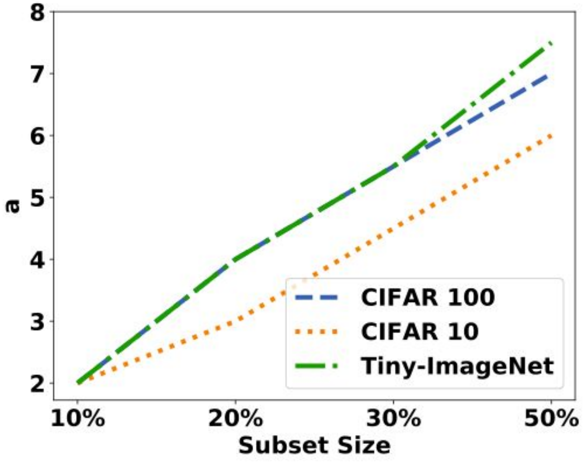
$$F((\mathbf{x}_j, \mathbf{y}_j)|\mathcal{S}) = F(\mathcal{S} \cup (\mathbf{x}_j, \mathbf{y}_j)) - F(\mathcal{S})$$

Empirical Analysis

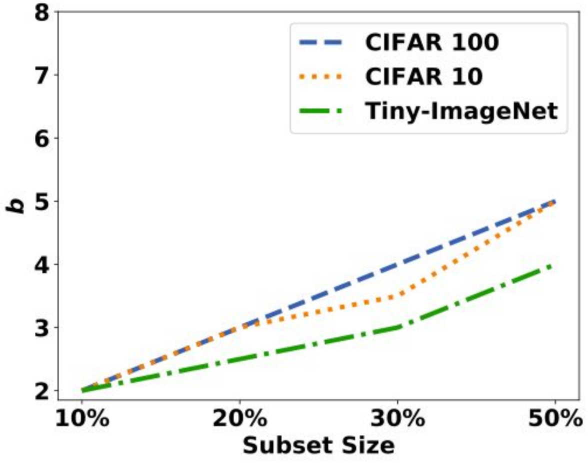
Dataset	Subset	Random	CRAIG	GradMatch	Adacore	LCMAT	Moderate	CCS	BOSS(Ours)
Tiny ImageNet	10%	24.11	24.61	23.68	24.12	23.26	24.16	29.59	**32.54**
	20%	37.67	37.76	38.20	37.94	36.71	37.57	40.42	**44.49**
	30%	45.12	44.63	44.93	44.72	44.06	45.30	47.11	**51.21**
	50%	53.07	53.03	53.81	53.37	53.10	53.31	55.11	**57.77**
CIFAR 100	10%	37.35	38.67	36.68	37.65	37.23	37.76	40.26	**46.54**
	20%	51.55	51.44	53.16	52.79	53.11	50.90	55.48	**61.76**
	30%	62.89	62.92	63.02	62.28	62.25	62.55	64.61	**67.73**
	50%	70.67	70.69	70.68	71.19	70.53	71.13	71.53	**73.93**
CIFAR 10	10%	70.69	70.96	72.26	72.65	71.03	72.04	74.78	**78.27**
	20%	83.27	83.36	84.30	84.30	83.98	83.64	86.45	**88.14**
	30%	88.89	88.98	88.47	88.37	88.54	88.46	91.49	**92.14**
	50%	92.69	92.75	91.89	92.67	92.58	92.61	93.45	**94.46**
SVHN	8%	84.98	84.30	84.31	82.31	84.05	84.51	86.69	**88.83**
	12%	87.16	88.49	88.99	88.41	87.49	88.97	92.16	**93.16**
	16%	90.47	89.92	90.42	90.34	90.16	90.35	93.87	**94.51**
	20%	91.64	92.13	91.56	91.95	91.36	91.30	94.38	**95.15**

Empirical Analysis

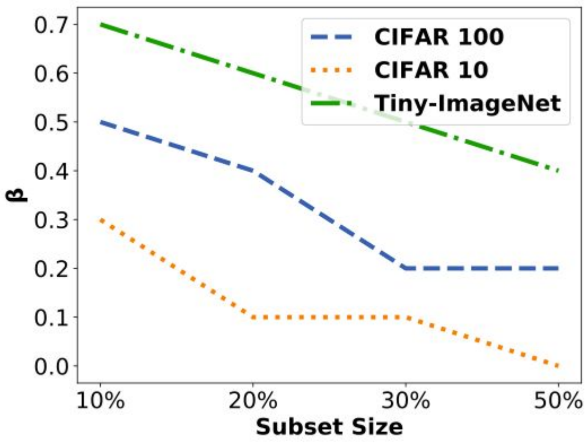
- **Optimal values** of parameters ***a*** and ***b***
- **Cutoff rate parameter *β*** to ensure **robust selection**



(a) Parameter *a*



(b) Parameter *b*



(c) Parameter *β*

Conclusion

- Subset selection is an **important direction** to alleviate the **resource consumption**
- Existing techniques do not consider the **joint distribution** of diversity and difficulty
- **We propose a novel strategy** to balance diversity and difficulty for a subset size.
- We provide **theoretical analysis** leading to an **novel importance function**.
- The **empirical results** on real-world data show the **effectiveness of our method**.

Thank You!