Learning from Label Proportions with Prototypical Contrastive Clustering

Laura Elena Cué La Rosa[1], **Dário Augusto Borges Oliveira**[2]

[1]Electrical Engineering Department, Pontifical Catholic University of Rio de Janeiro, Brazil
[2]Data Science in Earth Observation, Technical University of Munich (TUM), Germany
lauracue@aluno.puc-rio.br, dario.oliveira@tum.de

Abstract

The use of priors to avoid manual labeling for training machine learning methods has received much attention in the last few years. One of the critical subthemes in this regard is Learning from Label Proportions (LLP), where only the information about class proportions is available for training the models. While various LLP training settings verse in the literature, most approaches focus on bag-level label proportions errors, often leading to suboptimal solutions. This paper proposes a new model that jointly uses prototypical contrastive learning and bag-level cluster proportions to implement efficient LLP classification. Our proposal explicitly relaxes the equipartition constraint commonly used in prototypical contrastive learning methods and incorporates the exact cluster proportions into the optimal transport algorithm used for cluster assignments. At inference time, we compute the clusters' assignment, delivering instance-level classification. We experimented with our method on two widely used image classification benchmarks and report a new state-of-art LLP performance, achieving results close to fully supervised methods.

Introduction

The performance of the firstly proposed deep learning methods was directly related to large amounts of annotated training samples (Russakovsky et al. 2015). However, annotating such large datasets promptly became a bottleneck in supervised learning, as it is a time-consuming and labor-intensive task. Additionally, various applied areas such as healthcare or democratic elections struggle with labels, which are often not available (Qi et al. 2016). In many scenarios, despite the unavailability of instance-level annotations, approximate group-level labels like class proportions are readily obtainable from other sources, like the census or even common knowledge. In this sense, efficient learning from group-level labels would have an important impact in many real-life applications, such as demographic classification (Ardehaly and Culotta 2017), presidential elections (Sun, Sheldon, and O'Connor 2017; Qi et al. 2016), remote sensing (Ding, Li, and Yu 2017), image analysis in medicine (Bortsova et al. 2018), activity recognition (Poyiadzi, Santos-Rodriguez, and Twomey 2018), and others.

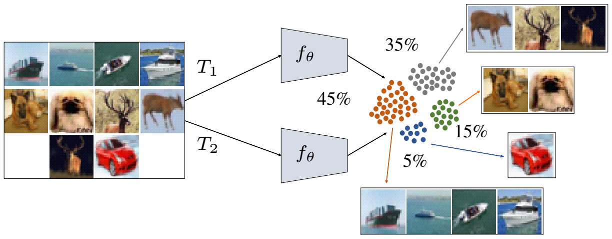

Figure 1: General overview of LLP-Co. Given an input bag of images, we generate two augmented views of each image. Then, we forward the full batch of image views into a network that assigns them to various clusters constrained to their a priori proportions. The colored arrows point to the images that were assigned to each cluster.

As a weakly supervised classification subtheme, Learning from Label Proportions (LLPs) has received much attention in the machine learning community in the last decade (Quadrianto et al. 2009; Yu et al. 2013, 2014; Qi et al. 2016; Dulac-Arnold et al. 2019; Shi et al. 2020; Scott and Zhang 2020). Standard LLP approaches split training samples into groups of bags where only label proportions are known and target to learn individual sample class. Many different ways for implementing that were proposed, including those based on support vector machine (Yu et al. 2013), Bayesian models (Hernández-González, Inza, and Lozano 2013), convolutional neural networks (Ardehaly and Culotta 2017), generative models (Liu et al. 2019) and clustering (Stolpe and Morik 2011). In (Yu et al. 2014), the authors presented the *Empirical Proportion Risk Minimization* (EPRM) algorithm to infer when and why bag-level proportions help predict individual labels. The study concluded that LLP performance strongly depends on the bag size and label proportions and that a predictor with a reasonable bag proportion estimate warrants a good instance-level prediction. EPRM and others LLP methods optimize the learning process by minimizing a bag-level loss that typically uses the Kullback-Leibler (KL) divergence to measure how much the predicted proportion distribution differs from the known distribution. Despite the recent advances observed in the field (Ardehaly and Culotta 2017; Liu et al. 2019; Dulac-Arnold et al. 2019), learning instance-label classifiers solely from bag-level label proportions with KL is still a challenge, as many valid hypotheses can match the know distribution and still lead to suboptimal solutions.

Recently, Caron et al. (Caron et al. 2020) proposed the Swapping Assignments between multiple Views (SwAV) method to perform unsupervised clustering. SwAV combines contrastive learning and Optimal Transport (OT) to cluster data while enforcing consistency between cluster assignments produced from different views of the same input image. The clustering implies that augmented views of a given sample belong to the same class and that different samples belong to different classes. An Optimal Transport solver assigns samples to cluster prototypes (or centroids) and computes pseudo cluster labels to guide the clustering process. At convergence, each prototype represents a group of semantically similar samples, and the ultimate training goal is to find the network's parameters that best describe the overall training set distribution. Li et al. (Li et al. 2020b) also explored contrastive learning and clustering with the Prototypical Contrastive Learning (PCL) method, which implements clustering from unsupervised representation learning.

This paper builds on this literature and proposes a method that embeds priors about class proportions to a contrastive cluster assignment framework to handle LLP. A general overview of the proposed framework is depicted in Figure 1. Specifically, our proposal, Learning from Label Proportions with Prototypical Contrastive Clustering (LLP-Co), relaxes the SwAV method's equipartition constraint by incorporating the exact cluster proportions into the Optimal Transport module. Consequently, our method inherently performs LLP classification from the clustering, and at inference time, we use the clusters' assignment to perform instance-level classification.

We experimented on two standard image classification benchmarks (CIFAR-10 and CIFAR-100 (Krizhevsky, Nair, and Hinton 2012)) using a ResNet18 architecture to evaluate our method and report state-of-art results compared to current multi-class LLP methods, achieving similar performance to that observed in fully supervised counterparts. Our work implements the following contributions:

- A new method combining class proportion priors and prototypical contrastive clustering to tackle LLP.
- State-of-the-art LLP performance on widely used public computer vision benchmark datasets.

Related Works

Our work is closely related to three research topics versing weakly-supervised/unsupervised learning: multi-class learning from label proportions, deep unsupervised clustering, and contrastive learning. We present a summary of related works to contextualize our paper.

Multi-class Learning from Label Proportions Current state-of-the-art methods to solve multi-class LLP problems use deep learning models. Ardehaly and Culotta introduced deep learning for LLP in (Ardehaly and Culotta 2017) with application in demographic attribute classification. The deep LLP (DLLP) approach incorporates a regularization layer to a deep neural network for averaging probability outputs towards the bag proportion using the KL divergence loss to train the network. In (Dulac-Arnold et al. 2019), the author investigates two loss functions to solve multi-class LLP: a modified KL divergence and a function based on balanced Optimal Transport with entropic regularization (ROT). The authors concluded that such models perform close to supervised models for bags of up to 16 samples, but both loss functions degrade as the bag size increases, even if the ROT loss presents higher robustness to big bags. More recently, Liu et al. (Liu et al. 2019) introduced adversarial learning to classification based on label proportions. LLP-GAN significantly improves previous work (Yu et al. 2014; Ardehaly and Culotta 2017) achieving SOTA performance in several computer vision benchmark datasets. Despite the success of these methods, their performance is still far behind supervised counterparts. We believe that using solely bag-level proportions to perform instance-level classification is a flawed approach. Concurrent to our work, (Liu et al. 2021) proposed the use of Optimal Transport (OT) to obtain noisy pseudo-labels that meet the exact proportions in an LLP problem. The methodology consists of a two-stage training process that employs LLP models based on KL-divergence as the first stage and supervised learning using cross-entropy loss with the pseudo-labels generated by OT as the second stage.

Deep Unsupervised Clustering Combining clustering and feature representation learning has emerged as a promising approach for unsupervised learning. DeepCluster (Caron et al. 2018) and Deep *k*-means (Fard, Thonet, and Gaussier 2020) are two approaches that jointly optimize representation learning and clustering in an end-to-end framework. DeepCluster groups the features with *k*-mean and uses the assignment as pseudo-labels to update the weights of a convolutional neural network. Deep *k*-means replaces the cluster assignments with soft-assignments and proposes a clustering loss that is jointly minimized over the network's parameters and the centers of the clusters, using stochastic gradient descent (SGD). Genevay et al. (Genevay, Dulac-Arnold, and Vert 2019) proposed a differentiable deep clustering method with cluster size constraints. The main contribution of the study is the rewriting of the *k*-mean clustering algorithm as an OT with entropic regularization task. The authors report promising results outperforming Deep *k*-means and the multi-class learning from label proportions (Dulac-Arnold et al. 2019) approach for large bag sizes. However, this approach strongly depends on the *k*-means algorithm cluster initialization.

Prototype Learning and Contrastive Learning In prototype learning methods (Asano, Rupprecht, and Vedaldi 2019; Caron et al. 2020; Li et al. 2020b,a), prototypes are defined as the centroid of a cluster formed by semantically similar instances. In this setup, the embedding of a neural network is the input to a clustering algorithm that performs prototype assignments, which are subsequently used as "pseudo-labels" to supervise a self-representation learning process. Asano et al. (Asano, Rupprecht, and Vedaldi 2019) impose the constraint that the labels must force equipartition of the samples and proposed to use a fast version of the Sinkhorn-Knopp algorithm (Cuturi 2013) to find an approximate solution to the OT problem. The Swapping Assignments between multiple Views (SwAV) method (Caron et al. 2020) proposes combining contrastive learning and clustering, reporting impressive results in self-supervised learning. Contrastive learning methods usually perform data transformations to create instances of the same input and increase the similarity of positive sample pairs while augmenting the distance of negative sample pairs (Wu et al. 2018). SwAV method proposes to use prototypical cluster assignment that disregards pairwise comparisons. More recently, (Regatti et al. 2021) proposed the Consensus Clustering using Unsupervised Representation Learning (ConCURL) method, which introduces consensus consistency into the SwAV by defining random transformations to the feature vector and codes.

This paper proposes to embed cluster proportion priors to the prototypical contrastive cluster assignment used in SwAV to solve LLP efficiently. Different from (Liu et al. 2021), our proposal consists of a unique training stage and does not employ KL-divergence. Our efficient and straightforward approach set a new boundary for LLP classification, outperforming the existing state-of-art methods in the field.

Preliminaries

Before detailing our method, we present preliminaries related to our contribution, including formally stating the LLP problem and the OT algorithm.

Reminders on Optimal Transport

Using the notation of Cuturi in (Cuturi 2013), let $\mathbf{r}$ and $\mathbf{a}$ be two probability vectors in the simplex $\sum_d := \{\mathbf{r} \in \mathbb{R}^d_+ : \mathbf{r}^T\mathbf{1}_d = 1\}$, where $\mathbf{1}_d$ is a d-dimensional vector with all elements equal to one in order to to satisfy the marginal constrains. Then, consider $U(\mathbf{r}, \mathbf{c})$ as a set of $d \times d$ matrices namely the transportation polytope of $\mathbf{r}$ and $\mathbf{a}$,

$$U(\mathbf{r}, \mathbf{a}) := \{\mathbf{P} \in \mathbb{R}_+^{d\times d} | \mathbf{P}\mathbf{1}_d = \mathbf{r}, \mathbf{P}^T\mathbf{1}_d = \mathbf{a}\}. \quad (1)$$

All $d \times d$ matrices in $U(\mathbf{r}, \mathbf{a})$ are non-negative such that $\mathbf{r}$ and $\mathbf{a}$ are their row and column marginals, respectively. For two multinomial random variables X and Y with distribution $\mathbf{r}$ and $\mathbf{a}$ respectively, $U(\mathbf{r}, \mathbf{a})$ contains all possible joint probabilities of (X, Y) (Cuturi 2013). Hence, any $\mathbf{P} \in U(\mathbf{r}, \mathbf{a})$ is a joint probability matrix of (X, Y) such that $p(X = i, Y = j) = \mathbf{P}_{i,j}$. With this notation, the entropy h of the joint probabilities $\mathbf{P}$ and their marginals $\mathbf{r} \in \sum_d$, can be formalized as

$$h(\mathbf{r}) = -\sum_{i=1}^{d} r_i \log r_i,\ h(\mathbf{P}) = -\sum_{i,j=1}^{d} P_{i,j} \log P_{i,j}. \quad (2)$$

Now, considering $\mathbf{M} \in \mathbb{R}^{d\times d}$ as the cost matrix, the cost of transport $\mathbf{r}$ to $\mathbf{a}$ using the joint probability $\mathbf{P}$ is formulated as the Frobenius dot-product $\langle \mathbf{P}, \mathbf{M} \rangle$, and the optimal transport (OT) problem for $\mathbf{r}$ and $\mathbf{a}$ given M is defined as

$$d_{\mathbf{M}}(\mathbf{r}, \mathbf{a}) := \min_{\mathbf{P} \in U(\mathbf{r}, \mathbf{a})} \langle \mathbf{P}, \mathbf{M} \rangle = \sum_{i,j} P_{i,j} M_{i,j} \quad (3)$$

where $d_{\mathbf{M}}(\mathbf{r}, \mathbf{a})$ is a distance between $\mathbf{r}$ and $\mathbf{a}$ (Cuturi 2013).

Entropic Constraints The OT solution in equation 3 is solved on the vertices of the polytope $U(\mathbf{r}, \mathbf{a})$ which lead to undesirable sparse solutions. Moreover, solving equation 3 is computationally expensive as it requires solving a linear equation that scales quadratically with the size of the sample. To address this issue, Cuturi (Cuturi 2013) proposed an entropic regularization term that smooth the prediction and allows for an efficient solver using the Sinkhorn-Knopp algorithm. The entropic function of the joint probability matrix $h(\mathbf{P})$ is strongly concave and subject to $h(\mathbf{P}) \leq h(\mathbf{r}) + h(\mathbf{a}) = h(\mathbf{r}\mathbf{a}^T)$. Hence, the authors use $-h(\mathbf{P})$ as a regularization function to obtain an approximate solution as follows

$$d^{\varepsilon}_{\mathbf{M}}(\mathbf{r}, \mathbf{a}) := \min_{\mathbf{P} \in U(\mathbf{r}, \mathbf{a})} \langle \mathbf{P}, \mathbf{M} \rangle - \varepsilon h(\mathbf{P}), \quad (4)$$

where ε is a trade-off parameter that controls the smoothness of the prediction. This modification pushes away the solution from the vertex towards an "entropic center" (Peyré, Cuturi et al. 2019). For more details about entropic regularization of OT we refer the reader to (Peyré, Cuturi et al. 2019).

Learning from Label Proportions

In the standard LLP formulation, training samples are split into bags where only the label proportions inside each bag is known, and used to obtain the instance-level label using a solver of choice. Following previous works, we assume that the training data is composed of N disjoint bags. Let B_i be the i^{th} bag with a set of randomly generated samples $\mathcal{B}_i = \{(\mathbf{x}_{i,j})\}_{j=1}^{n_i}$, where $\mathbf{x}_{i,j}$ is the sample j within bag i, and n_i is the total amount of samples in the bag. Then, the training set can be expressed as $\mathcal{D} = \{(\mathcal{B}_i, \mathbf{w}_i)\}_{i=1}^{N}$, where $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset, \forall i \neq j$. For a multi-class problem with K classes, let $\mathbf{w}_i \in \Delta_K$ be the vector of label proportion for the bag $\mathcal{B}_i$, where the $\mathbf{w}_i^k$ element is the proportion of instances that belong to class k subject to $\sum_{k=1}^{K} \mathbf{w}_i^k = 1$.

Method

This section details the proposal for embedding label proportion priors to prototypical cluster assignments. We first explain the link between OT and LLP and then describe our prototypical contrastive cluster method incorporating learning from label proportions.

LLP as an Instance of the OT Problem

In a standard deep LLP setting, the network commonly implements a feature extractor followed by a classification head that maps the features to a probabilities vector $\tilde{\mathbf{p}}_{i,j} = p_\theta(\mathbf{y}|\mathbf{x}_{i,j})$ using a softmax operator, where $\mathbf{x}_{i,j}$ is the j^{th} sample of bag B_i, θ are the network parameters, p_θ is the probability assigned to a given class by the network and $\mathbf{y}$ is a cluster assignment vector of size K (Liu et al. 2019). Then the estimated bag-level label proportion for a given bag can be calculated as the summation of element-wise posterior probability for a given bag:

$$\hat{\mathbf{w}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\mathbf{p}}_{i,j} = \frac{1}{n_i} \sum_{j=1}^{n_i} p_\theta(\mathbf{y}|\mathbf{x}_{i,j}). \quad (5)$$

Then, given the known labels proportion $\mathbf{w}_i = (\frac{m_1}{n_i}, \frac{m_2}{n_i}, ..., \frac{m_K}{n_i}) \in \Delta_K$, the bag-level loss function boils down to a standard cross-entropy loss function

$$L(\hat{w}, w) = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_i \log \hat{\mathbf{w}}_i. \quad (6)$$

To effectively link LLP with OT we build on (Asano, Rupprecht, and Vedaldi 2019; Liu et al. 2021) and reformulate equation 6 by encoding the labels proportions $\mathbf{w}_i$ as posterior distributions $q(y^k|\mathbf{x}_{i,j})$:

$$L(p, q) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{K} \frac{q(y^k|\mathbf{x}_{i,j})}{n_i} \log p_\theta(y^k|\mathbf{x}_{i,j}). \quad (7)$$

where K is the number of classes and p is the model output probability. With the addition of the proportion constraint, the function objective reads:

$$\min_{(p,q)} L(q, p), \quad s.t. \quad \forall y : q(y^k|\cdot) \in [0, 1] \quad (8)$$

and

$$\sum_{j=1}^{n_i} q(y^k|\mathbf{x}_{i,j}) = m_k. \quad (9)$$

Where the proportion constraint ensures that each label $k = \{1, ..., K\}$ contains overall m_k samples.

As discussed in (Asano, Rupprecht, and Vedaldi 2019) and (Liu et al. 2021), the objective in equation 8 is combinatorial in q which can be difficult to optimize. However, it is also an instance of the OT problem and can be solved efficiently using the Sinkhorn-Knopp algorithm.

Following (Asano, Rupprecht, and Vedaldi 2019) we can writing in terms of OT. Let $\mathbf{P}^y_{i,j} = p_\theta(y|\mathbf{x}_{i,j})\frac{1}{n_i}$ be the $K \times n_i$ joint probabilities matrix estimate by the model and $\mathbf{Q}^y_{i,j} = q(y|\mathbf{x}_{i,j})\frac{1}{n_i}$ be the $K \times n_i$ matrix of assigned joint probabilities for bag $\mathcal{B}_i$. In our case, we want that $\mathbf{Q}_i$ split the data non-uniformly within the bag, i.e. constrained to prior information about the labels proportions. Hence, we add the constraint to the transportation polytope as follows

$$U(\mathbf{w}, \mathbf{a})_i := \{\mathbf{Q}_i \in \mathbb{R}_+^{K \times n_i} | \mathbf{Q}_i \mathbf{1}_{n_i} = \mathbf{w}_i, \mathbf{Q}_i^T \mathbf{1}_K = \mathbf{a}\}. \quad (10)$$

where, as previously stated, $\mathbf{w}_i$ is the vector of known labels proportion for bag $\mathcal{B}_i$ and $\mathbf{a} = \frac{1}{n_i}\mathbf{1}_{n_i}$. From equation 4, the objective function in equation 8 for the i^{th} bag can be written as an OT solver

$$L(q, p)_i + \log n_i = \langle \mathbf{Q}_i, -\log \mathbf{P}_i \rangle, \quad (11)$$

and with the addition of the entropic regularization term, the objective function for the i^{th} bag is

$$\min_{\mathbf{Q}_i \in U(\mathbf{w}, \mathbf{a})_i} \langle \mathbf{Q}_i, -\log \mathbf{P}_i \rangle + \varepsilon h(\mathbf{Q}_i). \quad (12)$$

The advantage of adding the regularization term is that the minimization problem can now be written as a normalized exponential matrix. The next section give more details on how this optimization problem is solved.

Learning from Label Proportions with Prototypical Contrastive Clustering

The above formulation assigns discrete labels to samples and, therefore, can also be interpreted as clustering. To perform online cluster assignment, we employ the Swapping Assignments between multiple Views (SwAV) method (Caron et al. 2020). SwAV is an online clustering-based self-supervised method that trains a convolutional neural network to learn an embedding that delivers consistent cluster assignments between codes, i.e., the cluster assignments from different views (i.e., augmentations) of the same input image are consistent. The method built on contrastive learning methods (Wu et al. 2018) to learn semantic representations by comparing the images cluster assignment instead of their features. The clustering uses an augmented view of a given sample to compute targets using an OT solver and other augmented views of the same sample to predict these targets using the cross-entropy loss function. As in Asano et al. (Asano, Rupprecht, and Vedaldi 2019), the authors impose an equipartition constraint to avoid all samples mapped to the same cluster. Conversely, in our implementation, we substitute the equipartition condition by cluster size constraints and solve the OT problem using equation 12.

Online clustering Let the number of clusters be equal to the number of classes K, and $\mathbf{v}_k \in \mathbb{R}^d$ the prototype vector associated with cluster k. Given a bag $\mathcal{B}_i$ of images, each image j within the bag is transformed into two augmented views $\mathbf{x}^s_{i,j}$ and $\mathbf{x}^t_{i,j}$, and fed to an encoder network f_θ to extract the two corresponding set of features $\mathbf{z}^s_{i,j}, \mathbf{z}^t_{i,j} \in \mathbb{R}^m$. The features are then projected to the unit sphere (Caron et al. 2020) and mapped to a K trainable prototypes vectors $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_k]$ deriving the codes $\mathbf{c}^s_{i,j}$ and $\mathbf{c}^t_{i,j}$. The loss function performs a "swapped" procedure that predicts the assignment of one feature from the code of the other one. Hence, the feature extractor network and prototypes weights are jointly trained, minimizing the subsequent loss for all samples j within bag i:

$$L_{swap}(\mathbf{z}^s_{i,j}, \mathbf{z}^t_{i,j}) = \ell(\mathbf{z}^s_{i,j}, \mathbf{c}^t_{i,j}) + \ell(\mathbf{z}^t_{i,j}, \mathbf{c}^s_{i,j}), \quad (13)$$

where each term represents the cross-entropy loss between the code and the probability obtained as the softmax function of the dot product between the features and all the prototypes in $\mathbf{V}$:

$$\ell(\mathbf{z}^t_{i,j}, \mathbf{c}^s_{i,j}) = -\sum_k \mathbf{c}^{s(k)}_{i,j} \log \mathbf{p}^{t(k)}_{i,j}, \quad (14)$$

where

$$\mathbf{p}^{t(k)}_{i,j} = \frac{\exp(\frac{1}{\tau}(\mathbf{z}^t_{i,j})^\intercal \mathbf{v}_k)}{\sum_{k'} \exp(\frac{1}{\tau}(\mathbf{z}^t_{i,j})^\intercal \mathbf{v}_{k'})}, \quad (15)$$

and τ is the temperature parameter of the softmax function.

Computing codes with proportions constraint We solve the cluster assignment using the entropic regularized OT, which implies the samples in the bag are partitioned accordingly to the bag-level cluster/label size proportions. For the i^{th} bag, let $\mathbf{Z}_i = [\mathbf{z}_{i,1}, ..., \mathbf{z}_{i,n_i}]$ be the feature vectors that we want to map to the prototypes $\mathbf{V}$ and let $\mathbf{Q}_i = [\mathbf{c}_{i,1}, ..., \mathbf{c}_{i,n_i}]$ be the codes that perform the transportation, restricted to the

Algorithm 1: LLP-Co training loop using two views

```
Input: $\mathcal{D} = \{(\mathcal{B}_i, \mathbf{w}_i)\}_{i=1}^{N}$, $\varepsilon > 0$, epochs
Initialize: $f_\theta$ and prototypes $\mathbf{V}$
for i = 1 to epochs do
    for each $\mathcal{B}_i$ in $\mathcal{D}$ do
        Generate two random views $\mathbf{X}_i^{t,s}$
        Obtain the feature vectors $\mathbf{Z}_i^{t,s}$
        Compute the prototype scores $\mathbf{V}^\top \mathbf{Z}_i^{t,s}$
        Compute the codes $\mathbf{Q}_i^{t,s}$ through Sinkhorn constrained to $\mathbf{w}_i$
        Convert prototype scores to probabilities $\mathbf{P}_i^{t,s}$
        Compute loss using the swap prediction problem:
        loss $= -0.5 * mean(\mathbf{Q}_i^t * log(\mathbf{P}_i^s) + \mathbf{Q}_i^s * log(\mathbf{P}_i^t)$
        Update $\theta$ and $\mathbf{V}$ with a gradient step
    end for
end for
```

proportion constraints presented in equation 10. Using the notation of (Caron et al. 2020), we optimize $\mathbf{Q}_i$ in order to maximize the similarity between $\mathbf{Z}_i$ and $\mathbf{V}$ as follows

$$\max_{\mathbf{Q}_i} \mathbf{Tr}(\mathbf{Q}_i^\top \mathbf{V}^\top \mathbf{Z}_i) + \varepsilon h(\mathbf{Q}_i). \tag{16}$$

This formulation is equivalent to the learning objective in equation 12. To add the cluster size proportions, we introduce the marginals constraint as in equation 10. Using the regularization term allows writing the optimization problem as a normalized exponential matrix (Caron et al. 2020):

$$\mathbf{Q}_i^* = \operatorname{diag}(\alpha) \exp\left(\frac{\mathbf{V}_i^\top \mathbf{Z}_i}{\varepsilon}\right) \operatorname{diag}(\beta), \tag{17}$$

where α and β are renormalization vectors to ensure that the resulting matrix $\mathbf{Q}_i^*$ is a probability matrix. These vectors can be easily computed throughout iterative matrix multiplication using the Sinkhorn-Knopp algorithm (Cuturi 2013).

After computing the codes, the loss for updating the network weights s is computed using the cross-entropy loss probabilities between one view and assigned codes of the other view, and vice versa as shown equation 13. We outline the learning procedure for two random views in Algorithm 1. For more information about the SwAV method and the Sinkhorn-Knopp algorithm, we refer the readers to (Cuturi 2013; Caron et al. 2020).

Experiments

We empirically assess the performance of our method using two standard image classification benchmarks (CIFAR-10 and CIFAR-100 (Krizhevsky, Nair, and Hinton 2012)) and a ResNet18 architecture (He et al. 2016). CIFAR-10 and CIFAR-100 datasets are released under the MIT licenses. We implemented our method upon the SwAV (Caron et al. 2020) algorithm that is released under the Creative Commons Attribution-NonCommercial 4.0 International, introducing the cluster size constraint into the Sinkhorn-Knopp. However, since we provided the exact cluster size, we are not restricted to minimum batch size constraints. We compare our method with the LLP-GAN (Liu et al. 2019), and LLP-GAN-PLOT (Liu et al. 2021) methods, which are considered the state-of-the-art in LLP.

Experimental Details

Bag-level label proportions generation For a given bag size n_i, we create the training bag $\mathcal{B}_i$ by randomly selected n_i samples from the training set, such as each sample within the bag is unique. Following previous works (Liu et al. 2019; Dulac-Arnold et al. 2019; Liu et al. 2021), we defined four experiments with different bag sizes $n_i = [16, 32, 64, 128]$.

Architecture and training We implemented our method using a configuration similar to the one in (Caron et al. 2020). We used a ResNet18 as backbone architecture followed by a projection head that projects the output of the ResNet18 to a 1024-D space. All the experimented models were trained using stochastic gradient descent (SGD), with a weight decay of 1×10^{-6} and an initial learning rate of 0.1. We warmed up the learning rate during five epochs and then used the cosine learning rate decay (Loshchilov and Hutter 2016) with a final value of 0.0001. As in (Caron et al. 2020), the softmax temperature τ was set to 0.1, and the prototypes were frozen during the first epoch. All our models were trained for 500 epochs. The input images size is 32×32, and we used the same augmentations strategy that Caron et al. (Caron et al. 2020) to obtain four different image views, two standard resolution views, and two low-resolution views. However, we did not employ blur data augmentation.

Hyperparameters for the Cluster Assignment Using OT The weight of the entropy term ε was set to 0.05, and we stopped the Sinkhorn iterations when the element-wise error between the marginal r and the know proportion $\mathbf{w}_i$ was less than $(1/K) * 0.1$. As in previous works, we tested two cluster assignment strategies, hard- and soft-assignment (Li et al. 2020a; Caron et al. 2020; Liu et al. 2021). In the soft-assignment, we used the assignments $\mathbf{Q}$ obtained by the entropic regularized OT (a continuous solution), and for the hard-assignment, we converted the assignment to a binary output using a rounding procedure. In preliminary experiments, we found that in both datasets, hard-assignment delivers the best results for bag sizes 16 and 32, while soft-assignment delivers the best results for bag sizes 64 and 128. As previously observed in (Regatti et al. 2021), there is not a global hyperparameters configuration that holds the same performance across different datasets and experimental setups. On CIFAR-100 experiment with bag size 16 we needed to modify the gradient clipping to 0.1 and reduce the number of OT iterations to a maximum of 5 to avoid model collapse. With the standard parameters, the training reaches an inflection point at epoch 100 and then degrade slowly until the end of the training, and we observed that all centroids collapsed into a single region. We present more details on hyperparameter sensitivity in the supplementary material.

Evaluation Metrics For evaluating our results, we propose using the metric $\mathsf{Acc_H}$ that takes the cluster assignment as the prediction and computers the optimal matching between clusters and labels using the Hungarian algorithm (Kuhn

Dataset	LLP Methods	Bag Size				SwAV (kNN)	ConCURL	Supervised
		16	32	64	128			
CIFAR-10	DLLP-KL	86.0	72.0	56.0	41.0	80.0	84.6	93.6
	DLLP-ROT	78.0	65.0	53.0	40.0			
	LLP-GAN	86.3	83.8	79.0	72.6			
	LLP-GAN-PLOT	89.3	88.2	84.1	79.1			
	LLP-Co ($\mathsf{Acc_H}$) **(ours)**	**90.0**	**89.8**	**90.9**	**86.2**			
	LLP-Co ($\mathsf{Acc_A}$) **(ours)**	**90.0**	**89.8**	**90.9**	72.1			
CIFAR-100	DLLP-KL	58.0	38.0	24.0	14.0	45.8	47.9	78.3
	DLLP-ROT	51.0	37.0	24.0	14.0			
	LLP-GAN	49.1	43.6	35.6	15.0			
	LLP-GAN-PLOT	**65.4**	61.7	55.7	43.4			
	LLP-Co ($\mathsf{Acc_H}$) **(ours)**	59.5	**65.9**	**66.5**	**64.7**			
	LLP-Co ($\mathsf{Acc_A}$) **(ours)**	59.4	65.7	**66.5**	62.0			

Table 1: Test accuracy rates (%) on CIFAR-10 and CIFAR-100 datasets with different bag sizes.

1955). We also report the classification accuracy $\mathsf{Acc_A}$, which takes the exact cluster assignment as the predicted label, i.e., if a test sample was assigned to the prototype $\mathbf{v}_1$ the corresponding label will be 1. Since we expect the addition of the proportions constraint to improve the quality of the learned features, we also evaluate our models by performing k-nearest neighbor (kNN) classification (Wu et al. 2018). For a feature $\mathbf{z}$ in the test set, we take the top 25 nearest neighbors from the training set and perform majority voting to assign the label.

Results and Analysis

Comparison with the State-of-the-Art Models Table 1 provides the accuracy for our proposed method and for three baseline LLP approaches: DLLP (Dulac-Arnold et al. 2019), LLP-GAN (Liu et al. 2019) and LLP-GAN-PLOT (Liu et al. 2021). In Table 1, DLLP-KL and DLLP-ROT stand for the KL-divergence and the ROT results reported in (Dulac-Arnold et al. 2019). In addition, we also compared our results with SwAV and ConCURL (Regatti et al. 2021), both methods considered SOTA in unsupervised clustering approaches. For the SwAV experiments, we used the same network configuration and augmentation strategy used for the LLP-Co and set the Sinkhorn iterations to 5 and batch size 1024. For the others baseline models, we used the results reported by the authors. As an additional reference, we also provide the fully supervised learning results for both datasets using a ResNet50 offered by (Chen et al. 2020).

As observed, our method delivers definite improvements compared to the baseline LLP methods, more significantly observed for large bag sizes (64 and 128). In particular, we observed that our approach is robust to the four analyzed bag sizes for CIFAR-10, reporting overall very similar values for all sizes and with similar behavior observed for CIFAR-100. However, the experiment with bag size 16 for CIFAR-100 reports worse results than larger bags and the LLP-GAN-PLOT baseline method. To better understand that, Figure 2 presents the convergence curve in terms of Acc_A for this bag size. The curve shows the model reached a peak of 65% accuracy at epoch 190 and then degraded until it converged to an accuracy of around 59%. While we envisage that early stopping could potentially circumvent that undesired behavior, we did not implement it in our experiments. The $\mathsf{Acc_H}$ of LLP-Co with bag size 128 achieved 64.7% on CIFAR-100, which is 21% superior to the LLP-GAN-PLOT result. Analyzing CIFAR-10, LLP-Co reached similar performance to the fully supervised scenario for all bag sizes. Compared with the unsupervised clustering method, LLP-Co ($\mathsf{Acc_H}$) method outperformed ConCURL by only 1.6% for bag size 128, and up to 6% for 16, 32, and 64 on CIFAR-10 dataset. In contrast, more significant improvements were observed for CIFAR-100, achieving close to 15% for bag size 32, 64, and 128.

Concerning $\mathsf{Acc_A}$, as expected, the models learned overall the correct label of each prototype. Results for $\mathsf{Acc_A}$ match almost perfectly the accuracy using the Hungarian algorithm, with an exception for bag size 128 in CIFAR-10. That effect is somewhat expected since CIFAR-10 had only ten classes with an equal number of samples per class when bag size $n_i \longrightarrow \infty$ the distribution of each label inside the bag converges to $\mathbf{1}_{n_i}/n_i$, which can lead to a cluster swapping at some point of the learning process.

Convergence Curves Figure 3 provides the convergence curves for the training loss (Figure 3 top) and test accuracy using a standard kNN classifier (Figure 3 bottom) for different bag sizes. As expected for both datasets, the models require more epochs to achieve convergence as the bag size

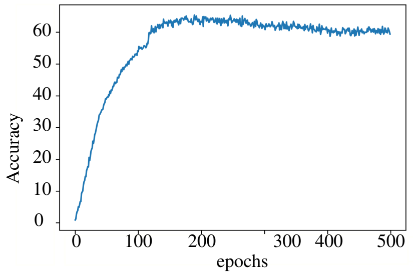

Figure 2: The Acc_A convergence curve for the test set for bag size 16 on CIFAR-100 dataset.

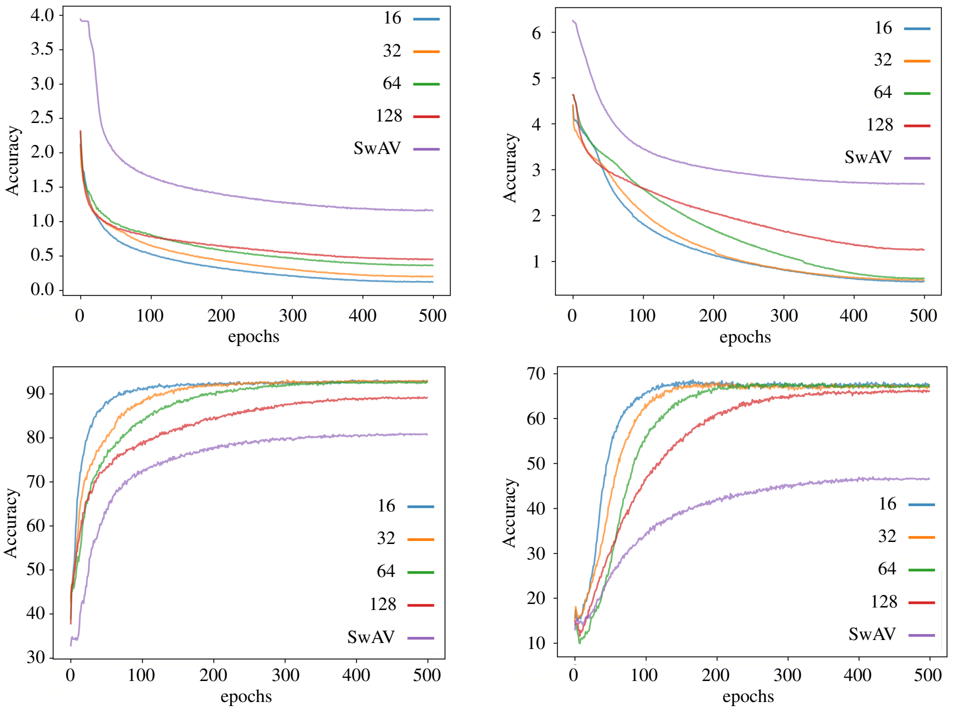

Figure 3: Training loss (top) and test accuracy using a kNN classifier (bottom) for different bag sizes for CIFAR-10 (left) and CIFAR-100 (right).

increases. We also observed that the LLP-Co loss converged faster than SwAV. Furthermore, using the kNN classifier, all models converged to similar values for the four bag sizes in both datasets, above 88% for CIFAR-10 and above 65% for CIFAR-100. For the CIFAR-10 dataset, the kNN classifier achieved 92% for bag sizes 16, 32, and 64, which is only 1.6% lower than the ResNet50 supervised counterpart (see Table 1). Considering the more challenging problem presented by CIFAR-100, the supervised ResNet50 (78.3% in Table 1) is only 12% higher than LLP-Co. Notice that ResNet50 is a much deeper network than ResNet18. Nonetheless, our proposal achieved competitive performances. Finally, contrary to other LLP methods that suffer degradation in accuracy as the bag size increases, results in Table 1 and Figure 3 indicate that our model converges at bag size relatively large (i.e., 64 and 128) for both datasets. It reached similar or better accuracy than the best values obtained with bag 16 in the baseline methods, close to the fully supervised models.

Feature Visualization In Figure 4 we visualize the learned representation projected to the unit sphere as well as the cluster centroids for CIFAR-10 e CIFAR-100 using t-SNE (Van der Maaten and Hinton 2008). As observed, the learned representations form distant clusters, which suggest the features have discriminative power, beneficial for various downstream tasks.

Conclusion

This paper proposed a method to address Learning from Label Proportions (LLP) from a new perspective using contrastive cluster assignment with class proportion constraints. Bag-level LLP approaches focusing on the classification task suffer from degradation as the bag size increases. Considering this, we propose to solve the LLP problem by combining prototypical contrastive cluster assignment and cluster size constraint in an end-to-end framework. To this end, we use a prototypical learning approach with an entropic regularized OT algorithm to solve the cluster assignment and strictly match the proportional information within a bag. Our model significantly improves the performance compared to previous SOTA works on LLP, achieving results close to those observed in fully supervised counterparts and presenting higher robustness to big bag sizes. We highlight the possible societal impacts related to eventual biases inherently used in the context of class proportions and reinforce the ethical use of this solution in this concern.

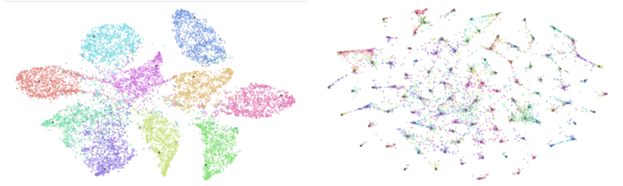

Figure 4: t-SNE plots considering the projection vector on the unit sphere for bag size 64 for CIFAR-10 (left) and CIFAR-100 (right). The gray points stands for the prototypes (i.e., the cluster centers).

References

Ardehaly, E. M.; and Culotta, A. 2017. Co-training for demographic classification using deep learning from label proportions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1017–1024. IEEE.

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2019. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*.

Bortsova, G.; Dubost, F.; Ørting, S.; Katramados, I.; Hogeweg, L.; Thomsen, L.; Wille, M.; and de Bruijne, M. 2018. Deep learning from label proportions for emphysema quantification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 768–776. Springer.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33: 9912–9924.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26: 2292–2300.

Ding, Y.; Li, Y.; and Yu, W. 2017. Learning from label proportions for SAR image classification. *Eurasip Journal on Advances in Signal Processing*, 2017(1): 1–12.

Dulac-Arnold, G.; Zeghidour, N.; Cuturi, M.; Beyer, L.; and Vert, J.-P. 2019. Deep multi-class learning from label proportions. *arXiv preprint arXiv:1905.12909*.

Fard, M. M.; Thonet, T.; and Gaussier, E. 2020. Deep k-means: Jointly clustering with k-means and learning representations. *Pattern Recognition Letters*, 138: 185–192.

Genevay, A.; Dulac-Arnold, G.; and Vert, J.-P. 2019. Differentiable deep clustering with cluster size constraints. *arXiv preprint arXiv:1910.09036*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hernández-González, J.; Inza, I.; and Lozano, J. A. 2013. Learning bayesian network classifiers from label proportions. *Pattern Recognition*, 46(12): 3425–3440.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2012. CIFAR-10 (Canadian Institute for Advanced Research). *University of Toronto*.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Li, C.; Li, X.; Zhang, L.; Peng, B.; Zhou, M.; and Gao, J. 2020a. Self-supervised Pre-training with Hard Examples Improves Visual Representations. *arXiv preprint arXiv:2012.13493*.

Li, J.; Zhou, P.; Xiong, C.; Socher, R.; and Hoi, S. C. 2020b. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Liu, J.; Wang, B.; Qi, Z.; Tian, Y.; and Shi, Y. 2019. Learning from Label Proportions with Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 32: 7169–7179.

Liu, J.; Wang, B.; Shen, X.; Qi, Z.; and Tian, Y. 2021. Two-stage Training for Learning from Label Proportions. *arXiv preprint arXiv:2105.10635*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Poyiadzi, R.; Santos-Rodriguez, R.; and Twomey, N. 2018. Label propagation for learning with label proportions. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

Qi, Z.; Wang, B.; Meng, F.; and Niu, L. 2016. Learning with label proportions via NPSVM. *IEEE transactions on cybernetics*, 47(10): 3293–3305.

Quadrianto, N.; Smola, A. J.; Caetano, T. S.; and Le, Q. V. 2009. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(10).

Regatti, J. R.; Deshmukh, A. A.; Manavoglu, E.; and Dogan, U. 2021. Consensus clustering with unsupervised representation learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Scott, C.; and Zhang, J. 2020. Learning from Label Proportions: A Mutual Contamination Framework. *Advances in neural information processing systems*.

Shi, Y.; Liu, J.; Wang, B.; Qi, Z.; and Tian, Y. 2020. Deep learning from label proportions with labeled samples. *Neural Networks*, 128: 73–81.

Stolpe, M.; and Morik, K. 2011. Learning from label proportions by optimizing cluster model selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 349–364. Springer.

Sun, T.; Sheldon, D.; and O'Connor, B. 2017. A Probabilistic Approach for Learning with Label Proportions Applied to the US Presidential Election. In *2017 IEEE International Conference on Data Mining (ICDM)*, 445–454.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Yu, F. X.; Choromanski, K.; Kumar, S.; Jebara, T.; and Chang, S.-F. 2014. On learning from label proportions. *arXiv preprint arXiv:1402.5902*.

Yu, F. X.; Liu, D.; Kumar, S.; Jebara, T.; and Chang, S.-F. 2013. $\propto$SVM for learning with label proportions. arXiv:1306.0886.