ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

ProtoNCD: Prototypical Parts for Interpretable Novel Class Discovery

Tomasz Michalski[1,2], Dawid Rymarczyk[2],
Daniel Barczyk[2], Bartosz Zieliński[2,3] *

1- Jagiellonian University, Doctoral School of Exact and Natural Sciences, Poland

2- Jagiellonian University, Faculty of Mathematics and Computer Science, Poland

3- IDEAS NCBR, Poland

Abstract. In this work, we introduce ProtoNCD, a novel approach to novel class discovery (NCD) that leverages prototypical parts for enhanced interpretability. ProtoNCD extends the ProtoPool methodology to the NCD setting, employing techniques such as knowledge distillation and specialized prototypical parts initialization. Through comprehensive experiments on the CUB-200-2011 dataset, we demonstrate the efficacy of ProtoNCD and its pivotal role in explaining how the reasoning of known classes influences predictions for those newly discovered.

1 Introduction

Novel Class Discovery (NCD) addresses the challenge of requiring large annotated datasets for deep learning models. It aims to train a network capable of classifying known classes and leveraging this supervision to identify and categorize new classes within unlabeled data [1]. Existing NCD methods [2, 3] are effective, but their lack of interpretability can undermine trust, especially in high-stake applications such as criminal justice and medicine [4]. Transparency can be achieved through post-hoc methods [5], which explain decisions after deployment, or ante-hoc methods, which embed interpretability directly into the model design [6, 7]. However, the latter is preferable because they are more precise [4], and among them, those based on prototypical parts [7, 8, 9, 10, 11] due to the ease of interpreting their predictions.

Nevertheless, although interpretable methods have been extensively developed, they have not yet been applied to enhance the interpretability of NCD setups. In this paper, we fill this gap by introducing ProtoNCD, an interpretable architecture for novel class discovery. Interpretability is achieved by applying prototypical parts to both labeled and unlabeled classes. First, the model learns prototypical parts for labeled classes through standard supervised training. Then, in the discovery stage, the model can use previously trained prototypical parts or learn new ones for the unlabeled classes. As a result, we can

*This work was funded by the National Science Centre (Poland) grant no. 2022/47/B/ST6/03397. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016555. Some experiments were performed on servers purchased with funds from a grant from the Priority Research Area (Artificial Intelligence Computing Center Core Facility) under the Strategic Programme Excellence Initiative at Jagiellonian University.

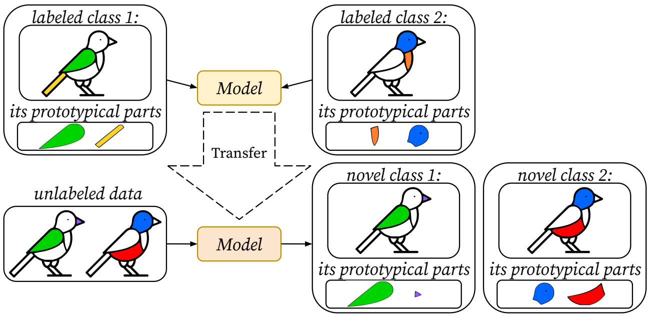

Fig. 1: In ProtoNCD, predictions for novel classes rely on prototypical parts generated for labeled and novel classes. In this example, *novel class 1* and *novel class 2* use two prototypical parts each. Two of them are reused from prototypical parts of labeled classes (green wing and blue head), and two others are prototypical parts created in the discovery phase (purple beak and red belly). This way, we can discover similarities between known and new classes and, in consequence, detect biases in the model or discover new knowledge.

interpret the decisions of the model but also discover similarities and differences between known and novel classes (see Fig. 1).

Our contributions can be summarized as follows:

- We introduce the ProtoNCD, the first approach to interpretable novel category discovery.
- ProtoNCD explains how the reasoning of known classes influences predictions for those newly discovered.
- We conduct extensive experiments on CUB-200-2011 to demonstrate the effectiveness of our method.

2 Method

Preliminaries. In the Novel Class Discovery (NCD) task, the training dataset is divided into two parts: a labeled set $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{|\mathcal{D}^l|}$ containing images with known labels, and an unlabeled set $\mathcal{D}^u = \{x_j^u\}_{j=1}^{|\mathcal{D}^u|}$ used to identify C^u novel classes. The labels y_i^l belong to a set $\mathcal{Y}^l = \{1, \ldots, C^l\}$, and it is assumed that the labeled and unlabeled classes are disjoint. The objective is to learn a mapping that classifies images into a complete set of labels $\mathcal{Y} = \{1, \ldots, C^l, C^l + 1, \ldots, C^l + C^u\}$, which includes both known and novel classes.

ProtoNCD. We utilize a standard framework for NCD that includes an encoder, represented as f, and two classifier heads: c_l for existing classes and c_u for novel

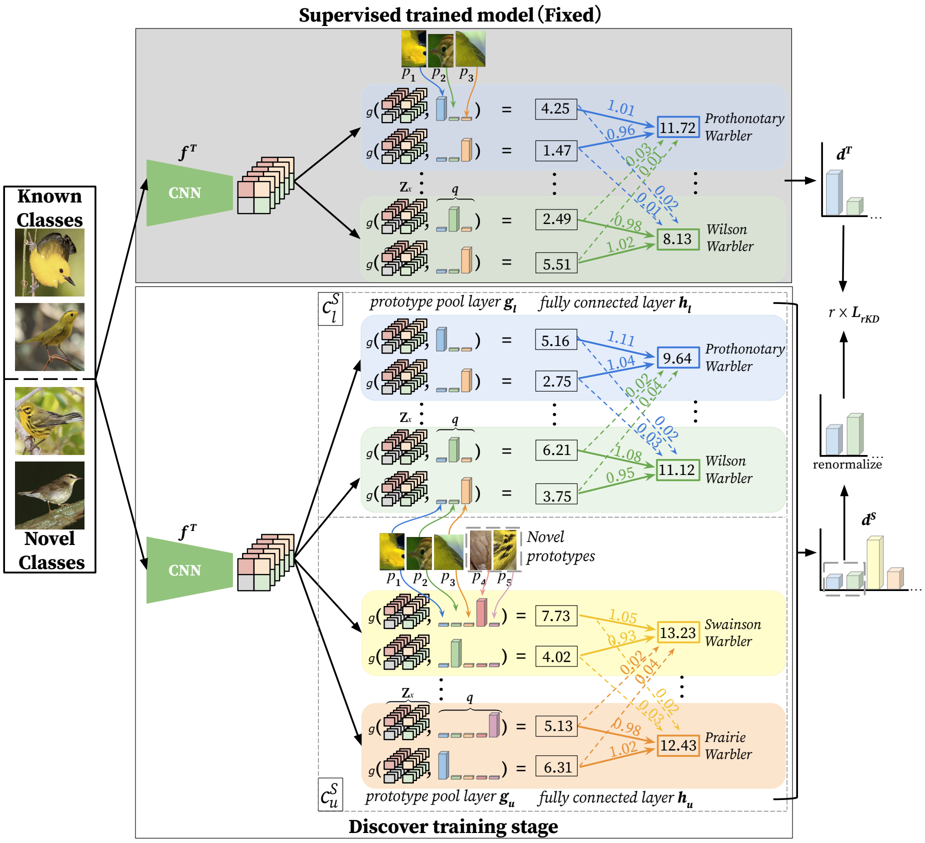

Fig. 2: Overview of the ProtoNCD with the upper part (grey) representing a fixed pretrained model used for distillation, and the lower part corresponding to two trainable heads with known (c_l^S) and novel (c_u^S) classes.

classes (Fig.2). Each classifier head consists of a prototypical parts pool layer g, introduced in [10], and a fully connected layer h. The representation from the last convolution layer of the encoder is compared with prototypical parts designated for specific classes. This comparison is used to calculate the similarities that are then passed to a fully connected layer to generate logits. The logits from both classifier heads are combined to form the final prediction distribution

$$d^S(y \mid x) = \mathrm{Softmax}\left(\left(h_l^S \circ g_l^S \circ f^S(x) \oplus h_u^S \circ g_u^S \circ f^S(x)\right)/\tau\right) \in \mathbb{R}^{C^l + C^u},$$

where the superscript S indicates the model during its discovery training phase and τ refers to the temperature.

The training process consists of two stages. In the first stage, supervised training is performed on known classes, optimized with standard cross-entropy, and three other losses: clustering, separation, and orthogonality that encourage proper prototype creation and assignment [8]. In the second phase, the novel classes are assigned to pseudo labels by solving an optimal transport problem [2]. The knowledge distillation loss (L_{rKD}) is also used to distill the knowledge from the model prepared in the previous stage with additional adaptive regularization term (r) that stabilizes noisy predictions of the novel class data [3].

3 Experiments and results

Experimental setup. The method is evaluated on the fine-grained dataset CUB-200-2011 [12]. The dataset's classes are equally split into known and novel, according to [13]. The task-agnostic protocol is used [2], where the prediction is indicated by the highest value after concatenating logits from labeled and unlabeled heads. Predicted labels are assigned with the Hungarian method [14]. The ResNet50 [15] pretrained on iNaturalist [16] is used as the backbone, and the supervised training is performed according to [10] with the following exceptions. Since the model is trained on 100 classes, only 101 prototypes are used. Also, as suggested by [17], the last layer remains unaltered after the prototype projection phase. In the discovery stage, 101 additional prototypes are introduced to account for the characteristics present in novel classes. These prototypes, their probability distributions, and the fully connected layer in the unlabeled head are the only newly initiated objects in the discovery setting (see Fig.2). All the others are initiated by the corresponding values from the model prepared in the initial stage. Furthermore, the model parameters used for distillation are fixed during training. The code is available online[1].

Table 1: Comparison with state-of-the-art methods on the CUB dataset for novel class discovery in task-agnostic evaluation protocol. ProtoNCD is the first interpretable NCD method with performance at the level of the leading non-interpretable approaches.

Method	Interpretable	Known	Novel	All
NCL[18]	✗	79.8	13.1	46.3
RankStats+[1]	✗	80.7	51.8	66.1
UNO[2]	✗	78.7	62.1	70.3
CRKD[3]	✗	80.5	66.1	73.3
ProtoNCD (ours)	✓	74.2	44.7	59.4

ProtoNCD performance. In Table 1, we evaluate ProtoNCD for novel class discovery against leading non-interpretable NCD methods. ProtoNCD achieves an accuracy of 74.2% for known classes and 44.7% for novel classes, resulting in an overall accuracy of 59.4%. While ProtoNCD surpasses method NCL [18] in novel and overall accuracy, it achieves worse results than the remaining methods, such as CRKD [3]. The emphasis on interpretability in model design causes

[1] https://github.com/michalskit/ProtoNCD

Table 2: Ablation study with the first row corresponding to ProtoNCD. Note that "Random" and "Copy" refer to initializing novel prototypical parts randomly or as a copy of known prototypical parts, "GN" corresponds to Gaussian noise, and "ICICLE" to similarity regularization [17].

Distillation	Frozen Pretrained Prototypes	Extra Prototypes Initialization	Accuracy		
			Known	Novel	All
KD	✓	Copy+GN	74.2	44.7	59.4
KD	✓	Random	71.5	40.8	56.1
KD	✗	Copy+GN	73.3	37.7	55.4
KD	✓	Copy	73.7	36.4	54.9
✗	✓	Copy+GN	69.2	36.1	52.6
KD+ICICLE	✗	Copy+GN	75.6	34.7	55.0

this performance trade-off, particularly in novel class identification. By prioritizing understandability and transparency, our model sacrifices some performance compared to purely performance-focused, non-interpretable methods.

In our ablation study, presented in Table 2, we delve into the rationale behind our design decisions. ProtoNCD performs best across all accuracy metrics (the first row). Conversely, random initialization of additional prototypical parts, as in [10], yields notably inferior results. Similar declines are observed when pretrained prototypical parts are not frozen or when distillation techniques are omitted. Furthermore, we observe that using interpretability distillation (KD+ICICLE) from [17] is detrimental.

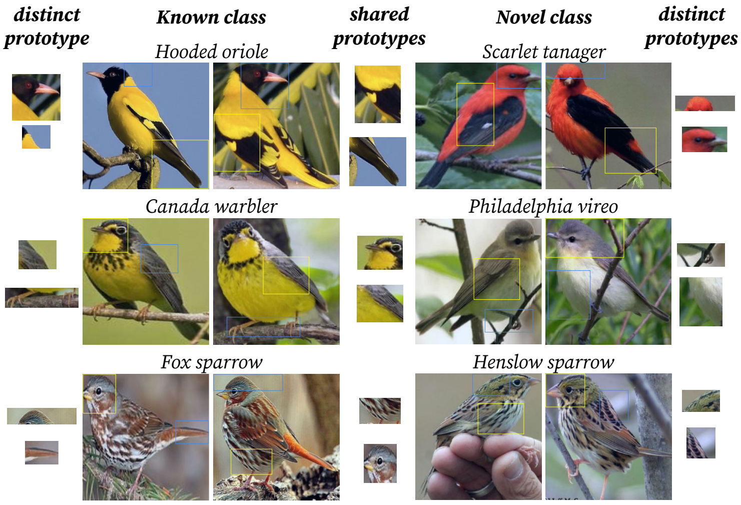

Fig. 3: Sample known and novel classess with shared and distinct prototypical parts.

Interpretability results. ProtoNCD identifies which prototypical parts are shared between the known and novel classes. Therefore, it is possible to identify similar characteristics between them. As illustrated in Fig.3, the novel class, the *Scarlet tanager*, is characterized by shared prototypical parts with the known class, the *Hooded oriole*, including the black wings and forked tail. Simultaneously, distinct prototypical parts capture class-specific features, such as the striking gradient transition from the bright orange of its back to the deep black of its upper head and the rich redhead with a sharp, conical bill of the *Scarlet tanager*. Such observations can be crucial to detect biases in the model but also can be used for knowledge discovery.

4 Conclusions

In conclusion, our work introduces ProtoNCD, an interpretable approach to NCD that effectively identifies new classes within an unlabeled subset of the CUB-200-2011 dataset. Despite a slight accuracy trade-off compared to non-interpretable counterparts, the interpretability of its classification process justifies its performance and paves the way for future advancements in interpretable NCD methods.

References

[1] Kai Han et al. Learning to discover novel visual categories via deep transfer clustering, 2019.

[2] Enrico Fini et al. A unified objective for novel class discovery, 2021.

[3] Peiyan Gu et al. Class-relation knowledge distillation for novel class discovery, 2023.

[4] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. 2019.

[5] Been Kim et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors, 2018.

[6] Moritz Böhle et al. B-cos networks: Alignment is all we need for interpretability, 2022.

[7] Chaofan Chen et al. This looks like that: deep learning for interpretable image recognition. 2019.

[8] Jiaqi Wang et al. Interpretable image recognition by constructing transparent embedding space. 2021.

[9] Dawid Rymarczyk et al. Protopshare: Prototype sharing for interpretable image classification and similarity discovery. 2020.

[10] Dawid Rymarczyk et al. Interpretable image classification with differentiable prototypes assignment, 2022.

[11] Meike Nauta et al. Pip-net: Patch-based intuitive prototypes for interpretable image classification. 2023.

[12] Catherine Wah et al. The caltech-ucsd birds-200-2011 dataset. 2011.

[13] Sagar Vaze et al. Open-set recognition: a good closed-set classifier is all you need?, 2022.

[14] Harold W Kuhn. The hungarian method for the assignment problem. 1955.

[15] Kaiming He et al. Deep residual learning for image recognition, 2015.

[16] Grant Van Horn et al. The inaturalist species classification and detection dataset, 2018.

[17] Dawid Rymarczyk et al. Icicle: Interpretable class incremental continual learning, 2023.

[18] Zhun Zhong et al. Neighborhood contrastive learning for novel class discovery, 2021.