Class-Balanced Active Learning for Image Classification

Javad Zolfaghari Bengar\textsuperscript{1,2} Joost van de Weijer\textsuperscript{1,2} Laura Lopez Fuentes\textsuperscript{1} Bogdan Raducanu\textsuperscript{1,2}

Computer Vision Center (CVC)\textsuperscript{1}, Univ. Autònoma of Barcelona (UAB)\textsuperscript{2}

\{jzolfaghari, joost, llopez, bogdan\}@cvc.uab.es

Abstract

Active learning aims to reduce the labeling effort that is required to train algorithms by learning an acquisition function selecting the most relevant data for which a label should be requested from a large unlabeled data pool. Active learning is generally studied on balanced datasets where an equal amount of images per class is available. However, real-world datasets suffer from severe imbalanced classes, the so called long-tail distribution. We argue that this further complicates the active learning process, since the imbalanced data pool can result in suboptimal classifiers. To address this problem in the context of active learning, we proposed a general optimization framework that explicitly takes class-balancing into account. Results on three datasets showed that the method is general (it can be combined with most existing active learning algorithms) and can be effectively applied to boost the performance of both informative and representative-based active learning methods. In addition, we showed that also on balanced datasets our method\textsuperscript{1} generally results in a performance gain.

1. Introduction

Neural networks obtain state-of-the-art results on several computer vision tasks such as large-scale object detection \cite{45} or VQA \cite{53}. However, the training of these often very large networks requires large-scale labeled datasets, that are labor intensive and expensive to construct. Generally, in real-world the amount of data that could be labeled is literally unlimited (e.g. in autonomous driving, or robotics applications). Given an initial labeled dataset, deciding what new data to label from the unlabeled data pool is a relevant research question addressed by active learning. It aims to minimize the labeling effort while maximizing the obtained performance of the machine learning algorithm. Active learning has successfully been shown to reduce the labeling effort for image classification \cite{3, 48}, object detection \cite{61}, regression \cite{33}, and semantic segmentation \cite{52, 25}.

Several query strategies have been proposed for sample selection. The most popular ones are those based on informativeness \cite{58} and representativeness \cite{48} which demonstrated to be efficient for the task of selecting the most valuable samples. The informativeness criteria is responsible for selecting those samples which are the most uncertain (usually characterized by high-entropy) because they affect the generalization capability of the model (they are the ones which are mostly confusing the classifier, especially at the start of the active learning process when the number of labeled samples is small), while representativeness guarantees a diversity of the samples, following the underlying data distribution of the unlabeled data pool.

Visual recognition datasets are often almost uniformly distributed (e.g. CIFAR \cite{35} and ILSVRC \cite{56}). However, for many real-world problems data follows a long-tail distribution \cite{42}, meaning that a small number of head-classes are much more common than a large number of tail-classes (e.g. iNaturalist \cite{51}, landmarks \cite{43}). Classification on such imbalanced dataset is an important research topic \cite{50, 12, 46}. However, active learning is mostly studied on curated close to uniform datasets. Given the predominance of long-tail distributions, especially for real-world applications in which active learning is a crucial capability, we here study active learning for imbalanced datasets. The aim is to minimize the labeling effort, while maximizing performance when measured on a balanced test set.

Closely related to the class-imbalance dataset problem, is the sampling bias problem which is a well-documented drawback of active learning \cite{41, 14}. Datasets collected by active learning algorithms break the assumption that the data is identically and independently distributed (i.i.d), since the active learning algorithm might be biased towards particular regions of the unlabeled data manifold. One possible consequence of the sampling bias can be that the distribution over the classes does no longer follow that of the unlabeled data pool. Several papers have investigated this aspect of active learning however it remains not fully understood \cite{6, 20}. To mitigate the problems caused by the sampling bias and imbalanced datasets, in the current paper

\textsuperscript{1}Our code is available at: https://github.com/Javadzb/Class-Balanced-AL.git
we introduce an optimization framework which corrects the class-imbalance presented in the unlabeled data pool, and aims to bias instead our selected samples to resemble the uniform distribution of the test set. The overview of the proposed approach is depicted in figure 1. Since we have no access to the class labels of the unlabeled data, we propose to trust the predicted labels, and use them to select a set of class-balanced images. This combination leads to a minimization problem, which can be formalized as a binary programming problem. We show that our optimization scheme is efficient, boosting the performance of both informativeness and representativeness methods. In summary, the main contributions of this paper are:

- We propose a novel active learning method for imbalanced unlabeled dataset that encourages the selection of class-balanced samples.
- The proposed optimization method is general and can be applied to both informativeness and representativeness based methods.
- Extensive experiments show that our method improves performance of active learning on imbalanced datasets. We show that even for balanced datasets the proposed method can lead to improvements, mostly by countering the sampling bias introduced by active learning.

2. Related Work

Active Learning. Active Learning has been widely studied in various applications such as image classification [34, 24, 22] (including medical image classification [47] and scene classification [40]), image retrieval [60], image captioning [15], object detection [61], and regression [21, 33]. Strategies can be divided in three main categories: informativeness [57, 23, 26, 9, 4], representativeness [47, 48] and hybrid approaches [31, 56]. A comprehensive survey of these frameworks can be found in [49].

Among all the aforementioned strategies, the informativeness-based approaches are the most successful ones, with uncertainty being the most used selection criteria used in both bayesian [23] and non-bayesian frameworks [39]. In [23], they obtain uncertainty estimates through multiple forward passes with Monte Carlo Dropout, but it is computationally inefficient for recent large-scale learning as it requires dense dropout layers that drastically slow down the convergence speed. More recently, [2] measures the uncertainty of the model by estimating the expected gradient length. On the other hand, [58, 38] employ a loss module to learn the loss of a target model and select the images based on their output loss.

Representativeness-based methods rely on selecting examples by increasing diversity in a given batch [18]. The Core-set technique [48] selects the samples by minimizing the Euclidian distance between the query data and labeled samples in the feature space. The Core-set technique is shown to be an effective method, however, its performance is limited by the number of classes in the dataset. Furthermore, it is less effective due to feature representation in high-dimensional spaces since p-norms suffer from the curse of dimensionality [17]. In a different direction, [50] uses an adversarial approach for diversity-based sample query, which samples the data points based on the dis-

Figure 1. Overview of our active learning framework. The unlabeled samples are sorted by their uncertainty from green to yellow in ascending order. Given the uncertainty of unlabeled samples and class distribution at cycle $c$, we propose to solve an optimization problem ($\lambda > 0$) yielding samples that are simultaneously informative and form a balanced class distribution for training. Our sampling selects samples with lower uncertainty (in green) in addition to high uncertainty to improve class-balanced profile. In contrast, classical AL methods ($\lambda = 0$) selects the most uncertain samples (in yellow) that result in an informative yet imbalanced training set.
Figure 2. Biased sampling across four AL cycles. In the background, the imbalanced dataset is illustrated in yellow. The class distributions for two active learning approaches and random sampling are shown. Similar to Random sampling (in cyan), samples selected by active learning algorithms follow the biased distribution. Results are on imbalanced CIFAR10 (IF=0.3).

3. Class Imbalance in Active Learning

3.1. Active Learning Setup

Given a large pool of unlabeled data $D_U$ and a total annotation budget $B$, the goal is to select $b$ samples in each cycle to be annotated to maximize the performance of a classification model. In general, AL methods proceed sequentially by splitting the budget in several cycles. Here we consider the batch-mode variant [49], which annotates $b$ samples per cycle, since this is the only feasible option for CNN training. At the beginning of each cycle, the model is trained on the labeled set of samples $D_L$. After training, the model is used to select a new set of samples to be annotated at the end of the cycle via an acquisition function. The selected samples are added to the labeled set $D_L$ for the next cycle and the process is repeated until the annotation budget $b$ is spent. The acquisition function is the most crucial component and the main difference between AL methods in the literature. In the remainder of this section, we describe the motivation behind our proposed acquisition function.

3.2. Motivation

Most active learning methods propose efficient sampling methods that are class agnostic. The underlying assumption is that the distribution of train and test datasets are uniform. However, in real world scenarios, where the datasets might be heavily imbalanced, the methods suffer from biased sampling towards the majority class. AL methods tend to sample more from frequent classes and less from minority classes which consequently leads to biased predictions and a performance drop. Fig. 2 presents an example of a such dataset with various AL methods (see Suppl. Mat. J). As it can be seen, the distribution of samples selected by both informative and representative based methods follow the distribution of the unlabeled dataset. More-
over the imbalance of selected samples grows across the cycles. It is known that when we aim for good performance on all classes these imbalanced training sets are suboptimal [30-12]. We tackle the problem of class imbalance in the remainder of this section.

3.3. Reducing Class Imbalance

A balanced set of samples requires an equal number of samples per class. Since we have no access to the class labels, we make an estimate of distribution of samples by using a probability matrix. Assume we have $|D_U| = N$ unlabeled samples in $C$ categories. We use the classifier to output the softmax probability matrix $P$ on the unlabeled samples:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1C} \\ p_{21} & p_{22} & \cdots & p_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NC} \end{bmatrix} \in \mathbb{R}^{N \times C}$$ (1)

where each row sums to 1. Similar to [19], we use variable $z_i \in \{0, 1\}$ associated to sample $i$ to indicate whether a sample $i$ is selected or not. To measure the distance between the estimated distribution and the desired distribution we employ $\ell 1$ norm as:

$$\ell 1(\Omega, P^T z) = \|\Omega(c) - P^T z\|_1. \quad (2)$$

Here $\Omega(c)$ is vector with components specifying the number of required samples from each class in order to achieve balance at cycle $c$. Given the labels of samples selected in previous cycles, it is straightforward to compute the samples required at cycle $c$:

$$\Omega(c) = [\omega_1, \omega_2, \ldots, \omega_C], \quad (3)$$

where,

$$\omega_i = \max \{c b + b_0 \frac{C}{n_i} - n_i, 0\}, \quad (4)$$

$b$ is the budget per cycle, $b_0$ is the size of the initial labeled set, $c \in \{1, 2, 3, \ldots\}$ denotes the cycle, and $n_i$ is the number of samples selected from class $i$ in previous cycles. Condition 2 avoids oversampling from a particular class. To obtain $\Omega$ at cycle $c = 1$ for instance, given that we start the AL cycles from uniform initial set with $n_i = b_0/C$ we have:

$$\Omega(1) = \frac{b}{C} \mathbb{1}_{C \times 1} \quad (5)$$

In the following, we will minimize Eq.2 to encourage the selection of class-balanced samples.

4. Class Balanced Active Learning

In this section, we introduce the Class Balanced Active Learning (CBAL) formulation for classification.

4.1. Informativeness

Entropy We describe our optimization framework that selects the most uncertain samples while seeking to balance the number of samples over classes. Based on informativeness approach, given the probability matrix the goal is to find samples that are most uncertain for the model. To measure the uncertainty we use Entropy [13] as an information theory measure that captures the average amount of information contained in the predictive distribution, attaining its maximum value when all classes are equiprobable. Given the softmax probabilities, the entropy of a sample is computed as:

$$H = - \sum_{i=1}^{C} p_i \log p_i. \quad (6)$$

We aim to select samples with maximum entropy. Consequently, the sum of candidates’ entropy should also be maximized. In matrix notation form, this is expressed as:

$$\sum_{i \in \{j | z_j = 1\}} H(x_j) = -z^T (P \odot \log (P)) 1_{C \times 1}, \quad (7)$$

where $z$ is all-ones column vector and $\odot$ denotes element wise multiplication. $1_{C \times 1}$ is an all-ones column vector. In our objective we will minimize the negative entropy, which is equal to maximizing the entropy.

Finally, we combine the informative and balancing objectives in a single optimization problem given as:

$$\min_z z^T (P \odot \log (P)) 1_{C \times 1} + \lambda \|\Omega(c) - P^T z\|_1 \quad \text{s.t.} \quad z^T 1_{N \times 1} = b, \quad z_i \in \{0, 1\}, \quad \forall i = 1, 2, \ldots, N \quad (8)$$

where $\lambda$ is a parameter that regularizes the contribution of the balancing term in the objective. Minimizing the cost in Eq.8 encourages to select sufficient number of samples per class while choosing the most informative ones. The cost function consists of an affine term and a $\ell 1$ norm that are both convex, and subsequently their linear combination is also convex. However, as the constraint is non-convex the optimization problem becomes non-convex. The underlying problem is Binary Programming that can be optimally solved by an off-the-shelf optimizer using LP relaxation and the branch and bound method. Algorithm 4.1 presents the AL cycles using our approach.

Regularizer $\lambda$. Next, we analyze the effect of varying parameter $\lambda$ on the cost function. We start with a model trained on initial labeled samples of CIFAR100 dataset. Then, for every $\lambda$ in range $(0, 3)$ the cost function in Eq.8 is minimized. Fig.4 illustrates the changes in entropy loss and the $\ell 1$ loss as the components of the cost function with respect to $\lambda$. For comparison purposes, the horizontal lines represent the same losses measured on samples given by standard entropy and entropy L1-pseudo label methods.
The latter uses the hard labels given by the model to unlabeled samples also known as “Pseudo Labels” for balancing (see Suppl. Mat. H for more details and performance evaluation of Entropy-L1-Pseudo Label). As can be seen, greater λ reduces entropy, l1, and L1score (introduced in 5.1). It is notable that the samples selected with greater λ are more balanced but at the cost of lower entropy. As a result, there is a trade-off between balancedness and entropy of samples.

Variational Adversarial Active Learning (VAAL) [50] is considered to be one of the current state-of-the-art algorithms on active learning. This model uses a variational autoencoder to map the distribution of labeled and unlabeled data to a latent space. A binary adversarial classifier (analogous to a GAN discriminator) is trained to predict if an image belongs to the labeled or the unlabeled pool. The unlabeled images which the discriminator classifies with lowest certainty as belonging to the labeled pool are considered to be the most representative with respect to other samples which the discriminator thinks belong to the labeled pool. Thus, the images labeled by the discriminator with lower certainty are sampled to be labeled in the next cycle. Considering the uncertainty estimate u of the discriminator, we can encourage finding a balanced sample set by minimizing:

$$
\min_z z^T u + \lambda \| \Omega(c) - P^T z \|_1
$$

s.t. $z^T 1_{N \times 1} = b, \quad z_i \in \{0, 1\}, \quad \forall i = 1, 2, ..., N$  \hspace{1cm} (9)

Bayesian Active Learning with Disagreement (BALD) method chooses samples that are expected to maximise the information gained about the model parameters. In particular, it select samples that maximise the mutual information between predictions and model posterior [23]. It approximates Bayesian inference by drawing Monte Carlo sampling via dropout. Similar to our previous approach, we

**Algorithm 1** Class Balancing AL

**Input:** Unlabeled Pool $\mathcal{D}_U$, Total Budget $B$, Budget Per Cycle $b$.

**Initialize:** Initial labeled pool $|\mathcal{D}_L| = b_0, c = 1$

1: while $|\mathcal{D}_L| < B$

2: Train CNN classifier $\Theta$ on $\mathcal{D}_L$

3: Use $\Theta$ to compute probabilities for $x \in \mathcal{D}_U$

4: Compute $\Omega(c)$ from Eq. 3

5: Solve for Algorithm 2 for greedy, to obtain $\mathcal{D}_L$, $\mathcal{D}_U$

6: Query $\mathcal{D}_L \cup \mathcal{D}_U$, $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{D}_L$

7: $c \leftarrow c + 1$

8: end while

9: return $\mathcal{D}_L, \Theta$

The effect of $\lambda$ on L1 and entropy losses in the cost function \hspace{1cm} (8)

Figure 4. **The effect of $\lambda$ on L1 and entropy losses in the cost function**
summarize the mutual information assigned to samples into a vector and incorporate into our optimization problem.

4.2. Representativeness

Representativeness-based methods aim to increase the diversity of the selected batch \([48]\). These active learning approaches select the samples iteratively one at a time. In fact, every selected sample influences the next one. Therefore, a method that integrates greedy selection while maintaining the class balance of samples is of great interest. For this reason, we present the greedy class balancing algorithm that incorporates balancing in the sample selection.

We focus on a prominent method of this approach namely KCenterGreedy, which is a greedy approximation of KCenter problem also known as min-max facility location problem \([55]\). Our aim is to find \(b\) samples having maximum distance from their nearest labeled samples while keeping the samples class-balanced. Similar to \([48]\), we compute the embeddings for unlabeled samples via a deep neural network. Specifically, we employ the model for inference on unlabeled samples and consider the penultimate fully connected layer as the visual embedding. Then, we compute the geometrical distances between the representations in the embedding space and construct the distance matrix \(D\). Given \(N\) unlabeled and \(L\) labeled samples, \(d_{ij} \in D_{N \times L}\) is the euclidean distance between the embeddings of unlabeled sample \(i\) to labeled sample \(j\). The algorithm \([42]\) presents the KCenterGreedy sample selection combined with class balancing. We propose similar cost function to \([8]\) for the greedy sampling. In the algorithm \(P^Tz\) represents the cost of already selected samples and matrix \(Q\) represents the unlabeled samples to choose from. The broadcasting within the L1 norm is for the consistency across dimensions of labeled samples, unlabeled samples and thresholds. Although here we integrated the balanced sampling with KCenterGreedy, our method is general and applicable to any greedy acquisition method.

5. Experiments

5.1. Experimental Setup

We evaluate our method on three image classification benchmarks and the imbalanced variants. The initial labeled set \(D_C\) consists of 10\% of the training dataset that is uniformly selected from all classes at random. At each cycle we start with our base model either from scratch or, in case of Tiny-imagenet, we start from a pretrained imagenet model. We train the model in \(c\) cycles until the budget \(B\) is exhausted. The budget per cycle for all experiments is 5\% of the original dataset.

Datasets. To evaluate our method, we use CIFAR10 and CIFAR100 \([35]\) datasets with 50K images for training and 10K for test. CIFAR10 and CIFAR100 have 10 and 100 object categories respectively and an image size of 32×32.

| Algorithm 2 Greedy Class Balancing Selection |
|-----------------------------------------------|
| **Input:** Softmax output \(P_{N \times C}\), Distance Matrix \(D_{N \times L}\), Balancing threshold \(\Omega_{C \times 1}\), Regularizer \(\lambda\), Budget Per Cycle \(b\) |
| **Initialize:** \(z^{(0)} = 0_{N \times 1}\), \(Q = P\) |
| 1: for \(i = 0 : b - 1\) do |
| 2: \(d_{N \times 1} \leftarrow \min(D, \text{axis} = 1)\) \(\triangleright\) for each unlabeled sample find the nearest labeled sample |
| 3: \(\psi \leftarrow \arg\min(-d^{(i)}_{(N-i) \times 1}) + \lambda\|\Omega(c) - Q_{C \times (N-i)}^T z^{(i)}_{1 \times (N-i)}\|_1^T\) |
| 4: \(z^{(i+1)}(\psi) \leftarrow 1\) \(\triangleright\) select the sample |
| 5: \(Q \leftarrow P(z^{(i)} = 0,:)\) \(\triangleright\) keep the remaining unlabeled samples in \(Q\) |
| 6: \(D \leftarrow D_{(N-i) \times (L+i)}\) \(\triangleright\) update \(D\) by removing a row and adding a column correspond to newly selected sample |
| 7: end for |
| 8: return \(z^{(b)}\) |

To evaluate the scalability of our method we evaluate on Tiny ImageNet dataset \([37]\) with 90K images for training and 10K for testing. There are 200 object categories in Tiny ImageNet with an image size of 64×64.

Long-Tailed Datasets. To verify our approach on imbalanced datasets, we make the CIFAR10, CIFAR100 and Tiny ImageNet class-imbalanced. Again, we reserve 10\% of samples of the three datasets for initial labeled set. As in \([11]\) we create long-tailed datasets with the remaining 90\% by randomly removing training examples. In particular, the number of samples drops from \(y\)-th class is \(n_y\). IF where \(n_y\) is the original number of training samples in class \(y\) and the imbalance factor \(IF \in (0, 1)\). For the construction of long-tailed datasets we apply IF to half of the classes, and use IF \(\in \{0.1, 0.3\}\).

Baselines. We compare our method with Random sampling and several informative and representative-based approaches including Entropy sampling, KCenterGreedy, VAAL and BALD (See also suppl. mat. \([\blacksquare]\) for Coreset).

In order to make a fair comparison with the baselines, we used their official code and adapted them into our code to ensure an identical setting.

Performance Evaluation. We measure the accuracy on the test set to evaluate the performance. Results for all experiments are averaged over 3 runs. For each method we plot the average performance for all runs with vertical bars to represent the standard deviation. To measure the balancedness of selected samples, we use \(L1\_score\) by computing \(\ell_1\) distance between samples’ distribution and uniform distribution. In order to have a measure ranging from 0 to 1, we normalize \(\ell_1\) with the factor obtained as follows:
5.2. Experimental Results

Performance on CIFAR10. Fig. 5 presents the performance on CIFAR10. In Fig. 5a, the class balanced methods improve baselines marginally even though the dataset is balanced (IF=1). The improvements of class-balanced methods improve for the lower IF values (see Fig 5b and c). Notably, VAAL-CB achieves 3% improvement on average over the VAAL baseline in Fig. 5b. See Table 2 for a detailed gain analysis. To put these improvements into perspective, the gain obtained by class balancing methods over the baselines is comparable to the improvement of those methods over Random. Specifically, Entropy-CB after 4 cycles achieves over 1% improvement over the Entropy baseline regardless of imbalance factor of the dataset.
Figure 6. Performance evaluation. Results for several active learning methods on CIFAR100 for different imbalance factors (IF).

| Imbalance Factor | Methods                  | Cycles |
|------------------|--------------------------|--------|
|                  |                          | 1 2 3 4|
| IF=0.1           | Entropy CB(%)            | 0.54 0.86 0.57 0.27 |
|                  | KcenterGreedy CB(%)      | 0.84 0.44 0.77 0.77 |
|                  | VAAL CB(%)               | 0.57 0.85 1.08 1.19 |
| IF=0.3           | Entropy CB(%)            | 0.40 1.11 0.31 0.08 |
|                  | KcenterGreedy CB(%)      | 0.31 0.47 0.53 0.34 |
|                  | VAAL CB(%)               | 0.28 0.24 0.42 0.91 |
| IF=1             | Entropy CB(%)            | 0.00 0.027 -0.15 -0.12 |
|                  | KcenterGreedy CB(%)      | 0.19 0.14 0.03 0.05 |
|                  | VAAL CB(%)               | 1.08 0.68 0.47 0.29 |

Table 1. Performance gain over AL baselines on CIFAR 10.

| Imbalance Factor | Methods                  | Cycles |
|------------------|--------------------------|--------|
|                  |                          | 1 2 3 4|
| IF=0.1           | Entropy CB (%)           | 1.28 2.23 1.50 1.43 |
|                  | KcenterGreedy CB (%)     | 1.03 0.92 1.04 0.93 |
|                  | VAAL CB (%)              | 0.37 1.86 2.32 2.23 |
| IF=0.3           | Entropy CB (%)           | 1.16 1.76 1.44 1.63 |
|                  | KcenterGreedy CB (%)     | 0.28 0.76 1.52 1.70 |
|                  | VAAL CB (%)              | 2.47 2.42 3.23 3.29 |
| IF=1             | Entropy CB (%)           | 1.37 1.40 1.96 1.82 |
|                  | KcenterGreedy CB (%)     | 0.55 1.15 1.03 0.48 |
|                  | VAAL CB (%)              | 1.01 0.11 0.86 1.53 |

Table 2. Performance gain over AL baselines on CIFAR 100.

| Imbalance Factor | Methods                  | Cycles |
|------------------|--------------------------|--------|
|                  |                          | 1 2 3 4|
| IF=0.1           | Entropy CB (%)           | 0.48 0.63 0.21 0.58 |
|                  | BALD CB (%)              | 0.31 0.10 0.08 0.21 |
| IF=0.3           | Entropy CB (%)           | 0.34 -0.04 0.52 0.19 |
|                  | BALD CB (%)              | 0.07 0.07 0.36 0.19 |
| IF=1             | Entropy CB (%)           | 0.35 0.62 0.56 0.74 |
|                  | BALD CB (%)              | 0.10 0.21 1.11 0.51 |

Table 3. Performance gain over baselines on Tiny ImageNet.

Performance on Tiny ImageNet. Tiny ImageNet is a challenging large scale dataset which we use to evaluate the scalability of our approach. Also to evaluate the generality of our approach we show the performance of class balancing applied to BALD as a Bayesian approach. Table 3 shows evidently the addition of class balancing to Entropy and BALD boost their performance on both balanced and imbalance datasets. (See suppl. mat. G for a detailed performance evaluation on Tiny ImageNet).

6. Conclusions

We have investigated the influence of class-imbalance on active learning performance. Class-imbalance can be caused by an imbalanced unlabeled data pool or by the sampling bias present in active learning algorithms. When aiming for good performance of the final classifier on all classes, class-imbalance has a detrimental effect. Therefore, to address class-imbalance we proposed an optimization-based method that aims to balance classes. The method is general and can be combined with both the informativeness and representativeness criteria often used in active learning. Extensive experiments, on several datasets show that our method improves results of existing active learning methods. Our results suggests that class-balancing should be an important criteria when selecting samples, and that it should be considered next to the long-standing active learning criteria of informativeness and representativeness.

Acknowledgements We acknowledge the support of the project PID2019-104174GB-I00 (MINECO, Spain), the CERCA Programme of Generalitat de Catalunya, the EU project CybSpeed MSCA-RISE-2017-777720 and CYTED Network (Ref. 518RT0559).

2Representativeness-based methods are infeasible on large datasets.
[35] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. PhD thesis, University of Toronto, 2012.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[37] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7:7, 2015.

[38] Minghan Li, Xialei Liu, Joost van de Weijer, and Bogdan Raducanu. Learning to rank for active learning: A listwise approach. In *ICPR*, pages 5587–5594, 2020.

[39] Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *CVPR*, pages 859–866, 2013.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[41] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[42] Mark EJ Newman. Power laws, pareto distributions and zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.

[43] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.

[44] Junran Peng, Xingyuan Bu, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, and Junjie Yan. Large-scale object detection in the wild from imbalanced multi-labels. In *CVPR*, pages 9709–9718, 2020.

[45] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, pages 4334–4343, 2018.

[46] P.T. Saito, C.T. Suzuki, J.F. Gomes, P.J. de Rezende, and A.X. Falcão. Robust active learning for the diagnosis of parasites. *Pattern Recognition*, 48(11):3572–3583, 2015.

[47] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *IJCV*, 113(2):113–127, 2015.

[48] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, pages 93–102, 2019.

[49] Hualong Yu, Xibei Yang, Shang Zheng, and Changyin Sun. Active learning from imbalanced data: A solution of online weighted extreme learning machine. *IEEE transactions on neural networks and learning systems*, 30(4):1088–1103, 2018.

[50] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, pages 5972–5981, 2019.

[51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[52] Jun Wang, Shaoguo Wen, Kaixing Chen, Jianghua Yu, Xin Zhou, Peng Gao, Changsheng Li, and Guotong Xie. Semi-supervised active learning for instance segmentation via scoring predictions. In *Proc. of BMVC*, 2020.
Supplementary Materials for Class-Balanced Active Learning for Image Classification

G. Performance on Tiny ImageNet dataset

Fig. 7 illustrates the performance of class balanced (CB) methods and AL baselines. As can be seen, both Entropy-CB and BALD-CB outperform the corresponding baselines. Notably in Tiny ImageNet, Random sampling serve as a competitive baseline. Nevertheless the addition of class balancing made Entropy-CB superior in almost all active learning cycles across different imbalance factors.

![Figure 7. Performance evaluation. Results for active learning methods on Tiny ImageNet with different imbalance factors (IF).](image)

H. Pseudo Label balancing

Fig. 8 presents the performance of another Entropy variation on CIFAR100 for comparison. Among them, "Entropy L1 Pseudo Label" benefits from "pseudo labels" defined as the most probable labels that the model assigns to unlabeled samples (the prediction of the model is then converted to a one-hot vector). This method utilizes the pseudo labels to balance the distribution of samples and select certain number of samples (specified by $\Omega$ in Eq.3) from each class with maximum entropy. The experiments show that Entropy-CB outperforms Entropy L1 Pseudo Label both in terms of active learning performance (see Fig. 8) and the ability of class balancing (see $\lambda$ tuning in Section 4).

I. CoreSet performance

The performance of CoreSet on CIFAR10 and CIFAR100 is shown in Fig. 9 and Fig. 10 respectively. In our experiments CoreSet and KCenterGreedy-CB perform similarly on the balanced dataset (IF=1). However, when the dataset is imbalanced (IF=0.3 and IF=0.1) the performance of CoreSet degrades compared to KCenterGreedy-CB. As CoreSet is a MIP (Mixed Integer Programming) problem, our technique cannot be applied to this method.
Figure 8. **Performance evaluation.** Comparing Entropy standard, Entropy balanced by Pseudo Labels against the proposed Entropy CB.

Figure 9. **Performance evaluation.** CoreSet compared to active learning methods on CIFAR10 with different imbalance factors (IF).

Figure 10. **Performance evaluation.** CoreSet compared to active learning methods on CIFAR100 with different imbalance factors (IF).
J. Distribution of selected samples in CIFAR100

Fig. 11, 12 and 13 show the distribution of samples selected by AL methods on original (IF=1), imbalanced (IF=0.3) and (IF=0.1) respectively. The L1 score above the distributions (introduced in Section 5.1) measures the $\ell_1$ distance from uniform distribution in the corresponding cycle. As can be seen, CB methods are remarkably effective in balancing the distribution of selected samples regardless of imbalance factor. It is worth mentioning in Fig 11 although the dataset is balanced, AL baselines (Entropy and KCenterGreedy) result in biased sampling. In contrast, CB methods provide more balanced samples across all cycles and imbalance factors.

Figure 11. Distribution of samples selected by our proposed method (CB) compared to baselines on CIFAR100 with IF=1.
Figure 12. Distribution of samples selected by our proposed method (CB) compared to baselines on CIFAR100 with IF=0.3.

Figure 13. Distribution of samples selected by our proposed method (CB) compared to baselines on CIFAR100 with IF=0.1.