BASIL: Balanced Active Semi-supervised Learning for Class Imbalanced Datasets

Suraj Kothawade1, Pavan Reddy2, Ganesh Ramakrishnan2, and Rishabh Iyer1

1 University of Texas at Dallas, USA
2 Indian Institute of Technology, Bombay, India
suraj.kothawade@utdallas.edu

Abstract. Current semi-supervised learning (SSL) methods assume a balance between the number of data points available for each class in both the labeled and the unlabeled data sets. However, there naturally exists a class imbalance in most real-world datasets. It is known that training models on such imbalanced datasets leads to biased models, which in turn lead to biased predictions towards the more frequent classes. This issue is further pronounced in SSL methods, as they would use this biased model to obtain pseudo-labels (on the unlabeled data) during training. In this paper, we tackle this problem by attempting to select a balanced labeled dataset for SSL that would result in an unbiased model. Unfortunately, acquiring a balanced labeled dataset from a class imbalanced distribution in one shot is challenging. We propose BASIL (Balanced Active Semi-supervised Learning), a novel algorithm that optimizes the submodular mutual information (SMI) functions in a per-class fashion to gradually select a balanced dataset in an active learning loop. Importantly, our technique can be efficiently used to improve the performance of any SSL method. Our experiments on Path-MNIST and Organ-MNIST medical datasets for a wide array of SSL methods show the effectiveness of BASIL. Furthermore, we observe that BASIL outperforms the state-of-the-art diversity and uncertainty based active learning methods since the SMI functions select a more balanced dataset.

1 Introduction

Deep neural networks (DNNs) have proven to be successful on a variety of machine learning tasks. However, they are mostly fueled by large amounts of data. These DNNs can be trained using multiple objective functions that require labeled or unlabeled data. Towards this end, we face multiple data related challenges in training DNNs. Firstly, most real-world datasets have a natural class-imbalance. For instance, in the medical imaging domain, cancerous images are rare in comparison to non-cancerous images. Secondly, obtaining labeled data is notoriously time-consuming and expensive. This issue is highly pronounced in the biomedical domain where the annotators need to be well compensated since they are experts like doctors, radiologists, etc. Thirdly, in many scenarios, even procuring unlabeled data is challenging. For example, acquiring few samples of medical data of
a new disease is rare and involves several privacy constraints. Hence, it is critical to use the unlabeled data even when a few labeled data points are available.

The above introduced data related issues are well-known, and the community has devised several techniques to tackle these issues separately. For mitigating labeling costs, active learning (AL) \[2, 13, 11, 12, 23, 24\] is an established paradigm that samples uncertain or diverse data points from an unlabeled set. The goal is to acquire a subset that entails the largest improvement in performance of the model. Another technique, called semi-supervised learning (SSL) \[16, 20, 15, 3, 26, 28\] leverages the unlabeled data when only a small amount of labeled data is available. Lastly, several subset selection techniques that tackle class imbalance \[12, 14, 13\] have also been proposed. Evidently, these techniques revolve around the idea of obtaining the best possible model at a minimum cost. However, individually, each of these techniques suffer from limitations that the other ones do not. For example, existing AL and SSL techniques are known to suffer from class imbalance, thereby leading to learning biased models. Also, existing AL and subset selection methods do not leverage the remaining unlabeled data, and simply discard it. To bridge these gaps in existing methods, we propose BASIL, a unified framework that actively samples data points per-class to create a balanced labeled set followed by SSL to make the most of the remaining unlabeled data.

1.1 Related work

**Active Learning (AL).** Uncertainty based methods aim to select the most uncertain data points for labeling. The most common technique is entropy \[24\] that aims to select data points with maximum entropy. The main drawback of uncertainty based methods is that they lack diversity within the acquired subset. To mitigate this, a number of approaches have proposed to incorporate diversity. A recent approach called Badge \[2\] uses the last linear layer gradients to represent data points and runs K-means++ \[1\] to obtain centers that have a high gradient magnitude. The centers being representative and having high gradient magnitude ensures uncertainty and diversity at the same time. However, for batch active learning, this diversity and uncertainty are limited within the batch and not across all batches. Another method, BatchBald \[11\] requires a large number of Monte Carlo dropout samples to obtain significant mutual information which limits its application to medical domains where data is scarce. Recently, \[12\] proposed the use of submodular information measures for active learning in realistic scenarios, while \[13\] used them to find rare objects in an autonomous driving object detection problem. However, they focus on acquiring data points only from the rare classes or slices. Our proposed method maximizes the per-class mutual information, thereby selecting data points for each class to obtain a balanced labeled set, which is critical for training unbiased models.

**Semi-supervised Learning (SSL).** The goal of SSL methods is to leverage unlabeled data alongside the labeled data to obtain a better representation of the dataset than supervised learning \[24\]. The most basic SSL method, pseudolabeling \[10\] uses model predictions as target labels as a regularizer, and a
standard supervised loss function for the unlabeled dataset. Some SSL methods such as Π-Model [15, 22] and Mean Teacher [26] use consistency regularization, by using data augmentation and dropout techniques. Mean Teacher obtains a more stable target output by using an exponential moving average of parameters across previous epochs. Virtual Adversarial Training (VAT) [20] uses an effective regularization technique that involves slight perturbations such that the prediction of the unlabeled samples is affected the most. ICT [28] encourages the prediction at an interpolation of unlabeled points to be consistent with the interpolation of the predictions at those points. More recent techniques like FixMatch [25], MixMatch [3] and UDA [29] use data augmentations like flip, rotation, and crops to predict pseudo-labels. All the above methods depend on the model trained using a small labeled set. Hence, they are susceptible to using a biased model if the labeled set is randomly sampled from an unlabeled set with class imbalance. In this paper, we study the effect of selecting a balanced seed set using Basil for a wide array of SSL techniques.

1.2 Our contributions

We summarize our contributions as follows: 1) We emphasize on the need of jointly addressing multiple real-world data related problems such as class imbalance, expensive labeling costs, and leveraging the unlabeled data. Particularly, we show that these problems co-exist in the medical domain. 2) We propose Basil, a novel algorithm that can tackle these problems in an end-to-end manner. Concretely, we acquire a balanced subset of the unlabeled data by maximizing per-class instantiations of submodular mutual information functions in an active learning loop followed by semi-supervised learning (see Fig. 1). 3) Basil can leverage any SSL method and yield improved performance over the vanilla SSL approach. 4) We evaluate the effectiveness of Basil on two diverse modalities of medical data. Namely, histopathology (Path-MNIST [9]) and Abdominal CT (Organ-MNIST [10]). 5) We conduct rigorous experiments on 6 AL strategies and 8 SSL techniques, and show that balanced labeled set selection using Basil outperforms existing AL methods and obtains larger gain in performance for various SSL techniques (see Tab. 1 and Tab. 2).

2 Preliminaries

**Submodular Functions:** We let \( V \) denote the ground-set of \( n \) data points \( V = \{1, 2, 3, ..., n\} \) and a set function \( f: 2^V \to \mathbb{R} \). The function \( f \) is submodular [4] if it satisfies the diminishing marginal returns, namely \( f(j|\mathcal{X}) \geq f(j|\mathcal{Y}) \) for all \( \mathcal{X} \subseteq \mathcal{Y} \subseteq V, j \notin \mathcal{Y} \). Facility location, graph cut, log determinants, etc. are some examples [8].

**Submodular Mutual Information (SMI):** Given a set of items \( \mathcal{A}, \mathcal{Q} \subseteq V \), the submodular mutual information (MI) [5, 7] is defined as \( I_f(\mathcal{A}; \mathcal{Q}) = f(\mathcal{A}) + f(\mathcal{Q}) - f(\mathcal{A} \cup \mathcal{Q}) \). Intuitively, this measures the similarity between \( \mathcal{Q} \) and \( \mathcal{A} \) and we refer to \( \mathcal{Q} \) as the query set.
Kothawade et al. [14] extend SMI to handle the case when the target can come from a different set $V'$ apart from the ground set $V$. In the context of imbalanced medical image classification, $V$ is the source set of images and the query set $Q$ is the target set containing the rare class images. To find an optimal subset given a query set $Q \subseteq V'$, we can define $g_Q(A) = I_f(A; Q)$, $A \subseteq V$ and maximize the same.

### 2.1 Examples of SMI functions

For balanced subset selection via BASIL we use the recently introduced SMI functions in [7, 5] and their extensions introduced in [14] as acquisition functions. Note that we only use a subset of functions presented in [14], that are the most scalable, for per-class selection of data points. For any two data points $i \in V$ and $j \in Q$, let $s_{ij}$ denote the similarity between them.

**Graph Cut MI (GCMI):** The SMI instantiation of graph-cut (GCMI) is defined as: $I_{GC}(A; Q) = 2 \sum_{i \in A} \sum_{j \in Q} s_{ij}$. Since maximizing GCMI maximizes the joint pairwise sum with the query set, it will lead to a summary similar to the query set $Q$. In fact, specific instantiations of GCMI have been intuitively used for query-focused summarization for videos [27] and documents [18, 17].

**Facility Location MI (FLMI):** We consider two variants of FLMI. In the first variant of facility location which is defined over $V$ (FLVMI), the SMI instantiation can be defined as: $I_{FLV}(A; Q) = \sum_{i \in V} \min(\max_{j \in A} s_{ij}, \max_{j \in Q} s_{ij})$. The first term in the min(.) of FLVMI models diversity, and the second term models query relevance.

For the second variant, which is defined over $Q$ (FLQMI), the SMI instantiation can be defined as: $I_{FLQ}(A; Q) = \sum_{i \in Q} \max_{j \in A} s_{ij} + \sum_{i \in A} \max_{j \in Q} s_{ij}$. FLQMI is very intuitive for query relevance as well. It measures the representation of data points that are the most relevant to the query set and vice versa. It can also be thought of as a bidirectional representation score.

![Fig. 1: Architecture of BASIL](image)

Fig. 1: Architecture of BASIL. We select a balanced labeled set $A \subseteq U$ by maximizing the per-class submodular mutual information in an active learning loop. The balanced labeled set is used to train an unbiased model which better leverages the remaining unlabeled data $U \setminus A$ using semi-supervised learning.
3 BASIL: Our Active Semi-supervised Learning framework

In this section, we present BASIL, a unified framework that jointly tackles data problems of class imbalance, high labeling costs and leveraging unlabeled data. We do so by gradually acquiring balanced subsets in an active learning loop, followed by training the final model using the balanced labeled set and the remaining unlabeled set (see Fig. 1).

Algorithm 1 BASIL: Balanced Active Semi-supervised Learning

Require: Labeled set of data points: \( \mathcal{L} \), large unlabeled dataset: \( \mathcal{U} \), Loss function defined over \( \mathcal{L} \): \( \mathcal{H}_\mathcal{L} \), Loss function defined over \( \mathcal{U} \): \( \mathcal{H}_\mathcal{U} \), batch size: \( B \), number of selection rounds: \( N \), number of classes: \( C \)

1: for selection round \( i = 1 : N \) do
2: \( \mathcal{G}_\mathcal{L} \leftarrow \{ \nabla \theta \mathcal{H}_\mathcal{L}(x, \hat{y}, \theta) : \forall y \in \mathcal{L} \} \) \{Compute gradients of \( \mathcal{L} \) using true labels\}
3: \( \mathcal{G}_\mathcal{U} \leftarrow \{ \nabla \theta \mathcal{H}_\mathcal{L}(x, \hat{y}, \theta) : \forall y \in \mathcal{U} \} \) \{Compute gradients of \( \mathcal{U} \) using hypothesized labels\}
4: \( \mathcal{X} \leftarrow \text{Cosine Similarity}(\mathcal{G}_\mathcal{L}, \mathcal{G}_\mathcal{U}) \) \{\( \mathcal{X} \in \mathbb{R}^{C \times |\mathcal{U}|} \}\}
5: for class \( c \in \{1 \cdots C\} \) do
6: Instantiate a SMI function \( I_f \) based on \( \mathcal{X}^c \) \{\( \mathcal{X}^c \) contains rows from \( \mathcal{X} \) corresponding to class \( c \)\}
7: \( \mathcal{A}^c_i \leftarrow \text{argmax}_{A^c_i \subseteq \mathcal{U}, |A^c_i| \leq (B/C)I_f(A^c_i; Q^c)} \{\text{Q}^c \subseteq \mathcal{L} \text{ contains data points from class } c\}\}
8: \( A_i \leftarrow A_i \cup A^c_i \)
10: end for
11: Get labels \( L(A_i) \) for batch \( A_i \) and \( \mathcal{L} \leftarrow \mathcal{L} \cup L(A_i) \), \( \mathcal{U} \leftarrow \mathcal{U} - A_i \)
12: end for
13: Train final model \( M_\theta \) with loss \( \mathcal{H}_\mathcal{L} + \mathcal{H}_\mathcal{U} \) \{Train model on balanced \( \mathcal{L} \) and remaining \( \mathcal{U} \) using SSL\}
14: Return trained model \( M \) and parameters \( \theta \).

The main idea in BASIL is to maximize per-class instantiations of submodular mutual information (SMI) functions to obtain a balanced labeled set. Concretely, we formulate an SMI function using a query set \( Q^c \) that is a subset of the current labeled set \( \mathcal{L} \) with data points from class \( c \). The SMI functions are instantiated using a similarity kernel \( \mathcal{X} \), where \( X_{ij} \) is the pairwise similarity between data points \( i \) and \( j \), represented by gradients computed using \( M \). Specifically, we define \( X_{ij} = \langle \nabla \theta \mathcal{H}_\mathcal{L}(x_i, \hat{y}_i, \theta), \nabla \theta \mathcal{H}_\mathcal{L}(x_j, \hat{y}_j, \theta) \rangle \), where \( \mathcal{H}_\mathcal{L}(\theta) = \mathcal{H}_\mathcal{L}(x_i, \hat{y}_i, \theta) \) is the labeled loss on the \( i \)th data point. Note that we use hypothesized labels, i.e., the label with maximum class probability, for computing gradients of \( \mathcal{U} \). Next, we optimize the SMI function for class \( c \) using a greedy strategy [19] with a constraint that the budget is equally divided across all classes:

\[
\max_{A^c_i \subseteq \mathcal{U}, |A^c_i| \leq (B/C)} I_f(A^c_i; Q^c) 
\]
Hence, BASIL gradually builds a balanced labeled set $\mathcal{L}$ by accumulating smaller balanced sets $\mathcal{A}_i$ in every AL round. Finally, we can use any SSL technique to train the model $\mathcal{M}$ with the balanced labeled set $\mathcal{L}$ and the remaining unlabeled set $\mathcal{U}\setminus\mathcal{L}$. We summarize our method in Algo. 1 illustrate the architecture in Fig. 1 and discuss its scalability in Appendix. C.

4 Experiments

In this section, we evaluate the effectiveness of BASIL on two modalities of medical data, viz., histopathology (see Sec. 4.1) using the Path-MNIST dataset [9, 30] and Abdominal CT (see Sec. 4.2) using the Organ-MNIST dataset [10, 30]. For evaluation, we compare the test accuracy of the model obtained after training via a semi-supervised learning algorithm on a combination of labeled set (selected using active learning) and the unlabeled set. We also compare the imbalance ratio (IR) of all AL methods in Fig. 2. The IR is computed as follows: $IR(\mathcal{A}) = \frac{|\mathcal{A}_F|*|\mathcal{R}|}{|\mathcal{A}_R|*|\mathcal{F}|}$, where $\mathcal{R}$ contains class indices of the rare classes and $\mathcal{F}$ contains class indices of the remaining frequent classes. Note that $IR(\mathcal{A})=1$ when $\mathcal{A}$ is perfectly balanced.

Our results show that selecting a balanced labeled set using BASIL (see Fig. 2) outperforms existing AL baselines (supervised row in Tab. 1, 2). Importantly, we observe that, independent of the SSL algorithm used, using a balanced labeled set selected using BASIL helps leverage the remaining imbalanced unlabeled set better.

Baseline AL methods and SSL techniques. We compare the performance of BASIL against an uncertainty based AL method (ENTROPY), and a diversity based AL method (BADGE). We discuss the details of all baselines in Sec. 1.1. Lastly, we compare with random sampling (RANDOM). We evaluate the subset selected by BASIL and the AL baselines on a wide array of SSL algorithms such as Psuedo-Label (PL), Mean Teacher (MT), II-Model, MixMatch (MM), ICT, VAT, VAT + Entropy Minimization (EM) on two medical imaging datasets which are described in Sec. 4.1 and Sec. 4.2.

Experimental setup: We use the same training procedure and hyperparameters for all AL methods to ensure a fair comparison. For the first AL round, we randomly sample data points for labeling from the unlabeled set $\mathcal{U}$. We do so in order to obtain meaningful model parameters for the AL acquisition functions in the next round of AL. For all experiments, we train a Wide ResNet (WRN) [6] model using an Adam optimizer with an initial learning rate of 3e-4. For each AL round, the weights are reinitialized using Xavier initialization and the model is trained for 100K iterations. After obtaining the labeled set using AL, we initialize the WRN model and train it using SSL for 500K iterations. We run each experiment $3\times$ on a V100 GPU and provide the error bars (std deviation). We discuss dataset splits and hyperparameters below and provide more details in Appendix. B.

4.1 Analysis on Histopathology data

Path-MNIST Dataset: Path-MNIST [9] is a dataset based on a prior study for predicting survival from colorectal cancer histology slides. It includes a training dataset of 100,000 non-overlapping image patches from hematoxylin and eosin
stained histological images, and a test dataset of 7,180 image patches from a different clinical center. These patches are categorized into 9 types of tissues, resulting in a multi-class classification task. For our experiments, we use a preprocessed version of Path-MNIST [30], where the original images are resized to $3 \times 28 \times 28$. We consider a subset of 26K data points to create the initial imbalanced unlabeled set $U$. The unlabeled set is imbalanced by randomly choosing 4 rare classes and randomly selecting 250 data points for each class. For the remaining 5 classes, we randomly select 5000 data points for each class.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{SSL} \setminus \text{AL} & \text{Random} & \text{Badge} & \text{Entropy} & \text{Flqmi} & \text{Gcmi} & \text{Flvmi} \\
\hline
\text{Supervised} & 69.37 \pm 1.33 & 76.56 \pm 2.4 & 70.54 \pm 1.76 & 78.51 \pm 0.58 & 78.76 \pm 1.84 & 77.68 \pm 0.58 \\
\text{PL} [16] & 69.38 \pm 2.19 & 80.11 \pm 1.68 & 66.16 \pm 1.8 & 76.66 \pm 2.3 & 81.24 \pm 0.91 & 78.51 \pm 0.58 \\
\text{ICT} [28] & 75.45 \pm 0.28 & 80.25 \pm 0.17 & 74.91 \pm 0.78 & 82.29 \pm 0.96 & 82.58 \pm 0.13 & 82.08 \pm 0.66 \\
\text{Π-Model} [15] & 64.59 \pm 0.19 & 78.63 \pm 0.03 & 67.29 \pm 0.73 & 80.62 \pm 1.72 & 80.59 \pm 1.19 & 80.77 \pm 0.71 \\
\text{MT} [26] & 70.98 \pm 1.76 & 79.62 \pm 0.85 & 74.19 \pm 1.56 & 82.14 \pm 0.91 & 82.43 \pm 1.44 & 83.36 \pm 0.52 \\
\text{MM} [3] & 67.13 \pm 1.78 & 67.77 \pm 1.43 & 67.01 \pm 0.44 & 73.78 \pm 1.22 & 76.08 \pm 0.09 & 73.42 \pm 1.54 \\
\text{VAT} [20] & 80.27 \pm 1.95 & 83.05 \pm 1.72 & 77.49 \pm 1.8 & 85.72 \pm 1.83 & 83.7 \pm 0.23 & 84.38 \pm 0.81 \\
\text{VAT+EM} [20] & 81.48 \pm 1.1 & 84.11 \pm 0.06 & 82.6 \pm 0.75 & 85.46 \pm 1.54 & 84.12 \pm 1.94 & 85.22 \pm 2.15 \\
\hline
\end{array}
\]

Table 1: Active SSL on Path-MNIST. Test accuracies in bold (one in each row) show the best performing AL acquisition functions for a particular SSL method. We observe that Flqmi which models query-relevance and representation obtains the highest gains for VAT and VAT+EM. Whereas, Flvmi acquires the best labeled set for PL, Π-Model and MT.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{SSL} \setminus \text{AL} & \text{Supervised} & \text{PL} & \text{ICT} \\
\hline
\text{Supervised} & 69.37 \pm 1.33 & 76.56 \pm 2.4 & 70.54 \pm 1.76 \\
\text{PL} [16] & 69.38 \pm 2.19 & 80.11 \pm 1.68 & 66.16 \pm 1.8 \\
\text{ICT} [28] & 75.45 \pm 0.28 & 80.25 \pm 0.17 & 74.91 \pm 0.78 \\
\text{Π-Model} [15] & 64.59 \pm 0.19 & 78.63 \pm 0.03 & 67.29 \pm 0.73 \\
\text{MT} [26] & 70.98 \pm 1.76 & 79.62 \pm 0.85 & 74.19 \pm 1.56 \\
\text{MM} [3] & 67.13 \pm 1.78 & 67.77 \pm 1.43 & 67.01 \pm 0.44 \\
\text{VAT} [20] & 80.27 \pm 1.95 & 83.05 \pm 1.72 & 77.49 \pm 1.8 \\
\text{VAT+EM} [20] & 81.48 \pm 1.1 & 84.11 \pm 0.06 & 82.6 \pm 0.75 \\
\hline
\end{array}
\]

Table 1: Active SSL on Path-MNIST. Test accuracies in bold (one in each row) show the best performing AL acquisition functions for a particular SSL method. We observe that Flqmi which models query-relevance and representation obtains the highest gains for VAT and VAT+EM. Whereas, Flvmi acquires the best labeled set for PL, Π-Model and MT.

Fig. 2: Comparison of Imbalance Ratio (IR) on Path-MNIST (Left) and Organ-MNIST (Right) datasets (Lower the better). We observe that the SMI functions select more data points from the rare classes, thereby having a $\approx 4 \times$ to $5 \times$ lower IR than the best performing baseline and $\approx 10 \times$ to $20 \times$ lower IR than Random.

**Results:** We present results for Active SSL on Path-MNIST in Tab. [1] We observe that the SMI based AL acquisition functions outperform existing AL methods in the supervised and across all semi-supervised learning methods. This is due to the fact that per-class selection using SMI functions in Basil results in a more balanced labeled set (see Fig. 2). This reinforces the need for a framework like Basil for training models in a supervised or semi-supervised manner in class
imbalance scenarios. Interestingly, we observe that the choice of SMI function depends on the modality of medical data and the SSL method. For histopathology, we see that PL, II-Model and MT methods perform the best when an acquisition function like Flvmi that balances between query-relevance and diversity is used. For ICT and MM, Gcmi shows the best results, which models query relevance only. Whereas, on VAT and VAT+EM, Flqmi shows the best results which models query-relevance and representation.

4.2 Analysis on Abdominal CT data

Organ-MNIST Dataset: Organ-MNIST [10] is a dataset based on 3D computed tomography (CT) images from Liver Tumor Segmentation Benchmark (LiTS). For our experiments, we use a pre-processed version of Organ-MNIST [30], where images are cropped using bounding-box annotations of 11 body organs. The images are resized into $1 \times 28 \times 28$ to perform multi-class classification of 11 body organs. We consider a subset of 21.6K data points to create the initial imbalanced unlabeled set $U$. The unlabeled set is imbalanced by randomly choosing 4 classes and randomly selecting 150 data points for each class. For the remaining 5 classes, we randomly select 300 data points for each class.

Results: We present results for Active SSL on Organ-MNIST in Tab. 2. Similar to our results on Path-MNIST, we observe that the SMI based AL acquisition functions outperform existing AL methods across all supervised and SSL methods. We observe that the facility location based SMI functions dominate for the Abdominal CT modality. Particularly, Flqmi which models query-relevance and representation, outperforms other baselines and SMI functions for all SSL methods except VAT. The Flvmi functions performs slightly better than Flqmi when the SSL method is VAT or VAT+EM.

| SSL \ AL   | RANDOM BADGE | ENTROPY | Flqmi | Gcmi | Flvmi |
|------------|--------------|---------|-------|------|-------|
| Supervised | 60.63±0.73   | 64.32±0.29 | 61.03±0.36 | 65.58±0.28 | 62.15±0.63 | 64.41±0.66 |
| PL [16]    | 62.83±1.79   | 64.43±0.47 | 64.27±0.16 | 67.68±1.44 | 64.48±1.3 | 65.56±0.35 |
| ICT [28]   | 61.1±0.57    | 60.29±1.52 | 64.01±1.27 | 65.32±0.86 | 60.33±1.37 | 62.22±1.24 |
| II-Model [15] | 64.61±1.32 | 65.61±1.73 | 62.67±1.83 | 65.94±1.09 | 61.85±1.62 | 65.43±0.94 |
| MT [26]    | 64.64±0.63   | 66.49±1.18 | 65.53±0.34 | 66.89±0.01 | 60.81±0.63 | 64.81±0.54 |
| MM [3]     | 53.62±1.28   | 50.34±1.4 | 51.35±1.51 | 58.57±0.51 | 55.86±0.73 | 56.08±1.07 |
| VAT [20]   | 71.82±1.98   | 70.48±0.53 | 72.17±0.97 | 75.25±0.73 | 72.51±1.62 | 76.17±0.1 |
| VAT+EM [20] | 72.67±0.28 | 73.52±0.22 | 71.95±0.43 | 75.57±1.12 | 73.17±1.73 | 76.58±0.69 |

Table 2: Active SSL on Organ-MNIST. Test accuracies in bold (one on each row) show the best performing AL method for each SSL. We see that Flqmi which models query-relevance and representation obtains the highest gains for most SSL methods except VAT, where Flvmi consistently performs better.
5 Conclusion

We demonstrate the effectiveness of a unifying algorithm like BASIL for selecting a balanced labeled set in scenarios with class imbalanced data. Through rigorous experiments on diverse modalities of medical datasets, we show that BASIL selects a more balanced labeled set than other AL acquisition functions, resulting in relatively unbiased models leading to better performance for supervised and semi-supervised learning.
Bibliography

[1] Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007)

[2] Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. In: ICLR (2020)

[3] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)

[4] Fujishige, S.: Submodular functions and optimization. Elsevier (2005)

[5] Gupta, A., Levin, R.: The online submodular cover problem. In: ACM-SIAM Symposium on Discrete Algorithms (2020)

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[7] Iyer, R., Khargoankar, N., Bilmes, J., Asnani, H.: Submodular combinatorial information measures with applications in machine learning. arXiv preprint arXiv:2006.15412 (2020)

[8] Iyer, R.K.: Submodular optimization and machine learning: Theoretical results, unifying and scalable algorithms, and applications. Ph.D. thesis (2015)

[9] Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine 16(1), e1002730 (2019)

[10] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5), 1122–1131 (2018)

[11] Kirsch, A., Van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. arXiv preprint arXiv:1906.08158 (2019)

[12] Kothawade, S., Beck, N., Killamsetty, K., Iyer, R.: Similar: Submodular information measures based active learning in realistic scenarios. Advances in Neural Information Processing Systems 34 (2021)

[13] Kothawade, S., Ghosh, S., Shekhar, S., Xiang, Y., Iyer, R.: Talisman: Targeted active learning for object detection with rare classes and slices using submodular mutual information. arXiv preprint arXiv:2112.00166 (2021)

[14] Kothawade, S., Kaushal, V., Ramakrishnan, G., Bilmes, J., Iyer, R.: Prism: A rich class of parameterized submodular information measures for guided subset selection. arXiv preprint arXiv:2103.00128 (2021)
[15] Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
[16] Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in LRtation learning, ICML. vol. 3 (2013)
[17] Li, J., Li, L., Li, T.: Multi-document summarization via submodularity. Applied Intelligence 37(3), 420–430 (2012)
[18] Lin, H.: Submodularity in natural language processing: algorithms and applications. Ph.D. thesis (2012)
[19] Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J., Krause, A.: Lazier than lazy greedy. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
[20] Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence 41(8), 1979–1993 (2018)
[21] Oliver, A., Odena, A., Raffel, C., Cubuk, E.D., Goodfellow, I.J.: Realistic evaluation of deep semi-supervised learning algorithms. arXiv preprint arXiv:1804.09170 (2018)
[22] Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems 29, 1163–1171 (2016)
[23] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018)
[24] Settles, B.: Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2009)
[25] Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
[26] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)
[27] Vasudevan, A.B., Gygli, M., Volokitin, A., Van Gool, L.: Query-adaptive video summarization via quality-aware relevance estimation. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 582–590 (2017)
[28] Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. arXiv preprint arXiv:1903.03825 (2019)
[29] Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)
[30] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. arXiv preprint arXiv:2008 (2021)
Supplementary Material

A Summary of Notations

| Topic        | Notation | Explanation                                                                 |
|--------------|----------|-----------------------------------------------------------------------------|
| BASIL (Sec. 3) | $\mathcal{U}$ | Unlabeled set of $|\mathcal{U}|$ instances                             |
|              | $A$      | A subset of $\mathcal{U}$                                                   |
|              | $S_{ij}$ | Similarity between any two data points $i$ and $j$                         |
|              | $f$      | A submodular function                                                       |
|              | $\mathcal{L}$ | Labeled set of data points                                                |
|              | $\mathcal{Q}$ | Query set                                                                 |
|              | $\mathcal{Q}^c$ | Query set containing data points from class $c$ for per-class SMI selection, $\mathcal{Q}^c \subseteq \mathcal{L}$ |
|              | $\mathcal{M}$ | Deep model                                                                 |
|              | $B$      | Active learning selection budget                                            |
|              | $\mathcal{H}_\mathcal{L}$ | Labeled Loss function used to train model $\mathcal{M}$ and compute gradients |
|              | $\mathcal{H}_\mathcal{U}$ | Unlabeled Loss function used for semi-supervised learning                  |
|              | $\mathcal{X}'$ | Pairwise similarity matrix computed using gradients                         |
|              | $G_A$    | Gradients of some subset $A$                                                 |

Table 3: Summary of notations used throughout this paper

B Details of Datasets and Experimental setting

B.1 Datasets

In this section, we will describe details of the datasets Path-MNIST (Histopathology data) and Organ-MNIST (Abdominal CT data).

Path-MNIST A dataset based on a prior study for predicting survival from colorectal cancer histology slides, which provides 100,000 non-overlapping image patches from hematoxylin and eosin stained histological images, and a test dataset of 7,180 image patches from a different clinical center. 9 types of tissues are involved, resulting a multi-class classification task. We resize the source images of $3 \times 224 \times 224$ into $3 \times 28 \times 28$.

Out of 100K images we take 5K images each from classes 0,1,4,6,8 and 250 images each from classes 2,3,5,7 and form a unlabeled train dataset of 26K images and for
validation set we had 10 samples each from all classes thus forming a validation set of size 90. For the test dataset we used default one which consists of 7180 samples. For Path-MNIST we had active learning selection budget \( B \) as 900.

**Organ-MNIST** A dataset based on 3D computed tomography (CT) images from Liver Tumor Segmentation Benchmark (LiTS). Hounsfield-Unit (HU) of the 3D images are transformed into grey scale with a abdominal window; we then crop 2D images from the center slices of the 3D bounding boxes in axial views (planes). The images are resized into 1 x 28 x 28 to perform multi-class classification of 11 body organs.

There are total of 34581 training samples out of which we pick 3000 samples from classes 4,5,6,7,8,9,10 and 150 samples from classes 0,1,2,3 and form an unlabeled train dataset of 21.6K images. And for validation set we have selected 10 points each from a class to form a validation set of size 110. And the test set consists a total of around 17K images. For Organ-MNIST we had active learning selection budget \( B \) as 990.

### B.2 Experimental Setting

Given an unlabeled dataset \( U \) and an active learning selection budget \( B \), we have to select \( B/N \) points in each round of active learning. In the first round, we select \( B/N \) points randomly from the unlabeled dataset and label those points. Then in further rounds we train the Wide-ResNet model using an Adam optimizer with an initial learning rate of \( 3e^{-3} \) (same across different active learning methods to make better comparison) for 100K iterations using the labeled set formed till then and extract the gradients of remaining unlabeled dataset from model and use the gradients to perform per-class selection from SMI functions to get a Balanced Dataset.

After forming a labeled set of size \( B \), we perform Semi-Supervised learning using labeled set \( L \) and unlabeled dataset \( U \). We use many SSL algorithms to evaluate our selected balanced datasets such as Psuedo-Label, ICT, \( II \)-Model, Mean-Teacher, MixMatch, VAT, VAT+EM. For each of these methods the loss
component consists of two components, they are supervised loss (loss corresponding to labeled dataset) and SSL loss (loss corresponding to unlabeled dataset). Supervised loss is common to all, whereas SSL loss depends on the algorithm which we are using. We use the same parameters for one SSL method across different selections of active learning for better comparison.

Parameters used for each SSL method

- **Supervised**: "lr": $3e^{-3}$
- **PL**: "threshold": 0.95, "lr": $3e^{-4}$, "consistency-coefficient": 1
- **ICT**: "ema_factor": 0.999, "lr": $4e^{-4}$, "consistency-coefficient": 100, "alpha": 0.1
- **II-Model**: "lr": $3e^{-4}$, "consistency-coefficient": 20.0
- **MT**: "ema_factor": 0.95, "lr": $4e^{-4}$, "consistency-coefficient": 8
- **MM**: "lr": $3e^{-3}$, "consistency-coefficient": 100, "alpha": 0.75, "T": 0.5, "K": 2
- **VAT**: "xi": $1e^{-6}$, "lr": $3e^{-3}$, "consistency-coefficient": 0.3, "eps": 6
- **VAT+EM**: "xi": $1e^{-6}$, "lr": $3e^{-3}$, "consistency-coefficient": 0.3, "eps": 6, "em": 0.06

Details on the computation of Imbalance Ratio: The IR is computed as follows: $IR(A) = \frac{|A_R||R|}{|A_F||F|}$, where $R$ contains class indices of the rare classes and $F$ contains class indices of the remaining frequent classes. For example, for a dataset with 5 total classes, assume it has 2 rare classes, $R = \{1, 2\}$, then size of $|R| = 2$. The remaining classes are frequent classes, $F = \{3, 4, 5\}$ and, $|F| = 3$. Further, $A_R$ denotes the set of data points that belong to the rare classes. Similarly, $A_F$ denotes the set of data points that belong to the frequent classes. Note that $IR(A)=1$ when $A$ is perfectly balanced.

C Scalability of BASIL

Below, we provide a detailed analysis of the complexity of creating and optimizing the different SMI functions. Denote $|X|$ as the size of set $X$. Also, let $|U| = n$ (the ground set size, which is the size of the unlabeled set in this case).

- **Facility Location**: We start with FLVMI. The complexity of creating the kernel matrix is $O(n^2)$. The complexity of optimizing it is $O(n^2)$ (using memoization)\(^3\) if we use the stochastic greedy algorithm \(^{19}\) and $O(n^2k)$ with

\(^3\) $O$: Ignoring log-factors
the naive greedy algorithm. The overall complexity is \( \tilde{O}(n^2) \). For FLQMI, the cost of creating the kernel matrix is \( O(n|Q|) \), and the cost of optimization is also \( \tilde{O}(n|Q|) \) (with naive greedy, it is \( O(nB|Q|) \)).

- **Graph-Cut:** For GCMI, we require a \( O(n|Q|) \) kernel matrix, and the complexity of the stochastic greedy algorithm is also \( \tilde{O}(n|Q|) \).

We end with a few comments. First, most of the complexity analysis above is with the stochastic greedy algorithm \[19\]. If we use the naive or lazy greedy algorithm, the worst-case complexity is a factor \( B \) larger. Secondly, we ignore log-factors in the complexity of stochastic greedy since the complexity is actually \( O(n \log 1/\epsilon) \), which achieves an \( 1 - 1/e - \epsilon \) approximation.