Spacing Loss for Discovering Novel Categories

K J Joseph\(^1,2\)  Sujoy Paul\(^2\)  Gaurav Aggarwal\(^2\)  Soma Biswas\(^3\)  Piyush Rai\(^2,4\)
Kai Han\(^2,5\)  Vineeth N Balasubramanian\(^1\)
\(^1\)Indian Institute of Technology Hyderabad  \(^2\)Google Research  \(^3\)Indian Institute of Science
\(^4\)Indian Institute of Technology Kanpur  \(^5\)The University of Hong Kong

\{cs17m18p100001, vineethnb\}@iith.ac.in, somabiswas@iisc.ac.in,
\{sujoyp, gauravaggarwal, piyushrai, kaihanx\}@google.com

Abstract

Novel Class Discovery (NCD) is a learning paradigm, where a machine learning model is tasked to semantically group instances from unlabeled data, by utilizing labeled instances from a disjoint set of classes. In this work, we first characterize existing NCD approaches into single-stage and two-stage methods based on whether they require access to labeled and unlabeled data together while discovering new classes. Next, we devise a simple yet powerful loss function that enforces separability in the latent space using cues from multi-dimensional scaling, which we refer to as Spacing Loss. Our proposed formulation can either operate as a standalone method or can be plugged into existing methods to enhance them. We validate the efficacy of Spacing Loss with thorough experimental evaluation across multiple settings on CIFAR-10 and CIFAR-100 datasets.

1. Introduction

Availability of large amount of annotated data has fueled unprecedented success of deep learning in various machine learning tasks \([5, 9, 18, 30, 36, 37]\). Though human learners also require various levels of supervision throughout their lifetime, we make use of the bulk of knowledge acquired so far to make intelligent choices, which guides effective learning. Drawing a parallel to the machine learning problem of image classification, it is natural to expect a model trained on a huge number of labeled classes (e.g., 1000 classes in ImageNet dataset \([34]\)) to give meaningful representations to identify and differentiate instances of novel categories. This is the basis for the research efforts in Novel Class Discovery (NCD) setting \([10, 11, 13, 15, 16, 44, 45]\). Given access to labeled training data from a set of classes, an NCD model identifies novel categories from an unlabeled pool containing instances from a disjoint set of classes.

As the nascent field of Novel Class Discovery continues to evolve, we introduce a categorization of existing NCD methods based on the data that is required to train them. Single-stage NCD models can access labeled data and unlabeled data together while discovering novel categories from the latter. Two-stage NCD models can access labeled and unlabeled data only in stages. Each of these settings has a wide practical applicability. Consider a marine biologist who studies about various kinds of organisms in the ocean, from images captured by under-water vehicles \([19, 20]\). While analysing these images for novel categories in their lab, it would be ideal to make use of any annotated data that they might have already collected overtime. Hence, a single-stage NCD methods would be ideal for their setting. Contrastingly, consider an autonomous robot that can assist the visually impaired \([24, 25]\). While being operational, it would be great for the robot to discover and identify instances of novel categories in the environment, so that it can alert its users. In this scenario, it is not practical to reuse all labeled instances that the robot was trained on in its factory, while discovering novel categories. A two-stage NCD method is more desired in this setting.

A common theme in most NCD methodologies is to learn a feature extractor using the labeled data and use clustering \([13, 15, 16]\), psuedo-labelling based learning \([10, 11]\) or contrastive learning \([17, 45]\) to identify classes in the unlabeled pool. In contrast, we propose a novel Spacing Loss which ensures separability in the latent space of feature extractor, for the labeled and unlabeled classes. This is achieved by transporting semantically dissimilar instances to equidistant areas in the latent space, identified via multi-dimensional scaling \([40]\). We note that our proposed loss formulation is orthogonal to the existing methodologies, and can easily complement these methods. Our experimental evaluation on CIFAR-10 \([23]\) and CIFAR-100 \([23]\) datasets suggests that the models trained with the proposed Spacing Loss achieve state-of-the-art performance when compared to two-stage NCD methods. Further, when combined with single-stage methodologies, our loss formulation improves each of them consistently.

The standard strategy to evaluate NCD methods is to train the model on a subset of classes from a classification dataset and evaluate its performance on the remaining classes. Complementing existing protocols, we introduce a new split where the number of classes in the labeled pool is significantly lower than the number of classes in the unlabeled pool. Such a protocol aligns more closely with the
real-world scenarios, where the number of classes in the labeled and unlabeled pool might be heavily imbalanced.

To summarize, the key contributions of our work are:

- We propose Spacing Loss, which enforces separability in the latent space, for the challenging problem of novel category discovery.
- We evaluate our proposed approach on benchmark datasets for novel category discovery, under both single- and two-stage settings, consistently outperforming existing methods.

## 2. Novel Class Discovery Methods

### Two-stage Methods

Early methods in Novel Class Discovery [13, 15, 16] operate in a phased setting. In the first phase, the model learns from the labeled data, and in the subsequent phase, it discovers novel categories from the unlabeled pool. MCL [16] and KCL [15] learn a binary similarity function using meta-learning in the first phase, and use this in the category discovery phase. DTC [13] first learns a feature extractor on the labeled data. In the next stage, these features are used to initialise a clustering algorithm [42], which further fine-tunes these representations using the unlabeled data, thereby improving class discovery.

### Single-stage Methods


## 3. Spacing Loss

Learning to adapt the latent representations of a model, such that semantically identical samples would share nearby locations in the latent manifold, while semantically dissimilar samples are spaced apart, would be ideal for discovering novel classes. Such a subspace shaping should evolve as latent representations mature. Two characteristics would be ideal in such a setting: 1) the ability to transport similar samples to locations equidistant from other dissimilar samples in the latent manifold, 2) the datapoints having the ability to refresh their associativity to a group as the learning progresses. We propose a simple yet effective methodology

---

1As NCD is a nascent field, we will maintain an updated list of methods here: https://github.com/JosephKJ/Awesome-Novel-Class-Discovery.
\[ \delta_{ij}, d_{ij}(P^c) \text{ corresponds to the distance between } p_i^c \text{ and } p_j^c \text{ in euclidean space. We can formulate the objective to learn } P^c \text{ as follows:} \]
\[ \sigma(P^c) = \sum_{i<j} w_{ij}(d_{ij}(P^c) - \delta_{ij})^2, \tag{1} \]
where \( W \) is a symmetric, non-negative and hollow matrix of weights \( w_{ij}, \) which captures the relative importance. For simplicity, we weigh each \( P^c \) equally. As finding an analytical solution to minimize Eq. (1) is intractable, an iterative majorization algorithm [4, 7] is used. We seek to find a manageable surrogate function \( \tau(P^c, Y), \) which majorizes \( \sigma(P^c), \) i.e., \( \tau(P^c, Y) > \sigma(P^c), \) with the initial supporting points \( Y. \) We can rewrite Eq. (1) as follows:
\[ \sigma(P^c) = \sum_{i<j} d_{ij}^2(P^c) + \sum_{i<j} \delta_{ij}^2 - 2 \sum_{i<j} \delta_{ij}d_{ij}(P^c). \tag{2} \]
The first term is a quadratic in \( P^c \) and can be expressed as \( \text{Tr} \ P^c \ T \ B(P^c) \ P^c, \) where \( V \) has \( v_{ij} = -w_{ij} \) and \( v_{ii} = \sum w_{ij} \) [7]. The second term is a constant, say \( k, \) and the third term can be bounded as follows:
\[ \sum_{i<j} \delta_{ij}d_{ij}(P^c) = \text{Tr} \ P^c T B(P^c) P^c \]
\[ \geq \text{Tr} \ P^c T B(Y) Y, \tag{3} \]
where \( B(Y) \) has
\[ b_{ij} = \begin{cases} \delta_{ij}, & \text{for } d_{ij}(Y) \neq 0, i \neq j \\ 0, & \text{for } d_{ij}(Y) = 0, i \neq j \end{cases} \]
\[ b_{ii} = -\sum_{j=1,j\neq i}^c b_{ij}. \tag{4} \]
The proof of this inequality follows [4, 7]. Hence, the surrogate function that majorizes \( \sigma(P^c) \) is as follows:
\[ \tau(P^c, Y) = \text{Tr} \ P^c T V P^c + k - 2 \text{Tr} \ P^c T B(Y) Y. \tag{5} \]

**Algorithm 1 GETEQUIDISTANTPOINTS**

**Input:** Prototype vectors: \( P = \{p_0, \cdots, p_c\}, \) Small constant \( \epsilon. \)

**Output:** Equidistant points: \( P^c. \)

1. \( p_{\text{dist}} \leftarrow \) maximum distance between all prototypes in \( P. \)
2. Compute \( \Delta \) from \( p_{\text{dist}}. \)
3. Initialize \( P^c \) randomly.
4. do
5. \( Y \leftarrow P^c \)
6. \( P^c \leftarrow \arg \min_{P^c} \tau(P^c, Y) \quad \triangleright \text{Defined in Eq. (5)} \)
7. while \( (Y - P^c) > \epsilon \)
8. return \( P^c \)

Algorithm 1 summarizes how \( P^c \) are computed by optimizing Eq. (5). In Line 2, we compute the dissimilarity matrix \( \Delta \) by using the maximum distance between the prototype vectors \( P. \) First \( P^c \) is randomly initialised. Until there is negligible change \( \epsilon \) in \( P^c, \) we update \( P^c \) to optimize the surrogate function \( \tau(P^c, Y). \) The resulting vectors in \( P^c \) are guaranteed to be equidistant from each other [4].

### 3.2. Learning Separable Latent Space

Once the equidistant locations in the latent space \( P^c \) are identified, they can be used to enforce separation in the latent representations of images from different classes. As each latent representation matures with training, it might need to change its associativity with its initial group. We propose a novel formulation in Algorithm 2 that would allow for this flexibility during learning. The training essentially alternates between learning with pseudo-labels derived from class prototypes (Lines 6 - 8) and modifying the class prototypes themselves (Lines 11 - 15). In Line 1, we initialize the class prototypes \( P \) as the centroids of latents from \( \Phi_{\theta} \) using \( k\)-means [29]. Based on the closeness to these prototypes, the class associativity of each image in a mini-batch is determined in Line 7. The feature extractor is updated to make the latent representations closer to these prototypes in Line 8. Using these newer features, the assignment is recomputed and the prototypes themselves are updated in Line 15. For each data-point \( z_i, \) its corresponding prototype \( p_{sz_i} \) is moved closer to the equidistant point \( p_{sz_i}^c \) and its current representation, controlled by a momentum parameter \( \eta. \) The parameter \( \eta \) dampens with more instances of the specific class seen during training.

**Algorithm 2 LEARNINGWITHSPACING**

**Input:** Feature extractor: \( \Phi_{\theta}, \) Data: \( D = \{X_i\}, \) # of epochs: \( e. \)

1. Initialize class prototypes \( P = \{p_0, \cdots, p_c\}. \)
2. Identify equidistant points \( P^c = \{p_0^c, \cdots, p_c^c\} \) using Algo. 1.
3. Initialize assignment frequency \( v \leftarrow 0; |v| = c. \)
4. for each epoch \( e \) do
5. \quad for each minibatch \( C \subset D \) do
6. \quad \quad \( Z \leftarrow \Phi_{\theta}(X) \)
7. \quad \quad \( A \leftarrow \text{assign the nearest prototype from } P \text{ for each } Z. \)
8. \quad \quad Update \( \theta \) with MeanSquaredError(\( Z, A). \)
9. \quad \quad \( Z \leftarrow \Phi_{\theta}(X) \quad \triangleright \text{Recompute } Z \text{ with updated } \theta \)
10. \quad \quad \( A \leftarrow \text{recompute prototype assign. for each new } Z. \)
11. \quad \quad for \( z_i \in Z \) do
12. \quad \quad \quad \( c_{sz_i} \leftarrow \text{retrieve assignment index of } z_i \text{ from } A. \)
13. \quad \quad \quad \( v[c_{sz_i}] \leftarrow v[c_{sz_i}] + 1 \)
14. \quad \quad \quad \( \eta \leftarrow \frac{\eta}{v[c_{sz_i}]} \)
15. \quad \quad \quad \( p_{sz_i} \leftarrow (1 - \eta)p_{sz_i} + \eta(z_i + p_{sz_i}^c) \)

### 3.3. Overall Objective

So far, we have explained how the feature extractor \( \Phi_{\theta} \) is adapted by Spacing Loss. Our complete model extends this backbone with one head for the labeled data \( F_{Lab} = \Phi_{Lab} \circ \Phi_{\theta} \) and another for the unlabeled data \( F_{Ulab} = \Phi_{Ulab} \circ \Phi_{\theta}. \) \( F_{Lab} \) is learned with the labeled examples. \( F_{Ulab} \) is learned with pairwise pseudo labels derived from cosine-similarity [45] between its latent repre-
sentations. We also enforce consistency in prediction with an augmented view of each image \([11,13,44,45]\) to enhance learning. While learning a two-stage model, we first learn \(F_{Lab}\) using cross entropy loss with labeled data and then learn \(F_{U,lab}\) with these auxiliary losses and Spacing Loss operating in the latent space. Labeled and unlabeled data, along with all the losses are used to learn the single-stage model. During inference, we do a \(k\)-means \([29]\) on the latent representations from the backbone network, to discover novel categories.

### 4. Experiments and Results

Following existing NCD methods \([10,11,13,15,16,44,45]\), we define splits on CIFAR-10 and CIFAR-100 to evaluate the efficacy of our method. Clustering Accuracy \([11]\) and NMI \([39]\) are used as the evaluation criteria. We use ResNet-18 \([14]\) backbone and closely follow the hyper-parameter settings in this table. Our proposed loss formulation can act as an add-on to existing methods, effectively enhancing their class discovery capability, even for severely skewed class distributions.

<table>
<thead>
<tr>
<th>Setting →</th>
<th>Imbalanced Class Split</th>
<th>Balanced Class Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Splits →</td>
<td>CIFAR-100-80-20</td>
<td>CIFAR-100-20-80</td>
</tr>
<tr>
<td>Method</td>
<td>CA</td>
<td>NMI</td>
</tr>
<tr>
<td>RS ([11])</td>
<td>69.39</td>
<td>0.6934</td>
</tr>
<tr>
<td>RS + Spacing loss</td>
<td>73.16</td>
<td>0.7252</td>
</tr>
<tr>
<td>NCL ([45])</td>
<td>81.01</td>
<td>0.7883</td>
</tr>
<tr>
<td>NCL + Spacing loss</td>
<td>85.11</td>
<td>0.7896</td>
</tr>
</tbody>
</table>

Table 1. We study the class discovery performance of single-stage NCD models across multiple settings in this table. Our proposed loss formulation can act as an add-on to existing methods, effectively enhancing their class discovery capability, even for severely skewed class distributions. We showcase this capability while evaluating in single-stage setting. In Tab. 1, we organise different dataset splits based on the balance between the number of classes in labeled and unlabeled pool. The concise notation in Row 2 can be expanded as: dataset—total_class_count—labeled_classes—unlabeled_classes. The latent space separation induced by Spacing Loss helps to improve the class discovery capability on all settings. It is interesting to note that the improvement is more pronounced in the more pragmatic setting, where the split of classes between the labeled and unlabeled pool is skewed. t-SNE \([38]\) visualization of backbone features in Fig. 2 shows good separation in these latent representations of novel categories in CIFAR-10-5-5 setting.

### 5. Enhancing Continual Learning with NCD

Continual learning setting aims to learn a single model which can incrementally accumulate knowledge across multiple tasks, without forgetting. Main-stream efforts in Continual Learning \([1,2,6,8,21,26–28,28,31–33,33,35,41]\) assume that the data which is introduced in each incremental task is fully annotated. Efforts in Novel Class Discovery can help to relax this requirement, where the model could be tasked to identify classes from the instances of a new task automatically, based on the learnings that it already had. Then, these identified novel categories may be incrementally learned. We hope that the unification of these two streams of research would lead to a more pragmatic problem setting by building on their complementary characteristics.

### 6. Conclusion

We characterise research efforts in the nascent Novel Class Discovery setting into single-stage and two-stage methods, based on their data requirement during training. We further propose a simple yet effective method which enhances both these settings by enforcing separability in the latent representations. Our experimental analysis on multiple settings on two benchmark datasets corroborates with our assertions. Advancements in NCD can help continual learning models to operate in an open-world \([3,18]\), where it can automatically identify novel categories and then incrementally learn them. We hope this pragmatic setting would be extensively explored in follow-up works.
References


[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 4


