From coarse to fine-grained open-set recognition

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
From Coarse to Fine-Grained Open-Set Recognition

Nico Lang\textsuperscript{1} Vésteinn Snæbjarnarson\textsuperscript{1} Elijah Cole\textsuperscript{2}  
Oisin Mac Aodha\textsuperscript{3} Christian Igel\textsuperscript{1} Serge Belongie\textsuperscript{1}  
\textsuperscript{1}University of Copenhagen \textsuperscript{2}Altos Labs \textsuperscript{3}University of Edinburgh

Abstract

Open-set recognition (OSR) methods aim to identify whether or not a test example belongs to a category observed during training. Depending on how visually similar a test example is to the training categories, the OSR task can be easy or extremely challenging. However, the vast majority of previous work has studied OSR in the presence of large, coarse-grained semantic shifts. In contrast, many real-world problems are inherently fine-grained, which means that test examples may be highly visually similar to the training categories. Motivated by this observation, we investigate three aspects of OSR: label granularity, similarity between the open- and closed-sets, and the role of hierarchical supervision during training. To study these dimensions, we curate new open-set splits of a large fine-grained visual categorization dataset. Our analysis results in several interesting findings, including: (i) the best OSR method to use is heavily dependent on the degree of semantic shift present, and (ii) hierarchical representation learning can improve coarse-grained OSR, but has little effect on fine-grained OSR performance. To further enhance fine-grained OSR performance, we propose a hierarchy-adversarial learning method to discourage hierarchical structure in the representation space, which results in a perhaps counter-intuitive behaviour, and a relative improvement in fine-grained OSR of up to 2\% in AUROC and 7\% in AUPR over standard training. Code and data are available: langnico.github.io/fine-grained-osr.

1. Introduction

The goal of open-set recognition (OSR) is to distinguish familiar or closed-set recognition that are seen during training from novel or open-set categories that were not seen\textsuperscript{45}. In other words, OSR focuses on detecting semantic shifts between the training and test categories\textsuperscript{1}. However, not all semantic shifts are equal. Different open-set categories will have different degrees of visual similarity to the familiar closed-set categories. Recently, Vaze et al.\textsuperscript{53} showed that OSR is harder for categories exhibiting smaller semantic shifts relative to the training categories. Intuitively, novel categories that are highly similar to the training categories tend to be confused with familiar ones with high confidence. We refer to this phenomenon as the “familiarity trap”. Note that this is the opposite of what we expect to see in zero-shot classification\textsuperscript{42} and transfer-learning\textsuperscript{12}, where smaller semantic shifts make the problem easier, not harder (see illustration in Figure 1). The vast majority of the OSR literature has focused on investigating coarse open-set splits\textsuperscript{53} while the impact of finer-grained splits is underexplored.

Unlike traditional (coarse-grained) OSR, fine-grained OSR reflects realistic category discovery challenges in fine-
grained domains like medicine [43], art [11], and the natural world [51]. For example, in species monitoring [3, 27, 39, 51], we might expect to see new species that are highly visually similar to known species. While existing unlabeled data might contain such examples, their similarity makes it difficult to detect them. In citizen science applications, automated vision systems can be used to distinguish poisonous mushrooms from similar-looking edible ones [41]. Ensuring that a novel, potentially dangerous mushroom species is not confused with a previously observed safe one is crucial. Fine-grained OSR problems may also be encountered when digitizing museum collections, screening potentially counterfeit goods, or identifying fashion trends.

Natural world image collections from citizen science platforms like iNaturalist [1] are particularly well-suited to studying fine-grained OSR thanks to their size, diversity, and high-quality taxonomic structure (which provides a proxy for semantic similarity). In this work, we use the iNat2021 [52] dataset to investigate the challenges in fine-grained OSR in terms of semantic similarity, supervision granularity, and hierarchical representations. We explore the role of semantic shift by studying progressively more difficult open-set splits (Figure 1). In contrast to previous work that uses fine-grained datasets with a limited number of categories [53], we work with a large-scale benchmark containing 10,000 categories and 2.7 million images. This allows us to study diverse real-world OSR phenomena. By training models from scratch on large splits of iNat2021, we avoid the risk of leaking knowledge from pretraining datasets like ImageNet [44] that might contain semantic overlap with the open-set [7].

While the use of hierarchical labels has been studied in closed-set recognition to reduce the severity of mistakes [6, 15, 17] in fashion [31], crop mapping [50], food [58] and species recognition [14], the role of hierarchical representations for fine-grained OSR is underexplored. We find that encouraging hierarchical structure in the representation space can be beneficial for closed-set classification and coarse-grained open-set recognition, but surprisingly it has a limited effect on fine-grained OSR. Motivated by this finding, we explore whether implicit hierarchical structure should be reduced for fine-grained OSR. We propose a hierarchy-adversarial learning approach that can improve fine-grained OSR performance by discouraging hierarchical structure. By adapting the gradient reversal layer developed for unsupervised domain adaptation [20, 21], our approach learns representations that reward linear separability of fine-grained categories while discouraging linear separability of coarser-grained granularities. As we discourage hierarchical structure, representations of instances from categories that share similar features are pushed away from each other. To put it differently, we hypothesize that the hierarchy-adversarial approach discourages features shared across classes in favour of class-specific features. Thus, fine-grained novel categories are less likely to get caught by the “familiarity trap”, i.e., less likely to be misclassified as a semantically similar category with high confidence. We make the following contributions:

1. We demonstrate that the choice of the best scoring rule for OSR depends on the semantic similarity between the closed- and open-set and that familiarity scores perform best in fine-grained OSR.
2. We investigate the impact of supervision granularity and show that fine-grained supervision improves OSR performance even for large semantic shifts.
3. We explore the role of hierarchical representations and introduce hierarchy-adversarial learning to improve fine-grained OSR by discouraging hierarchical structure.
4. We introduce iNat2021-OSR, a benchmark with curated open-set splits for the iNat2021 dataset [52] for two taxa: birds and insects. This enables the study of OSR along seven discrete “hops” that encode the semantic distance from coarse-grained (7-hop) to fine-grained (1-hop).

2. Related work

2.1. Open-set recognition

Open-set recognition (OSR) [45] is concerned with the problem of identifying novel categories, i.e., semantic shifts. A large number of OSR-specific deep learning methods have been proposed including OpenMax [4], OSRCI [37], ARPL [8], and OpenHybrid [60]. However, recently Vaze et al. [53] demonstrated that simple baselines based on classifier confidence (i.e. logits) perform on par and that the improvement in OSR performance is not necessarily solely attributed to dedicated OSR approaches, but can instead be explained by improvements in the closed-set recognition accuracy, as the result of using more sophisticated deep neural network architectures and training strategies. In parallel, Dietterich and Guyer [18] proposed the ‘familiarity hypothesis’ to explain the observation that deep networks can be interpreted as detecting the absence of familiarity, instead of the presence of novelty.

In this work, we are concerned with fine-grained OSR, a setting which has been partially explored in existing work [2, 13, 22, 48, 53]. Our goal is to investigate the impact in performance resulting from changes in the semantic similarity between the closed and open-sets. Many existing fine-grained approaches construct random data splits, which will thus contain a mixture of hard and easy categories in the open-set [22], or they use splits with very large semantic differences (e.g., using dogs as familiar and cars as novel categories) Dai et al. [13]. Using manually annotated attributes, Vaze et al. [53] constructed easy, medium, and hard open-set splits of three popular fine-grained datasets (i.e., CUB-Birds [55], Stanford Cars [30],
and FGVC-Aircraft [36]). We build on this work, by constructing dataset splits leveraging taxonomies to allow us to explore the impact of a larger variety of split types.

There are two major limitations in prior OSR work that can yield an overly optimistic impression of model performance. First, the small size of existing OSR benchmark datasets necessitates the use of pretrained models. This potentially results in a semantic overlap between the pretraining categories (e.g., from ImageNet [44]) and the open-set categories [25]. Hence, the open-set categories may already be well separated from the closed-set ones. Second, the experimental evaluation on random data splits can lead to coarse open-set splits, which are not representative of real-world open-set applications that can be finer-grained [53]. This work aims to overcome both limitations by conducting experiments on large-scale image datasets across a spectrum of coarse to fine open-set splits.

Related to OSR is the more general task of out-of-distribution (OOD) detection, also referred to as an anomaly or novelty detection [33, 47], which encompasses a set of methods that explore problems such as detecting covariate or distributional shifts [23, 29, 40, 62], semantic shifts [25], and combinations of shifts [26, 40, 49]. OSR is concerned with semantic shifts, and can thus be interpreted as a particular case of OOD detection [59]. Different methods have been explored for OOD detection, including score-based [24–26, 34], ensemble/disagreement-based [32, 32, 40], and distance metric-based [28]. In Sec. 5, we evaluate these three families of methods in the context of OSR.

2.2. Granularity, similarity, and hierarchy

Our work is concerned with exploring OSR through the lens of granularity, similarity, and hierarchy, we briefly summarize relevant related works on these topics below.

Label granularity encodes how detailed a category label is, and can range from coarse-grained (e.g., “animal”) to fine-grained (e.g., “blue jay”). Granularity has been widely studied in closed-set recognition [57]. Multiple granularity supervision has been used to implicitly focus on different parts of an image to improve fine-grained categorization [56]. The granularity of supervision can impact performance in different ways. Van Horn et al. [51] showed that fine-grained training labels can hurt performance on coarse-grained object detection. Cole et al. [9] demonstrated that there is a “sweet” spot of granularity supervision for weakly supervised object localization (i.e., not too fine and not too coarse). Furthermore, it has been demonstrated that the accuracy gap between self-supervised and supervised methods grows when evaluated using finer-grained labels [10].

Semantic similarity describes the distance between two concepts (e.g., categories) in terms of their meaning. Visual similarity (or perceptual similarity [61]) can serve as a proxy for semantic similarity. Selecting visually similar pretrained data has been shown to be effective for improving transfer learning performance [12]. In fine-grained domains, samples of the same category can be visually very different (i.e., exhibit high intra-class variance) and samples of different categories can be very visually similar (i.e., exhibit low inter-class variance) [57]. Depending on the domain, nuisance variables such as object pose and scene illumination [46, 62], or sex and age [51] can amplify the discrepancy between visual and semantic similarity. Additional metadata such as attributes or taxonomy has been studied as an alternative proxy for semantic similarity, for instance in zero-shot learning in ImageNet categories [19, 38] or for iNaturalist species [42]. Vaze et al. [53] used labeled attributes to demonstrate that open-set recognition performance depends on the semantic similarity between the training and test categories. Interestingly, the difficulty of a task is not consistently coupled with granularity or semantic similarity. While OSR becomes harder for small semantic shifts, in contrast, zero-shot and transfer learning becomes easier.

Label hierarchy has been incorporated in closed-set recognition to reduce the severity of mistakes [6, 15, 17] in fashion [31], crop mapping [50], food [58] and species recognition [14]. Similarly, in hierarchical OSR [5, 16, 33, 47, 54] the aim is to find the closest ancestor, the ‘super-category’ of a novel category. This task turns the open-set problem into a closed-set one by coarsening the granularity of the predicted category. Our goal is not to advance hierarchical OSR, but to instead study the role of granularity, semantic similarity, and hierarchical structure in OSR, with the aim of understanding how to advance performance in the presence of very small semantic shifts, i.e., fine-grained OSR.

3. Methodology

3.1. Scores for open-set recognition

For the task of binary OSR, i.e., familiar vs. novel, we model a score \( S(x) \) to indicate if a test sample \( x \) belongs to the set of familiar training categories \( F = \{1, \ldots, K\} \). The score \( S(x) \) should increase as \( p(y \in F|x) \) increases, i.e., there should be a high rank correlation between the two. We summarize three score methods from the OOD and OSR literature that cover three families of methods [32, 49, 53].

For a given input image, we extract representations using the encoder \( f_\theta \), e.g., a trained deep neural network, parameterized by \( \theta \). A linear decoding function \( h_\omega \) with parameters \( \omega \) maps the representations to a \( K \)-dimensional vector, the logits, and the softmax function \( \sigma \) maps these logits to probabilities. That is, the predicted probability of an input image \( x \) belonging to class \( y \) is given by \( p(y|x; \theta, \omega) = \sigma_y(h_\omega(f_\theta(x))) \). We write \( p(y|x) \) for \( p(y|x; \theta, \omega) \) if the model parameters are clear from the context.
Familiarity scores. These methods use either the logits of a classifier or the softmax confidence scores. It has been empirically shown that the unnormalized maximum logit score (MLS) can outperform the maximum softmax probability (MSP) as an open-set scoring function [53]. The MLS score is defined as:

\[ S_{\text{MLS}}(x) \triangleq \max_{y \in \mathcal{Y}} [h_{\omega}(f_{\theta}(x))]_y, \]  

(1)

where \([.]_y\) denotes the \(y\)-th component of a logit vector.

Disagreement scores. We consider an ensemble of \(M\) classifiers with parameter \(\theta_m\) and \(\omega_m\) for \(m = 1, \ldots, M\), and write \(p_m(y|x) = p(y|x; \theta_m, \omega_m)\). We can compute not only the average confidence \(p_E(y|x) = \frac{1}{M} \sum_m p_m(y|x)\) and average logits \(l_E(y|x) = \frac{1}{M} \sum_m h_{\omega_m}(f_{\theta_m}(x))\), but also an alternative score that quantifies the disagreement (i.e., variance) between the individual member outputs. The negative KL-disagreement [32], \(S_{\text{KLD}}\) is used as an indication that a high score denotes familiar categories:

\[ S_{\text{KLD}}(x) \triangleq -\sum_{m=1}^{M} \text{KL}(p_E(\cdot|x)||p_m(\cdot|x)), \]  

(2)

where \(\text{KL}\) represents the Kullback-Leibler divergence between the predicted class distribution of the ensemble \(p_E(y|x)\), i.e., the average confidence over the \(M\) members, and the predicted class distribution of each member \(p_m(y|x)\).

Distance metric scores. Finally, distance in representation space can be combined with k-nearest neighbours (KNN) to formulate an OSR score. We adopt the nearest-neighbour approach from [49], originally proposed for general OOD detection. The method takes the L2-normalized representations of the encoder \(\tilde{f}_{\theta}(x) = f_{\theta}(x)/\|f_{\theta}(x)\|_2\) and computes the shortest distance to the images in the training data set \(\{(x_1, y_1), \ldots, (x_N, y_N)\}\) in the normalized representation space:

\[ S_{\text{NN}}(x) \triangleq \min_{i=1, \ldots, N} -\|\tilde{f}_{\theta}(x) - \tilde{f}_{\theta}(x_i)\|_2. \]  

(3)

3.2. Hierarchy-supportive learning

One of our objectives is to study the role of label training granularity and hierarchy in OSR. Among many existing approaches [6, 14, 15, 17, 31, 50], we chose a simple hierarchical multi-task approach [58]. This provides us with a framework for exploring representations where the hierarchical structure is either encouraged or discouraged. We follow the notation in [58] and define a hierarchy of a given label space as \(\mathcal{T} = \{\mathcal{Y}_l\}_{l=0}^T\), where \(\mathcal{Y}_l\) defines the set of categories at the \(t\)-th level in the hierarchy and the leaf nodes at \(t = 0\) correspond to the fine-grained categories. From this, we can formulate a hierarchy-supportive multi-task loss that optimizes the cross-entropy at every level \(t\) in the hierarchy, where \(y^t_i\) denotes the true label for sample \(i\) at the \(t\)-th hierarchy level and \(p^t(y^t_i|x_i) = p(y^t_i|x_i; \theta, \omega^t)\):

\[ \mathcal{L}_{\text{HS}} = \sum_{t=0}^{T} \sum_{i=1}^{N} -\log p^t(y^t_i|x_i). \]  

(4)

To optimize this training loss, we append additional classification heads \(h_{\omega^t}\) with parameters \(\omega^t\) to the encoder, one for each level \(t\) in the hierarchy as illustrated in Figure 2b.

3.3. Hierarchy-adversarial learning

Similarly, we can formulate a multi-task loss for fine-grained classification that discourages hierarchical structure in the representation space. We are interested in learning an encoder \(f_{\theta}\) that yields a representation suitable for classifying fine-grained categories – but is agnostic to coarser hierarchical labels at levels \(t > 0\) in the hierarchy. We hypothesize that this encourages salient features specific for fine-grained categories rather than relying on coarse-grained features for distinguishing groups of fine-grained categories. This can be achieved by minimizing the cross-entropy for the finest granularity \(\mathcal{L}^0\), but maximizing the cross-entropy...
for the coarser granularities $\mathcal{L}^t$ with $t > 0$,

$$
\mathcal{L}_{HA} = N \sum_{i=1}^{N} - \log p^0(y_i^0|x_i) - \lambda \sum_{t=1}^{T} N \sum_{i=1}^{N} - \log p^t(y_t^t|x_i).
$$

To ensure that the $\mathcal{L}^t$ for $t > 0$ are not maximized by the decoding functions, but by the representation, we solve a min-max optimization problem:

$$
\min_{\theta} \max_{\omega^0, \omega^1, \ldots, \omega^T} \mathcal{L}_{HA}(\theta, \omega^0, \omega^1, \ldots, \omega^T).
$$

That is, we obtain the parameters of the encoder $\theta$ and the fine-grained decoder $\omega^0$ that minimize $\mathcal{L}^0$ (i.e., the representation is discriminative for fine-grained labels). At the same time, for $t > 0$, we obtain the decoders $\omega^t$ that maximize $\mathcal{L}^t$, i.e., the representation is regularized to not capture coarser-grained hierarchical labels. To train a deep neural network with this min-max objective using stochastic gradient descent (SGD), we adapt a gradient reversal layer-based approach inspired by the domain adaptation literature [20, 21]. By swapping the sign of the gradient corresponding to the $\mathcal{L}^t$ terms before back-propagating into the encoder (see Figure 2c), all learnable parameters can be updated in an otherwise standard forward- and backward-pass.

4. Experimental setup

4.1. The iNat2021-OSR dataset

While existing OSR benchmarks for coarse-grained labels contain many samples, the fine-grained ones are small in size [53]. This limitation of the fine-grained datasets makes it impossible to train models from scratch on them, which we require to have full control over a realistically plausible setting for OSR. We are also interested in having a granular label hierarchy that can serve as a proxy for semantic similarity. Motivated by this, we look to the iNat2021 dataset [52] which contains over 2.5 million images of 10,000 different plant and animal species (i.e., categories) as a benchmark for fine-grained OSR.

Our experimental setup is inspired by the zero-shot learning work of Rodríguez et al. [42], who use the taxonomy of iNat2021 to evaluate zero-shot performance at different semantic distances to the training data, measured in the number of edges to the lowest common ancestor in the hierarchical graph, which we simply refer to this distance as hops (see Figure 1). We develop two curated open-set data splits of iNat2021 (see Table 1) and group open-set categories by the smallest hop distance to any sample in the training data, ranging from 1–7 hops. The two splits provide a closed-set for the super-categories “birds” (Aves in biological classification) and “insects” (Insecta) at the hierarchy level $t = 5$ (i.e., “class”). The closed-set consists of a subset of species from the respective super-category, i.e., the familiar categories. The remaining categories are used as the open-set test sets with semantic distances ranging from 1-hop (fine-grained OSR) to 7-hop (coarse-grained OSR). While hops 1–4 correspond to relationships within the super-categories (i.e., novel bird or insect species), hops > 4 correspond to open-set categories outside these super-categories. Full details about the iNat2021-OSR dataset are given in the supplementary material.

4.2. Implementation details

Training. We use a ResNet-50 backbone and train all models from scratch for 100 epochs using SGD with a base learning rate set to 0.1, which is multiplied by factor 0.1 every 30 epochs, and weight decay of $1e-4$. We train five models for the ensemble disagreement score (Sec. 3.1) starting from different random initializations following prior work [40]. To train the hierarchy-adversarial approach (Sec. 3.3), we gradually increase the weight of the adversarial gradient $\lambda$ starting from zero to $\alpha$ by the end of the training process similar to [20].

Evaluation. Open-set performance is reported using an equal number of samples in the closed-set and open-set by sub-sampling the larger set. For all evaluation metrics, a higher score means better performance. More details can be found in the supplementary material. Closed-set accuracy of coarser granularities is evaluated using hard pooling, based on a single predicted category and the given hierarchy. We did not observe a difference using soft pooling, i.e., aggregating the probabilities of multiple categories.

5. Results

Our experiments investigate the role of semantic similarity, supervision granularity, and hierarchical structure in OSR.

5.1. What is the role of semantic similarity in OSR?

Fine-grained OSR is hardest. We observe a large performance gap between fine-grained (1-hop) and coarse-grained OSR (7-hop) (Figure 3). Hence, OSR performance correlates positively with semantic distance. One exception is the performance drop from 6-hop to 7-hop, where the
familiar categories from “birds” and “insects” are tested against “plants” and “fungi”. A possible explanation for this drop could be that plants are often seen during training in the background for the closed-set categories. This may increase familiarity scores, making these samples harder to detect as novel categories. Similar issues have been found to impair the evaluation of OOD detection [7].

The choice of OSR score depends on semantic similarity. All scores follow the same general trend that OSR performance improves with larger semantic distance (Figure 3). However, none of the scoring rules consistently performs best across the spectrum of semantic similarity. For fine-grained OSR (1-hop) and up to 6-hop, the maximum logit score (MLS) performs best. However, for coarse-grained OSR, the KL-disagreement outperforms the logit score in our experiments. Moreover, as birds and insects contain category relationships up to 4-hops, open-sets with hops >4 are completely outside the classifier’s domain expertise. This explains the discontinuity from 4- to 5-hop, where the performance gap between the KL-disagreement and the logit score is reduced. The nearest neighbour approach performs worst for both extremes. Interestingly, the L2-distance on unnormalized representations yields a lower performance on 1-hop, but higher performance on 7-hop OSR (see supplementary Figure A6). As the MLS performs best in fine-grained OSR, we will focus our analyses on this score.

5.2. What is the impact of supervision granularity?

Closed-set accuracy increases for coarser granularities.

By pooling both the fine-grained prediction and the target label to coarser granularities, the closed-set accuracy improves (Figure 4b). This confirms the observation that the closed-set problem is easier for coarser granularities with fewer categories [10].

Overly fine-grained supervision improves closed-set accuracy. Closed-set accuracy is highest when we pool the model trained on the finest granularity to the coarser test granularity. Training on the matching granularity deteriorates performance (Figure 4b). This contradicts results in both object detection, where overly fine supervision deteriorates performance for coarser super-categories [51], and weakly supervised localization where a “sweet spot” of medium-grained supervision performs best [9]. The iNat2021 dataset is approximately class-balanced at the finest granularity, but is imbalanced at coarser levels as the number of leaf nodes varies in the hierarchy. Therefore, we average the per-category accuracies, which shows a substantial performance gap between the matching granularity and the pooling of the finest supervision granularity.
Overly fine-grained supervision improves OSR. Like the closed-set accuracy, OSR performance is higher for models trained on the finer granularities (Figure 4a). Even if the familiar and novel data is separated at coarser granularities, i.e., hop distances > 1, training models on finer granularity is still beneficial. It is also best for finest-granularity supervision across the entire spectrum of shifts. This behaviour can be explained by the phenomena observed by Vaze et al. [53] whereby better closed-set accuracy improves OSR performance.

5.3. How does hierarchical structure affect OSR?

Hierarchical representations can improve coarse-grained OSR. First, we look at the closed-set accuracy when using hierarchy-supportive learning, which is unaffected at the finest granularity (Figure 5 and Table 2), likely because the coarse supervision does not help the learning of features to resolve confusions within very similar categories. However, closed-set accuracy improves for coarser granularities (see supplementary), which means that the severity of mistakes is reduced and is in line with previous work [6, 15, 17]. Furthermore, the hierarchical structure can also improve coarse-grained OSR performance (see Figure 5), which might be explained by the correlation between closed-set and open-set performance [53].

Hierarchical adversarial learning can improve fine-grained OSR. Based on these observations, we study an alternative direction and turn to the hierarchy-adversarial training strategy, which aims to reduce hierarchical structure in the learned representation. This approach improves fine-grained OSR for 1-hop open-sets across several categories. However, closed-set accuracy improves for coarser granularities (see supplementary), which means that the coarse supervision does not help the learning of features to resolve confusions within very similar categories. However, closed-set accuracy improves for coarser granularities (see supplementary), which means that the severity of mistakes is reduced and is in line with previous work [6, 15, 17]. Furthermore, the hierarchical structure can also improve coarse-grained OSR performance (see Figure 5), which might be explained by the correlation between closed-set and open-set performance [53].

metrics for both the MLS and NN score, but not for the KL-disagreement (see Figure 5 and Table 2). However, it consistently impairs coarser OSR performance (Figure 5). Qualitative examples where the hierarchy-adversarial approach improves the “familiarity trap” w.r.t. the baseline are included in the supplementary material.

To empirically analyse this strategy’s effect, we compute the L2-distance between category centroids from the 1-hop open-set categories and the training categories in the learned representation space (Figure 6a,b). We observe that hierarchy-adversarial learning pushes fine-grained open-set categories away from their most related training categories (1-hop). This contrasts hierarchy-supportive learning, which pulls the fine-grained open-set closer to its most related training categories but pushes categories at larger distances even further away.

We study the sensitivity of the hyperparameter $\lambda$ from Equation 5, which controls the weight of the adversarial gradient and gradually increases from zero to $\alpha$. We demonstrate that this parameter controls both the closed-set accuracy and the open-set performance (Figure 6c). First, by increasing $\alpha$, we see an improvement in open-set performance, but at the cost of closed-set accuracy. This behaviour contradicts the observation from Vaze et al. [53] as it shows that there is an underlying role of hierarchical structure in OSR that cannot necessarily be explained by improved closed-set accuracy. However, open-set performance drops at a point, most likely due to the drop in closed-set accuracy.
Figure 6. Hierarchy-adversarial learning analyses. To analyze how hierarchy-aware training affects the relationship between classes in feature space, we plot (a) the L2-distance between novel and familiar category centroids in representation space and (b) the change of the L2-distance w.r.t. the cross-entropy baseline (derived from a). We expect that the hierarchy-supportive learning pulls fine-grained open-set categories (1-hop) closer to the familiar categories and amplifies the distance between coarse-grained categories. In contrast, the hierarchy-adversarial learning is expected to push fine-grained categories away from familiar categories, but roughly preserve the distance of coarse-grained categories. Our empirical analyses confirm these expectations on the Aves dataset. (c) Hyperparameter study for the weight of the hierarchy-adversarial gradient on the fine-grained open-set (1-hop). With increasing \( \alpha \), open-set recognition performance improves at the cost of closed-set accuracy. However, there is a tipping point when the weight \( \alpha \) is too high, and open-set performance drops. This behaviour contradicts with the observation of Vaze et al. [53] as closed-set and open-set performance do not correlate.

5.4. Limitations

While our study presents a step towards understanding OSR under challenging fine-grained shifts, some limitations exist. First, we rely on a taxonomy as a proxy for concept granularity and semantic similarity to study our research questions. To fit our needs, we have curated a new large-scale open-set benchmark dataset using “birds” and “insects” from iNat2021 as the training domains. It remains to be tested if our observations hold for other domains, such as food, art, or medicine. However, Cole et al. [9] have demonstrated that the existing ImageNet hierarchy is not a good proxy for concept granularity. Thus, ImageNet is not a likely candidate for our OSR setting either. An additional challenge is the irregular depth of the ImageNet categories.

Second, controlling the degree of hierarchical structure in learned representations is non-trivial. Some hierarchical structure is essential to solving fine-grained problems and is learned implicitly, as shown in our experiments (see Figure 4). We study the “opposite” of hierarchical structure with a hierarchy-adversarial approach. Our hyperparameter study shows that there is a sweet spot for selecting the weight of hierarchy-adversarial gradients. If chosen too high, closed-set and consequently open-set performance is impaired. One approach to selecting this weight, without an open-set validation, is to select a conservatively small \( \alpha \) that still preserves closed-set validation accuracy.

Third, existing fine-grained OSR benchmarks are based on attribute labels [53]. Our presented hierarchy-adversarial approach relies on hierarchical labels and cannot make use of attribute labels. A more general formulation of a “similarity-adversarial” approach would benefit from supporting any proxy for semantic similarity, e.g., attributes.

Finally, the benefit of pretrained models for fine-grained OSR requires further investigation. While prior work pre-trained on different domains [53], pretraining on ImageNet would require the removal of 269 wild animal categories [35] that potentially overlap with iNat2021.

6. Conclusion

While previous work has mostly addressed coarser OSR we propose a new large-scale OSR benchmark which allows us to study OSR across a range of increasingly challenging semantic shifts, from coarse to fine-grained. Our results demonstrate that fine-grained OSR remains a very challenging task for current methods. We show that there are major underlying disruptive factors that hinder improved OSR performance, e.g., supervision granularity and the amount of semantic shift present in the open-set. Furthermore, hierarchical structure in the learned representation plays an important, and possibly counter-intuitive, role. We propose a hierarchy-adversarial learning strategy to discourage hierarchical structure which improves fine-grained OSR by attempting to address the “familiarity trap”. However, our analysis shows that there is still more to be done to close the gap between coarse and fine-grained OSR performance.

Acknowledgment. NL, VS, CI, SB acknowledge support by the Pioneer Centre for AI, DNRF grant number P1, and the Edinburgh–Copenhagen Strategic Partnership Award for OMA. CI is supported by “Risk-assessment of Vector-borne Diseases Based on Deep Learning and Remote Sensing” (grant number NNF21OC0069116) by the Novo Nordisk Foundation. Thanks to iNaturalist for the open data policy and to Stella Frank for valuable feedback on the manuscript.
References


[12] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 1, 3


for enabling artificial intelligence in the ocean. Scientific Reports, 12(1), 2022. 2


[31] Suren Kumar and Rui Zheng. Hierarchical category detector for clothing recognition from visual data. In IEEE International Conference on Computer Vision Workshops, 2017. 2, 3, 4


[37] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In European Conference on Computer Vision (ECCV), 2018. 2


[43] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pavlovski, Ian Freyberg, Yuan Liu, Zach Beaver, Nam Vo, Peggy Bui, Samantha Winter, Patricia MacWilliams, Greg S. Corrado, Umesh Telang, Yun Liu, Taylan Cemgil, Alan Karthikesalingam, Balaji Lakshminarayanan, and Jim Winkens. Does your dermatology classifier know what it doesn’t know? detecting the long-tail of unseen conditions. Medical Image Analysis, 75, 2022. 2


[51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The unnaturalist species classification and detection dataset. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2, 3, 6

[52] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking

[53] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In International Conference on Learning Representations (ICLR), 2022. 1, 2, 3, 4, 5, 6, 7, 8


[56] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In International Conference on Computer Vision (ICCV), 2015. 3


