Beyond Learning Features: Training a Fully-functional Classifier with Zero Instance-Level Labels

Deepak Babu Sam*, Abhinav Agarwalla*, R. Venkatesh Babu

Video Analytics Lab, Department of Computational and Data Sciences,
Indian Institute of Science, Bangalore, India
deepaksam@iisc.ac.in, agarwallaabhinav@gmail.com, venky@iisc.ac.in

Abstract

We attempt to train deep neural networks for classification without using any labeled data. Existing unsupervised methods, though mine useful clusters or features, require some annotated samples to facilitate the final task-specific predictions. This defeats the true purpose of unsupervised learning and hence we envisage a paradigm of 'true' self-supervision, where absolutely no annotated instances are used for training a classifier. The proposed method first pretrains a deep network through self-supervision and performs clustering on the learned features. A classifier layer is then appended to the self-supervised network and is trained by matching the distribution of the predictions to that of a predefined prior. This approach leverages the distribution of labels for supervisory signals and consequently, no image-label pair is needed. Experiments reveal that the method works on major nominal as well as ordinal classification datasets and delivers significant performance.

Introduction

One major reason for the practical success of deep learning is unarguably the use of large human annotated datasets. These huge collections contain labeled data to directly serve the fully supervised training for the task of interest. Though this paradigm has enabled excellent performance for nearly all problems, creating such datasets is laborious and expensive. There is a mounting, but difficult requirement to include more and more diversity in datasets for better generalization. Moreover, many applications deal with data that changes rapidly, where the annotation process itself is almost impossible. For instance, social media data is very dynamic with new categories being introduced quite often and manual supervision is challenging given the scale of operation. These issues have accentuated the need for unsupervised training schemes, where the requirement of large labeled datasets is mitigated.

The existing approaches to unsupervised learning broadly rely on either representation learning or clustering. There are several methods for learning features, starting from autoencoders (Hinton and Salakhutdinov 2006; Vincent et al. 2008;

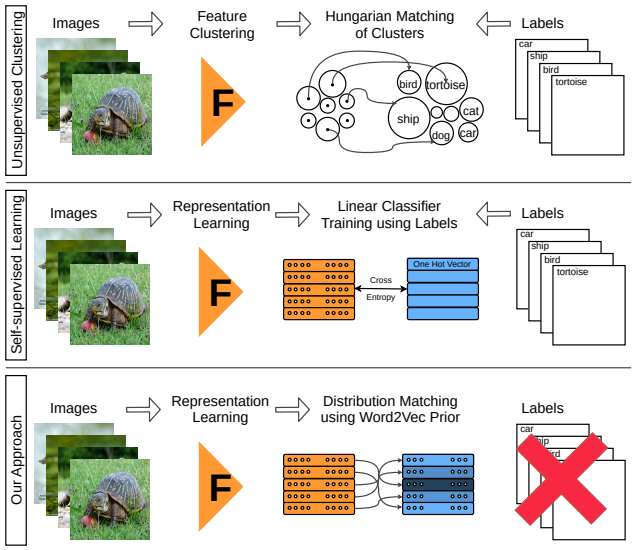

Figure 1: Comparison of related learning paradigms. Both clustering and self-supervision methods utilize labels; former in the cluster assignment step while the latter to train the classifier layer. Our approach does not require labels at any stage of the training.

Kingma and Welling 2013; Makhzani and Frey 2015) to recent self-supervision (Zhang, Isola, and Efros 2016; Pathak et al. 2016; Gidaris, Singh, and Komodakis 2018; Noroozi and Favaro 2016; Kolesnikov, Zhai, and Beyer 2019). Autoencoders, in general, are optimized to predict back their inputs often through a representational bottleneck, thereby learning useful features. Self-supervision takes it further and trains the model for some pseudo label prediction tasks, where the labels could be easily obtained from the input. Consider self-supervision with colorization (Zhang, Isola, and Efros 2016; Larsson, Maire, and Shakhnarovich 2016, 2017), where the objective is to predict the color image given its grayscale version. Note that the grayscale image is generated from the color input free of any human annotation cost. One could formulate several tasks like solving jumbled images (Noroozi and Favaro 2016), inpainting (Pathak et al. 2016), classifying the angle of image rotation (Gidaris, Singh, and Komodakis 2018; Feng, Xu, and Tao 2019),

*These authors contributed equally.

etc. In contrast, clustering methods focus on assigning data points to appropriate centroids. The centroids are expected to separate the data based on high-level semantics.

Though these unsupervised methods could learn fairly generic representations, they cannot be directly employed for the downstream task of interest without labeled supervision. Self-supervision simply returns a feature vector for every input, which is not useful per se for image classification. One needs to train at least a linear layer on the feature representations to map to the category labels. Unfortunately, this training requires image-label pairs defeating the real purpose of self-supervision. Similarly, clustering approaches assign input images to one of the cluster centers. But the clusters themselves do not have any task-specific identity and again require labeled data to associate clusters with corresponding classification labels. The final performance depends on this labeled training stage, though one might get good results with less amount of data. As summarized in Figure 1, existing methods have a mandatory training requirement of labeled data.

In contrast, we propose a 'true' self-supervision paradigm that do not require even a single annotated sample for training. The current focus is only on image classification tasks and the key difference is the ability to train a fully-functional classifier without providing any image-category label pairs. The only information necessary, other than plenty of images, is the approximate description of how labels are distributed in the dataset. This statistics on the labels covers the entire collection of given training images and is not per instance-level. We leverage the word vectors of the classification category names to form this label distribution and hence do not require human supervision even here. The word vectors describe the semantic similarity between the categories and are used to match the distribution of model predictions. Our main contributions can be summarized as:

- The new paradigm of 'true' self-supervision that does not necessitate even a single instance-level annotation, but can work with the description of label distribution.
- The first fully-functional classifier trained without any image-label pairs and yet provide significant classification performance.
- A novel formulation loss for distribution matching via the cluster-guided optimal transport objective.
- A useful extension of the proposed approach to ordinal classification without any labels, which delivers competitive performance compared to supervised works.

Related Works

Self-supervision: Early works in unsupervised feature learning employ some variants of autoencoder (Hinton and Salakhutdinov 2006; Vincent et al. 2008; Kingma and Welling 2013; Makhzani and Frey 2015), where discriminatory features are acquired by learning to predict back the input or its enhanced version. The more successful self-supervision methods use different pretext tasks for which obtaining labels is trivial. They usually leverage some structural properties of the unlabeled data to formulate suitable pretext task. For instance, works like (Agrawal, Carreira, and Malik 2015; Jayaraman and Grauman 2015; Pathak et al. 2017; Wang and Gupta 2015; Misra, Zitnick, and Hebert 2016) utilize the structure in the form of motion cues and temporal information in videos to construct a self-supervision objective. Other pretext tasks include predicting the angle of object rotation (Gidaris, Singh, and Komodakis 2018; Feng, Xu, and Tao 2019), colorizing a grayscale image (Zhang, Isola, and Efros 2016; Larsson, Maire, and Shakhnarovich 2016, 2017), inpainting missing regions (Pathak et al. 2016) and learning spatial context (Noroozi and Favaro 2016; Doersch, Gupta, and Efros 2015; Nathan Mundhenk, Ho, and Chen 2018). A comprehensive study on these popular self-supervision methods can be found in (Kolesnikov, Zhai, and Beyer 2019). These pretext tasks could introduce some bias in the acquired representations and contrastive learning is one way forward. Contrastive learning approaches (Chen et al. 2020a; He et al. 2020; Dosovitskiy et al. 2014; Oord, Li, and Vinyals 2018; Bachman, Hjelm, and Buchwalter 2019a,b) enforce consistency of learned features under various data augmentations. This framework makes use of contrastive loss (Hadsell, Chopra, and LeCun 2006; Wu et al. 2018) to minimize the distance between the different augmentations of the same image and maximize for other images in a latent space. The issue of learning pretext-specific features is mitigated to a certain extent and results in extracting more generic useful features. However, all these pretext-based and contrastive-based self-supervision, focus on just learning representations and requires mandatory supervised training for the final task of classification.

Clustering: An alternate line of works use clustering for unsupervised classification. There are many methods that either directly learns through a clustering-like loss or acquire representations that need to be post-processed using a clustering method (Haeusser et al.; Caron et al. 2018; Chang et al. 2017; Xie, Girshick, and Farhadi 2016; Yang, Parikh, and Batra 2016). IIC (Ji, Henriques, and Vedaldi 2019) is an unsupervised clustering method that maximizes the mutual information between augmented versions of the same image to form clusters. SeLa (Asano, Rupprecht, and Vedaldi 2019), on the other hand, simultaneously evolve features through clustering by alternating between self-labeling and representation learning. In contrast, SCAN (Van Gansbeke et al. 2020) decouples representation learning and clustering into a multi-stage process. It utilizes a self-supervision method such as SimCLR (Chen et al. 2020a) or MocoV2 (He et al. 2020) to learn representations and then employs a self-training based method to form clusters. All of these approaches require access to instance-level labels for labeling clusters to enable final target class prediction. This cluster mapping is typically done using Hungarian assignment (Kuhn 1955).

Other related paradigms: A class of zero-shot learning methods (Norouzi et al. 2013; Zhang, Gong, and Shah 2016; Al-Halah, Tapaswi, and Stiefelhagen 2016) encodes label descriptions into a semantic embedding space of word vectors such as Word2Vec (Mikolov et al. 2013). These works relate the semantic similarity of label descriptions of unseen classes to seen classes for classification on unseen classes.

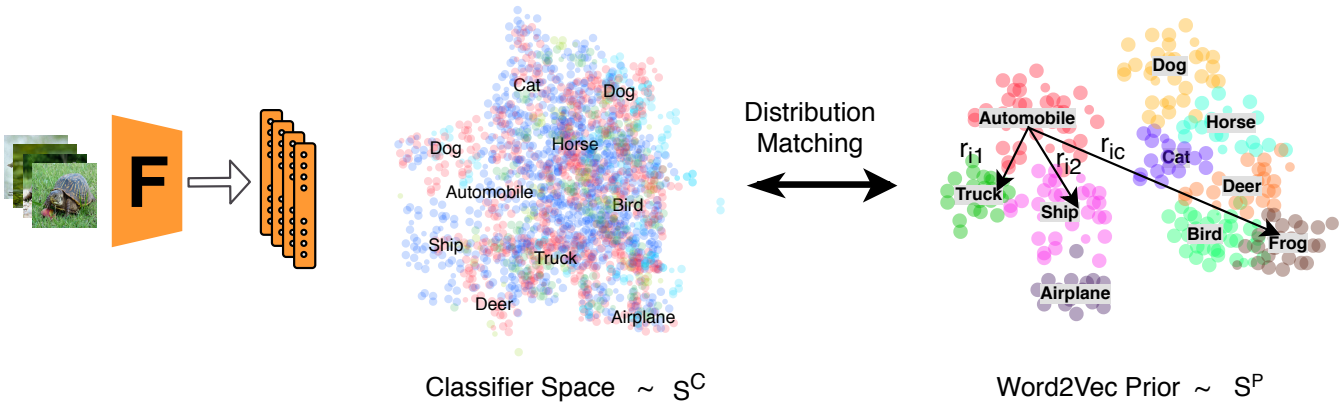

Figure 2: Depiction of the distribution matching process. The prior follows certain semantic similarity relations (extracted using word vectors of labels) across categories, which is enforced on the distribution of logits given by the classifier under training.

However, this framework requires access to the labeled set of seen classes and aims to learn the correct mapping to the unseen classes. We focus to completely eliminate the necessity for any labeled data whatsoever.

Ordinal classification: Ordinal classification is an actively researched problem in machine learning and computer vision communities, where there is some inherent order evident in the labels. It is commonly applied to computer vision problems such as age estimation (Liu, Wang, and Kong 2019; Diaz and Marathe 2019; Beckham and Pal 2017; Niu et al. 2016; Liu, Kong, and Goh 2018), depth estimation (Fu et al. 2018; Diaz and Marathe 2019), diagnosing diabetic retinopathy (Beckham and Pal 2016, 2017) etc. Almost all research in this area has been in the supervised setting. But our 'true' self-supervision framework naturally fits for ordinal classification as it can effectively leverage the inherent structure in the label space.

Our Approach

Distribution of Labels

We aim to train a classifier with zero annotated images. But realizing such a system is hard as some supervisory signal should exist for connecting the learned features or clusters to the category labels. This is thought to be an essentially unavoidable bottleneck in unsupervised learning. Interestingly, we tackle this issue by deriving supervisory signals from a measure of mismatch between the distribution of model predictions with the prior structure of the labels. We exploit the preexisting structure in labels by enforcing the relationships of the labels in a representational label space onto the classifier space. This relationship is formed using a semantic word embedding of the label names.

Observe that the predictions of trained classifier follow certain semantic characteristics. A model assigning higher confidence to 'dog' for a given image, is expected to label 'cat' as the next best category rather than 'car'. In fact, one could specify a ranking or ordinal relations among labels in terms of similarity. This semantic similarity can be easily computed with the word vector of the category labels. For instance, the term 'dog' is closer to 'cat' than 'car' or 'car' is nearer to 'ship' compared to 'bird'. The high-level semantic ordering from word vectors might not always correspond to visual similarity. But since typical image classifiers are trained on categories that have significant associations with each other, the correspondence holds in practice.

We construct a distribution over labels using the ordinal relationships of word vector similarity of label names. Let c be the total number of target classes for the classifier. Now for each label l_i, we obtain its word vector $\boldsymbol{v}_i$ and compute cosine similarity to other word vectors $\{\boldsymbol{v}_j\}_{j\neq i}$. Next, the label set $\mathbb{L} = \{1, 2, ...c\}$ is ranked according to its similarity to category l_i. The ranking is specified by,

$$r_{ij} = \text{rank}(\{\boldsymbol{v}_i^T \boldsymbol{v}_j\}_{j\in\mathbb{L}}), \tag{1}$$

where r_{ij} indicate the jth similar class to label l_i.

Since the classifier outputs are real-valued, we transform the obtained rank vector $\boldsymbol{r}_i$ to a vector $\boldsymbol{t}_i$ in the logit space. The elements of $\boldsymbol{t}_i$ are fixed conforming to $\boldsymbol{r}_i$. We set the entry for the target class as $\boldsymbol{t}_i[r_{i1}] = 1$ and fix others relative to it. As far as the target class discrimination is considered, the relations towards the most similar (r_{i2}) and the most dissimilar (r_{ic}) categories matter the most. As observed from the logits, categories typically have a unique signature in terms of the neighbouring classes in the logit space. For instance, one could say 'cat' is that category which closer to 'dog' and 'deer', but far away from 'truck' (see Figure 2). The most far away class gives valuable information in a negative discriminatory sense, while remaining categories do not provide much characterization to the target class. With this intent, we set the values for most similar classes as $\boldsymbol{t}_i[r_{i2}] = 0.5$ and $\boldsymbol{t}_i[r_{i3}] = 0.2$. The most dissimilar one $\boldsymbol{t}_i[r_{ic}]$ is fixed to -0.5. Other entries are set such that it forms an arithmetic series from $\boldsymbol{t}_i[r_{i4}]$ to $\boldsymbol{t}_i[r_{ic}]$. Finally, we unit normalize $\boldsymbol{t}_i$ to get $\hat{\boldsymbol{t}}_i$. The exact magnitude of values does not seem to have any significant effect, but the rank order matters.

We form a prior P over the logits using the vector set $T = \{\hat{\boldsymbol{t}}_0, \hat{\boldsymbol{t}}_1, ...\hat{\boldsymbol{t}}_c\}$. If the dataset is class-balanced, then P

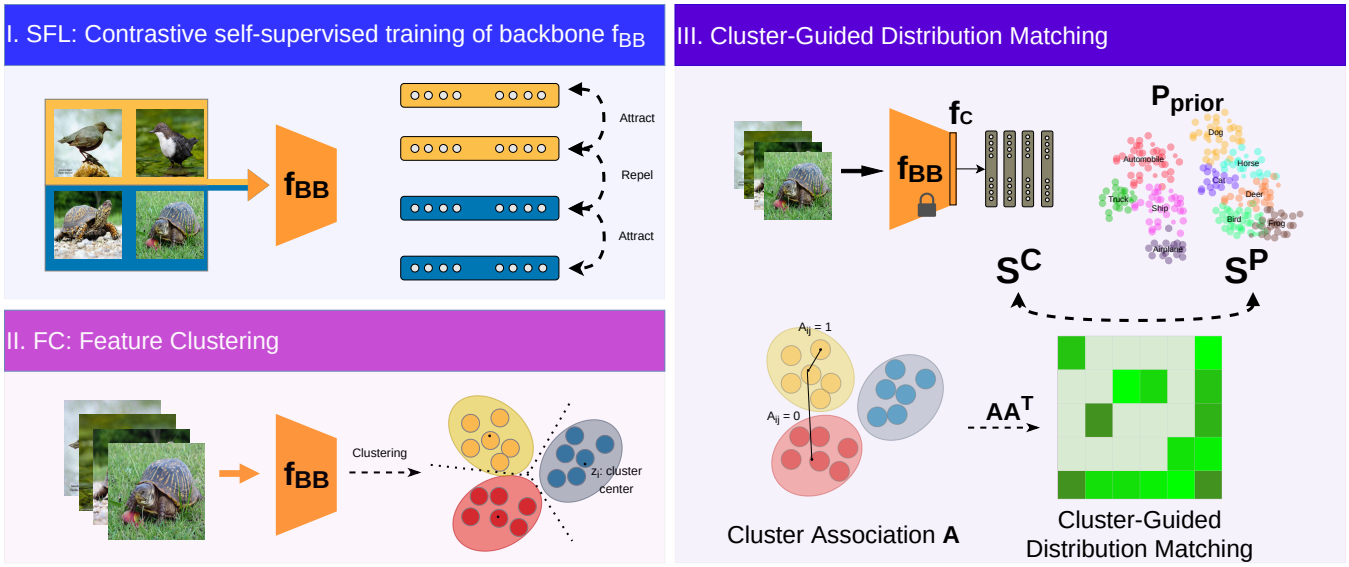

Figure 3: Training stages: I. A backbone network f_{BB} is trained to extract features using self-supervision, II. Features from f_{BB} are clustered together and III. The classifier layer f_C is updated through the cluster-guided distribution matching process.

reduces to a uniform distribution as,

$$P \sim \mathcal{U}(T); Pr(\hat{\boldsymbol{t}}_i) = \alpha_i = \frac{1}{c}. \tag{2}$$

Appropriate non-uniform distributions are chosen for unbalanced datasets with α_i set according to class size. Given this prior distribution on the labels, a classifier could be trained by enforcing the predicted logits to follow the prior. The error signal is derived from a measure of how well the prediction distribution conforms with the prior constructed from the word vectors. This matching process is illustrated in Figure 2. To further enforce the importance of most discriminatory classes (most similar and most dissimilar), we define weight vector $\boldsymbol{w}_i$ corresponding to the ith label as,

$$\boldsymbol{w}_i = [w_{i1}, ..w_{ic}]; w_{ij} = \begin{cases} e^{-r_{ij}} & \text{if } r_{ij} < \lfloor \frac{c}{2} \rfloor \\ e^{r_{ij}-c} & \text{otherwise.} \end{cases} \tag{3}$$

This exponential nature of the weights gives prominence to the characterizing categories, when used to compute distance between logit vectors.

Distribution Matching with Sinkhorn Loss

We have two distributions of vectors: one from the prior P and other formed by the logit predictions of the classifier under training, denoted by f_C. Our distributions are in the form of empirical measures (an array of samples), for which a measure of how close the underlying statistics that generated these samples needs to be defined. Optimal transport (OT) framework suits such scenarios and the distance is related to the amount of probability mass that should be transported to make the distributions similar. The standard Earth Mover's Distance (EMD) (Rubner, Tomasi, and Guibas 2000) can be used here, but is not differential as such. However, Cuturi et al. (Cuturi 2013) formulates the Sinkhorn distance between two empirical measures to be an upper bound for EMD and more importantly enables differentiability, while being computationally tractable.

We sample logit vectors by passing B images through the classifier. The resultant set of logit vectors, denoted with S^C, acts as empirical measures from the underlying distribution that the classifier has learned. Let S^P stand for samples randomly drawn the prior P. The Sinkhorn matching is performed between S^C and S^P, which essentially tries to match the underlying distributions. Now consider a transport plan $\boldsymbol{R}$, where the ijth entry indicate the likelihood of assigning ith logit in S^C to the jth sample in S^P. Every valid assignment $\boldsymbol{R}$ has a corresponding cost $\boldsymbol{M}$ defined as,

$$M_{ij} = \boldsymbol{w}_j \odot (S^C_i - S^P_j)^2, \tag{4}$$

where the weight vector $\boldsymbol{w}_j$ (computed as in equation 3) gives more consideration to the most discriminatory categories. The transport cost is simply specified as the Frobenius inner product $\langle \boldsymbol{R}, \boldsymbol{M} \rangle_F$; closer the two distribution under consideration, lower would be this cost. The Sinkhorn loss $\mathcal{L}_{sk}$ is formulated as the cost incurred for the best transport plan with an additional regularization term. Mathematically,

$$\begin{aligned} \mathcal{L}_{sk}(S^C, S^P) \quad = \quad & \arg\min_{\boldsymbol{R}} \langle \boldsymbol{R}, \boldsymbol{M} \rangle_F - \tfrac{1}{\beta} E(\boldsymbol{R}), \\ \text{s.t.} \quad & \boldsymbol{R}\mathbf{1} = \tfrac{1}{B}\mathbf{1} \\ & \boldsymbol{R}^T\mathbf{1} = \tfrac{1}{B}\mathbf{1}, \end{aligned} \tag{5}$$

where $E(\boldsymbol{R})$ is the entropy associated with $\boldsymbol{R}$ and β is a regularization constant. More details on the general formulation is available in (Cuturi 2013). So updating the classifier parameters by minimizing $\mathcal{L}_{sk}$, transforms the distribution of the predicted logits to resemble the prior.

Method	Labels	CIFAR-10	STL-10	CIFAR100-20	ImageNet-10
SimCLR (Chen et al. 2020a)	✓	91.4*	83.8*	76.1*	$91.1^{\perp}$
DeepCluster (Caron et al. 2018)	✓	37.4	33.4	18.9	-
ADC (Haeusser et al.)	✓	32.5	53.0	16.0	-
IIC (Ji, Henriques, and Vedaldi 2019)	✓	57.6 ± 5.0	59.8 ± 0.8	25.5 ± 0.5	-
SCAN (Van Gansbeke et al. 2020)	✓	87.6 ± 0.4	76.7 ± 1.9	45.9 ± 2.7	-
Random (self-supervised f_{BB} & random f_C)	✗	11.2	12.8	4.5	6.6
Ours (ZERO labels)	✗	42.1 ± 5.5	36.5 ± 3.3	13.6 ± 3.2	47.3 ± 8.6

Table 1: Percentage accuracy on various nominal classification datasets. Our approach performs significantly better over baselines not using labels and fares competitive to earlier methods requiring annotated data. * denotes our implementation and $^{\perp}$ indicates use of MocoV2 (He et al. 2020).

Self-supervision and Clustering

In Figure 3, the complete training pipeline of our method is illustrated. There are three parts, of which the first stage is *self-supervised feature learning* (SFL), where discriminatory features are extracted. We employ the SimCLR (Chen et al. 2020a) framework for training the backbone network f_{BB}. SimCLR uses a contrastive learning approach of learning representations that are close for augmented versions of the same image, but far apart for others. It extracts features that are agnostic to a variety of augmentations, thereby obtaining more object discriminative representations. We use ResNet50 (He et al. 2016) as our backbone network f_{BB} and train the model parameters till saturation under contrastive self-supervision.

The second stage runs *feature clustering* (FC) on the representations obtained from self-supervision inorder to be used as an additional guiding signal for $\mathcal{L}_{sk}$. We utilize optimal transport to minimize the distance between the features obtained from f_{BB} and the cluster centers ($\boldsymbol{z}_i$s). The number of cluster centers is set to the number of target classes. From the obtained cluster centers and the corresponding transport plan, we compute A_{ij} which represents if image i belongs to jth cluster or not (value 1 asserts the assignment, otherwise 0). Note that the FC step is completely unsupervised and the cluster centers need not correspond to the target classes. The cluster assignments in $\boldsymbol{A}$ are not to the actual target categories. At this point, unsupervised clustering works generally utilize labeled data to map the clusters to target categories, but is not plausible in our scenario. Hence, the cluster assignments are simply leveraged for better supervision via $\mathcal{L}_{sk}$.

Cluster Guided Distribution Matching

Given the pretrained backbone network f_{BB} and the cluster labels $\boldsymbol{A}$, we append a linear classifier layer f_C to f_{BB}. The prediction logits are taken from f_C to obtain the final classification scores. In the third stage, the f_{BB} is frozen, and only f_C is open for training. We modify the $\mathcal{L}_{sk}$ loss to take advantage of the signals from the cluster labels.

Since clusters are formed using f_{BB} trained in a self-supervised manner, the clustered samples share semantic characteristics. Samples in a cluster might not belong to a single category, but can have visual similarity. We infuse this semantic affinity to the optimal transport plan. The matrix $\boldsymbol{A}\boldsymbol{A}^T$ captures the relationships among the given sample set in terms of the cluster affinity. Now if $\boldsymbol{R}^*$ stand for the optimal transport plan obtained from the optimization in equation 5, then the cluster-guided Sinkhorn loss $\mathcal{L}_{cgsk}$ is formulated as,

$$\mathcal{L}_{cgsk}(S^C, S^P) = \left\langle \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{R}^*, \boldsymbol{M} \right\rangle_F . \tag{6}$$

By applying the cluster correlation matrix $\boldsymbol{A}\boldsymbol{A}^T$ on the optimal transport plan, a solution that regards both the transportation probabilities and feature semantics evolves.

We emphasize that our final loss $\mathcal{L}_{cgsk}$ is devoid of any requirement of image-label pairs. For training, a batch of images is sampled from the dataset to form the predicted logit set S^C and ordinal relationships are sampled from the prior S^P. The $\mathcal{L}_{cgsk}$ is computed and backpropagated to update the classifier layer weights. The value of the loss is monitored for saturation and the training is continued till the mean loss over a window improves. No labeled data is employed even to validate and complete the training. After training, the final classification accuracy on the test set is evaluated with the best model selected based on $\mathcal{L}_{cgsk}$.

Experiments

Implementation Details

We utilize ResNet-50 (He et al. 2016) as the base network f_{BB} for running all our experiments. For the self-supervised feature learning (SFL), a *projection head* maps average pooled features from f_{BB} as in SimCLR (Chen et al. 2020a). We use the Adam (Kingma and Ba 2014) optimizer with learning rate of 10^{-3} and a batch size of 128 for all datasets except CIFAR100-20 (Krizhevsky, Hinton et al. 2009). We use a batch size of 512 for CIFAR100-20. For ImageNet-10, we use MocoV2 (Chen et al. 2020b; He et al. 2020) because of its superior performance than SimCLR. After SFL, the projection head is dropped and clustering is performed on features obtained from f_{BB}. Here the sinkhorn clustering employs Adam optimizer with learning rate of 0.1 for a total of $5K$ steps. For obtaining Word2Vec embeddings, we utilize publicly available spaCy [1] library.

Since the loss defined in equation 6 is unbounded, we apply gradient clipping so as to fix the magnitude of the gradients. We clip the gradient norms above the value of 100.

[1] https://github.com/explosion/spaCy

Method	Label	Adience	Aesthetic	DR
Niu *et al*. (Niu et al. 2016)	✓	56.7 ± 6.0	68.96	-
CNN-POR (Liu, Kong, and Goh 2018)	✓	57.4 ± 5.8	70.05	-
SORD (Diaz and Marathe 2019)	✓	59.6 ± 3.6	72.03	-
SimCLR (Chen et al. 2020a)	✓	49.7 ± 2.7	69.87	74.3
Beckham *et al*.(Beckham and Pal 2017)	✓	55.0	-	77.0
Random (self-supervised f_{BB} & random f_C)	✗	13.2	10.73	39.8
Ours (ZERO labels)	✗	32.5 ± 8.1	57.93	57.7

Table 2: Classification accuracy (%) on ordinal classification datasets. Our approach has better accuracy than unsupervised baselines and stands comparable to methods using labeled data.

To stabilize the training, the learning rate is set to 10^{-5}. For distribution matching, we use a batch size of 500, entropy regularization constant β as 0.01 and training step as $10K$. The three most negative classes are also randomly shuffled while sampling from the prior P.

We find that initialization strategy of f_C directly influences the performance of distribution matching. Specifically, random initialization results in too much variance across different runs. This variance is reduced by a data-dependent initialization strategy developed by (Coates and Ng 2012). It utilizes spherical K-means (Buchta et al. 2012) to avoid degeneracy or empty clusters as is common with K-means based initialization. On the other hand, we just use cluster centers ($\boldsymbol{z}_i$) obtained in *feature clustering* (FC) stage using sinkhorn clustering which also avoids degenerate solutions. Moreover, a clustering based initialization has the added advantage of capturing the modes of the dataset and hence reduces degeneracy or mode collapse in the classifier space.

Nominal Classification

Datasets: We evaluate our approach on standard classification datasets employed by the unsupervised learning community. First is the CIFAR-10 (Krizhevsky, Hinton et al. 2009) dataset, which comprises of 10 classes. STL-10 (Coates, Ng, and Lee 2011) is another dataset with 5K labeled training images and an additional 100K unlabeled images, both of which are used for training. CIFAR-100-20 (Krizhevsky, Hinton et al. 2009) dataset is adapted from CIFAR-100 (Krizhevsky, Hinton et al. 2009) by grouping together 100 classes into 20 superclasses. We also test our approach on ImageNet (Krizhevsky, Sutskever, and Hinton 2012), by forming 10 superclasses from the 1000 classes present in ImageNet. We denote the resulting dataset as ImageNet-10.

The performance of our approach along with the baselines is reported in Table 1. We extensively evaluate the method through 50 training runs, each initialized with a different seed. To the best of our knowledge, there are no existing completely unsupervised methods. Due to the lack of any competing method, we consider the random accuracy as our baseline (*Random*). Random accuracy is computed by using self-supervised f_{BB} backbone with a random f_C layer. We observe that our approach performs considerably better than the baseline. The performance scores for unsupervised clustering works are taken from (Ji, Henriques, and Vedaldi 2019). It is important to note that they are not directly comparable to our approach since they utilize instance-level labels at some stage in the training or testing. Interestingly, the performance is competent with earlier unsupervised clustering based works like DeepCluster (Caron et al. 2018) and ADC (Haeusser et al.) and even outperforms them on CIFAR-10, without using any label whatsoever.

Ordinal Classification

Datasets: We utilize datasets from common applications of ordinal classification or regression in age estimation, image quality estimation and diagnosing diabetic retinopathy. The Adience (Eidinger, Enbar, and Hassner 2014) dataset is a real-world dataset of facial images collected for age and gender classification. It sources 26K phone clicked photos from Flickr with 2.2K unique faces. The images vary drastically in terms of lighting conditions, appearance, pose, etc. Age annotations are grouped into 8 ordinal classes: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53 and 60+ years old. The ordinal nature of the label set is clear and directly fixes the ranks $\boldsymbol{r}_i s$. We report the performance using the cross-validation splits released with the data. The Aesthetics (Schifanella, Redi, and Aiello 2015) dataset is collection of 15K Flickr images along with annotated aesthetic scores. All images are assigned an aesthetic score ranging from 1-5, with higher the score, the more aesthetically pleasing it is. We split the dataset into training and testing set in an 80-20 ratio. EyePACS-Kaggle Diabetic Retinopathy (Cuadros and Bresnick 2009) dataset, collected by EyePACS, is a set of 35K fundus images (*i.e*. retinal images) along with annotations for diagnosing the extent of diabetic retinopathy. The images provided are of high-resolution and are taken under a variety of lighting conditions. All images are labeled as one of *No DR*, *Mild DR*, *Moderate DR*, *Severe DR* and *Proliferative DR* categories. Following (Beckham and Pal 2017), we process the dataset using (Graham 2015), divide the dataset into training and validation images (using a 90-10 split) and report the results on the validation split.

The performance of our approach on ordinal classification datasets is reported in Table 2. As in the case of nominal classification, we consider random accuracy as our baseline in the absence of any unsupervised methods. Our approach performs considerably better. We borrow the performance numbers for ordinal classification/regression works

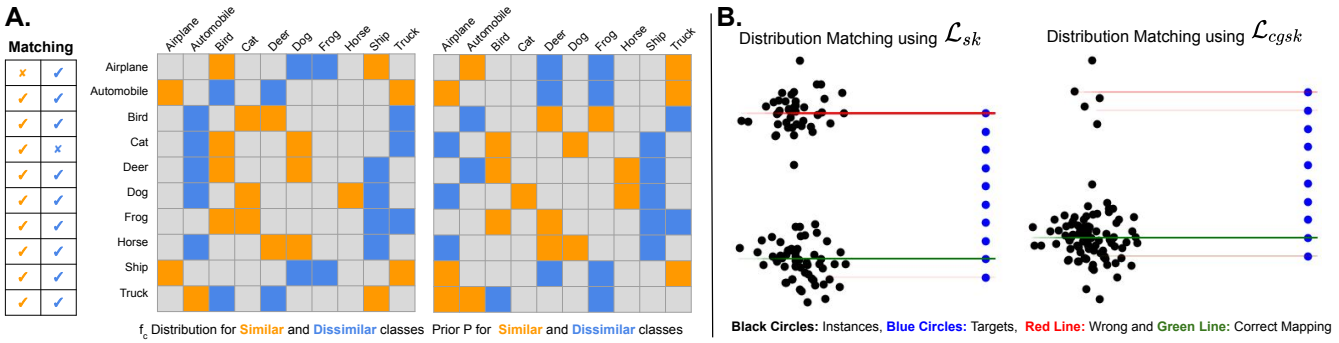

Figure 4: **A**. Visualizing alignment between f_C and prior P logits. Similar classes are colored orange while dissimilar with blue. The distributions are fairly aligned. **B**. Action of cluster-guided distribution matching. $\mathcal{L}_{cgsk}$ forces instances (black circles) belonging to a single cluster to conform to the same target (blue circles). For $\mathcal{L}_{sk}$, wrong mappings (red lines) can be seen.

Method	Accuracy
(A) No SFL, No FC Stage	9.9 ± 0.5
(B) No FC Stage ($\mathcal{L}_{sk}$)	9.8 ± 0.2
(C) No FC Stage, No Cluster Init	10.5 ± 1.9
(D) All Stages, No Cluster Init	39.4 ± 14.3
(E) All Stages	42.1 ± 5.5

Table 3: Architectural ablations on CIFAR-10. All stages (SFL $\rightarrow$ FC $\rightarrow$ Distribution Matching) are necessary.

from (Diaz and Marathe 2019). Since we consider the problem as ordinal classification and not regression as f_C directly returns the predicted class, we omit using mean-absolute-error (MAE) to measure performance. Since the backbone f_{BB} is trained using SimCLR, linear training on SimCLR using labels is an upper bound for our approach. Our approach performs well and successfully avoids degenerate solutions for datasets with high imbalance like Aesthetic and DR.

Architecture Ablations

We evaluate the effectiveness of various components of our approach and report the results on CIFAR-10 dataset in Table 3. We train f_C layer by obtaining features from a random f_{BB} network (denoted by (A)) and observe degraded performance as expected. This highlights the importance of self-supervised learning for training the backbone f_{BB}. (B) and (C) refer to ablations where the FC stage is skipped (*i.e.* using only $\mathcal{L}_{sk}$). For (C), we randomly initialize f_C without the cluster-based initialization strategy. All the components are seen to directly improve the classification performance. We also independently test the cluster initialization scheme and demonstrate its usefulness, denoted by (D). Random initialization performs slightly worse than cluster-based initialization on average with a high variance in the obtained results. Moreover, it sometimes results in models that do not predict certain classes (*i.e.* degeneracy). Cluster-based initialization strategy avoids degeneracy and enables learning all the categories for any seed value.

Analysis of Distribution Matching

We study the degree of alignment between Word2Vec prior P and the logit distribution of f_C. For this, the logit vectors from f_C corresponding to images in CIFAR-10 dataset are collected and a mean logit vector is computed for each category using the ground truth label. The mean vectors are examined to extract the similarity ranks (r_{ij}s in equation 1) with respect to other categories. A vector sampled from prior P is considered to match with that of an f_C mean vector if both contain a common class in the most similar set (*i.e.* $\{r_{i2}, r_{i3}\}$) or in the most dissimilar set (*i.e.* $\{r_{ic}, r_{i,c-1}\}$). This scheme is chosen since the proposed matching method is mostly sensitive to these extreme categories. As shown in Figure 4, we observe a good alignment between the two distributions, supporting our assumptions.

We further analyze the effect of cluster-guided distribution matching by studying the cluster assignments for $\mathcal{L}_{sk}$ and $\mathcal{L}_{cgsk}$ in Figure 4. After selecting samples from a particular cluster, the Sinkhorn assignments for these samples for a model trained with $\mathcal{L}_{sk}$ (left) and $\mathcal{L}_{cgsk}$ (right) are visualized. We find that using $\mathcal{L}_{cgsk}$ forces samples in a cluster to conform to the same output class, thus demonstrating the effectiveness of the cluster-guided matching process.

Conclusion

In this work, we presented a novel method to train an image classifier without using any instance-level labeled data. The key idea is to derive supervisory signal by matching the distribution of logit predictions to that of a prior formed from the word vector similarity of the target labels. The classification accuracy delivered by the approach is significant, considering the fact that not even a single image-label pair is required, making it highly useful for annotation intensive practical settings. However, there is a performance gap compared to fully supervised models, which should be addressed in future works. Another direction is to scale up the technique to classifiers with a large number of categories. Despite these shortcomings, our work substantiates that 'true' self-supervision could be realized and needs to be actively explored further.

Acknowledgement

This work was supported by Uchhatar Avishkar Yojana (UAY) project (IISC_010), Ministry of Human Resource Development (MHRD), Government of India.

References

Agrawal, P.; Carreira, J.; and Malik, J. 2015. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Al-Halah, Z.; Tapaswi, M.; and Stiefelhagen, R. 2016. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019a. Learning Representations by Maximizing Mutual Information Across Views. In *Advances in Neural Information Processing Systems*.

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019b. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*.

Beckham, C.; and Pal, C. 2016. A simple squared-error reformulation for ordinal classification. *arXiv preprint arXiv:1612.00775*.

Beckham, C.; and Pal, C. 2017. Unimodal Probability Distributions for Deep Ordinal Classification. In *International Conference on Machine Learning*.

Buchta, C.; Kober, M.; Feinerer, I.; and Hornik, K. 2012. Spherical k-means clustering. *Journal of Statistical Software*.

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*.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*.

Coates, A.; and Ng, A. Y. 2012. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*. Springer.

Cuadros, J.; and Bresnick, G. 2009. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*.

Diaz, R.; and Marathe, A. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Dosovitskiy, A.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*.

Eidinger, E.; Enbar, R.; and Hassner, T. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*.

Feng, Z.; Xu, C.; and Tao, D. 2019. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations.

Graham, B. 2015. Kaggle diabetic retinopathy detection competition report. *University of Warwick*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Haeusser, P.; Plapp, J.; Golkov, V.; Aljalbout, E.; and Cremers, D. ???? Associative deep clustering: Training a classification network with no labels. In *German Conference on Pattern Recognition*. Springer.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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*.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science*.

Jayaraman, D.; and Grauman, K. 2015. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ji, X.; Henriques, J. F.; and Vedaldi, A. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*.

Kolesnikov, A.; Zhai, X.; and Beyer, L. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2016. Learning representations for automatic colorization. In *Proceedings of the European Conference on Computer Vision*.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2017. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, Y.; Kong, A. W. K.; and Goh, C. K. 2018. A Constrained Deep Neural Network for Ordinal Regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, Y.; Wang, F.; and Kong, A. W. K. 2019. Probabilistic Deep Ordinal Regression Based on Gaussian Processes. In *Proceedings of the IEEE International Conference on Computer Vision*.

Makhzani, A.; and Frey, B. J. 2015. Winner-take-all autoencoders. In *Advances in Neural Information Processing Systems*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Misra, I.; Zitnick, C. L.; and Hebert, M. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nathan Mundhenk, T.; Ho, D.; and Chen, B. Y. 2018. Improvements to context based self-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; and Hua, G. 2016. Ordinal Regression With Multiple Output CNN for Age Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pathak, D.; Girshick, R.; Dollár, P.; Darrell, T.; and Hariharan, B. 2017. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*.

Schifanella, R.; Redi, M.; and Aiello, L. 2015. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. *arXiv preprint arXiv:1505.03358*.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning*.

Wang, X.; and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*.

Zhang, Y.; Gong, B.; and Shah, M. 2016. Fast zero-shot image tagging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.