NOVEL CLASS DISCOVERY: A DEPENDENCY APPROACH

Tanmoy Mukherjee[1,2], *and Nikos Deligiannis*[1,2]

[1]ETRO Department, Vrije Universiteit Brussel (VUB), Pleinlaan 2, B-1050 Brussels, Belgium
[2]imec, Kapeldreef 75, B-3001 Leuven, Belgium

ABSTRACT

Supervised and semi-supervised algorithms have been designed under a closed-world setting, with the assumption that unlabeled data consists of classes previously seen in labeled training data. However, real world is inherently open set where this assumption is often violated, and thus novel data may be encountered in test data. In this paper, we look at the problem where the model is required to discover novel classes never encountered in the labeled set. We propose a dependency measure based on Squared Mutual Information (SMI) where we simultaneously learn to classify and cluster the data. Our experiments show that our approach is able to achieve competitive performance on CIFAR and Imagenet datasets.

Index Terms— Novel class discovery, dependence measure, open set recognition.

1. INTRODUCTION

Recent advances in deep learning has resulted in significant improvement in various visual understanding tasks. Due to the availability of large datasets like Imagenet [1], the problem of visual recognition seems to be solved. Despite this impressive performance, visual recognition systems fail to account for the dynamic nature of classes or *novel classes* as it is practically impossible to collect training labels for all classes.

Semi-supervised learning (SSL) [2] aims at leveraging unlabeled data when labels are unavailable. However SSL also implicitly assumes that the unlabeled data and the labeled data share the same class labels making it unrealistic to be deployed in real world applications. To alleviate these issues *Novel Class Discovery* (NCD) [3, 4] has been recently proposed as a solution. NCD aims at training a network that can simultaneously classify a set of labeled classes while discovering new ones in an unlabeled image set. The motivation of this stems from the level of supervision which is easily available from supervised data that can be transferred to discover unknown classes. At training time, data is split into a set of labeled images and a set of unlabeled images. These two sets are *jointly* used to train a single network to classify both the known and unknown classes. This problem bears a strong resemblance towards *semi-supervised learning* (SSL) but differs as SSL assumes that both the *labeled* and *unlabeled* data share the same number of classes. NCD aims to transfer knowledge from labeled data of known classes to improve recognizing new and novel classes from unlabeled data. Identifying which knowledge to transfer from old classes while discovering *novel* classes remains a challenge.

We propose to address the challenge posed by NCD by combining a deep network trained with a novel *dependency maximization* term. Our contributions can be summarized as follows: (*i*) we use Squared Mutual Information (SMI) [5] as a criterion to bridge supervised and unsupervised learning objectives. We aim to *distill* knowledge from old classes for improved new class discovery. (*ii*) We demonstrate through experiments on publicly available datasets, outperforming competitors by a margin.

In the remainder of the paper, Section 2 reviews related work and Section 3 introduces our dependence measure. Section 4 describes the training method .Section 6 presents our experiments. Finally, Section 7 draws the conclusion.

2. RELATED WORK

In this section, we briefly review the representative works in these areas.

Semi-supervised learning (SSL) [2] and more recently deep semi-supervised learning [6] aim to utilize unlabeled examples which are easily available when labeled examples are often hard or expensive to obtain. The objective of SSL algorithms is to leverage both the labeled data and unlabeled data to make robust models which are prone to overfitting only on the labeled data. SSL assumes that that labeled data and unlabeled share the same class labels, an assumption that is often violated. In this work, we aim to leverage knowledge from old classes to discover new classes from unlabeled data.

Transfer Learning [7] is an effective way to reduce the amount of data annotations by pretraining on a labeled dataset like Imagenet [1] and then transferring the knowledge to target dataset. Traditional transfer learning settings expect the source and target data to be labelled which is in contrast to the current setting

Novel Class Discovery (NCD) is a recently tackled problem related to transfer learning and clustering methods. In contrast to clustering, NCD assumes prior knowledge is pro-

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

vided given a labeled dataset. The task is to cluster unlabeled dataset consisting of similar, but completely disjoint classes than those present in the labeled dataset. The motivation is to leverage knowledge from labeled dataset to improve discovery of novel classes. Some of the recent works in this direction are [3, 4] use a self supervised pretraining step and also maintain different classifiers to generate class/cluster assignment.

3. DEPENDENCE MEASURE

In this paper, we explore the squared-loss mutual information (SMI) [8, 9] as a statistical dependence measure. The SMI between two random variables is defined as [8]

$$\mathrm{SMI} = \iint \left(\frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} - 1 \right)^2 p(\mathbf{x})p(\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y},$$

which is the Pearson divergence [10] from $p(\mathbf{x}, \mathbf{y})$ to $p(\mathbf{x})p(\mathbf{y})$. The SMI is an f-divergence [11]; it is a non-negative measure and is zero only if the random variables are independent.

To estimate the SMI, a direct density ratio estimation approach is useful. The key idea is to approximate the true density ratio by the model:

$$r(\mathbf{x}, \mathbf{y}; \boldsymbol{\alpha}) = \sum_{\ell=1}^{n} \boldsymbol{\alpha}_\ell K(\mathbf{x}_\ell, \mathbf{x}) L(\mathbf{y}_\ell, \mathbf{y}),$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]^\top \in \mathbb{R}^n$ is the model parameter. Then, the model parameter is given by minimizing the error between true density-ratio and its model:

$$J(\boldsymbol{\alpha}) = \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - r(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\alpha}) \right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

By approximating the loss function by samples, we have the following optimization problem [8]

$$\min_{\boldsymbol{\alpha}} \;\; \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}} + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2,$$

where $\widehat{\boldsymbol{H}} = \frac{1}{n^2}(\boldsymbol{K}\boldsymbol{K}^\top) \circ (\boldsymbol{L}\boldsymbol{L}^\top)$, $\widehat{\boldsymbol{h}} = \frac{1}{n}(\boldsymbol{K} \circ \boldsymbol{L})\mathbf{1}_n$, $\lambda \geq 0$ is a regularization parameter and $\circ$ is the element-wise product.

The optimal solution of the above optimization problem can be analytically obtained as

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_n \right)^{-1} \widehat{\boldsymbol{h}}.$$

Then, the estimator of SMI can be given as [9, 12]:

$$\widehat{\mathrm{SMI}}(\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n) = \frac{1}{2n}\mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}\right)\boldsymbol{K}\boldsymbol{L}\right) - \frac{1}{2}, \qquad (1)$$

where $\mathrm{diag}\left(\boldsymbol{\alpha}\right) \in \mathbb{R}^{n \times n}$ is the diagonal matrix whose diagonal elements are $\boldsymbol{\alpha}$.

4. DEPENDENCE MAXIMIZATION VIEW OF NOVEL CLASS DISCOVERY

4.1. Problem Formulation

In this section, we formally introduce the problem of novel class discovery (NCD). We are provided with a labeled dataset $D^l = \{(\mathbf{x}_1^l, y_1^l), \cdots (\mathbf{x}_m^l, y_m^l)\}$ and an unlabeled dataset $D^u = \{\mathbf{x}_1^u, \cdots \mathbf{x}_n^u\}$ where x_i^l and x_j^u are images and $y_i^l \in \{1, \cdots C^l\}$ are categorical labels. NCD assumes $C^l \cap C^u = \phi$. The goal is to use D^u to find C^u clusters where C^l and C^u are disjoint. Our aim during testing is to classify the images from both the labeled and unlabeled classes.

4.2. Recognizing Seen and Discovering Novel Classes

We propose an approach based on Sec 3 for solving the novel class discovery problem. Given labeled samples $X^l = \{x_i \in \mathbb{R}^d\}_i^m$, $Y^l \in \mathbb{R}^{m \times m_l}$ and unlabeled samples $X^u = \{x_i \in \mathbb{R}^d\}_i^n$. We also maintain a deep neural network (DNN) which can be viewed as a *non-linear* feature extractor $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d_0}$, parameterized by θ.

In the Novel Class Discovery Using Dependency Maximization (NoDM), we propose an objective function that jointly solves (a) a supervised classification task and (b)an unsupervised clustering task. The key challenge is to *transfer* the bias from seen classes towards discovering novel classes. Formally our objective function can be written as

$$\mathcal{L} = \mathcal{L}_{unp} + \lambda \mathcal{L}_{sup} \qquad (2)$$

4.3. Supervised Objective

In the supervised learning objective, we use the labeled dataset $D^l = \{(\mathbf{x}_i^l, y_i^l\}_{i=1}^n$ and utilize SMI to *maximize* dependency between $f_\theta(\mathbf{x}_i^l)$ and y_i^l. We obtain the following optimization problem

$$\max_\theta \mathrm{D}(f_\theta(X^l), Y^l) \qquad (3)$$

where D is the SMI estimator described in Eq 1

4.4. Unsupervised Objective

Given the labeled and unlabeled data, we concatenate them to form $X \in \mathbb{R}^{n+m \times d}$. We also denote a latent embedding of X denoted by $U \in \mathbb{R}^{(n+m) \times c}$ where $c = m_l + m_u$ is the defined dimensionality of latent embedding. We learn the feature extractor and the latent embedding matrix by optimizing the following function

$$\max_\theta \mathrm{D}(f_\theta(X), U) \qquad (4)$$

5. PROPOSED APPROACH

Now that we have defined the supervised and unsupervised objectives, we elaborate on the training mechanism.

5.1. Pre-training stage

Firstly, given a labeled dataset X_l, pre-training involves training a feature extractor f_θ using stochastic gradient descent (SGD) over the cross-entropy loss.

5.2. Novel Class Discovery using SMI

Then, at the conclusion of pre-training stage, we have learned the feature extractor f_θ. We now utilize the learnt embedding to discover new classes. As explained in (2), we combine the supervised and unsupervised objective functions.

$$\begin{aligned} &\max_{\theta, U} \mathrm{D}(\theta, U) = \mathrm{D}(f_\theta(X), U) + \lambda \mathrm{D}(f_\theta(X^l), Y^l) \\ &\text{s.t} \quad U^T U = I \end{aligned} \tag{5}$$

To find new classes, we use $f_\theta(X^u)$ and perform K$-$means to obtain cluster assignments $\{y_j\}_{j=1}^{m} \in \{1, \cdots m_u\}^m$ where m_u constitutes our new classes.

5.3. Optimization

We give details of our optimization procedure.
Initialization We initialize θ in the supervised training phase. We initialize U by conducting spectral clustering on $f_\theta(X)$ [13]
Updating θ We adopt an alternating optimization strategy where we fix U and update θ using SGD. Since we need to update an unsupervised and supervised objective, we sample mini-batches for the unsupervised and supervised part respectively to avoid randomly sampling among all samples. Our update equations are

$$\theta = \theta + \psi \nabla \mathrm{D}(f_\theta(X_b), U_b) \tag{6}$$

and

$$\theta = \theta + \psi \nabla \mathrm{D}(f_\theta(X_b), Y_b) \tag{7}$$

where ψ is the learning rate, U_b is the corresponding latent clustering vector, Y_b is the label matrix for the labeled part.
Updating U With θ is fixed, we solve (5). The optimal solution for U is given by the top c eigenvectors of the following matrix $DL^{-\frac{1}{2}}K_{f_\theta(X)}L^{\frac{1}{2}}D$, where L is the degree matrix and D is the SMI term given by (1).

6. EXPERIMENTAL RESULTS

We evaluate the performance of our method on three established datasets for NCD, i.e., CIFAR10, CIFAR100 [14] and Imagenet [1]. Each of these datasets are controlled by a *labeled* set or known classes and an *unlabeled* set of unknown classes for which we have no level of supervision except for the number of classes. Table 1 reports the details of the split.

	Labeled set		Unlabeled set	
Dataset	Images	Class	Images	Class
CIFAR10	25K	5	25K	5
CIFAR100	40K	80	10K	20
Imagenet	1.25M	882	30K	30

Table 1. Statistics of the datasets and splits used for the novel class discovery task

	CIFAR-10		CIFAR-100		Imagenet	
Method	ACC	NMI	ACC	NMI	ACC	NMI
DEC [17]	0.62	0.67	0.21	0.17	0.21	0.12
DAC [18]	0.68	0.72	0.25	0.20	0.23	0.15
UCD-Knet [16]	0.72	0.74	0.26	0.21	0.24	0.19
CD-Knet [16]	0.73	0.75	0.27	0.23	0.24	0.20
SMI (ours)	**0.82**	**0.78**	**0.28**	**0.29**	**0.27**	**0.26**

Table 2. Results on CIFAR-10 and CIFAR-100 and Imagenet

We use a convolution neural network (VGG-16) [15] as our backbone model. We determine all the hyperaprameters of the model by holding 10% of the training data. We use the SGD optimizer for all three datasets with momentum of 0.9 on weight decay of $5e-4$ on CIFAR and $1e-4$ on Imagenet. We run our models for 200 epochs in pretraining stage with the labeled data and 200 epochs during the discovery stage with both the labeled and unlabeled datasets. We randomly sample training samples from both the labeled and unlabeled data where the batch size is set to 128 for CIFAR and 512 for Imagenet. Our methods are implemented using Pytorch.
Competing Methods To study the effectiveness of our SMI model we compare it with the kernel discovery network [16]. Since their models use a expansion network, we didn't use it to compare. We also compare to two state of the art clustering algorithms namely Deep embedding clustering (DEC) [17] and Deep Adaptive Clustering (DAC) [18]. To measure the

The results reported in Table 2 demonstrate that our SMI based novel class discovery outperforms all baselines on all three datasets.

7. CONCLUSION

In this paper, we introduced a dependency maximization perspective of novel class discovery using SMI. Our method is simple, effective and presents a unified objective between a supervised and unsupervised objective. We also do not use any self-supervision making class discovery a simple approach. We also show promising results on benchmark datasets. Although, we have primarily worked in vision problems, this is an important step where in future work, we will be able to work on other modalities.

8. REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[2] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2006.

[3] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman, "Autonovel: Automatically discovering and learning novel visual categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[4] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman, "Automatically discovering and learning new visual categories with ranking statistics," in *International Conference on Learning Representations (ICLR)*, 2020.

[5] Masashi Sugiyama and Makoto Yamada, "On kernel parameter selection in hilbert-schmidt independence criterion," *IEICE Trans. Inf. Syst.*, pp. 2564–2567, 2012.

[6] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Advances in Neural Information Processing Systems*. 2018, vol. 31, Curran Associates, Inc.

[7] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[8] Taiji Suzuki and Masashi Sugiyama, "Sufficient dimension reduction via squared-loss mutual information estimation," in *AISTATS*, 2010.

[9] Makoto Yamada, Leonid Sigal, Michalis Raptis, Machiko Toyoda, Yi Chang, and Masashi Sugiyama, "Cross-domain matching with squared-loss mutual information," *IEEE TPAMI*, vol. 37, no. 9, pp. 1764–1776, 2015.

[10] Karl Pearson F.R.S., "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[11] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[12] Makoto Yamada and Masashi Sugiyama, "Cross-domain object matching with model selection," in *AISTATS*, 2011.

[13] Donglin Niu, Jennifer Dy, and Michael I. Jordan, "Dimensionality reduction for spectral clustering," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Geoffrey Gordon, David Dunson, and Miroslav Dudík, Eds., Fort Lauderdale, FL, USA, 11–13 Apr 2011, vol. 15 of *Proceedings of Machine Learning Research*, pp. 552–560, PMLR.

[14] Alex Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[15] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[16] Zifeng Wang, Batool Salehi, Andrey Gritsenko, Kaushik Chowdhury, Stratis Ioannidis, and Jennifer Dy, "Open-world class discovery with kernel networks," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 631–640.

[17] Junyuan Xie, Ross Girshick, and Ali Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of The 33rd International Conference on Machine Learning*, Maria Florina Balcan and Kilian Q. Weinberger, Eds., New York, New York, USA, 20–22 Jun 2016, vol. 48 of *Proceedings of Machine Learning Research*, pp. 478–487, PMLR.

[18] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "Deep adaptive image clustering," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5880–5888.