Divide and Conquer: Object Co-occurrence Helps Mitigate Simplicity Bias in OOD Detection

Supplementary Material

7. Dempster-Shafer Theory

Let Ω be a finite set called the frame of discernment, which represents all possible states or hypotheses in a given context. The power set of Ω, denoted as 2^Ω, contains all possible subsets of Ω including the empty set $\emptyset$. A basic probability assignment (BPA) or mass function m is a mapping from 2^Ω to $[0, 1]$ that satisfies:

$$m(\emptyset) = 0 \text{ and } \sum_{\mathcal{A}\subseteq\Omega} m(\mathcal{A}) = 1 \tag{10}$$

For any subset $\mathcal{A} \subseteq \Omega$, $m(\mathcal{A})$ represents the degree of evidence supporting exactly $\mathcal{A}$, not including any of its proper subsets. Based on the mass function, the belief function Bel and plausibility function Pl are defined as:

$$Bel(\mathcal{A}) = \sum_{\mathcal{B}\subseteq\mathcal{A}} m(\mathcal{B}) \tag{11}$$

$$Pl(\mathcal{A}) = \sum_{\mathcal{B}\cap\mathcal{A}\neq\emptyset} m(\mathcal{B}) \tag{12}$$

where $Bel(\mathcal{A})$ represents the total belief committed to $\mathcal{A}$ and all its subsets, while $Pl(\mathcal{A})$ measures the total belief that does not contradict $\mathcal{A}$. The interval $[Bel(\mathcal{A}), Pl(\mathcal{A})]$ can be interpreted as the lower and upper bounds of the probability of $\mathcal{A}$. Given two pieces of evidence represented by mass functions m_1 and m_2 from independent sources, Dempster's rule of combination $\oplus$ defines their fusion as:

$$(m_1 \oplus m_2)(\mathcal{A}) = K^{-1} \sum_{\mathcal{B}\cap C=\mathcal{A}} m_1(\mathcal{B})m_2(\mathcal{C}) \tag{13}$$

where $K = 1-\sum_{\mathcal{B}\cap\mathcal{C}=\emptyset} m_1(\mathcal{B})m_2(\mathcal{C})$ is the normalization factor, and $K \neq 0$. This combination rule provides a formal mechanism for evidence fusion and belief updating.

In OCO, to reduce complexity, we only consider the categories predicted from aggregated features and slot predictions, rather than examining all 2^K possible combinations.

8. Experiment

8.1. Datasets

SSB-hard consists of 49,000 images covering 980 categories selected from ImageNet-21K. Classes outside ImageNet-1K but still within the ImageNet, making it semantically close and thus near OOD.
NINCO is a noise-free dataset of 5,879 images manually curated and verified by humans to ensure complete freedom from noise and contamination from ImageNet-1K classes.
iNaturalist offers natural world species images that differ significantly from ImageNet-1K's object categories.
Textures presents a fundamentally different challenge by focusing on textural patterns rather than object recognition.
OpenImage-O is carefully curated from the Open Images dataset, and provides diverse image content with 1,763 images specifically designated for validation.
ImageNet-c: it contains 15 corruption types (e.g., noise, blur, weather effects) each with 5 severity levels. The benchmark randomly samples 10K images across these 75 corruption-severity combinations for evaluation.
ImageNet-r: it tests generalization to artistic renditions, presenting ImageNet objects in various styles including sketches, paintings, and cartoons.
ImageNet-v2: it tests generalization under data collection bias. This dataset helps evaluate whether models are truly learning robust features or are overfitting to specific characteristics of the original ImageNet distribution.

8.2. Baselines

Energy is a post-processing method based on energy functions, which transforms the model's logits into energy scores to distinguish between ID and OOD samples.
MaxLogit utilizes the maximum logit value as the detection score, avoiding the smoothing effect introduced by the softmax function to achieve more discriminative OOD detection.
SHE integrates energy functions with feature prototypes, utilizing a hybrid approach that combines both energy-based scoring and prototype-based feature comparison.
NNguide presents a hybrid approach that combines energy-based scoring with feature-space nearest neighbor distance, leveraging both energy functions and feature similarities for detection.
SCALE implements a thresholded energy function approach that requires no access to training data, making it particularly practical for deployment scenarios.
NECO employs PCA reconstruction distance as its core mechanism for OOD detection, measuring the dissimilarity between original and reconstructed features.
FDBD identifies OOD samples by locating and leveraging the decision boundaries in the ID feature space, focusing on the geometric properties of the learned representations.
CoRP applies non-linear kernel methods to detect out-of-distribution data, using explicit feature mappings with co-

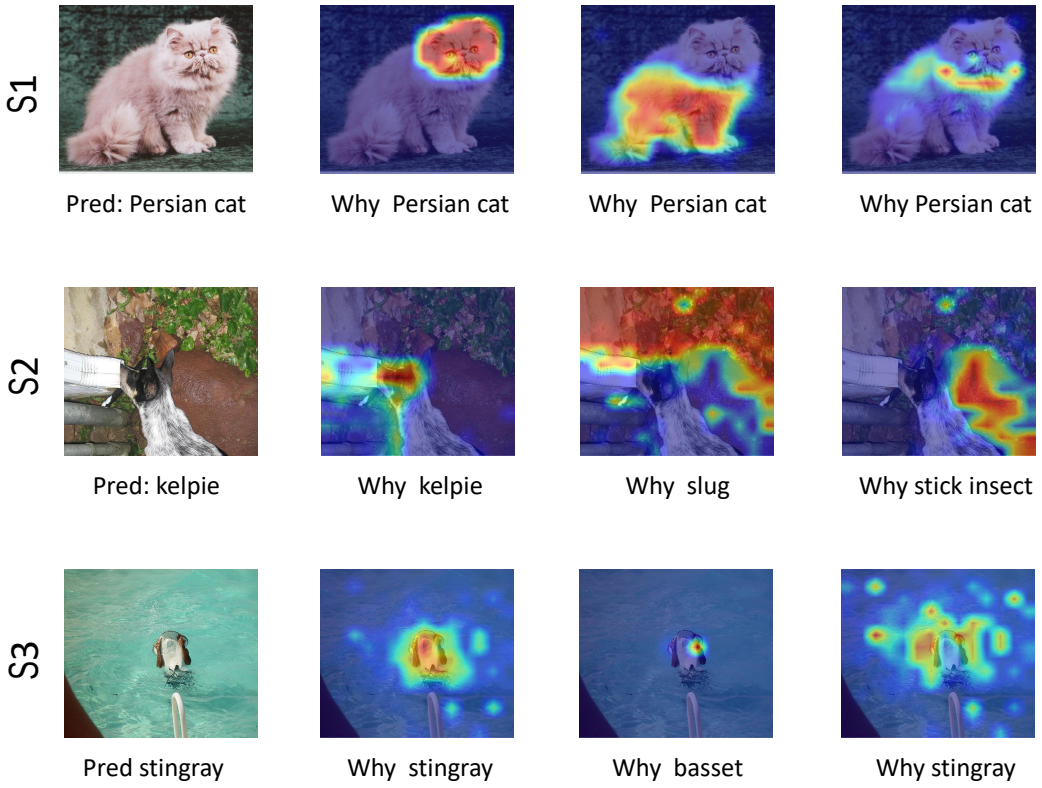

Figure 8. Attention visualization of ID.

sine and cosine-Gaussian kernels to improve detection performance while maintaining computational efficiency. Our method is primarily based on probabilistic scoring, leveraging Maximum Softmax Probability (MSP) to normalize all scores within the [0,1] interval. This probabilistic formulation enables a natural representation of OOD scores while ensuring consistent scaling across all three scenarios, thereby avoiding potential scale inconsistency issues that may arise in detection.

OODD is a test-time OOD detection method that constructs a dynamic OOD dictionary during inference, continuously collecting representative latent OOD features from test samples without requiring fine-tuning. It further combines informative inlier sampling with a priority queue-based update mechanism and a dual OOD stabilization strategy to calibrate OOD scores based on both ID-feature similarity and dynamically accumulated OOD-feature similarity, thereby improving detection performance under evolving test-time OOD scenarios.

8.3. Training Details

We present the training parameters in detail, as shown in Tab. 6. We employed vanilla slot attention in our implementation.

	ImageNet-1k
Fine-tune epochs	20
Batch size	64
Initial LR	4×10^{-4}
Final LR	4×10^{-5}
LR schedule	cosine
K in Eq. 1	6
Slot Dim.	256

Table 6. Configurations of OCO.

8.4. Quantitative Results

In this section, we visualized the regions of attention focus by plotting attention scores for each slot across different scenarios' ID images (See Fig. 8). When no object co-occurrence (first row), the sample contains only one object besides the background. Here, each slot attends to the holistic features of the cat, leading to the prediction of `Persian cat`. When ID object co-occurrence appears (second row), the scene is more complex. The slots attend to both the `kelpie's` features and the background, resulting in predictions of `slug` and `stick insect`.

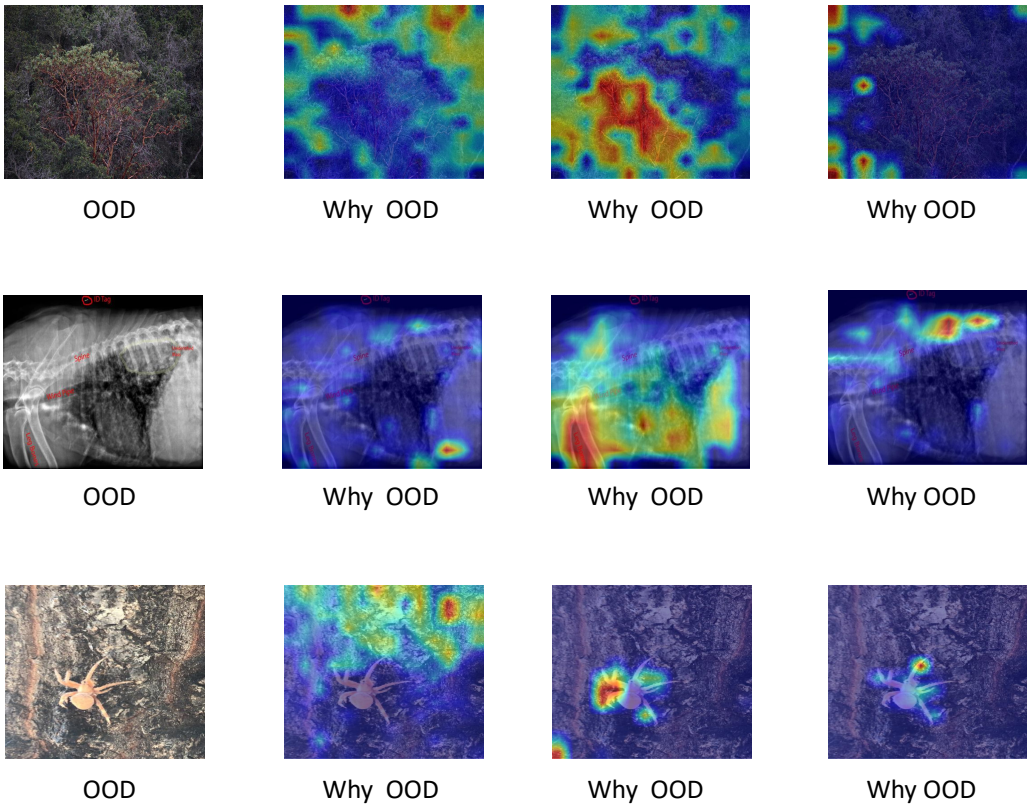

Figure 9. Attention visualization of OOD.

From a human visual perspective, the background indeed shares similar visual features with `slugs`. When OOD object co-occurrence appears (third row), the scene presents higher complexity with human arms intersecting with a `stingray`. Initially, the slots capture the human arm features and misidentify them as `basset`. However, the model correctly identifies the object as a `stingray` upon detecting the marine context and more comprehensive features. The attention scores for each slot in OOD scenarios are shown in Fig. 9. When processing OOD samples, the attention distribution across slots exhibits significantly higher dispersion.

8.5. Visualization Results for Each Scenario

Our OCO demonstrates a clear separation between OOD distributions, as illustrated in Fig. 10, especially in co-occurrence scenarios S2 and S3, where the advantages over traditional methods are more pronounced.

9. Limitation

In this section, we discuss the limitations of our method. Our model only represents object co-occurrence patterns based on slot attention. While slot attention is currently the state-of-the-art method for extracting object-centric representations, it is limited to a fixed number of slots. When the number of slots exceeds the number of objects, the representation of certain object edges may not be extracted effectively, resulting in average performance on small target objects. In the future, we can try to introduce a lightweight network to estimate the quantity, improving the dynamic number of slots, thereby making the object occurrence pattern more robust.

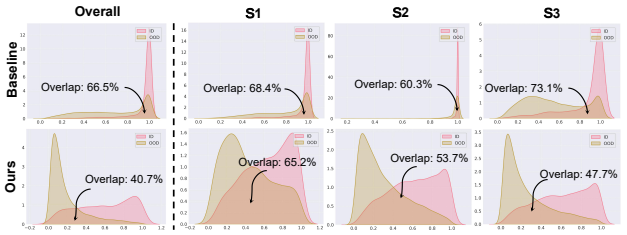

Figure 10. Score distributions for ViT model on ImageNet-200 (ID) and SSB-hard (OOD) (Left). Overall comparison of vanilla Maximum Softmax Probability (MSP) vs. OCO scores (Right).