Novel Class Discovery (NCD)

Noemi Canovi (230487), Kevin Depedri (229358), Jacopo Donà (229369), Mostafa Haggag (229674)
University of Trento
Assignment for Trends and Applications of Computer Vision

I. Introduction

Here, we present a short survey of Novel Class Discovery works. Firstly, we introduce the NCD task and similar settings. Then, we focus on some methods that are of great relevance to this topic.

II. Task

A. Novel Class Discovery (NCD)

In recent years, deep models have shown surprising results in various fields, even outperforming humans in some tasks. However, one of their biggest limitations is the amount of annotated data required for the training phase. In some cases, data can be expensive or infeasible to retrieve. This also means these types of models only recognize classes that they saw during training, and are unable to differentiate novel classes. Novel class discovery addresses this problem, with the objective of teaching deep models to recognize new classes in unlabeled data.

Here, we identify two types of data: labeled data (known old classes) and unlabeled data (novel classes). As the main assumption, these two sets are disjoint but have strong visual similarities, for instance, we can have labeled images of cats and dogs but unlabeled images of birds. This allows us to leverage knowledge in the labeled set to better identify the unlabeled set, i.e. learning models and features that can classify both types of classes. An example of the NCD problem with both labeled data and unlabeled data can be seen in fig.1.

B. Related Tasks

Most of the works we will present in this report take ideas from related tasks. For fairness, it is right to introduce them and highlight what common properties they share and what differences they present with respect to NCD.

First of all, *semi-supervised learning* 's goal is to solve a classification problem in which part of the data is labeled and the rest is not. It differs from NCD because it assumes both types of data to share the same categories (e.g., there will be some images of cats that are labeled and some others that are not). In our case, the disjoint sets make the problem more challenging as we do not have label information for the novel samples.

Secondly, in *transfer learning* a model is first trained on a labeled dataset and then fine-tuned on another labeled dataset with a different set of classes. We still want to transfer knowledge from one dataset to another, but in the NCD case, the target set is not annotated.

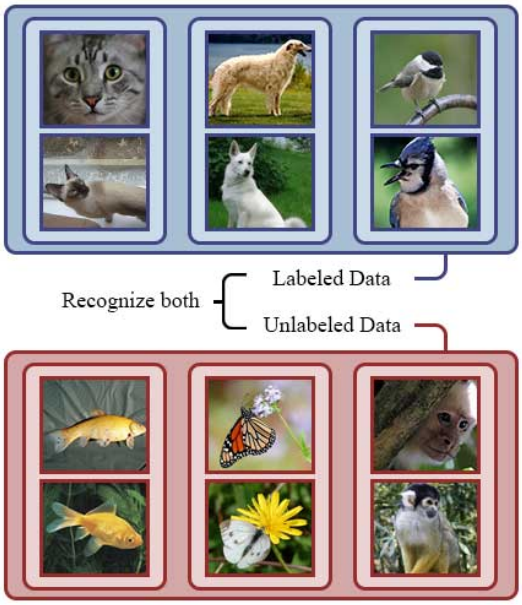

Fig. 1. An illustration of the labeled classes and the unlabeled classes and how we want to recognize both using the NCD method.

Lastly, *unsupervised clustering* aims at partitioning unlabeled samples into different categories. Since no label information is given, this family of methods relies on the notion of similarity between samples to aggregate data in groups called clusters. In NCD, we relax the assumptions relying on the labeled source data to retrieve informative features and patterns.

C. Task Definition

NCD assumes two disjoint datasets: a labeled dataset $D^l = \{(x_i^l, y_i^l), i = 1, ..., N\}$, where $y_i^l \in \{1, ..., C^l\}$ is the class label for image x_i^l and an unlabeled dataset $D^u = \{x_i^u, i = 1, ..., M\}$ of images x_i^u for which we know the number of classes C^u a priori. In addition, the set of C^l labeled classes is disjoint from the set of C^u unlabeled ones. The objective at test time is to have a model that is able to classify images from both labeled and unlabeled classes.

III. Related Works

In recent years, numerous papers have been published to tackle the task of NCD. This popularity is given by the fact that NCD, with other tasks like semi-supervised and unsupervised

learning, aims at broadening the spectrum to which deep learning can be applied.

The majority of NCD methods [1, 5, 6] typically involve a first supervised step in which the network learns from the labeled data, and a second unsupervised step in which the network constructs clusters for the unlabeled data. To an extent, this simple but effective pipeline is able to learn a good representation of the samples while transferring knowledge from the labeled dataset. These methods integrate two separate objectives: classification of the labeled classes through direct supervision and clustering to discover the unlabeled, novel classes. These objectives are often treated with independent losses such as cross-entropy (CE) and binary cross-entropy (BCE). For the latter, pseudo-labels are often used [3, 7]–[11]. Other methods try to construct more cohesive networks, in which the two objectives are partially or totally merged together [3, 4].

In this section, we will see more in detail three of these methods: DTC [1], AutoNovel [3], and UNO [4]. The first two proposals were among the first works created for solving NCD and are often cited and taken as references by newer works, while UNO takes a different and novel pipeline to solve the task, resulting in a more compact model.

A. Deep Transfer Clustering

The first paper presented is DTC [1]. It was one of the first architecture proposals for solving the Novel Class Discovery task and, despite having outdated results compared with more recent works, provided ideas that were later re-proposed and expanded. In particular, they decided to first initialize the network using the labeled data and then fine-tune on the unlabeled portion.

The main intuition of this method is to simultaneously cluster the data and learn good data representation.

They achieve this by adopting Deep Embedding Clustering [2], which is an iterative procedure in which two steps are alternated and repeated: first, the distribution is matched to a suitably shaped target distribution, second the target is constructed as a sharper version of the current distribution.

In more detail, let $p(k|i)$ be the probability of assigning data point $i \in \{1, ..., N\}$ to cluster $k \in \{1, ..., K\}$. DEC uses the following parameterization of this conditional distribution by assuming a Student's t distribution:

$$q(k|i) \propto (1 + \frac{\|z_i - \mu_k\|^2}{\alpha}) \tag{1}$$

In order to anneal to a good solution, instead of maximizing the likelihood of the model p directly, the authors match the model to a suitably-shaped distribution q. This is done by minimizing the KL divergence between joint distributions:

$$E(q) = KL(q\|p) \tag{2}$$

Then, the target distribution is constructed as a sharpened version of the current distribution p, to learn from high-confidence samples:

$$q(k|i) \propto \frac{p(k|i)^2}{\sum_{i=1}^{N} p(k|i)} \tag{3}$$

Hence, the target distribution is calculated by first raising $p(k|i)$ to the second power, which sharpens it, and then normalizing it by the frequency per cluster.

DEC requires an initial setting for the cluster centers. To obtain this initialization, the k-means algorithm is run on the set of features extracted from the unlabeled data. They also found this step performs better by introducing a PCA dimensionality reduction step to the feature representation.

In addition, they also added temporal ensembling, by maintaining an exponential moving average (EMA) of the previous distributions to compute the clustering models p, and consistency. This further enhances the performance of the model.

DTC has also a peculiarity: it encapsulates a method to estimate the number of novel classes. In fact, the majority of NCD models rely on the assumption that the number of novel classes in the dataset is known, requiring the users to possess a certain degree of knowledge of the samples. This is not always true in real applications and can pose difficulties, especially on big datasets.

Their estimation consists of first running k-means multiple times with a different number of clusters and selecting the one that achieves higher performance on the following metrics: average clustering accuracy (ACC), appliable to a validation probe set extracted from the labeled data, and cluster validity index (CVI), for instance, Silhouette index, that addresses the unlabeled data by capturing notions such as intra-cluster cohesion and inter-cluster separation.

Nonetheless, one of the main drawbacks of this method is the forgetting of old classes. In fact, the features obtained after the DEC step generalize well on the newly discovered classes but result in an accuracy drop for the known labeled classes. We can see the pipeline for DTC in fig.2.

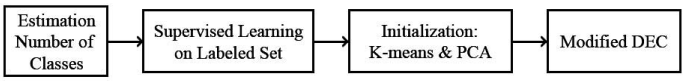

Fig. 2. DTC Pipeline

B. AutoNovel

In AutoNovel [3], a three-stage pipeline is used, in which both labeled and unlabeled data are exploited in different ways to aid the learning. In particular, in the first step, the network learns low-level features with a self-supervision module that is applied to both types of data. Then, only the labeled data are used to learn a classifier and to fine-tune the highest layer of the network, in which self-supervision was found not so effective. Finally, a joint optimization on both labeled and unlabeled data takes place, improving both the supervised classification of the labeled data and the clustering of the unlabeled data.

In comparison to previous models, they integrate three novel ideas. First of all, they use self-supervision, instead of the typical pre-training on labeled data, in order to avoid the representation being biased towards known classes.

Secondly, they transfer information by sharing the same representation between labeled and unlabeled images. In the case of novel data, the unknown classes are defined as a relation among pairs of unlabeled images (x_i^u, x_j^u), with the basic idea that similar images belong to the same class. In more detail, the representation vectors of the pairs are compared using robust rank statistics, i.e. ranking the values in the vectors by magnitude and testing if the subset of top-k ranked dimensions is shared between the images. The resulting similarity score s_{ij} is:

$$s_{ij} = \mathbb{1}\{(top_k(\Phi(x_i^u)) = top_k(\Phi(x_j^u))\} \tag{4}$$

The third idea is to minimize a joint objective function, in which cross entropy (CE) and labels will be used for labeled data, binary cross entropy (BCE) and pseudo-labels (s_{ij}) are exploited for unlabeled data and a regularization term (MSE) is introduced to improve consistency. The complete loss can be written as:

$$L = L_{CE} + L_{BCE} + \omega(t) L_{MSE} \tag{5}$$

where $\omega(t)$ is a ramp-up function.

This joint objective helps avoid the forgetting issue, i.e. prevents the model from the suffering of accuracy drops on known old classes once the Novel Class Discovery step is performed.

In addition, they saw that incremental learning was useful to improve performance, as it leads information to flow between the labeled and unlabeled data.

For all the above, AutoNovel is really good at discriminating both old and new classes and, compared with DTC, achieves better results on various datasets. However, the three-stage pipeline is complex and costly. We can see the pipeline for AutoNovel in fig.3.

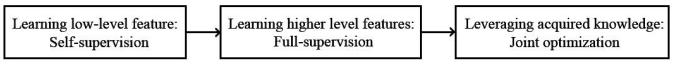

Fig. 3. AutoNovel Pipeline

C. Unified Objective for Novel Class Discovery

UNO (UNified Objective for NCD) [4] goal is to simplify NCD task by taking a more direct approach. The authors proposed to eliminate the supervised or self-supervised pre-training step, common in previous models, and to unify all the objectives through a single loss function.

To do so, they use a multi-view strategy in which the generated pseudo-labels and the ground truth labels are treated homogeneously, allowing a single loss function to drive the training. This allows better cooperation between supervised and unsupervised learning.

In particular, they first generated two different views v_1 and v_2 for each image in the batch, using random transformations such as cropping and color jittering. Then, the two views pass through a common encoder (a CNN), in which feature vectors representing the images are extracted. Attached to it there are two different classifier heads: the first addresses labeled classes and have C_l output neurons while the other addresses unlabeled classes and have C_u output neurons. The output logits generated by both heads are concatenated and fed into a shared softmax layer to produce a posterior distribution over all labels.

Here, we have two variants: if it is known which image is from labeled and unlabeled classes, i.e. we are in a task-aware setting, we consider only the corresponding classifier, zeroing the other logit. For instance, in the case of labeled images, all the predictions for unlabeled classes are set to zero, because the classes are disjoint. If we do not have this information, i.e. we are in a task-agnostic setting, the prediction results in the most likely output after the concatenation.

Then, the cross-entropy loss is used to train the network. In particular, if the image belongs to a labeled class, we simply associate the two views v_1 and v_2 with the same ground-truth label. Instead, if the image is unlabeled the two views are used to compute two corresponding pseudo-labels that are swapped to encourage the network to output consistent predictions between different views. The following equation, the swapped prediction task, is applied:

$$l(v_1, y_2) + (v_2, y_1) \tag{6}$$

where l is the cross-entropy loss. This work is superior in performance compared to previous methods and is easier to implement and train. We can see a simplified schema for UNO in fig.4.

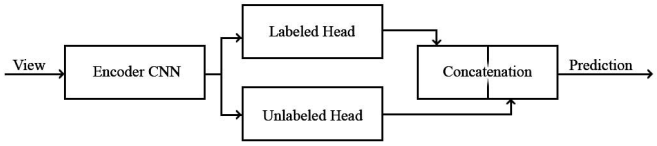

Fig. 4. Simplified UNO Schema

IV. Conclusion

Here, we have seen why novel class discovery is useful and which methods the community developed to address this task. Some rely on semi-supervised or unsupervised techniques, and others choose more novel approaches, but overall the objective is to transfer knowledge of 'what makes a good class' from known labeled samples to novel unlabeled classes. In this report, we focused on works regarding Computer Vision, but we believe this task could also be applied to other fields.

To conclude, we think that Novel Class Discovery will become increasingly popular in the future, as annotating large datasets will become more and more difficult. Also, we believe this task could bring contributions in settings where samples themselves may be difficult to annotate without extensive domain knowledge, such as fine-grained classification [12].

REFERENCES

[1] K. Han, A. Vedaldi, A. Zisserman - "Learning to Discover Novel Visual Categories via Deep Transfer Clustering", ICCV, 2019

[2] J. Xie, R. Girshick, A. Farhadi - "Unsupervised Deep Embedding for Clustering Analysis", ICML, 2016

[3] K. Han, S. Rebuffi, S. Ehrhardt, A. Vedaldi, A. Zisserman - "Automatically Discovering and Learning New Visual Categories with Ranking Statistics", ICLR, 2020

[4] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, E. Ricci - "A Unified Objective for Novel Class Discovery", ICCV, 2021

[5] Y.-C. Hsu, Z. Lv, Z. Kira. "Learning to cluster in order to transfer across domains and tasks", in Proc. ICLR, 2018

[6] Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, Z. Kira. "Multi-class classification without multi-class labels" In Proc. ICLR, 2019

[7] X. Jia, K. Han, Y. Zhu, B. Green. "Joint representation learning and novel category discovery on single- and multi-modal data", in Proc. ICCV, 2021.

[8] A. Iscen, G. Tolias, Y. Avrithis, O. Chum, "Label propagation for deep semi-supervised learning", in Proc. CVPR, 2019.

[9] B. Zhao, K. Han, "Novel visual category discovery with dual ranking statistics and mutual knowledge distillation" NeurIPS, 2021

[10] Z. Zhong, E. Fini, S. Roy, Z. Luo, E. Ricci, N. Sebe, "Neighborhood contrastive learning for novel class discovery", in Proc. CVPR, 2021

[11] Z. Zhong, L. Zhu, Z. Luo, S. Li, Y. Yang, N. Sebe, "Openmix: reviving known knowledge for discovering novel visual categories in an open world", in Proc. CVPR, 2021

[12] Yixin Fei, Zhongkai Zhao, Siwei Yang, Bingchen Zhao, "XCon: Learning with Experts for Fine-grained Category Discovery" BMVC, 2022