Fine-grain Inference on Out-of-Distribution Data with Hierarchical Classification

Randolph Linderman[1] **Jingyang Zhang**[1] **Nathan Inkawhich**[2] **Hai Li**[1] **Yiran Chen**[1]
[1] Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708
{first}.{last}@duke.edu

[2]Information Directorate
Air Force Research Laboratory
Rome, NY 13441
nathan.inkawhich@us.af.mil

Abstract

Machine learning methods must be trusted to make appropriate decisions in real-world environments, even when faced with out-of-distribution (OOD) samples. Many current approaches simply aim to detect OOD examples and alert the user when an unrecognized input is given. However, when the OOD sample significantly overlaps with the training data, a binary anomaly detection is not interpretable or explainable, and provides little information to the user. We propose a new model for OOD detection that makes predictions at varying levels of granularity—as the inputs become more ambiguous, the model predictions become coarser and more conservative. The code available at https://github.com/rwl93/hierarchical-ood.

1 Introduction

Recent studies have shown that fine-grained OOD samples are significantly more difficult to detect, especially when there is a large number of training classes [1, 5, 13, 15, 6]. We argue that the difficulty stems from trying to address two opposing objectives: learning semantically meaningful features to discriminate between ID classes while also maintaining tight decision boundaries to avoid misclassification on fine-grain OOD samples [1, 5]. We hypothesize that additional information about the relationships between classes could help determine those decision boundaries and simultaneously offer more interpretable predictions.

To address these challenges, we propose a new method based on hierarchical classification. The approach is illustrated in Figure 1. Rather than directly outputting a distribution over all possible classes, as in a flat network, hierarchical classification methods leverage the relationships between classes to produce conditional probabilities for each node in the tree. This can simplify the classification problem since each node only needs to distinguish between its children, which are far fewer in number [10, 12]. It can also improve the interpretability of the neural network [14]. For example, we leverage these conditional probabilities to define novel OOD metrics for hierarchical classifiers and make coarser predictions when the model is more uncertain.

By employing an inference mechanism that predicts at different levels of granularity, we can estimate how similar the OOD samples are from the ID set and at what node of the tree the sample becomes OOD. When outliers are encountered, predicting at lower granularity allows the system to convey imprecise, but accurate information.

2 Method

Hierarchical Classification. We define a hierarchy, $\mathcal{H}$, as a tree-structured directed acyclic graph so that there is a unique path from the root node to each leaf node. For notation, associate each node in

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

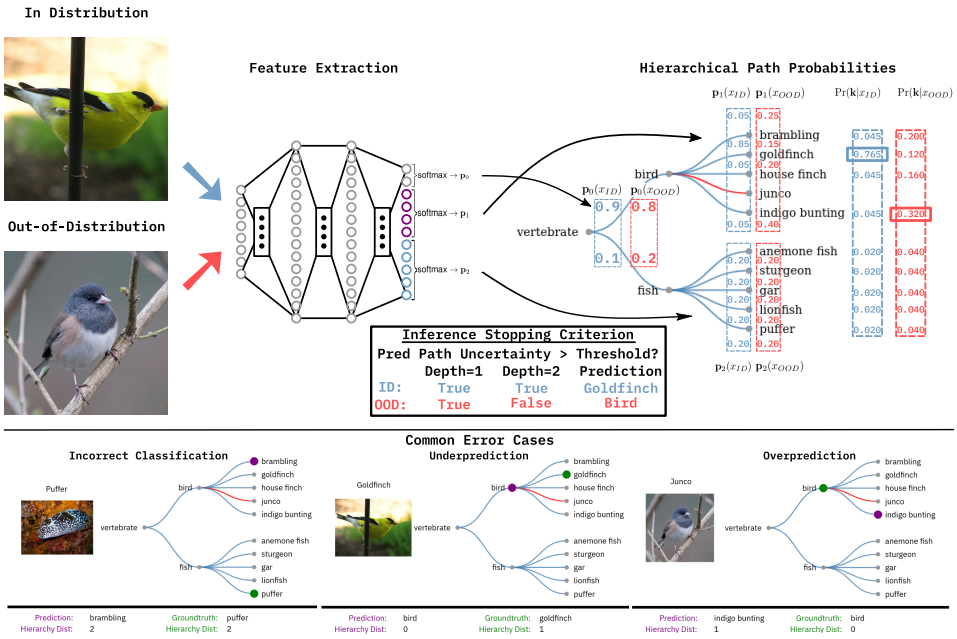


Figure 1: Method overview. *Top*: A ResNet50 extracts features from images and fully-connected layers output softmax probabilities $\mathbf{p}_n(x_i)$ for each set in the hierarchy $\mathcal{H}$. Path-wise probabilities are used for final classification. Path-wise probability and entropy thresholds generated from the training set $\mathcal{D}_{\text{train}}$ form stopping criterion for the inference process. *Bottom*: Common error cases encountered by the hierarchical predictor. From left to right: Standard error results from and incorrect intermediate or leaf decision, ID under-prediction where the network predicts at a coarse granularity due to high uncertainty, OOD over-prediction where the OOD sample is mistaken for a sibling node.

the tree with an integer $\{0, 1, \ldots, N\}$ where 0 denotes the root node. Let $\mathsf{par}(n) \in \{0, \ldots, n-1\}$ denote the parent of node n, let $\mathsf{anc}(n) \subset \{0, \ldots, n-1\}$ be the set of all ancestors of node n, and let $\mathsf{ch}(n) \subseteq \{n+1, \ldots, N\}$ denote the set of children of node n. Finally, let $\mathcal{Y} \subset \{0, 1, \ldots, N\}$ denote the set of leaf nodes (i.e. nodes for which $\mathsf{ch}(n) = \emptyset$) and $\mathcal{Z} = \{0, 1, \ldots, N\} \setminus \mathcal{Y}$ be the set of internal nodes.

Each training data point has an input $x_i \in \mathbb{R}^d$ and a label $y_i \in \mathcal{Y}$, which is associated with a leaf node of the hierarchy. The training distribution, $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}$, is comprised of tuples of input images, x_i, and associated leaf nodes, y_i. For each node n in the set of internal nodes $\mathcal{Z}$, we define $\mathcal{D}_n \subseteq \mathcal{D}_{\text{train}}$ to be all the samples (x_i, y_i) whose ancestors contain n. Likewise, define $\mathcal{D}_{\neg n}$ all the examples that whose ancestors do not contain n.

Given the input x_i, the network outputs probability distributions $\mathbf{p}_n = [p_{n,1}, \ldots, p_{n,|\mathsf{ch}(n)|}]$ for each internal node n, where $p_{n,j} \geq 0$ and $\sum_{j=1}^{|\mathsf{ch}(n)|} p_{n,j} = 1$. In practice, we model each $\mathbf{p}_n$ as a softmax function of the features in the penultimate layer of a neural network. We parameterize a distribution on leaf nodes as the product of probabilities associated with each node along that path, $\Pr(y_i = k \mid x_i) = \prod_{a \in \mathsf{anc}(k) \setminus 0} p_{\mathsf{par}(a),a}$.

Hierarchical OOD Loss. To achieve high ID accuracy and reliable OOD detection we propose a weighted multi-objective loss to optimize the hierarchical classifier. Formally, it is defined as,

$$\mathcal{L}_{\text{soft}} = \sum_{n \in \mathcal{Z}} W_n \cdot \sum_{(x,y) \in \mathcal{D}_n} H\left[\mathsf{onehot}_n(y), \mathbf{p}_n(x)\right] \tag{1}$$

$$W_n = \frac{|\{j \in \{1 \ldots N\} : n \in \mathsf{anc}(j)\}|}{N} \tag{2}$$

$$\mathcal{L}_{\text{other}} = \sum_{n \in \mathcal{Z}} \sum_{(x,y) \in \mathcal{D}_{\neg n}} H\left[\mathcal{U}(|\mathsf{ch}(n)|), \mathbf{p}_n(x)\right] \tag{3}$$

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{soft}} + \beta \cdot \mathcal{L}_{\text{other}}, \tag{4}$$

Table 1: Hierarchical softmax classifier (HSC) performance on the Imagenet-100 dataset.OOD performance is measured by AUROC scores for the fine-, medium- and coarse- OOD classes as well as the overall OOD performance. For ensemble OOD methods [7] cells follow the format: "mean$_{(\text{std})}$/ensemble". All numbers are percentages.

MODEL (METHOD)	ACCURACY	AUROC			
		FINE	MEDIUM	COARSE	OVERALL
MSP [4]	81.26(0.53)/82.75	72.47(0.31)/73.62	—	92.62(0.67)/94.38	90.25(0.62)/91.94
ODIN [9]	81.26(0.53)/82.75	72.93(1.87)/74.36	—	95.90(0.47)/**96.71**	93.20(0.36)/**94.08**
MAHALANOBIS [11]	81.26(0.53)	**78.05**(0.09)	—	91.34(0.62)	89.78(0.54)
MOS [5]	**82.41**(0.02)	70.00(0.72)	—	**96.66**(0.23)	**93.66**(0.22)
HSC ($\alpha = 1$, $\beta = 0$, PRED)	82.38(0.06)/83.25	76.78(3.38)/79.80	—	93.93(0.22)/95.08	91.33(0.28)/92.38
HSC ($\alpha = 1$, $\beta = 0$, H_{mean})	82.38(0.06)/83.25	77.27(3.83)/75.86	—	96.90(0.11) /96.89	93.92(0.20)/93.17
HSC ($\alpha = 1$, $\beta = 0.2$, PRED)	82.85(0.14)/83.33	79.40(0.76)/80.39	—	95.06(0.13)/95.48	92.29(0.15)/92.82
HSC ($\alpha = 1$, $\beta = 0.2$, H_{mean})	**82.85**(0.14)/83.33	**79.40**(0.67)/76.89	—	**97.23**(0.11)/96.79	**94.08**(0.13)/93.28

where $H[p, q]$ is the cross-entropy from p to q and $\mathsf{par}_n(y)$ is the one-hot vector for the ancestor of y corresponding to node n, $\mathsf{onehot}_n(y) = [\mathbb{1}(k \in \mathsf{anc}(y)) : k \in \mathsf{ch}(n)]$.

The first objective optimizes the network for ID classification accuracy by applying cross-entropy to the network's predictions $\Pr(\mathbf{k}|x_i) = [\Pr(y_i = k|x_i)]_{k \in \mathcal{Y}}$ for each sample in the training distribution, $\mathcal{D}_{\mathsf{train}}$ (Equation (1)) [1]. The second objective (Equation (3)) drives the probabilities at internal nodes that are not along the path from root to ground-truth node to the uniform distribution, parameterized by size, ($\mathcal{U}(s)$) with cross-entropy. This utilizes in-distribution data as outliers for all nodes in the hierarchy that are not one of its ancestors.

Prediction Path Entropy OOD Metric. We propose prediction path based OOD scoring functions for performing OOD detection with hierarchical classifiers. First, we propose using maximum prediction path probabilities calculated according to Section 2 which is hierarchical analog to max softmax probability for standard networks. Second, we propose mean path-wise entropy, $H_{\mathsf{mean}}(x_i) = \frac{1}{|\mathsf{anc}(\hat{y}_i)|} \sum_{n \in \mathsf{anc}(\hat{y}_i)} H[\mathbf{p}_n]$

3 Experiments

Fine-grain OOD datasets. Some applications may face more extreme OOD examples than others. To construct OOD detection tasks with varying degrees of difficulty, we leveraged the fact that the Imagenet-1K classes correspond to nouns in the WordNet hierarchy [2]. We generated OOD sets by holding out subsets of Imagenet-1K classes in entire subtrees of the WordNet hierarchy. Withholding large subtrees—those rooted at low depths of the hierarchy—leads to coarse-grained OOD detection tasks, since the held-out classes are very different from the training classes. Holding out small subtrees—those rooted at nodes deep in the hierarchy—leads to fine-grained OOD-detection tasks.

Results. We found that the hierarchical softmax classifier (HSC) outperformed baseline methods on the Imagenet-100 dataset (Table 1).In particular, the $\mathcal{L}_{\mathsf{other}}$ loss adds a regularization term that improves ID accuracy as well as improving OOD performance. We assessed the effect of holdout class granularity and found that the softmax-based OOD heuristics (MSP, ODIN, and prediction path probability) and Mahalanobis detectors are most sensitive to fine-grain OOD samples whereas MOS and path entropy metrics perform best on coarse-grain OOD as shown in Table 1. In Table 2, we evaluate the sensitivity to hierarchy depth and composition for Imagenet-100 datasets.We find that the performance across all OOD metrics introduced in Section 2 is comparable with no apparent benefit to visually-derived hierarchies ([14]) vs. human-defined semantic hierarchies. However, we believe that the hierarchy is a critical design choice and is likely application dependent. Specifically, the hierarchy's class balance, depth, and alignment with visual features are important characteristics to consider. In natural image classification domains, human-defined semantic structures may improve interpretability because they project image inputs into a human conceptual framework even though they may not perfectly represent the visual properties of the input.

[1] When $W_n = 1 \forall n \in \mathcal{Z}$ the form in Equation (1) is equivalent to the entropy over the leaf nodes $H[y, \Pr(k|\mathcal{D}_{\mathsf{train}})]$

4 Analysis

Hierarchical classifiers decompose the classification problem into simpler intermediate tasks. By analyzing the model's confidence at each intermediate decision, we can understand where the model becomes uncertain. Wan et al. [14] show that through analyzing intermediate decisions we can explain the model's decision process to understand where the model makes mistakes and how it behaves on ambiguous labels greatly improving the interpretability and explainability compared to softmax classifiers. We build off of this work by leveraging intermediate model confidence estimates to determine at what level of granularity to make a prediction.

First, we aim to understand the effects of OOD data on the hierarchical classifier's performance. We plot the micro-ROC curves (Figure 2) for 4 synsets each corresponding to a separate classification decision in the hierarchy. The "artifact", "dog" and "bird" synsets include one or more OOD samples and the "ball" synset does not have any corresponding OOD samples (see Figure 4). Notice in Figure 2 that when adding the activations of the OOD data ("OOD" curve) the number of false-positives increases and AUROC drops compared to the ID-only curve because the OOD data is being predicted more confidently than some ID data. This occurs across all synsets in the hierarchy even in the "ball" synset that does not contain any OOD descendants. However, when we employ a path-wise probability based threshold at 99% TNR on the training data ("THR" curves in Figure 2), the performance is recovered in all synsets. The micro-ROC curves for all synsets is displayed in Figure 10.

Next, we compare path-wise and node-wise thresholding hierarchy distance and accuracy performance to non-hierarchical OOD methods (Appendix A.3). We achieve 73% accuracy on the OOD samples while maintaining 74% ID accuracy using a path-wise probability threshold chosen at 95% TNR as witnessed by the blue line in Figure 8. However, when inspecting the hierarchy distance and accuracy vs TNR plots (Figures 3 and 8), we notice a stepped nature for the path-wise thresholding technique that underperforms and is less stable compared to the node-wise technique. The large step changes and deviation of path-wise thresholding in Figure 8 and Figure 3 reflect that the path-wise thresholds cause the network to predict at increasingly coarse nodes as the confidence degrades with increasing depth (i.e. specificity, see Section 2). When the distribution of OOD classes is balanced across granularity levels, the node-wise inference technique greatly outperforms the path-wise technique due to the stepped nature of the path-wise technique (Figures 6 and 7).

We show that OOD average hierarchy distance consistently decreases and the ID average hierarchy distance remains relatively constant (Figure 3, bottom). While the ID accuracy drops from 82.75% to 74.46% at the 95% TNR node-wise threshold, the average hierarchy distance decreases from 0.4045 to 0.4005 (Figure 3 bottom right). Therefore, by allowing the hierarchical classifier to predict with less specificity, we can improve the overall prediction quality by removing uncertain leaf node predictions. Our experiments show that by leveraging the intermediate predictions made by hierarchical classifiers, we can directly interpret, explain, and validate the model's decisions prior to deployment and improve performance on uncertain ID and OOD data.

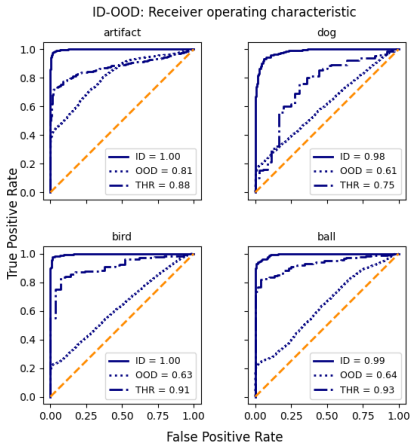


Figure 2: Micro-ROC curves for ID data, ID/OOD, and ID/OOD threshold TNR=0.95.

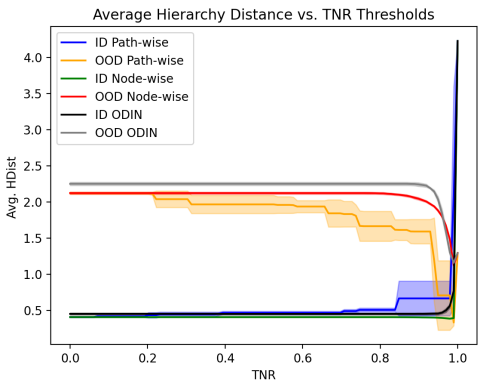


Figure 3: Imagenet-100 ID and OOD average hierarchy distance across TNR threshold values.

Acknowledgments

The views expressed in this article are those of the authors and do not reflect official policy of the United States Air Force, Department of Defense or the U.S. Government. PA# AFRL-2022-2046.

References

[1] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *AAAI*, pages 3154–3162, 2020.

[2] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[4] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[5] Rui Huang and Yixuan Li. MOS: towards scaling out-of-distribution detection for large semantic space. In *CVPR*, pages 8710–8719, 2021.

[6] Nathan A. Inkawhich, Eric K. Davis, Matthew J. Inkawhich, Uttam K. Majumder, and Yiran Chen. Training sar-atr models for reliable operation in open-world environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3954–3966, 2021. doi: 10.1109/JSTARS.2021.3068944.

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, pages 6402–6413, 2017.

[8] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177, 2018.

[9] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[10] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, pages 6517–6525, 2017.

[11] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. In *ICML'21 Workshop on Uncertainty & Robustness in Deep Learning*, 2021.

[12] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS*, 2021.

[13] Ryne Roady, Tyler L. Hayes, Ronald Kemker, Ayesha Gonzales, and Christopher Kanan. Are open set classification methods effective on large-scale datasets? *PLOS ONE*, 15(9):1–18, 2020.

[14] Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Suzanne Petryk, Sarah Adel Bargal, and Joseph E. Gonzalez. NBDT: neural-backed decision tree. In *ICLR*, 2021.

[15] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments, 2021. URL `https://arxiv.org/abs/2106.03917`.

A Appendix

A.1 Compute resources

All experiments were run on an internal compute cluster with nodes containing 8 NVIDIA RTX A5000 GPUs. Imagenet 100 experiments were trained on 1 GPU for 90 epochs which completed in $\sim$8 hours for the longest running experiments. Imagenet 1K experiments were trained on 2 GPUs with data parallelization. Training for the Imagenet 1K's longest running experiments lasted $\sim$52.5 hours.

A.2 Model Training

All models were trained from scratch as the available pretrained weights were trained on the fine-grained OOD holdout classes. We used a ResNet50 [3] backbone for all models and trained for 90 epochs of stochastic gradient descent (SGD). We used a learning rate of 0.1 with learning rate decay steps with a decay factor of 0.1 performed at epoch 30 and 60. The momentum and weight decay parameters were 0.9 and 10^{-4}, respectively. We standardized the training hyperparameters to avoid performance differences due to the optimization procedure.

A.3 Inference Stopping Criterion

Given a hierarchical classifier optimized over $\mathcal{D}_{\mathsf{train}}$, we define a stopping criterion utilizing the performance statistics on the validation data. Specifically, we select a true negative rate (TNR) on the ID data to decide our inference stopping threshold from the micro-averaged receiver operating characteristic (ROC) curve. In practice, this TNR threshold will be determined by the specific application's prediction fidelity requirements. Micro-averaged ROC curves are used to generate the TNR thresholds for each node in $\mathcal{Z}$. We utilize path probabilities $\Pr(n|x_i)$ as the threshold score.

During inference the leaf node prediction $\hat{y}$ is determined, then the prediction path $\mathsf{anc}(\hat{y})$ is traversed from root to leaf. If any of the nodes in the path do not meet the TNR threshold, the parent node is chosen as the prediction (fig. 1). Both global path probability and node-wise probability and mean-, min-entropy were explored as TNR threshold metrics.

A.4 Hierarchical Accuracy and Distance

We analyze the hierarchical classifier's inference on ID and OOD samples with top-1 accuracy, as well as, average hierarchical distance. The groundtruth for OOD samples is the closest ancestor that is contained within ID hierarchy. For example, the OOD node *junco* in 1 is assigned the ID groundtruth node *bird*.

Furthermore, we consider the inference procedure's failure modes by decomposing the hierarchy distance into two parts: (1) the prediction and (2) the groundtruth distance to their closest common parent. Hierarchy distance is defined as the number of edges in the hierarchy between two nodes. By recording the groundtruth and prediction distances to the closest common parent we can determine how frequently the model incorrectly predicts, overpredicts, and underpredicts for a set of inputs. fig. 1 (bottom) depicts common error cases that are encountered and their corresponding hierarchy distances.

A.5 Imagenet 100 Hierarchy Experiments

Table 2: ID and OOD sensitivity to hierarchy selection on the Imagenet-100 dataset. $\mathcal{H}$ type indicates whether the hierarchy is defined by human semantics or learned visual feature clustering. All numbers are percentages.

Hierarchy $\mathcal{H}$	$\mathcal{H}$ Type	Accuracy	Path Predition	Path Entropy		
				Mean	Max	Min
2 Lvl WN	Semantic	82.19(0.38)	91.73(0.17)	93.43(0.08)	92.12(0.04)	93.08(0.13)
Pruned WN	Semantic	82.38(0.06)	91.33(0.28)	93.92(0.20)	89.16(0.46)	93.70(0.13)
Binary NBDT [14]	Visual	81.28(0.48)	91.33(0.29)	92.92(0.14)	86.68(0.18)	93.15(0.24)

A.6 Hierarchy Statistics

Table 3: Imagenet dataset holdout set statistics. The number of leaf nodes that are held out due to trimmed branches at each level of granularity. The uniform probability used to choose the holdout nodes and the hierarchy depths for each granularity level are given for each dataset.

Dataset	Max Depth Internal Leafs	# Leaf Holdouts		
		Coarse	Medium	Fine
Imagenet 100	6	15	0	2
	28	—	—	—
	100	lvls 2	—	Leafs
Balanced Imagenet 100	6	15	5	10
	28	—	—	—
	100	lvls 2	4–5	Leafs

A.7 Imagenet 100 Hierarchy

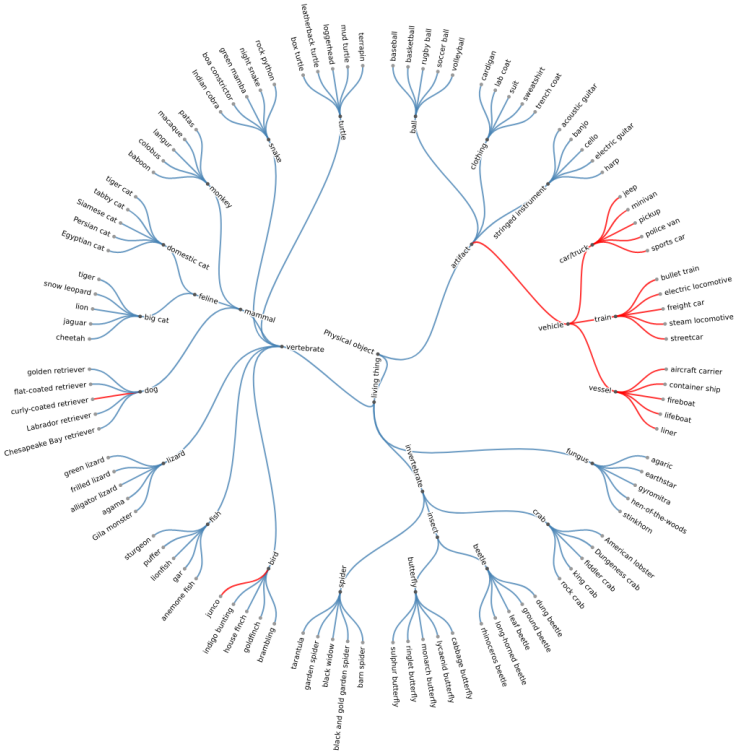


Figure 4: Imagenet-100 pruned WordNet hierarchy. Red edges correspond to OOD paths and blue to ID.

A.8 Balanced Imagenet 100 Results

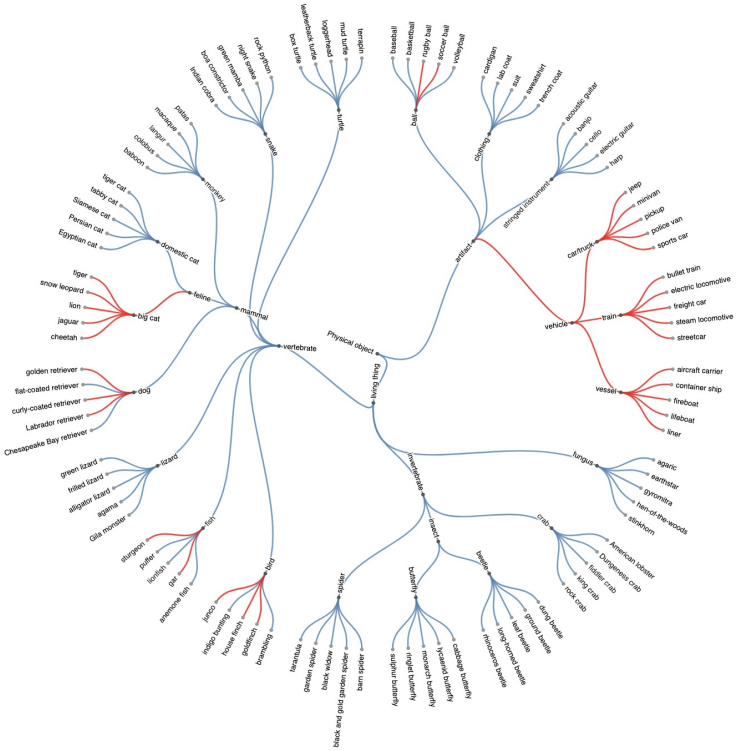


Figure 5: BALANCED IMAGENET 100 pruned WordNet hierarchy. Red edges correspond to OOD paths and blue to ID.

Table 4: Hierarchical softmax classifier (HSC) performance on the Balanced Imagenet 100 dataset. The $\mathcal{L}_{\text{soft}}$ and $\mathcal{L}_{\text{other}}$ weights (α, β) and the OOD metric are given in parenthesis for each HSC model. OOD performance is measured by AUROC scores for the fine-, medium- and coarse- OOD classes as well as the overall OOD performance. Each cell includes the performance statistics across 3 models trained with separate random seeds. For ensemble OOD methods [7] cells follow the format: "mean$_{(\text{std})}$/ensemble". Note that relative Mahalanobis [11] performance is reported as it outperformed the original method [8]. All models are ResNet50 architectures trained for 90 epochs. All numbers are percentages.

MODEL (METHOD)	ACCURACY	AUROC			
		FINE	MEDIUM	COARSE	OVERALL
BALANCED IMAGENET 100					
MSP [4]	80.85(0.23)/82.11	72.11(0.65)/73.91	71.07(0.58)/73.51	92.32(0.49)/93.66	82.04(0.28)/83.72
ODIN [9]	80.85(0.23)/82.11	79.16(0.56)/80.37	74.35(0.57)/75.84	96.09(0.63)/96.78	86.82(0.23)/87.82
MAHALANOBIS [11]	80.85(0.23)	83.07(0.80)	72.66(0.59)	91.11(0.89)	85.36(0.64)
MOS [5]	80.35(0.21)	81.49(0.65)	86.80(0.35)	74.23(1.05)	86.80(0.35)
HSC ($\alpha = 1$, $\beta = 0$, PRED)	81.19(0.26)/81.83	69.44(0.90)/71.25	71.57(1.44)/73.11	93.29(0.18)/94.49	81.72(0.44)/83.18
HSC ($\alpha = 1$, $\beta = 0$, H_{mean})	81.19(0.26)/81.83	69.81(1.01)/78.56	68.12(1.61)/72.75	96.46(0.11)/96.69	82.85(0.60)/86.66
HSC ($\alpha = 1$, $\beta = 0$, H_{max})	81.19(0.26)/81.83	66.57(0.76)	71.00(1.53)	89.46(0.23)	78.75(0.60)
HSC ($\alpha = 1$, $\beta = 0$, H_{min})	81.19(0.26)/81.83	70.72(1.85)/73.56	28.15(2.41)/71.45	95.21(0.79)/95.56	75.87(0.98)/84.21
HSC ($\alpha = 1$, $\beta = 0.2$, PRED)	81.83(0.10)/82.97	73.91(1.04)/75.52	73.88(1.19)/75.86	94.20(0.09)/95.32	84.05(0.46)/85.47
HSC ($\alpha = 1$, $\beta = 0.2$, H_{mean})	81.83(0.10)/82.97	74.23(1.01)/80.38	70.64(1.33)/74.54	96.65(0.03)/96.43	84.84(0.42)/87.43
HSC ($\alpha = 1$, $\beta = 0.2$, H_{max})	81.83(0.10)/82.97	71.29(0.89)/80.74	73.18(1.02)/74.23	92.68(0.14)/95.12	82.30(0.44)/86.84
HSC ($\alpha = 1$, $\beta = 0.2$, H_{min})	81.83(0.10)/82.97	73.72(1.98)/81.94	27.26(0.85)/75.26	95.76(0.37)/96.33	77.00(0.44)/88.02

A.9 Balanced 100 hierarchy distance and accuracy

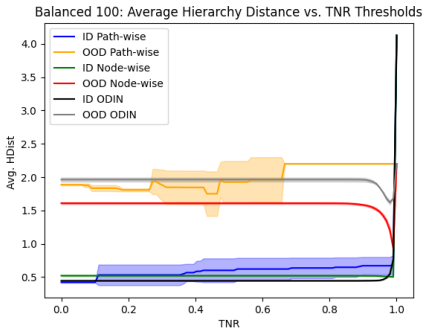


Figure 6: Balanced Imagenet 100 average hierarchy distance vs. TNR threshold values.

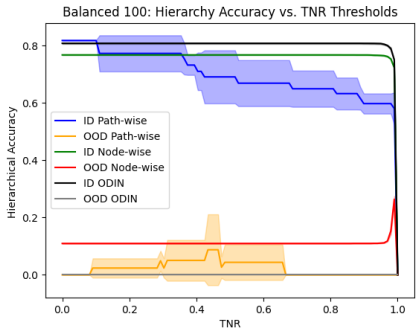


Figure 7: Balanced Imagenet 100 hieararchy accuracy vs. TNR threshold values.

A.10 ID and OOD Inference accuracy vs. TNR

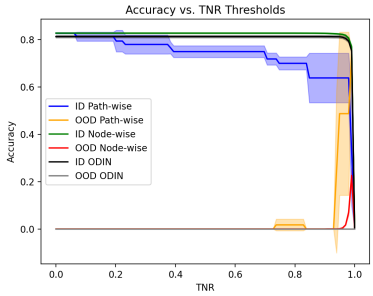
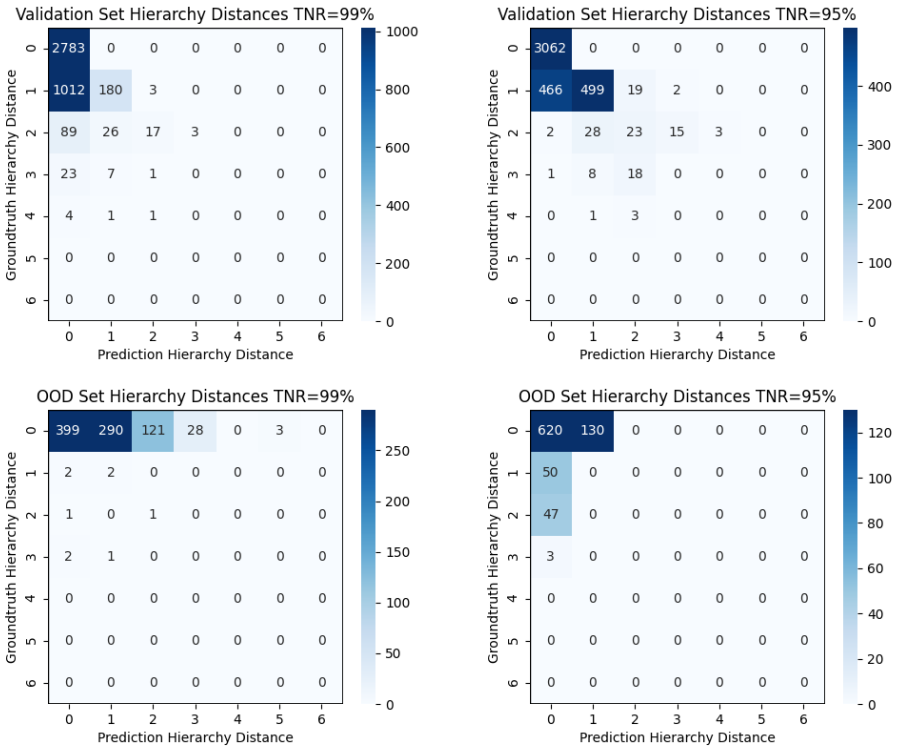


Figure 8: Imagenet-100 ID and OOD accuracy across TNR threshold values for path-wise and synset-wise threshold metrics with ODIN baseline.

A.11 Supplemental hierarchy distance confusion matrices



(a) Node-wise Inference (b) Path-wise Inference

Figure 9: Imagenet-100 path- and node-wise inference hierarchy distance confusion matrices on ID and OOD data.

A.12 Supplemental Micro-ROC curves

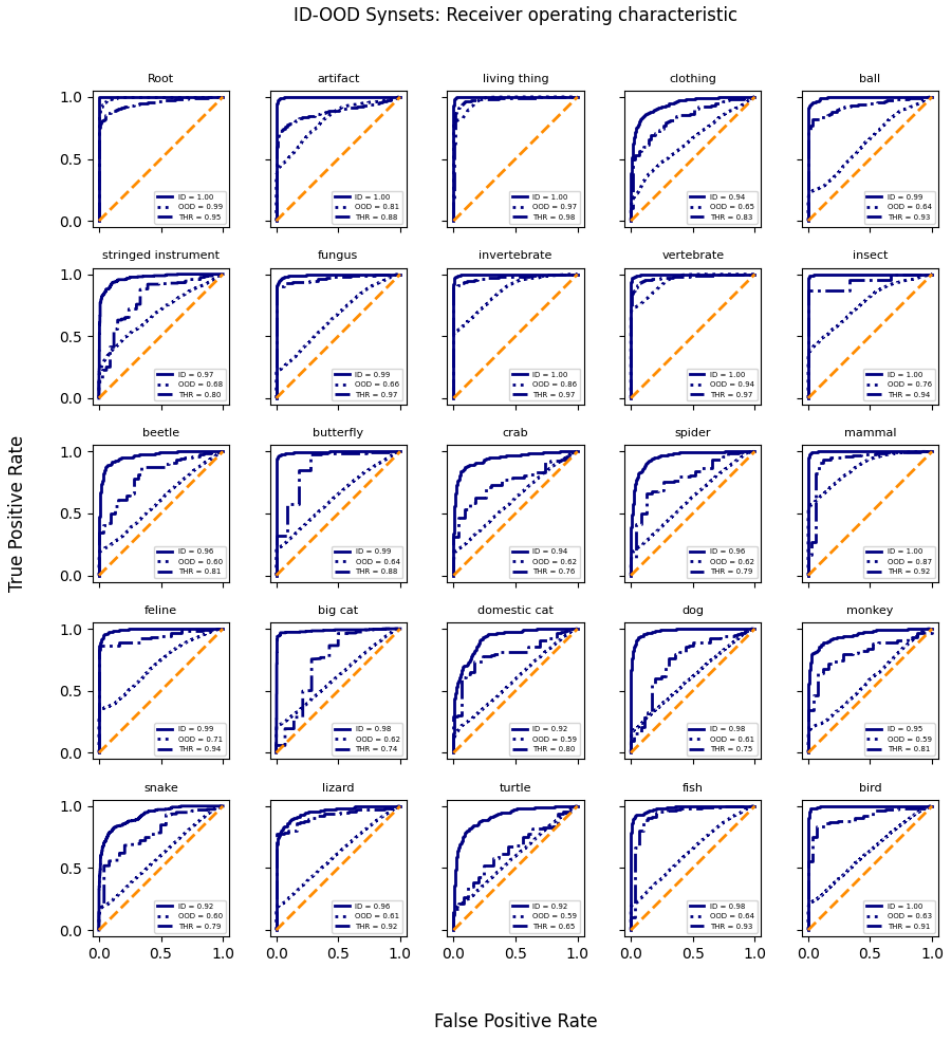


Figure 10: Imagenet-100 synset micro-ROC curves for ID data only, ID and OOD, and ID and OOD with a TNR=0.95 prediction path threshold.

A.13 Fully Connected Head Experiments

Table 5: Hierarchical softmax classifier (HSC) performance on the Imagenet-100 dataset when adding additions fully-connected (FC) layers to classification head. AUROC scores are provided for each OOD method: MSP [4], ODIN [9], MOS [14]. Note that the node-wise scaling of the HSC methods is different from table 1 causing the discripancy in HSC numbers. All numbers are percentages.

Model	Accuracy	Baseline AUROC	Path Prediction	Path Entropy	
				Mean	Min
		Imagenet 100			
MSP	81.26(0.53)	90.25(0.62)	—	—	—
ODIN	81.26(0.53)	93.20(0.36)	—	—	—
MOS	82.41(0.02)	93.66(0.22)	—	—	—
MSP FC3	82.12(0.29)	90.68(0.41)	—	—	—
ODIN FC3	82.12(0.29)	93.78(0.31)	—	—	—
MOS FC3	81.82(0.15)	93.49(0.49)	—	—	—
HSC ($\alpha = 1, \beta = 0$)	78.36(0.87)	—	89.09(0.65)	92.39(0.43)	92.68(0.24)
HSC ($\alpha = 1, \beta = 0.2$)	83.05(0.12)	—	92.16(0.26)	93.93(0.27)	94.29(0.27)
HSC FC3 ($\alpha = 1, \beta = 0$)	81.90(0.08)	—	91.51(0.35)	93.86(0.20)	93.46(0.18)
HSC FC3 ($\alpha = 1, \beta = 0.2$)	82.73(0.24)	—	91.80(0.46)	93.78(0.28)	94.57(0.20)

A.14 Supplemental OOD granularity performance

Table 6: Hierarchical classifier performance on the fine-grain Imagenet-1K datasets. AUROC scores are provided for each OOD method: (Ours) Hierarchical softmax classifier (HSC), MSP [4], ODIN [9], MOS [14]. All models are ResNet50 architectures trained for 90 epochs. FC: Number of fully-connected layers for classifier

Dataset	Model	AUROC		
		Fine	Medium	Coarse
Imagenet 100	MSP	72.47(0.31)	—	92.62(0.67)
	ODIN	72.93(1.87)	—	95.90(0.47)
	MOS	70.00(0.72)	—	96.66(0.23)
	MSP FC3	72.55(1.22)	—	93.09(0.39)
	ODIN FC3	69.05(0.75)	—	97.08(0.26)
	MOS FC3	69.78(2.65)	—	96.66(0.21)
	HSC ($\alpha = 1, \beta = 0$, pred)	74.63(0.92)	—	91.02(0.67)
	HSC ($\alpha = 1, \beta = 0$, H_{mean})	73.09(0.95)	—	94.96(0.46)
	HSC ($\alpha = 1, \beta = 0$, H_{min})	62.24(0.72)	—	96.74(0.25)
	HSC ($\alpha = 1, \beta = 0.2$, pred)	71.11(1.98)	—	94.96(0.04)
	HSC ($\alpha = 1, \beta = 0.2$, H_{mean})	70.01(1.93)	—	97.11(0.07)
	HSC ($\alpha = 1, \beta = 0.2$, H_{min})	65.14(1.40)	—	98.18(0.13)
	HSC FC3 ($\alpha = 1, \beta = 0$, pred)	72.22(0.56)	—	94.09(0.44)
	HSC FC3 ($\alpha = 1, \beta = 0$, H_{mean})	71.55(0.49)	—	96.83(0.28)
	HSC FC3 ($\alpha = 1, \beta = 0$, H_{min})	59.69(1.57)	—	97.96(0.01)
	HSC FC3 ($\alpha = 1, \beta = 0.2$, pred)	71.86(0.17)	—	94.46(0.53)
	HSC FC3 ($\alpha = 1, \beta = 0.2$, H_{mean})	70.67(0.25)	—	96.86(0.33)
	HSC FC3 ($\alpha = 1, \beta = 0.2$, H_{min})	68.87(0.11)	—	97.99(0.23)

A.15 Performance on far-OOD benchmarks

Table 7: Coarse-grain OOD dataset baseline performance. AUROC scores are provided for each OOD method: (Ours) Hierarchical softmax classifier (HSC), MSP [4], ODIN [9], MOS [14]. All models are ResNet50 architectures trained for 90 epochs. FC: Number of fully-connected layers for classifier

ID Dataset	Model	AUROC			
		iNaturalist	SUN	Places	Textures
iNaturalist 100	MSP	92.22(0.46)	93.62(0.23)	92.24(0.17)	88.60(0.57)
	ODIN	95.60(0.34)	97.30(0.19)	96.10(0.13)	94.85(0.35)
	MOS	93.50(0.04)	95.85(0.04)	94.72(0.12)	95.13(0.21)
	MSP FC3	93.04(0.23)	94.77(0.03)	93.43(0.13)	90.07(0.13)
	ODIN FC3	96.52(0.10)	98.00(0.07)	96.87(0.04)	95.89(0.12)
	MOS FC3	93.83(0.32)	95.89(0.20)	94.83(0.14)	94.78(0.18)
	HSC ($\alpha = 1$, $\beta = 0$, PRED)	91.22(0.27)	92.64(0.59)	91.25(0.76)	87.69(0.03)
	HSC ($\alpha = 1$, $\beta = 0$, H_{mean})	91.83(0.26)	93.40(0.55)	92.33(0.71)	89.83(0.04)
	HSC ($\alpha = 1$, $\beta = 0$, H_{min})	86.50(0.82)	94.19(0.68)	92.66(0.87)	92.88(0.06)
	HSC ($\alpha = 1$, $\beta = 0.2$, PRED)	94.05(0.40)	95.73(0.19)	94.59(0.03)	91.51(0.15)
	HSC ($\alpha = 1$, $\beta = 0.2$, H_{mean})	94.38(0.29)	96.31(0.18)	95.45(0.02)	93.44(0.14)
	HSC ($\alpha = 1$, $\beta = 0.2$, H_{min})	90.73(0.39)	96.30(0.19)	95.06(0.19)	95.22(0.22)
	HSC FC3 ($\alpha = 1$, $\beta = 0$, PRED)	93.25(0.48)	95.25(0.36)	93.80(0.36)	90.82(0.21)
	HSC FC3 ($\alpha = 1$, $\beta = 0$, H_{mean})	94.09(0.47)	96.08(0.29)	94.87(0.28)	92.79(0.17)
	HSC FC3 ($\alpha = 1$, $\beta = 0$, H_{min})	91.93(0.47)	96.79(0.12)	95.53(0.10)	95.29(0.19)
	HSC FC3 ($\alpha = 1$, $\beta = 0.2$, PRED)	93.83(0.20)	95.68(0.12)	94.32(0.18)	90.82(0.29)
	HSC FC3 ($\alpha = 1$, $\beta = 0.2$, H_{mean})	94.13(0.10)	96.19(0.14)	95.16(0.12)	92.69(0.22)
	HSC FC3 ($\alpha = 1$, $\beta = 0.2$, H_{min})	91.20(0.76)	96.42(0.14)	95.14(0.11)	94.93(0.06)