Hierarchical Novelty Detection for Visual Object Recognition
Supplementary Material

Kibok Lee* Kimin Lee† Kyle Min* Yuting Zhang* Jinwoo Shin† Honglak Lee‡*
*University of Michigan, Ann Arbor, MI, USA
†Korea Advanced Institute of Science and Technology, Daejeon, Korea
‡Google Brain, Mountain View, CA, USA

A. More on hierarchical novelty detection

A.1. Details about objectives

We present the exact objective functions we propose without notation abuse. Let $S(k) = S(y = k|x)$ be an unnormalized softmax score of the k-th class (which can be either known or novel), e.g., $S(k) = \exp\left(w_k^\top x + b_k\right)$.

Top-down. We note that there is a notation abuse in the objective function of the top-down method for simplicity; without notation abuse, the exact objective is

$$\min_\theta \mathbb{E}_{Pr(x,y|s)}\left[-\log Pr(y|x,s;\theta_{\mathcal{N}(s)\cup\mathcal{C}(s)})\right] + \mathbb{E}_{Pr(x,y|\mathcal{O}(s))}\left[D_{KL}\left(U(\cdot|s)\parallel Pr(\cdot|x,s;\theta_{\mathcal{N}(s)\cup\mathcal{C}(s)})\right)\right]. \tag{A.1}$$

The softmax probability used in this objective is

$$Pr(y|x,s;\theta_{\mathcal{N}(s)\cup\mathcal{C}(s)}) = \frac{S(y)}{S(\mathcal{N}(s)) + \sum_{y'\in\mathcal{C}(s)} S(y')}.$$

Relabel. Since super classes in taxonomy have training data by data relabeling, the objective is a standard cross entropy loss over all super and leaf classes:

$$\min_\theta \mathbb{E}_{Pr(x,y)}\left[-\log Pr(y|x;\theta_{\mathcal{T}})\right]. \tag{A.2}$$

The softmax probability used in this objective is

$$Pr(y|x;\theta_{\mathcal{T}}) = \frac{S(y)}{\sum_{y'\in\mathcal{T}} S(y')} = \frac{S(y)}{\sum_{l\in\mathcal{L}(\mathcal{T})} S(l) + \sum_{s\in\mathcal{T}\setminus\mathcal{L}(\mathcal{T})} S(\mathcal{N}(s))}.$$

Here, $\mathcal{T}\setminus\mathcal{L}(\mathcal{T})$ represents all super classes in $\mathcal{T}$.

LOO. We note that there is a notation abuse in the second term of the objective function of LOO for simplity; without notation abuse, the exact objective is

$$\min_\theta \mathbb{E}_{Pr(x,y)}\left[-\log Pr(y|x;\theta_{\mathcal{L}(\mathcal{T})}) + \sum_{a\in\mathcal{A}(y)} -\log Pr(\mathcal{N}(\mathcal{P}(a))|x;\theta_{\mathcal{N}(\mathcal{P}(a))\cup\mathcal{L}(\mathcal{T}\setminus a)})\right]. \tag{A.3}$$

The softmax probabilities are defined as:

$$Pr(y|x;\theta_{\mathcal{L}(\mathcal{T})}) = \frac{S(y)}{\sum_{l\in\mathcal{L}(\mathcal{T})} S(l)},$$

$$Pr(\mathcal{N}(\mathcal{P}(a))|x;\theta_{\mathcal{N}(\mathcal{P}(a))\cup\mathcal{L}(\mathcal{T}\setminus a)}) = \frac{S(\mathcal{N}(\mathcal{P}(a)))}{S(\mathcal{N}(\mathcal{P}(a)) + \sum_{l\in\mathcal{L}(\mathcal{T}\setminus a)} S(l)}.$$

A.2. Hyperparameter search

A difficulty in hierarchical novelty detection is that there are no validation data from novel classes for hyperparameter search. Similar to the training strategy, we leverage known class data for validation: specifically, for the top-down method, the novelty detection performance of each classifier is measured with $\mathcal{O}(s)$, i.e., for each classifier in a super class s, known leaf classes not belong to s are considered as novel classes.

$$\hat{y} = \begin{cases} \arg\max_{y'} Pr(y'|x,s;\theta_s) & \text{if } D_{KL}(U(\cdot|s) \parallel Pr(\cdot|x,s;\theta_s)) \geq \lambda_s, \\ \mathcal{N}(s) & \text{otherwise,} \end{cases}$$

where λ_s is chosen to be maximize the harmonic mean of the known class accuracy and the novelty detection accuracy. Note that λ_s can be tuned for each classifier.

For validating flatten methods, we discard logits of ancestors of the label of training data in a hierarchical manner. Mathematically, at the stage of removal of an ancestor $a \in \mathcal{A}(y)$, we do classification on $\theta_{\mathcal{T}\setminus a}$:

$$\hat{y} = \arg\max_{y'} Pr(y'|x;\theta_{\mathcal{T}\setminus a}),$$

where the ground truth is $\mathcal{N}(\mathcal{P}(a))$ at the stage. The hyperparameters with the best validation AUC are chosen.

Model-specific description. DARTS has an accuracy guarantee as a hyperparameter. We took the same candidate in the original paper, $\{$0 %, 10 %, $\ldots$, 80 %, 85 %, 90 %, 95 %, 99 %$\}$, and find the best accuracy guarantee, which turned out to be 90 % for ImageNet and CUB, and 99 % for AwA2. Similarly, for Relabel, we evaluated relabeling rate from 5 % to 95 %, and found that 30 %, 25 %, and 15 % are the best for ImageNet, AwA2, and CUB, respectively. For the top-down method and LOO, the ratio of two loss terms can be tuned, but it turned out that the performance is less sensitive to the ratio, so we kept 1:1 ratio. For TD+LOO, we extracted the multiple softmax probability vectors from the top-down model and then trained the following LOO.

There are some more strategies to improve the performance: The proposed losses can be computed in a class-wise manner, i.e., weighted by the number of descendant classes, which is helpful when the taxonomy is highly imbalanced, e.g., ImageNet. Also, the log of softmax and/or ReLU can be applied to the output of the top-down model. We note that stacking layers to increase model capacity improves the performance of Relabel, while it does not for LOO.

A.3. Experimental results on CIFAR-100

In this section, we provide experimental results on CIFAR-100 [3]. The compared algorithms are the same with the other experiments, and we tune the hyperparameters following the same procedure used for the other datasets described in Section A.2.

Dataset. The CIFAR-100 dataset [3] consists of 50k training and 10k test images. It has 20 super classes containing 5 leaf classes each, so one can naturally define the taxonomy of CIFAR-100 as the rooted tree of height two. We randomly split the classes into two known leaf classes and three novel classes at each super class, such that we have 40 known leaf classes and 60 novel classes. To build a validation set, we pick 50 images per known leaf class from the training set.

Preprocessing. CIFAR-100 images have smaller size than natural images in other datasets, so we first train a shallower network, ResNet-18 with 40 known leaf classes. Pretraining is done with only training images, without any information about novel classes. And then, the last fully connected layer of the CNNs is replaced with our proposed methods. We use 100 training data per batch. As a regularization, L2 norm weight decay with parameter 10^{-2} is applied. The initial learning rate is 10^{-2} and it decays at most two times when loss improvement is less than 2 % compared to the last epoch.

Experimental results. Table A.1 compares the baseline and proposed methods. One can note that the proposed methods outperform the baseline in both novel class accuracy and AUC. However, unlike the results on other datasets, TD+LOO does not outperform the vanilla LOO method, as one can expect that the vectors extracted from the top-down method might not be useful in the case of CIFAR-100 since its taxonomy is too simple and thus not informative.

Table A.1. Hierarchical novelty detection results on CIFAR-100. For a fair comparison, 50 % of known class accuracy is guaranteed by adding a bias to all novel class scores (logits). The AUC is obtained by varying the bias. Values in bold indicate the best performance.

Method	Novel	AUC
DARTS [2]	22.38	17.84
Relabel	22.58	18.31
LOO	**23.68**	**18.93**
TD+LOO	22.79	18.54

B. Sample-wise qualitative results

In this section, we show sample-wise qualitative results on ImageNet. We compared four different methods: DARTS [2] is a baseline method where we adapt their method to our task, and the others, Relabel, LOO, and TD+LOO, are our proposed methods. In Figure B.1–B.8, we put each test image at the top, a table of the classification results in the middle, and a sub-taxonomy representing the hierarchical relationship between classes appeared in the classification results at the bottom. In tables, we provide the true label of the test image at the first row, which is either a novel class (unseen during training) or a known leaf class. In the "Method" column in tables, "GT" is the ground truth label for hierarchical classification/novelty detection: if the true label of the test image is a novel class, "GT" is the closest known ancestor (super class) of the novel class, which is the expected prediction; otherwise, "GT" is the true label of the test image. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. "Word" is the English word of the predicted label. Each method has its own background color in both tables and sub-taxonomies. In sub-taxonomies, the novel class is shown in ellipse shape if exists, GT is double-lined, and the name of the methods is displayed below its prediction. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes: for example, a dashed edge labeled with 3 implies that two classes exist in the middle of the connection. Note that some novel classes have multiple ground truth labels if they have multiple paths to the taxonomy.

Figure B.1–B.2 show the hierarchical novelty detection results of known leaf classes, and Figure B.3–B.8 show the hierarchical novelty detection results of novel classes. In general, while DARTS tends to produce a coarse-grained label, our proposed models try to find a fine-grained label. In most cases, the prediction is not too far from the ground truth except some cases: for example, in Figure B.2 (g), LOO and TD+LOO attempt to predict the content in the object rather than the object itself.

(a) (b) (c) (d)

Known class: stingray

Method	ϵ	A	Word
GT			stingray
DARTS	1	Y	ray
Relabel	4	N	tiger shark
LOO	2	Y	elasmobranch
TD+LOO	1	Y	ray

Known class: hen

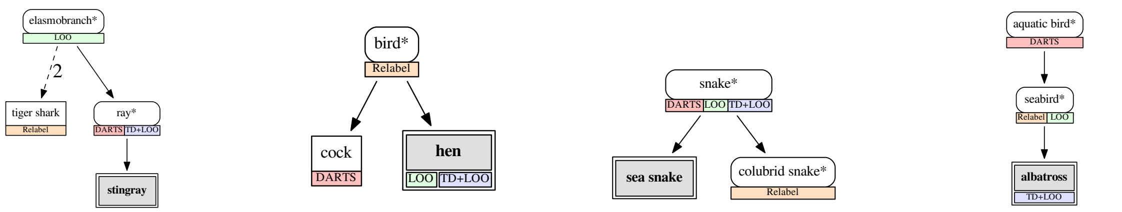
Method	ϵ	A	Word
GT			hen
DARTS	2	N	cock
Relabel	1	Y	bird
LOO	0	Y	hen
TD+LOO	0	Y	hen

Known class: sea snake

Method	ϵ	A	Word
GT			sea snake
DARTS	1	Y	snake
Relabel	2	N	colubrid snake
LOO	1	Y	snake
TD+LOO	1	Y	snake

Known class: albatross

Method	ϵ	A	Word
GT			albatross
DARTS	2	Y	aquatic bird
Relabel	1	Y	seabird
LOO	1	Y	seabird
TD+LOO	0	Y	albatross



(e) (f) (g) (h)

Known class: Maltese dog

Method	ϵ	A	Word
GT			Maltese dog
DARTS	5	N	Tibetan terrier
Relabel	4	N	terrier
LOO	0	Y	Maltese dog
TD+LOO	0	Y	Maltese dog

Known class: English foxhound

Method	ϵ	A	Word
GT			English foxhound
DARTS	4	N	Rhodesian ridgeback
Relabel	3	Y	hunting dog
LOO	1	Y	foxhound
TD+LOO	1	Y	foxhound

Known class: golden retriever

Method	ϵ	A	Word
GT			golden retriever
DARTS	2	Y	sporting dog
Relabel	1	Y	retriever
LOO	0	Y	golden retriever
TD+LOO	0	Y	golden retriever

Known class: Siberian husky

Method	ϵ	A	Word
GT			Siberian husky
DARTS	2	Y	working dog
Relabel	3	N	Eskimo dog
LOO	1	Y	sled dog
TD+LOO	1	Y	sled dog

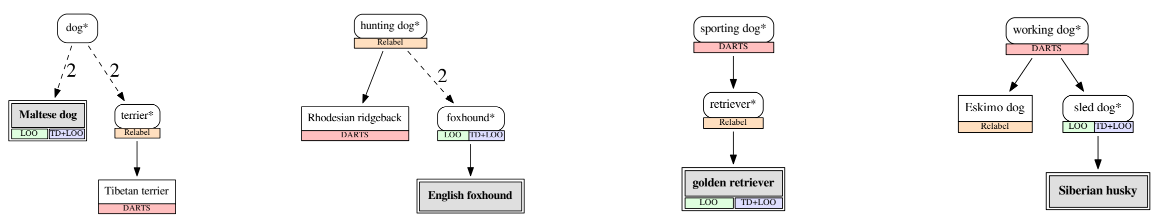


Figure B.1. Qualitative results of hierarchical novelty detection on ImageNet. "GT" is the true known leaf class, which is the expected prediction, "DARTS" is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

(a) (b) (c) (d)

(a) Known class: dingo

Method	ϵ	A	Word
GT			dingo
DARTS	5	N	shepherd dog
Relabel	3	N	dog
LOO	1	Y	wild dog
TD+LOO	0	Y	dingo

(b) Known class: Egyptian cat

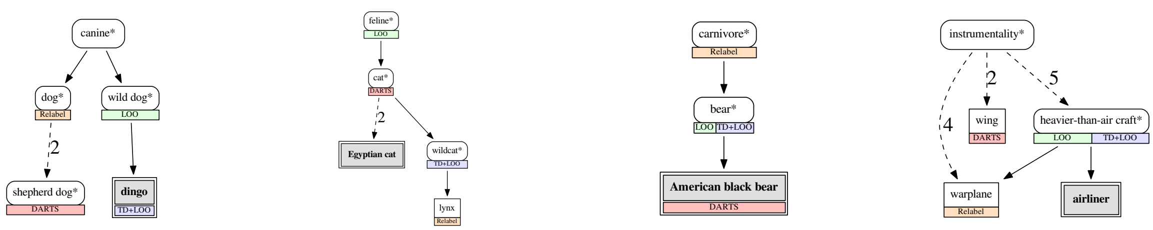
Method	ϵ	A	Word
GT			Egyptian cat
DARTS	2	Y	cat
Relabel	4	N	lynx
LOO	3	Y	feline
TD+LOO	3	N	wildcat

(c) Known class: American black bear

Method	ϵ	A	Word
GT			American black bear
DARTS	0	Y	American black bear
Relabel	2	Y	carnivore
LOO	1	Y	bear
TD+LOO	1	Y	bear

(d) Known class: airliner

Method	ϵ	A	Word
GT			airliner
DARTS	8	N	wing
Relabel	2	N	warplane
LOO	1	Y	heavier-than-air craft
TD+LOO	1	Y	heavier-than-air craft



(e) (f) (g) (h)

(e) Known class: digital clock

Method	ϵ	A	Word
GT			digital clock
DARTS	3	N	digital watch
Relabel	3	Y	measuring instrument
LOO	2	Y	timepiece
TD+LOO	0	Y	digital clock

(f) Known class: pitcher

Method	ϵ	A	Word
GT			pitcher
DARTS	7	N	drum
Relabel	1	Y	vessel
LOO	6	N	percussion instrument
TD+LOO	0	Y	pitcher

(g) Known class: soup bowl

Method	ϵ	A	Word
GT			soup bowl
DARTS	1	Y	bowl
Relabel	1	Y	bowl
LOO	11	N	punch
TD+LOO	11	N	punch

(h) Known class: toaster

Method	ϵ	A	Word
GT			toaster
DARTS	9	N	furniture
Relabel	7	N	instrumentality
LOO	1	Y	kitchen appliance
TD+LOO	0	Y	toaster

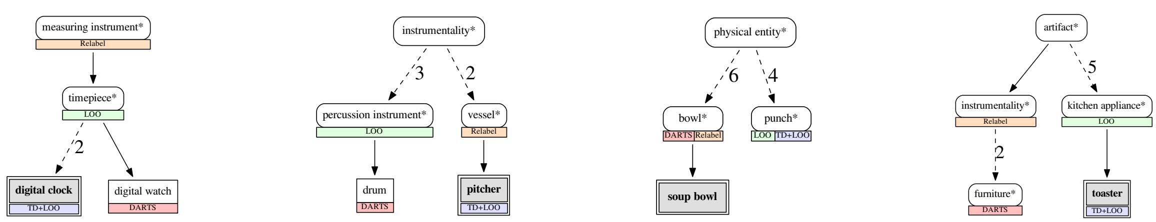


Figure B.2. Qualitative results of hierarchical novelty detection on ImageNet. "GT" is the true known leaf class, which is the expected prediction, "DARTS" is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

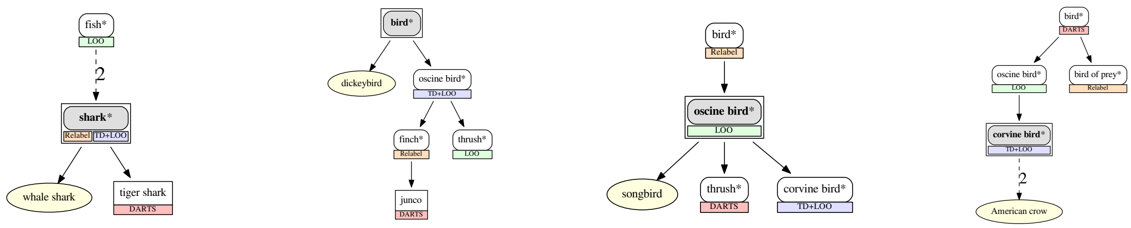
(a) (b) (c) (d)

Novel class: whale shark			
Method	ϵ	A	Word
GT			shark
DARTS	1	N	tiger shark
Relabel	0	Y	shark
LOO	2	Y	fish
TD+LOO	0	Y	shark

Novel class: dickeybird			
Method	ϵ	A	Word
GT			bird
DARTS	3	N	junco
Relabel	2	N	finch
LOO	2	N	thrush
TD+LOO	1	N	oscine bird

Novel class: songbird			
Method	ϵ	A	Word
GT			oscine bird
DARTS	1	N	thrush
Relabel	1	Y	bird
LOO	0	Y	oscine bird
TD+LOO	1	N	corvine bird

Novel class: American crow			
Method	ϵ	A	Word
GT			corvine bird
DARTS	2	Y	bird
Relabel	3	N	bird of prey
LOO	1	Y	oscine bird
TD+LOO	0	Y	corvine bird



(e) (f) (g) (h)

Novel class: raven			
Method	ϵ	A	Word
GT			corvine bird
DARTS	0	Y	corvine bird
Relabel	2	Y	bird
LOO	1	Y	oscine bird
TD+LOO	2	N	thrush

Novel class: swallow			
Method	ϵ	A	Word
GT			oscine bird
DARTS	0	Y	oscine bird
Relabel	1	Y	bird
LOO	1	N	finch
TD+LOO	3	N	kite

Novel class: sheldrake			
Method	ϵ	A	Word
GT			duck
DARTS	4	N	American coot
Relabel	2	Y	aquatic bird
LOO	1	Y	anseriform bird
TD+LOO	0	Y	duck

Novel class: scoter			
Method	ϵ	A	Word
GT			duck
DARTS	4	N	American coot
Relabel	2	Y	aquatic bird
LOO	1	Y	anseriform bird
TD+LOO	0	Y	duck

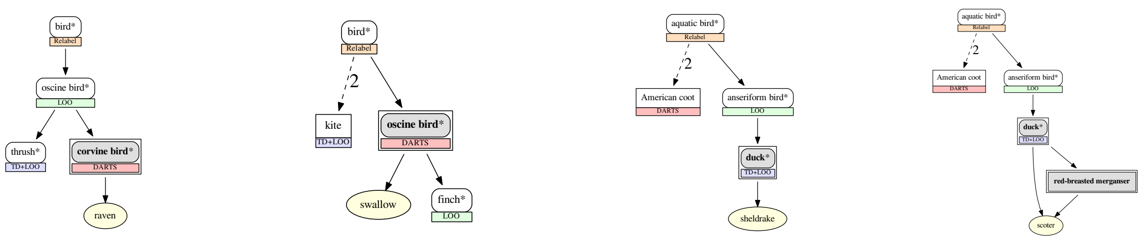


Figure B.3. Qualitative results of hierarchical novelty detection on ImageNet. "GT" is the closest known ancestor (super class) of the novel class, which is the expected prediction, "DARTS" is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

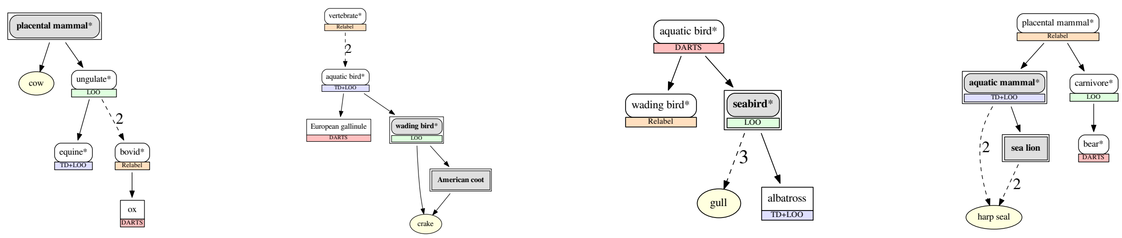
(a) (b) (c) (d)

Novel class: cow			
Method	ϵ	A	Word
GT			placental mammal
DARTS	4	N	ox
Relabel	3	N	bovid
LOO	1	N	ungulate
TD+LOO	2	N	equine

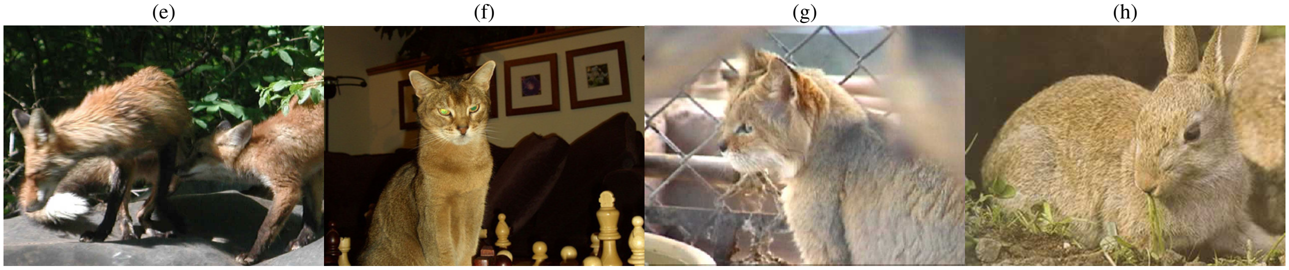
Novel class: crake			
Method	ϵ	A	Word
GT			wading bird
DARTS	2	N	European gallinule
Relabel	3	Y	vertebrate
LOO	0	Y	wading bird
TD+LOO	1	Y	aquatic bird

Novel class: gull			
Method	ϵ	A	Word
GT			seabird
DARTS	1	Y	aquatic bird
Relabel	2	N	wading bird
LOO	0	Y	seabird
TD+LOO	1	N	albatross

Novel class: harp seal			
Method	ϵ	A	Word
GT			aquatic mammal
DARTS	3	N	bear
Relabel	1	Y	placental mammal
LOO	2	N	carnivore
TD+LOO	0	Y	aquatic mammal



(e) (f) (g) (h)



Novel class: red fox, Vulpes fulva			
Method	ϵ	A	Word
GT			fox
DARTS	1	N	red fox, Vulpes vulpes
Relabel	1	Y	canine
LOO	0	Y	fox
TD+LOO	0	Y	fox

Novel class: Abyssinian cat			
Method	ϵ	A	Word
GT			domestic cat
DARTS	1	N	Egyptian cat
Relabel	0	Y	domestic cat
LOO	1	Y	cat
TD+LOO	0	Y	domestic cat

Novel class: sand cat			
Method	ϵ	A	Word
GT			wildcat
DARTS	2	Y	feline
Relabel	2	N	domestic cat
LOO	1	Y	cat
TD+LOO	0	Y	wildcat

Novel class: European rabbit			
Method	ϵ	A	Word
GT			rabbit
DARTS	1	Y	leporid mammal
Relabel	1	N	wood rabbit
LOO	0	Y	rabbit
TD+LOO	0	Y	rabbit

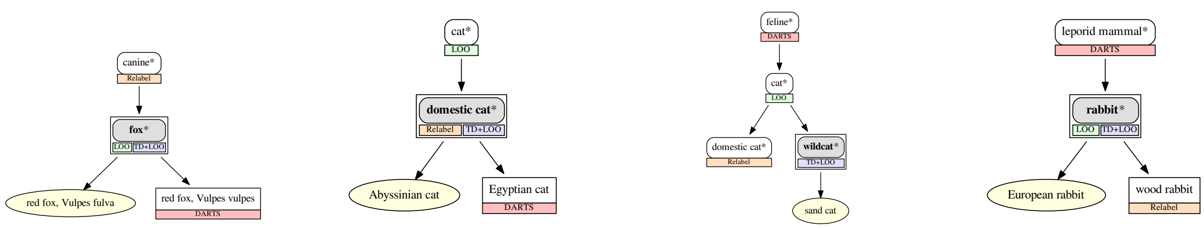


Figure B.4. Qualitative results of hierarchical novelty detection on ImageNet. "GT" is the closest known ancestor (super class) of the novel class, which is the expected prediction, "DARTS" is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

(a) (b) (c) (d)

Novel class: pika

Method	ϵ	A	Word
GT			placental mammal
DARTS	2	N	marmot
Relabel	1	N	rodent
LOO	1	N	rodent
TD+LOO	0	Y	leporid mammal

Novel class: Appaloosa

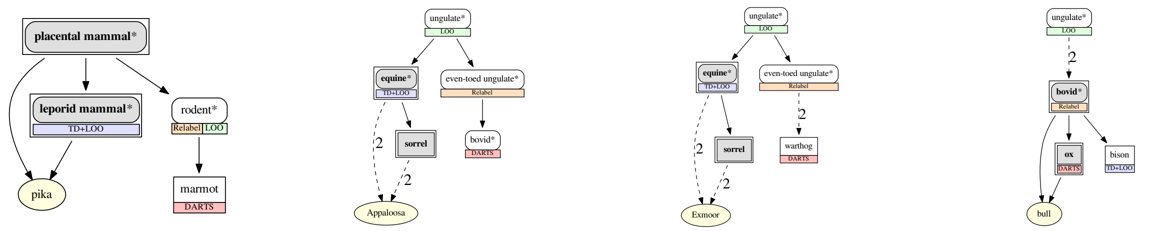
Method	ϵ	A	Word
GT			equine
DARTS	3	N	bovid
Relabel	2	N	even-toed ungulate
LOO	1	Y	ungulate
TD+LOO	0	Y	equine

Novel class: Exmoor

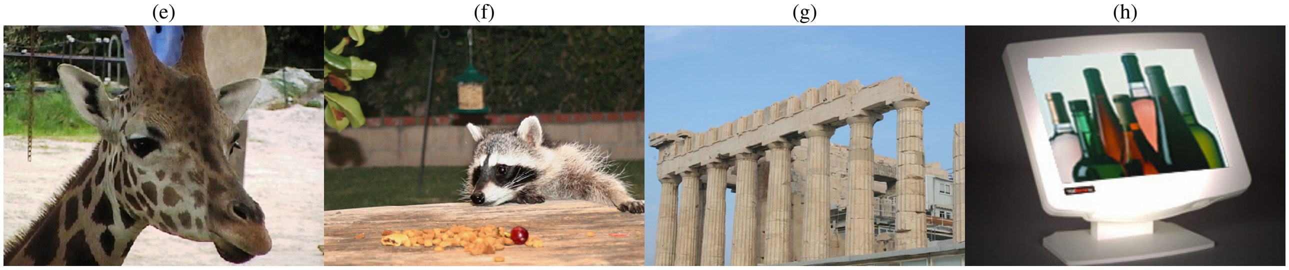
Method	ϵ	A	Word
GT			equine
DARTS	4	N	warthog
Relabel	2	N	even-toed ungulate
LOO	1	Y	ungulate
TD+LOO	0	Y	equine

Novel class: bull

Method	ϵ	A	Word
GT			bovid
DARTS	0	Y	ox
Relabel	0	Y	bovid
LOO	2	Y	ungulate
TD+LOO	1	N	bison



(e) (f) (g) (h)



Novel class: giraffe

Method	ϵ	A	Word
GT			even-toed ungulate
DARTS	1	N	antelope
Relabel	0	Y	even-toed ungulate
LOO	1	Y	ungulate
TD+LOO	2	N	equine

Novel class: raccoon

Method	ϵ	A	Word
GT			procyonid
DARTS	2	N	musteline mammal
Relabel	1	Y	carnivore
LOO	1	Y	carnivore
TD+LOO	0	Y	procyonid

Novel class: acropolis

Method	ϵ	A	Word
GT			castle
DARTS	2	N	dam
Relabel	0	Y	structure, construction
LOO	2	N	residence
TD+LOO	1	N	triumphal arch

Novel class: active matrix screen

Method	ϵ	A	Word
GT			electronic device
DARTS	4	N	personal computer
Relabel	2	Y	instrumentality
LOO	5	N	peripheral
TD+LOO	0	Y	electronic device

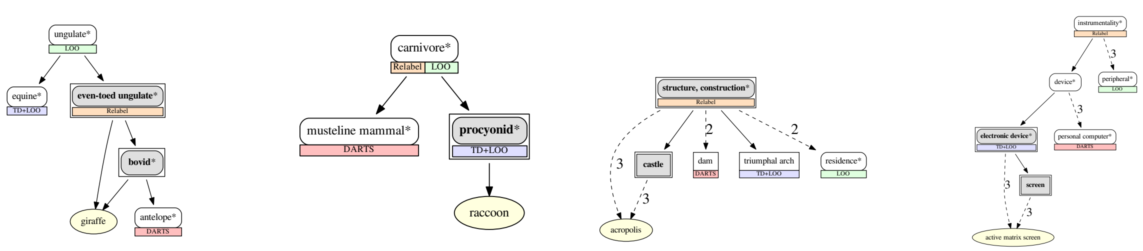


Figure B.5. Qualitative results of hierarchical novelty detection on ImageNet. “GT” is the closest known ancestor (super class) of the novel class, which is the expected prediction, “DARTS” is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. “ϵ” stands for the distance between the prediction and GT, and “A” indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

(a) (b) (c) (d)

Novel class: aisle

Method	ϵ	A	Word
GT			patio
DARTS	2	N	place of worship
Relabel	0	Y	structure, construction
LOO	3	N	church
TD+LOO	1	N	altar

Novel class: amphibian

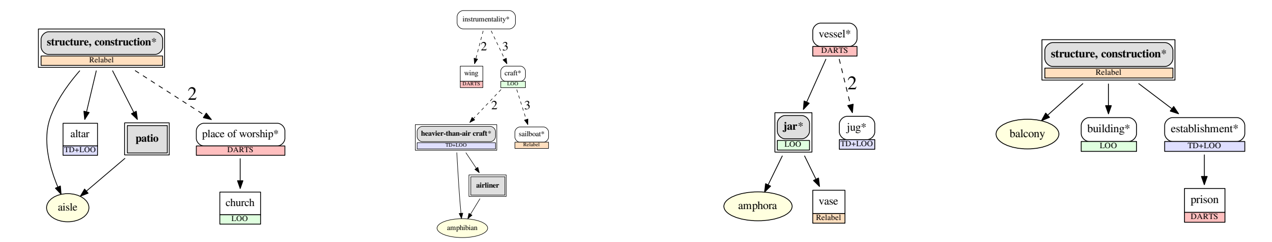
Method	ϵ	A	Word
GT			airliner
DARTS	7	N	wing
Relabel	5	N	sailboat
LOO	2	Y	craft
TD+LOO	0	Y	heavier-than-air craft

Novel class: amphora

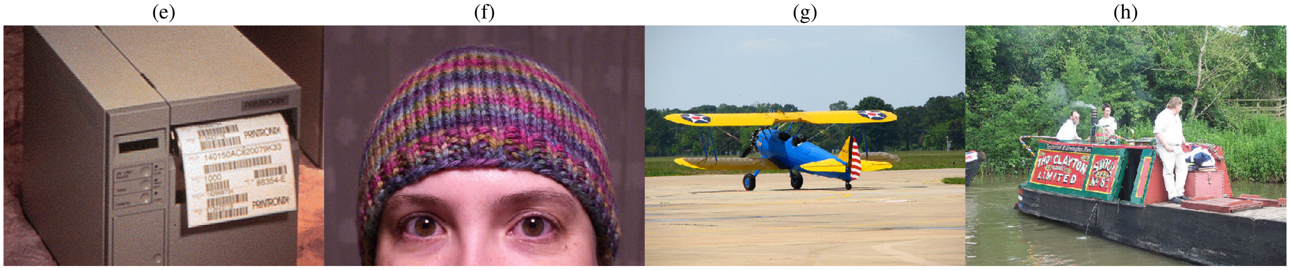
Method	ϵ	A	Word
GT			jar
DARTS	1	Y	vessel
Relabel	1	N	vase
LOO	0	Y	jar
TD+LOO	3	N	jug

Novel class: balcony

Method	ϵ	A	Word
GT			structure, construction
DARTS	2	N	prison
Relabel	0	Y	structure, construction
LOO	1	N	building
TD+LOO	1	N	establishment



(e) (f) (g) (h)



Novel class: bar printer

Method	ϵ	A	Word
GT			machine
DARTS	1	Y	peripheral
Relabel	2	Y	electronic equipment
LOO	0	Y	machine
TD+LOO	0	Y	printer

Novel class: beanie

Method	ϵ	A	Word
GT			cap
DARTS	6	N	wool
Relabel	2	N	hat
LOO	5	N	mask
TD+LOO	6	N	ski mask

Novel class: biplane

Method	ϵ	A	Word
GT			airliner
DARTS	7	N	wing
Relabel	7	N	parachute
LOO	1	Y	aircraft
TD+LOO	0	Y	heavier-than-air craft

Novel class: canal boat

Method	ϵ	A	Word
GT			boat
DARTS	3	Y	vehicle
Relabel	7	N	structure, construction
LOO	9	N	shed
TD+LOO	0	Y	boat

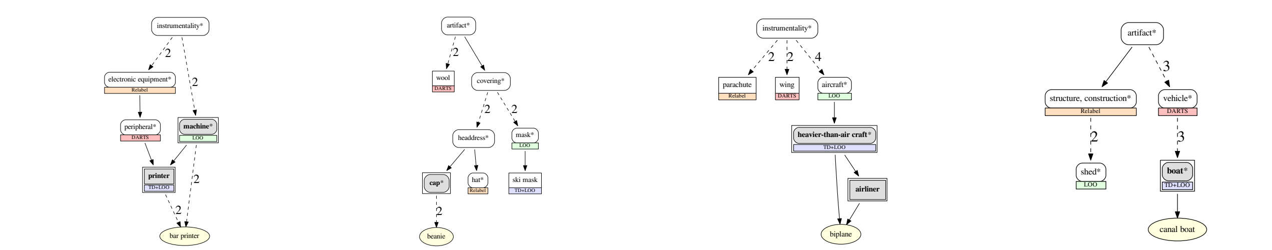


Figure B.6. Qualitative results of hierarchical novelty detection on ImageNet. "GT" is the closest known ancestor (super class) of the novel class, which is the expected prediction, "DARTS" is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

(a) (b) (c) (d)

(a)

Novel class: cassette tape			
Method	ϵ	A	Word
GT			device
DARTS	3	N	cassette
Relabel	1	Y	instrumentality
LOO	2	N	measuring instrument
TD+LOO	0	Y	hard disc

(b)

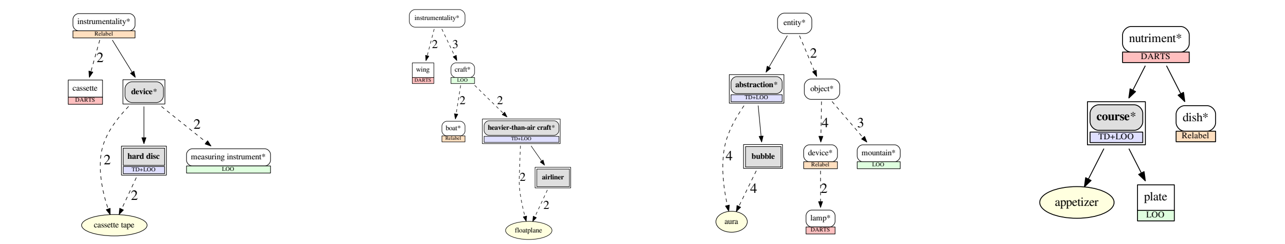
Novel class: floatplane			
Method	ϵ	A	Word
GT			airliner
DARTS	7	N	wing
Relabel	4	N	boat
LOO	2	Y	craft
TD+LOO	0	Y	heavier-than-air craft

(c)

Novel class: aura			
Method	ϵ	A	Word
GT			abstraction
DARTS	9	N	lamp
Relabel	7	N	device
LOO	6	N	mountain
TD+LOO	0	Y	abstraction

(d)

Novel class: appetizer			
Method	ϵ	A	Word
GT			course
DARTS	1	Y	nutriment
Relabel	2	N	dish
LOO	1	N	plate
TD+LOO	0	Y	course



(e) (f) (g) (h)

(e)

Novel class: hors d'oeuvre			
Method	ϵ	A	Word
GT			course
DARTS	1	N	plate
Relabel	2	N	dish
LOO	1	Y	nutriment
TD+LOO	0	Y	course

(f)

Novel class: BLT sandwich			
Method	ϵ	A	Word
GT			sandwich
DARTS	2	Y	nutriment
Relabel	1	N	cheeseburger
LOO	0	Y	sandwich
TD+LOO	0	Y	sandwich

(g)

Novel class: kale			
Method	ϵ	A	Word
GT			cruciferous vegetable
DARTS	0	Y	cruciferous vegetable
Relabel	1	Y	vegetable
LOO	1	Y	vegetable
TD+LOO	0	Y	head cabbage

(h)

Novel class: cranberry			
Method	ϵ	A	Word
GT			edible fruit
DARTS	0	Y	fruit
Relabel	0	Y	edible fruit
LOO	1	N	pomegranate
TD+LOO	0	Y	strawberry

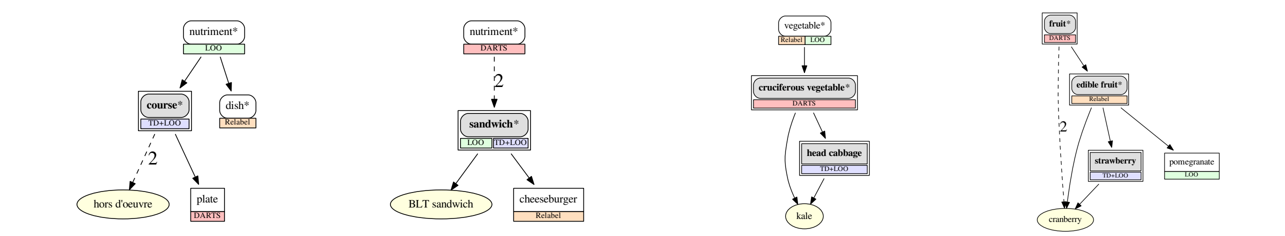


Figure B.7. Qualitative results of hierarchical novelty detection on ImageNet. "GT" is the closest known ancestor (super class) of the novel class, which is the expected prediction, "DARTS" is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

(a)	(b)	(c)	(d)

Novel class: cherry

Method	ϵ	A	Word
GT			edible fruit
DARTS	3	N	solanaceous vegetable
Relabel	1	N	Granny Smith
LOO	0	Y	fruit
TD+LOO	4	N	bell pepper

Novel class: cream sauce

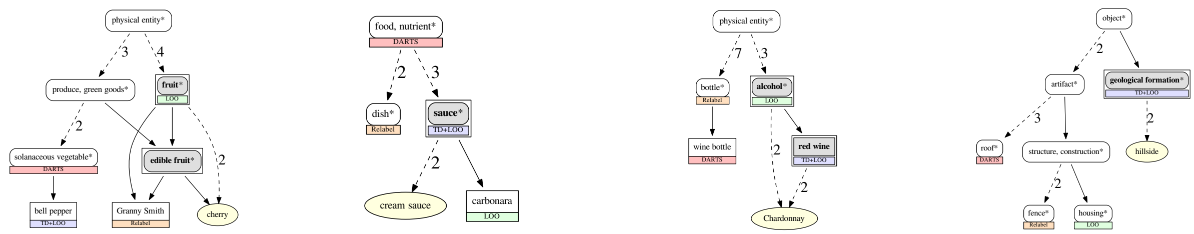
Method	ϵ	A	Word
GT			sauce
DARTS	3	Y	food, nutrient
Relabel	5	N	dish
LOO	1	N	carbonara
TD+LOO	0	Y	sauce

Novel class: Chardonnay

Method	ϵ	A	Word
GT			alcohol
DARTS	11	N	wine bottle
Relabel	10	N	bottle
LOO	0	Y	alcohol
TD+LOO	0	Y	red wine

Novel class: hillside

Method	ϵ	A	Word
GT			geological formation
DARTS	6	N	roof
Relabel	6	N	fence
LOO	5	N	housing
TD+LOO	0	Y	geological formation



(e)	(f)	(g)	(h)

Novel class: heliophila

Method	ϵ	A	Word
GT			flower
DARTS	3	N	earthstar
Relabel	8	N	vegetable
LOO	1	Y	organism, being
TD+LOO	0	Y	flower

Novel class: tangle orchid

Method	ϵ	A	Word
GT			flower
DARTS	6	N	pot, flowerpot
Relabel	1	N	daisy
LOO	8	N	vegetable
TD+LOO	0	Y	flower

Novel class: rose mallow

Method	ϵ	A	Word
GT			organism, being
DARTS	5	N	pot, flowerpot
Relabel	7	N	vegetable
LOO	0	Y	organism, being
TD+LOO	0	Y	flower

Novel class: jasmine

Method	ϵ	A	Word
GT			organism, being
DARTS	6	N	jar
Relabel	1	N	daisy
LOO	0	Y	organism, being
TD+LOO	0	Y	flower

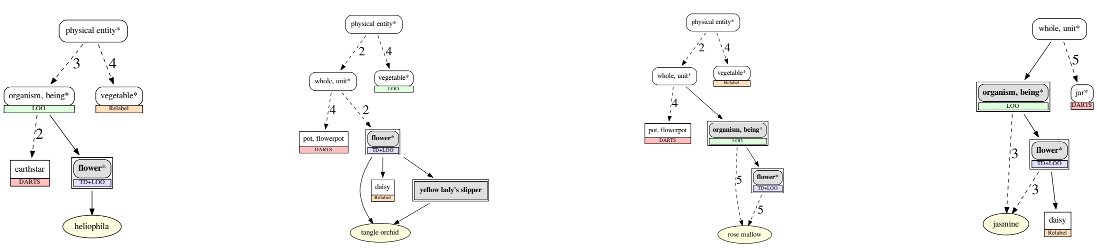
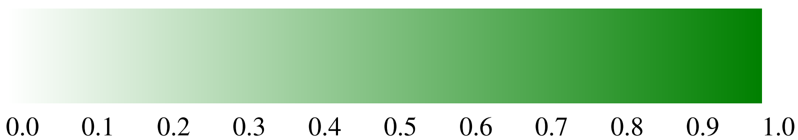


Figure B.8. Qualitative results of hierarchical novelty detection on ImageNet. "GT" is the closest known ancestor (super class) of the novel class, which is the expected prediction, "DARTS" is the baseline method proposed in [2] where we adapt their method to our task, and the others are our proposed methods. "ϵ" stands for the distance between the prediction and GT, and "A" indicates whether the prediction is an ancestor of GT. Dashed edges represent multi-hop connection, where the number indicates the number of edges between classes. If the prediction is on a super class (marked with * and rounded), then the test image is classified as a novel class whose closest class in the taxonomy is the super class.

C. Class-wise qualitative results

In this section, we show class-wise qualitative results on ImageNet. We compared four different methods: DARTS [2] is a baseline method where we adapt their method to our task, and the others, Relabel, LOO, and TD+LOO, are our proposed methods. In a sub-taxonomy, for each test class and method, we show the statistics of the hierarchical novelty detection results of known leaf classes in Figure C.1–C.2, and that of novel classes in Figure C.3–C.6. Each sub-taxonomy is simplified by only showing test classes predicted with a probability greater than 0.03 in at least one method and their common ancestors. The probability is represented in colored nodes as well as the number below the English word of the class, where the color scale is displayed in each page. Note that the summation of the probabilities shown may be less than 1, since some classes with a probability less than 0.03 are omitted. In the graphs, known leaf classes are in rectangle, and super classes are rounded and starred. If the prediction is on a super class, then the test image is classified as a novel class whose closest class in the taxonomy is the super class. We remark that most of the incorrect prediction is in fact not very far from the ground truth, which means that the prediction still provides useful information. While our proposed methods tend to find fine-grained classes, DARTS gives to more coarse-grained classes, where one can find the trend clearly in deep sub-taxonomies. Also, Relabel sometimes fails to predict the correct label but closer ones with a high probability which can be seen as the effect of relabeling.



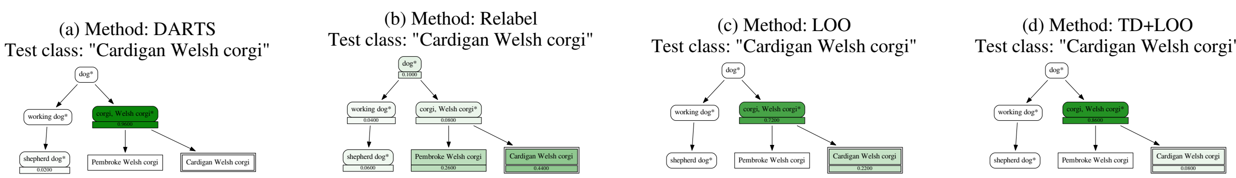


Figure C.1. Sub-taxonomies of the hierarchical novelty detection results of a known leaf class "Cardigan Welsh corgi." (Best viewed when zoomed in on a screen.)

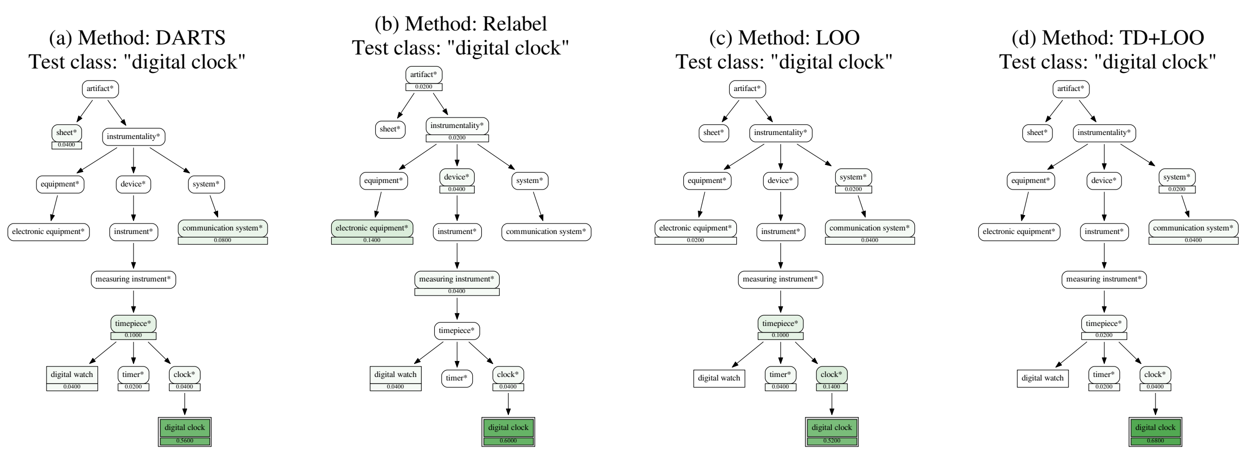


Figure C.2. Sub-taxonomies of the hierarchical novelty detection results of a known leaf class "digital clock." (Best viewed when zoomed in on a screen.)

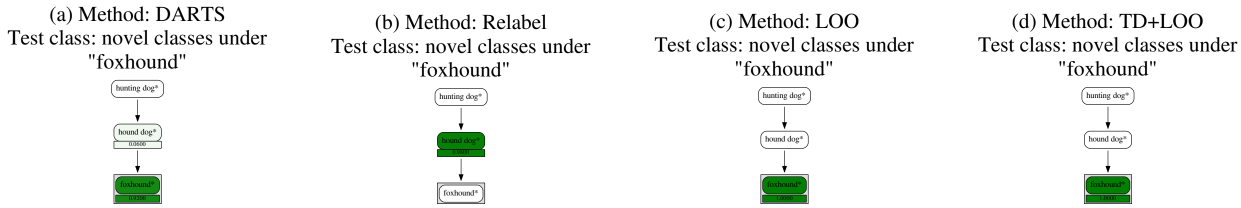


Figure C.3. Sub-taxonomies of the hierarchical novelty detection results of novel classes whose closest class in the taxonomy is "foxhound." (Best viewed when zoomed in on a screen.)

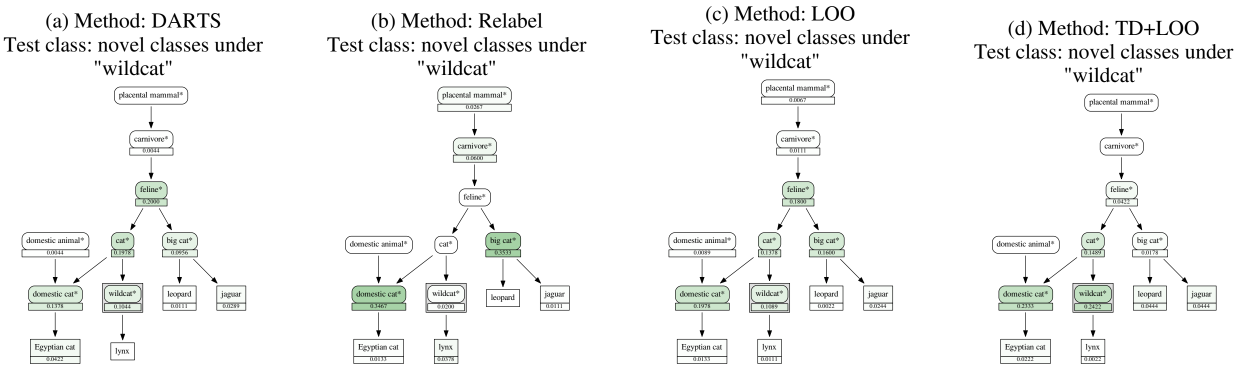


Figure C.4. Sub-taxonomies of the hierarchical novelty detection results of novel classes whose closest class in the taxonomy is "wildcat." (Best viewed when zoomed in on a screen.)

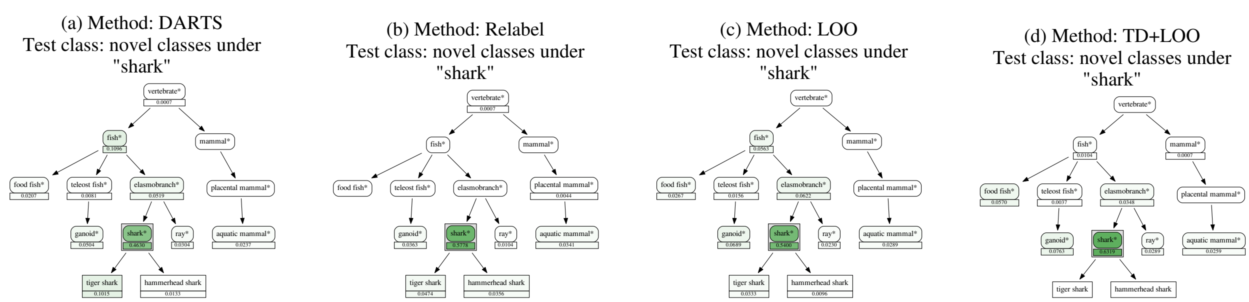


Figure C.5. Sub-taxonomies of the hierarchical novelty detection results of novel classes whose closest class in the taxonomy is "shark." (Best viewed when zoomed in on a screen.)

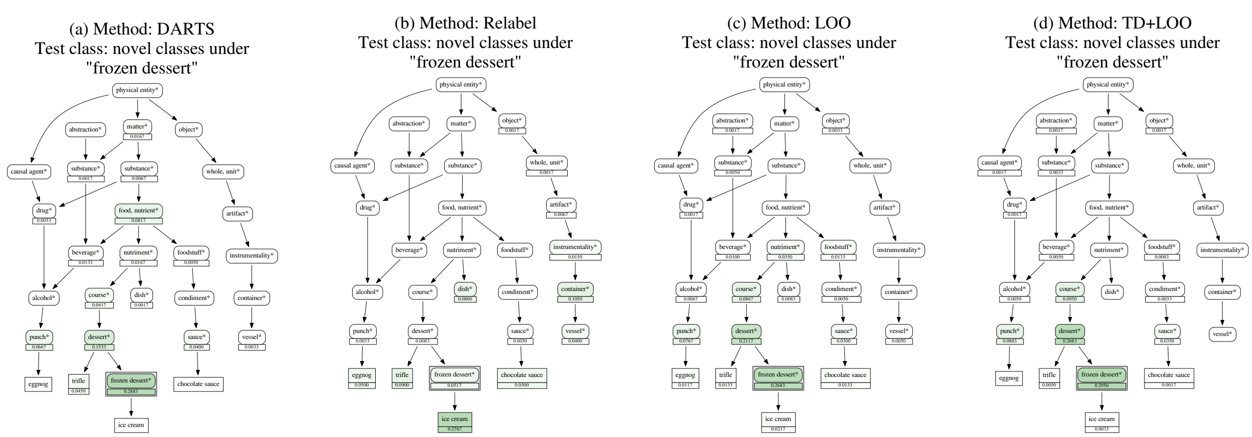
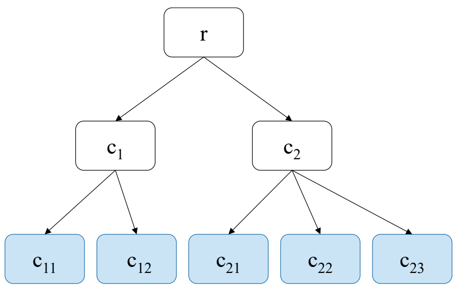


Figure C.6. Sub-taxonomies of the hierarchical novelty detection results of novel classes whose closest class in the taxonomy is "frozen dessert." (Best viewed when zoomed in on a screen.)

D. More on generalized zero-shot learning

D.1. Example of top-down embedding

Here we provide an example of the ideal output probability vector t^y in a simple taxonomy, where t^y corresponds to the concatenation of the ideal output of the top-down method when the input label is y.



$$
\begin{array}{rl}
t^y = & [\quad t^{(y,r)}, \qquad\qquad t^{(y,c_1)}, \qquad\qquad t^{(y,c_2)} \qquad] \\
\hline
t^r = & [1/2,\ 1/2,\ 1/2,\ 1/2,\ 1/3,\ 1/3,\ 1/3] \\
t^{c_1} = & [\ 1\ ,\ \ 0\ ,\ 1/2,\ 1/2,\ 1/3,\ 1/3,\ 1/3] \\
t^{c_2} = & [\ 0\ ,\ \ 1\ ,\ 1/2,\ 1/2,\ 1/3,\ 1/3,\ 1/3] \\
t^{c_{11}} = & [\ 1\ ,\ \ 0\ ,\ \ 1\ ,\ \ 0\ ,\ 1/3,\ 1/3,\ 1/3] \\
t^{c_{12}} = & [\ 1\ ,\ \ 0\ ,\ \ 0\ ,\ \ 1\ ,\ 1/3,\ 1/3,\ 1/3] \\
t^{c_{21}} = & [\ 0\ ,\ \ 1\ ,\ 1/2,\ 1/2,\ \ 1\ ,\ \ 0\ ,\ \ 0\ ] \\
t^{c_{22}} = & [\ 0\ ,\ \ 1\ ,\ 1/2,\ 1/2,\ \ 0\ ,\ \ 1\ ,\ \ 0\ ] \\
t^{c_{23}} = & [\ 0\ ,\ \ 1\ ,\ 1/2,\ 1/2,\ \ 0\ ,\ \ 0\ ,\ \ 1\ ]
\end{array}
$$

Figure D.1. An example of taxonomy and the corresponding t^y values.

D.2. Evaluation: Generalized zero-shot learning on different data splits

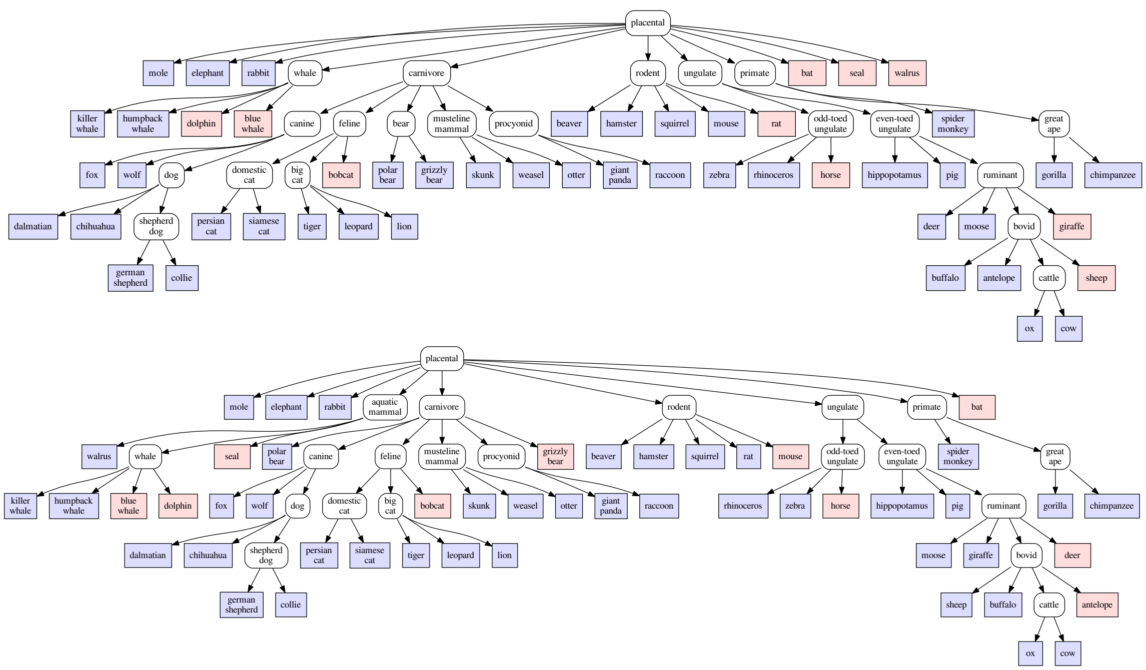


Figure D.2. Taxonomy of AwA built based on the split proposed in [5] (top) and the split we propose for balanced taxonomy (bottom). Taxonomy is built with known leaf classes (blue) by finding their super classes (white), and then novel classes (red) are attached for visualization.

We present the quantitative results on a different split of AwA1 and AwA2 in this section. We note that the seen-unseen split of AwA proposed in [5] has an imbalanced taxonomy as shown in the top of Figure D.2. Specifically, three classes belong to the root class, and another two classes belong to the same super class. To show the importance of balanced taxonomy, we make another seen-unseen split for balancing taxonomy, while unseen classes are ensured not to be used for training the CNN feature extractor. The taxonomy of new split is shown in the bottom of Figure D.2.

Table D.1 shows the performance of the attribute, word, and path embedding model, the hierarchical embedding model derived from the proposed top-down method, and their combinations on AwA1 and AwA2 with the split with imbalanced taxonomy [5] and the split with balanced taxonomy. Compared to the imbalanced taxonomy case, in the balanced taxonomy, the standalone performance of hierarchical embeddings has similar tendency, but the overall performance is better in all cases. However, in the combined model, while path embedding does not improve the performance much, top-down embedding still shows improvement on both ZSL and GZSL tasks. Note that the combination with the top-down model has lower ZSL performance than the combination without the top-down model, because only AUC is the criterion for optimization.

Compared to the best single semantic embedding model (with attributes), the combination with the top-down embedding leads to absolute improvement of AUC by 1.66 % and 4.85 % in the split we propose for balanced taxonomy on AwA1 and AwA2, respectively.

These results imply that with more balanced taxonomy, the hierarchy of labels can be implicitly learned without a hierarchical embedding such that the performance is generally better, but yet the combination of an explicit hierarchical embedding improves the performance.

Table D.1. (G)ZSL performance of semantic embedding models and their combinations on AwA1 and AwA2 in the split with imbalanced taxonomy [5] and the split with balanced taxonomy. "Att" stands for continuous attributes labeled by human, "Word" stands for word embedding trained with the GloVe objective [4], and "Hier" stands for the hierarchical embedding, where "Path" is proposed in [1], and "TD" is output of the proposed top-down method. "Unseen" is the accuracy when only unseen classes are tested, and "AUC" is the area under the seen-unseen curve where the unseen class score bias is varied for computation. The curve used to obtain AUC is shown in Figure D.3. Values in bold indicate the best performance among the combined models.

AwA1			Imbalanced		Balanced	
Att	Word	Hier	Unseen	AUC	Unseen	AUC
✓			65.29	50.02	65.86	54.18
	✓		51.87	39.67	54.29	42.40
✓	✓		67.80	52.84	**67.32**	55.40
		Path	42.57	30.58	53.40	41.63
✓		Path	67.09	51.45	65.86	54.18
	✓	Path	52.89	40.66	58.49	45.62
✓	✓	Path	68.04	53.21	**67.32**	55.40
		TD	33.86	25.56	40.38	31.39
✓		TD	66.13	54.66	65.86	54.18
	✓	TD	56.14	46.28	57.88	47.63
✓	✓	TD	**69.23**	**57.67**	66.41	**55.84**

(a) AwA1

AwA2			Imbalanced		Balanced	
Att	Word	Hier	Unseen	AUC	Unseen	AUC
✓			63.87	51.27	71.21	59.51
	✓		54.77	42.21	59.60	46.83
✓	✓		65.76	53.18	72.89	60.60
		Path	44.34	33.44	60.45	48.13
✓		Path	66.58	53.50	71.87	60.08
	✓	Path	55.28	42.86	66.83	53.05
✓	✓	Path	67.28	54.31	73.04	60.89
		TD	31.84	24.97	45.33	36.76
✓		TD	66.86	57.49	72.75	62.79
	✓	TD	59.67	49.39	65.29	53.40
✓	✓	TD	**68.80**	**59.24**	**75.09**	**64.36**

(b) AwA2

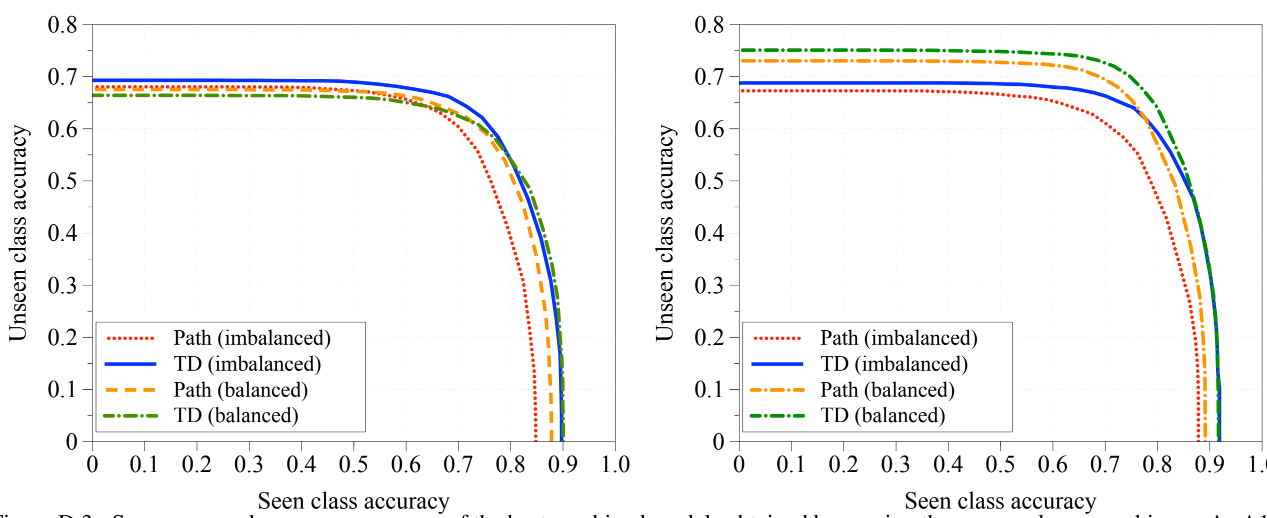


Figure D.3. Seen-unseen class accuracy curves of the best combined models obtained by varying the unseen class score bias on AwA1 and AwA2, with the split with imbalanced taxonomy [5] and the split with balanced taxonomy. "Path" is the hierarchical embedding proposed in [1], and "TD" is the embedding of the multiple softmax probability vector obtained from the proposed top-down method. We remark that if the dataset has a balanced taxonomy, the overall performance can be improved.

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 15

[2] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *CVPR*, pages 3450–3457. IEEE, 2012. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

[3] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. 2

[4] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014. 15

[5] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*, 2017. 14, 15