28th *Computer Vision Winter Workshop*
Graz, Austria, February 12–14, 2025

DOI: 10.3217/978-3-99161-022-9-005

Leveraging Intermediate Representations for Better Out-of-Distribution Detection

Gianluca Guglielmo[1] Marc Masana[1,2]

[1]Institute of Visual Computing, TU Graz
[2]SAL Dependable Embedded Systems, Silicon Austria Labs
{guglielmo, mmasana}@tugraz.at

Abstract

In real-world applications, machine learning models must reliably detect Out-of-Distribution (OoD) samples to prevent unsafe decisions. Current OoD detection methods often rely on analyzing the logits or the embeddings of the penultimate layer of a neural network. However, little work has been conducted on the exploitation of the rich information encoded in intermediate layers. To address this, we analyze the discriminative power of intermediate layers and show that they can positively be used for OoD detection. Therefore, we propose to regularize intermediate layers with an energy-based contrastive loss, and by grouping multiple layers in a single aggregated response. We demonstrate that intermediate layer activations improves OoD detection performance by running a comprehensive evaluation across multiple datasets.

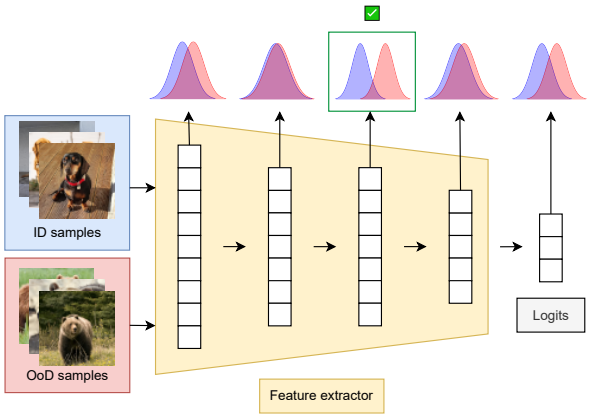

Figure 1. Intermediate representations are often more informative than the logits when dealing with OoD detection.

1. Introduction

When a model is exposed to data which does not belong to the distribution it was originally trained on, it is desirable that it can detect it and respond appropriately. Therefore, it is beneficial for a machine learning framework to include an *Out-of-Distribution* (OoD) detection mechanism, especially in real-world scenarios. Without it, the model might produce unreliable or even dangerous outputs when confronted with data from an unfamiliar distribution, leading to potential failures in critical applications such as autonomous driving, healthcare, or financial systems [1, 42]. Deep neural networks perform well in many applications but can be overly confident with unseen classes [30]. A key feature would be the ability to avoid providing (overconfident) predictions for unknown classes. Implementing this safety mechanism should not interfere with the intended tasks of the model, such as correctly classifying the samples from the *In-Distribution* (ID) data [42]. However, achieving a balance between ID performance and OoD detection, presents significant challenges. Furthermore, OoD detection mechanisms ought to perform efficiently, without imposing an excessive computational overhead or diminishing the capacity of the model when performing on the original task. Although recent advances have lead to promising strategies [25, 27, 42], developing methods that achieve this dual goal remains a pressing challenge for the design of trustworthy, scalable AI systems [7].

Building on these challenges, deep learning models have emerged as a powerful solution, becoming the preferred framework for constructing complex training pipelines [9]. These models address the need for effective OoD detection by leveraging their hierarchical architectures, which enable the learning and encoding of *mid-level features* [31], which are representations that bridge low-level patterns such as edges and textures to high-level abstract feature maps such as object parts and semantic categories [12]. In computer vision, these features are inherently diverse, capturing the hierarchical nature of the input data. This diversity not only makes them able to generalize to tasks within the original in-distribution but also highly transferable to new, related tasks. This has been shown within transfer learning scenarios [11].

For OoD detection, most methods rely on penultimate layer embeddings or on the logits of the model [42]. The potential of leveraging ensembles of intermediate layer embeddings remains under-explored [24]. We argue that

mid-level features alone can act as reliable stand-alone OoD-indicators. For instance, inputs with semantically unrelated characteristics compared to the training data may trigger unusual activations in specific layers, serving as an early warning of abnormality.

We pose that specific hidden layers can be effectively isolated and used to enhance OoD detection, performing better than the final layer. However, leveraging these intermediate representations may yield different results depending on the type of shift from ID data -— whether semantic or covariate. Also, this may result in varying effects depending on how close or distant the ID and OoD distributions are.

Building on this insight, we test with an aggregated approach, which leverages hidden-layer information in a layer-agnostic manner. This method avoids relying on specific layers, which enables a more robust and generalized use of the network's intermediate representations for OoD tasks. Further, we also propose an approach that regularizes selected hidden layers through an energy-based contrastive loss, improving OoD detection by leveraging their intermediate representations. The goal is to promote the information encoded in the hidden spaces to be distributed such that OoD detection is more efficient, and without disrupting the ID task performance.

Therefore, our contributions are summarized as:

- we establish that the embeddings of hidden layers are valuable for OoD detection,
- we introduce a layer-agnostic aggregated (Ag-EBO) approach that leverages intermediate representations,
- we propose a modular strategy to enhance robustness by regularizing specific layers (R-EBO).

The article is structured as follows: Sec. 2 presents a current overview of the OoD field, while Sec. 3 introduces the preliminaries needed for the study in Sec. 4, which shows that intermediate layers contain useful information for the OoD detection task. An overview of the proposed ways of exploiting this capabilities, with detailed results is presented in Sec. 5. Finally, in Sec. 6 we discuss some limitations of the proposed approaches and the main takeaways.

2. Related Work

In this paper, we concentrate on two main families of Out-of-Distribution methods: *post-hoc* and *training-based* [25, 42].

Post-hoc methods are applied after the model has been trained and typically involve analyzing its predictions or intermediate representations to identify whether an input is OoD [14, 17, 41]. These methods often focus on computational efficiency and adaptability to pre-trained models, as they avoid retraining [25].

Training-based methods modify the training process, sometimes completely restructuring the model to accommodate OoD detection [6, 20, 34]. These methods often come at the cost of higher training complexity, and might dilute the efforts to obtain an optimal ID training accuracy [25, 28]. Additionally, exposure to outliers (real or generated) can be done to improve generalization [39].

Baselines. A classic baseline for OoD is considered to be Maximum Softmax Probability (MSP) [17], a simple approach that relies on the logit scores to identify OoD samples. However, a major limitation of this approach is the tendency of models to produce overconfident predictions on anomalous data, leading to poor performance [14]. Temperature scaling [14] is a simple post-hoc way of tackling the overconfidence issue, where logits are scaled by a temperature T, but its results are not optimal [43].

OoD and intermediate layers. Some methods leverage intermediate embeddings within the network. However, most do it to refine the head's detection capabilities, rather than for direct OoD detection. ASH [8] enhances the network's OoD detection capabilities through activation masking of hidden layers. Similarly, ReAct [32] proposes to rectify the embeddings of the penultimate layer to reduce overconfidence. However, despite leveraging intermediate embeddings to an extent, the final detection decisions in both methods rely solely on the output logits. Mahalanobis distance-based method (MDSEns) [24] uses features from hidden layers to compute distances from the known distribution. However, this approach relies on the assumption that the class-conditional distributions of hidden layer features are Gaussian, which may not hold true for complex datasets and deep network architectures [36]. Head2Toe [11] leverages intermediate representations by training a classifier head on concatenated embeddings from multiple hidden layers to improve generalization during *transfer learning*. This enables the refinement of existing OoD detection techniques through the utilization of hidden layer structures.

3. Out-of-Distribution Detection

3.1. Problem statement

In Out-of-Distribution (OoD) detection, the objective is to differentiate between samples generated by the same distribution as the in-distribution dataset, $\mathcal{D}_{\text{in}}$, and those originating from a different, out-of-distribution dataset, $\mathcal{D}_{\text{out}}$. Due to the complexity and variance of image-based data, the concept of the amount of *out-of-distributioness* of samples is inherently challenging to define. However, two primary types of distributional shifts are commonly identified [35]:

- **Semantic (or Concept) shifts:** they arise when new classes appear at test time. For instance, encountering an image of a dog after the model has been trained on pictures of cats and mice.
- **Covariate shifts:** occur when the style or attributes of samples change within the same class. Examples include image corruptions [16], such as artifacts, blurs or noise, and domain changes [18, 38], such as shifting from natural photographs to artistic paintings.

Both semantic and covariate shifts can occur with varying levels of severity depending on the problem, and can also appear entangled within a distribution shift. Given a fixed $\mathcal{D}_{in}$, we refer to *near* and *far* OoD datasets as those that are semantically closer to or further from it, respectively.

Moreover, depending on the OoD detection application, different shifts might be considered within the spectrum that comprises between novelty and anomaly detection [28]. The first relates to distribution shift that might need to be explicitly added to the model, while the second is usually added in a more implicit way, in order to efficiently use the capacity of the model. In this paper, we do not distinguish samples based on the suitability for further learning, but instead aim to analyze these shifts from a perspective of distribution similarity.

Terminology. Consider a neural network $f(\mathbf{x};\theta)$ with input $\mathbf{x}$ and parameters θ, and trained to classify C classes. The architecture of the network is defined as a series of L layers with intermediate functions such that:

$$y = f(\mathbf{x};\theta) = (f_L^{\theta_L} \circ f_{L-1}^{\theta_{L-1}} \circ \cdots \circ f_1^{\theta_1})(\mathbf{x}),$$

where the output y is a vector of C logits representing the unnormalized prediction over the classes. Therefore, the intermediate representations or embeddings of a given layer l are defined as:

$$a_l = (f_l \circ \cdots \circ f_1)(\mathbf{x}).$$

To determine whether an input $\mathbf{x}$ belongs to $\mathcal{D}_{\text{in}}$ or $\mathcal{D}_{\text{out}}$, a score function $\mathcal{S}(\mathbf{x})$, is usually derived from the neural network. This score reflects the confidence of the model in the input belonging to the expected in-distribution. A threshold T is applied to classify the input such that:

$$g(\mathbf{x}) = \begin{cases} \mathbf{x} \in \mathcal{D}_{\text{in}} & \text{if } \mathcal{S}(\mathbf{x}) \geq T \\ \mathbf{x} \in \mathcal{D}_{\text{out}} & \text{if } \mathcal{S}(\mathbf{x}) < T. \end{cases}$$

The threshold can be adjusted depending on the desired balance between sensitivity and specificity for OoD detection.

Metrics. In order to evaluate the strength of a method, two essentials metrics are AUROC, the Area Under the Receiver Operating Characteristic (the higher the better) and FPR@TPR95, the False Positive Rate when the True Positive Rate is 95% (the lower the better).

3.2. Energy-based out-of-distribution detection.

Energy-based models [23] have demonstrated to be effective as post-hoc OoD detectors. The *free energy function* $E(\mathbf{x};\mathbf{f})$ is defined as:

$$E(\mathbf{x};\mathbf{f}) = -T \log \sum_{c=1}^{C} e^{f^c(\mathbf{x})/T}, \tag{1}$$

where T is the temperature, for temperature scaling [14]. When $T = 1$, it simplifies to the negative log of the denominator of the softmax function, which represents the normalization factor in the softmax computation. In this case, the energy function effectively captures the aggregate contribution of all logits, weighted by their exponential, to produce a measure of confidence over the entire output distribution. The Energy-Based OoD (EBO) [26] detection approach uses the free energy associated to each input to determine whether it is ID or OoD, where the higher the energy is, the more likely the sample is OoD. JEM [13] is another energy-based approach that improves the calibration (the mismatch between accuracy and confidence) of the model.

4. OoD with Intermediate Layers

4.1. Motivation

As data moves through the trained layers of the network, the represented features become more complex, from edges and simple texture patterns to higher-level representations or combinations of intermediate features [31]. Our assumption is that the use of these intermediate representations can improve out-of-distribution detection. Therefore, we take EBO [26] as a starting point and analyze how discriminative the different layers of the model are for OoD detection. To quantify the capacity of hidden layers in the OoD task, we introduce a hypothetical method called *Best Hidden Layer* (BHL), which utilizes an oracle to identify the optimal hidden layer for OoD detection. Therefore, since it requires access to the distribution ground truth, it is proposed as an a-posteriori analysis strategy.

Following classic setups, we train a classification model on $\mathcal{D}_{\text{in}}$, using the standard *cross-entropy* loss $\mathcal{L}_{\text{CE}}$. Then, we evaluate on test data from both $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$, extracting the embeddings from the intermediate layers for each sample. Here, the free energy from Eq. (1) is a natural candidate to use on the logits. However, the function can also take the embeddings $\mathbf{a}_l$ from any other layer l. Thus, we propose to extract the energy score:

$$E_l(\mathbf{x}) = -T \log \sum_i e^{a_l^i(\mathbf{x})/T}, \tag{2}$$

where the unit indices i correspond to the output of the l-th layer.

We extract and analyze the energy of each layer, regardless of its type, such as convolutional, batch normalization, or fully connected. We observe that certain intermediate layers consistently outperform the network logits from the original EBO approach. This effect is shown in Figure 2 for semantic shift, which presents the AUROC scores evaluated across all layers of a ResNet18 [15] for CIFAR-10 [22] as $\mathcal{D}_{\text{in}}$ and near and far OoD as $\mathcal{D}_{\text{out}}$. Some high-performing layers exhibit unexpected behavior by assigning lower energy values to $\mathcal{D}_{out}$ samples instead of $\mathcal{D}_{in}$ samples. This leads to two possibilities: assigning OoD to lower energy samples or to higher energy samples. Among the two, the "correct" possibility is reflected in the reported results.

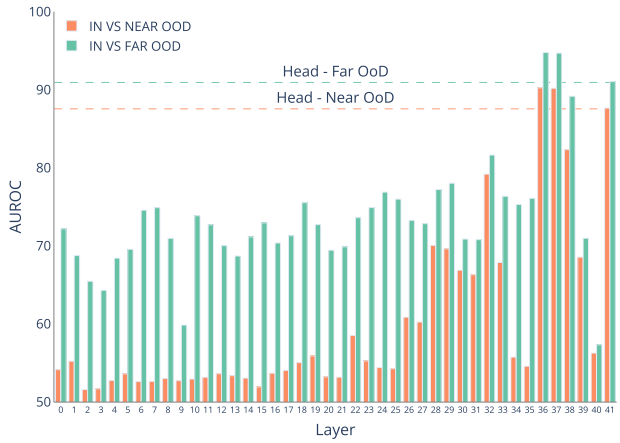

Figure 2. AUROC scores for OoD detection for each intermediate layer of ResNet18 are presented. The network is pretrained on CIFAR-10 ($\mathcal{D}_{\text{in}}$) and evaluated against the corresponding $\mathcal{D}_{out}^{near}$ and $\mathcal{D}_{out}^{far}$ datasets. Results are averaged across datasets in both categories.

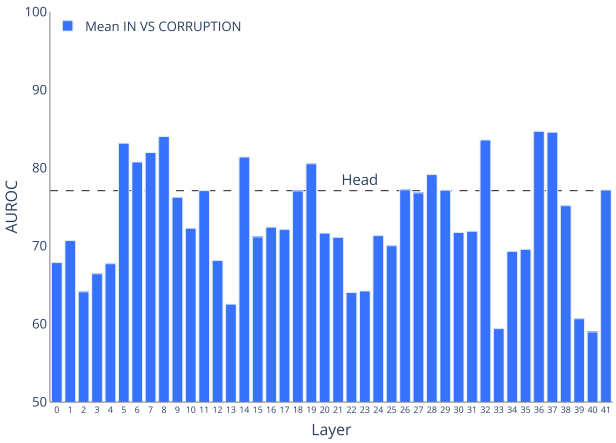

Figure 3. AUROC scores for each intermediate layer of ResNet18 pretrained on CIFAR-10 as $\mathcal{D}_{in}$ and evaluated against different corruptions (CIFAR-10-C). Results are averaged over all corruption types and seeds.

Covariate shift OoD detection also shows significant improvement when considering intermediate layers rather than relying solely on the network's output logits. To test it, we look at the performance of different layers when the OoD represents the in-distribution shifted by different corruptions (CIFAR-10-C [16], see Sec. 5.1). Figure 3 shows that throughout the depth of the network, several layers outperform yet again the head. Initial layers, which provide low-level features such as edges or local histogram projections, seem to be good candidates for OoD detection when covariate shift is present, since it represents a transformation on the in-distribution.

Despite the clear benefits from using some of the layers, determining which one to use for OoD detection under different shifts is still challenging due to different $\mathcal{D}_{out}$ distributions or modes having a tendency to elicit the strongest responses in different layers. This variability means that no single layer is universally optimal for detecting all types of OoD inputs effectively. It must be noted that, on average for semantic shifts, the optimal layers are observed to reside more towards the later layers of the network (see Fig. 2). However, this is not enough to identify a good one-fits-all layer, or to find a straightforward selection criteria. We try to circumvent this issue by proposing two strategies to leverage the information from the intermediate layer representation spaces:

- aggregating all intermediate responses into a single unified response (described in Sec. 4.2);
- strictly regularize selected layers to enforce generalization over different distributions (described in Sec. 4.3).

4.2. Energy aggregation (Ag-EBO)

To develop a fully layer-agnostic post-hoc method that leverages all the potential from intermediate embeddings, we propose to aggregate the energy values extracted from all L layers simultaneously. Thus, for each input $\mathbf{x}$, we construct a vector of energies:

$$\mathbf{E}(\mathbf{x}) = (E_1(\mathbf{x}), \ldots, E_L(\mathbf{x})),$$

which groups the energy contributions of each layer into a unified representation. The dimension of this vector is significantly smaller than the total hidden dimension of the network, making it scalable and suitable for use with most common OoD methods. However, for the intermediate layer to be considered, it is desirable that it offers better results than just relying on the logits or on the embeddings from the penultimate layer.

We tested with some straightforward approaches from literature, presented in the next paragraphs. Two of the following three methods need a reference for the ID data, therefore we use the set of energies $\tilde{E} = E_{\text{in}}^{\text{train}} = \{\mathbf{E}(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{D}_{\text{in}}^{\text{train}}\}$, extracted from $\mathcal{D}_{\text{in}}^{\text{train}}$.

Mahalanobis distance. The score $\mathcal{S}_{\text{MD}}(\mathbf{x})$ depends on the Mahalanobis distance [24] of $\mathbf{E}(\mathbf{x})$:

$$\mathcal{S}_{\text{MD}}(\mathbf{x}) = \min_{\boldsymbol{\mu}_c \in \tilde{E}} \sqrt{(\mathbf{E}(\mathbf{x}) - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1} (\mathbf{E}(\mathbf{x}) - \boldsymbol{\mu}_c)},$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean vector and covariance matrix of the energy vectors for class c in $\mathcal{D}_{\text{in}}^{\text{train}}$, respectively.

K-nearest neighbor. The score $\mathcal{S}_{\text{KNN}}(\mathbf{x})$ is based on the distance of $\mathbf{E}(\mathbf{x})$ to its K nearest neighbors [33]:

$$\mathcal{S}_{\text{KNN}}(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^{K} \left\| \mathbf{E}(\mathbf{x}) - \mathbf{E}_i \right\|_2,$$

where $\{\mathbf{E}_1, \ldots, \mathbf{E}_K\} \subset \tilde{E}$ are the K nearest neighbors of $\mathbf{E}(\mathbf{x})$ in the in-distribution training set, measured using the Euclidean distance.

Reconstruction Error. The score $\mathcal{S}_{\text{VAE}}(\mathbf{x})$ is computed as the reconstruction error of $\mathbf{E}(\mathbf{x})$ using a small Variational Autoencoder [21]:

$$\mathcal{S}_{\text{VAE}}(\mathbf{x}) = \left\| \mathbf{E}(\mathbf{x}) - \hat{\mathbf{E}}(\mathbf{x}) \right\|_2,$$

	CIFAR-10 [22]	**CIFAR-100** [22]	**ImageNet200** [4]	**ImageNet** [4]
Architecture	ResNet18 [30]	ResNet18	ResNet18	ResNet50 [30]
Input Size	$32\times32\times3$	$32\times32\times3$	$224\times224\times3$	$224\times224\times3$
Near-OoD	CIFAR-100 TinyImageNet [4]	CIFAR-10 TinyImageNet	SSB-Hard [43] NINCO [2]	SSB-Hard NINCO
Far-OoD	Texture [3] MNIST [5] SVHN [29] Places365 [44]	Texture MNIST SVHN Places365	Texture iNaturalist [19] OpenImageO [37] -	Texture iNaturalist OpenImageO -
Corruptions	CIFAR-10-C [16]	-	-	-

Table 1. Setup description for each ID dataset.

	CIFAR-100	**TIN**	**Near OoD**	**MNIST**	**Places365**	**SVHN**	**Texture**	**Far OoD**
EBO [26]	86.36	88.80	87.58	94.32	89.25	91.79	89.47	91.21
BHL	**88.23**	**92.26**	**90.25**	**99.89**	**92.13**	**98.46**	**93.5**	**96.00**
MDSEns [24]	61.29	59.57	60.43	99.17	66.56	77.40	52.47	73.90
Ag-EBO w/ MD	66.03	67.03	66.53	99.05	63.73	94.25	93.32	87.59
Ag-EBO w/ KNN	83.69	86.5	85.09	92.81	86.01	89.64	87.46	88.98
Ag-EBO w/ VAE	80.42	83.11	81.77	89.31	83.27	88.25	84.13	86.24

Table 2. AUROC scores of MDSEns, EBO, BHL and three aggregation methods with CIFAR-10 as $\mathcal{D}_{\text{in}}$, averaged over 3 runs.

where $\hat{\mathbf{E}}(\mathbf{x})$ is the reconstruction of $\mathbf{E}(\mathbf{x})$. Higher reconstruction error indicates that the input is likely to be out-of-distribution.

4.3. Energy regularization (R-EBO)

Regularizing intermediate layers directly provides an effective approach to addressing the intermediate layer selection problem. Ideally, by enforcing a strong energy-based discriminative behavior within the hidden layers, we promote their reliability, allowing them to be used confidently without additional selection mechanisms.

EBO [26] introduces an energy-bounded learning loss $\mathcal{L}_{\text{energy}}$ to push the network to assign low energy values to ID samples (and viceversa for OoD). Since their approach operates at the logits level, this loss is applied exclusively to the model's head. In contrast, our proposed strategy extends the scope of this loss by applying it to each hidden convolutional layer during training, computing and back-propagating all the losses simultaneously. Given an ID dataset $\mathcal{D}_{\text{in}}^{\text{train}}$ and an OoD seen dataset $\mathcal{D}_{\text{out}}^{\text{train}}$ (for outlier exposure), the energy regularization loss for the l-th hidden layer is defined as:

$$\begin{aligned}\mathcal{L}_{\text{energy},l} = &\ \mathbb{E}_{\mathbf{x}_{\text{in}}\sim\mathcal{D}_{\text{in}}^{\text{train}}}[\max(0, E_l(\mathbf{x}_{\text{in}}) - m_{\text{in}})]^2 \\ &+ \mathbb{E}_{\mathbf{x}_{\text{out}}\sim\mathcal{D}_{\text{out}}^{\text{train}}}[\max(0, m_{\text{out}} - E_l(\mathbf{x}_{\text{out}})]^2\,,\end{aligned} \tag{3}$$

where m_{in} and m_{out} are two margins, serving as the upper bound for the energy of the ID data and the lower bound for the energy of the seen OoD data, respectively. We define the total loss as:

$$\mathcal{L}_{\text{R-EBO}} = \sum_{l=1}^{L} \mathcal{L}_{\text{energy},l}\,, \tag{4}$$

where the same constant margin values for m_{in} and m_{out} are used across all layers, although each can be explored independently. In the original EBO paper [26], $\mathcal{L}_{\text{EBO}} = \mathcal{L}_{\text{energy},E_L}$, where L is the last layer of the network. Furthermore, the decision to reduce the free ID energy and increase the OoD energy in intermediate layers is a design choice. Alternative regularization strategies can also be considered.

5. Experimental results

5.1. Implementation details

Datasets. The datasets used in this study were selected based on the guidelines of the OpenOoD benchmark [43], which offers a comprehensive and well-documented collection of state-of-the-art (SoTA) methods across various OoD scenarios. Also, the results presented here have been extracted from its continuously updated report, to ensure alignment with the latest developments in the field. For each $\mathcal{D}_{in}$, the OpenOoD benchmark defines a set of semantically *near* and *far* OoD datasets from it Table 1. Additionally, we tested the response to covariate shift from CIFAR-10 with the corruptions dataset CIFAR-10-C [16]. This is a dataset consisting of corrupted versions of CIFAR-10 images, which serves as a

common benchmark for evaluating robustness to covariate shifts. It includes a variety of corruption types, such as noise, blur, and weather distortions, applied at varying levels of severity.

Architectures. To keep the consistency with OpenOoD evaluations, the main results have been calculated using the same architectures used in the benchmark, shown in Tab. 1. We also evaluate on a non-residual based convolutional neural network, EFFICIENTNET-B7, for which we select convolutional, fully-connected, batch normalization and average pooling layers. Finally, following recent trends in machine learning, we evaluate ViT-B-16 [10], a transformer-based [40] architecture. ViTs utilize multi-head self-attention layers, and their feed-forward sub-layers consist of fully-connected layers. Our experiments focus on the selection of these fully-connected layers for BHL.

Training. OpenOoD provides three pretrained ResNet18 checkpoints for CIFAR-10, CIFAR-100, and IMAGENET200 as $\mathcal{D}_{in}$, and a single pretrained ResNet50 checkpoint for IMAGENET, all trained using standard SoftMax loss. Additionally, we trained 3 checkpoints for both CIFAR-10 and CIFAR-100 as $\mathcal{D}_{in}$ using the hidden regularization approach.

5.2. Analysis of OoD with intermediate layers

In Table 2, CIFAR-10 is selected as $\mathcal{D}_{\text{in}}$. EBO refers to the standard energy-based OoD detection mechanism applied directly at the logit level, while BHL shows the energy-based OoD detection using the best performing hidden layer. The results presented for BHL are averaged across the best hidden layer identified in each run, which tends to slightly vary between runs. For every $\mathcal{D}_{\text{out}}$ the results are strongly improved by (at least) one hidden layer's response. It is important to mention that the results presented only consider the internal behavior of the network, while an algorithm which correctly weighs the importance of a layer for OoD detection would also take the head of the model into consideration, potentially merging the best results of the two rows.

Energy aggregation. The last rows of Table 2 present the results of the aggregation methods (Ag-EBO) proposed in Section 4. The row above displays the results of MDSEns [24], taken from the OpenOoD benchmark [43]. Each of our proposed aggregation methods achieves higher AUROC compared to MDSEns [24], an ensemble method that exploits Mahalanobis distance on hidden layers. The lower results for MDSEns might be related to their assumption of class-conditional distribution of the hidden features being Gaussian. $\mathcal{D}_{\text{out}}^{\text{far}}$ datasets, such as MNIST, SVHN, and TEXTURE, demonstrate improved performance with the KNN aggregation approach compared to EBO. However, none of these methods are robust enough on average to consistently outperform relying exclusively on the head logits. This indicates that the layer-selection problem remains unsolved and cannot yet be effectively simplified into an aggregation mechanism.

	CIFAR-10		CIFAR-100	
	Far	ID Acc.	Far	ID Acc.
EBO [26]*	84.86	**82.33**	67.86	**54.83**
BHL*	90.42	**82.33**	86.98	**54.83**
R-EBO*	**98.48**	78.2	**94.06**	50.05

Table 3. AUROC scores of EBO, BHL and R-EBO with CIFAR-10 as $\mathcal{D}_{\text{in}}$, averaged over multiple runs. EBO and BHL exploit identical checkpoints, retrained (*) for direct comparability with R-EBO.

Dataset	EBO [26]	BHL	R-EBO
BRIGHTNESS	56.51	**82.98**	79.8
CONTRAST	92.39	**99.92**	96.99
DEFOCUS BLUR	84.65	**97.26**	85.44
ELASTIC	73.24	**87.36**	85.11
FOG	71.3	**96.7**	94.38
FROST	76.83	**91.4**	91.08
GAUSSIAN BLUR	89.8	**98.86**	75.96
GAUSSIAN NOISE	84.39	**99.65**	99.16
GLASS BLUR	85.77	**88.45**	98.13
IMPULSE NOISE	89.04	**99.98**	97.71
JPEG	73.33	**87.87**	61.47
MOTION BLUR	75.78	**93.51**	72.77
PIXELATE	80.02	**94.17**	99.66
SATURATE	57.47	**90.97**	81.8
SHOT NOISE	84.93	**99.38**	98.78
SNOW	71.85	**89.11**	74.95
SPATTER	71.0	**90.33**	83.69
SPECKLE NOISE	85.29	**98.96**	98.66
ZOOM BLUR	79.36	**96.61**	66.11

Table 4. AUROC scores of EBO, BHL, and R-EBO with CIFAR-10 as $\mathcal{D}_{in}$ against corruption datasets.

Energy regularization. Table 3 presents the results of regularization against other SoTA methods that exploit $\mathcal{D}_{\text{out}}^{\text{seen}}$. The margin values are set to m_{in}=−25 and m_{out}=−7, following the original EBO setup [26]. In order to test the trade-off in hidden layer regularization compared to a completely post-hoc hidden layer analysis, we selected CIFAR-10 and CIFAR-100 as $\mathcal{D}_{\text{in}}$ and IMAGENET as $\mathcal{D}_{\text{out}}^{seen}$. We then trained 5 runs using only $\mathcal{L}_{CE}$, and 5 runs using $\mathcal{L}_{CE} + \mathcal{L}_{\text{R-EBO}}$. We opted not to use the checkpoints given by OpenOoD to guarantee a fair comparison between the two losses. Therefore, the EBO results are not comparable with the ones presented in other tables, and are marked with $(*)$ accordingly. Moreover, only results related to Far-OoD are presented, since Near-OoD includes TIN, which is based on IMAGENET.

As expected, the regularization of intermediate layers

		CIFAR-10		CIFAR-100		ImageNet-200		ImageNet-1K	
		Near	Far	Near	Far	Near	Far	Near	Far
ResNet18/50	EBO	87.58	91.21	**80.91**	79.77	82.50	**90.86**	75.89	89.47
	BHL	**90.25**	**96.00**	71.57	**86.08**	**86.72**	76.13	**79.04**	**89.75**
EfficientNet-B7	EBO	**97.39**	98.91	**87.46**	86.91	75.02	86.53	65.16	81.65
	BHL	87.43	**99.74**	84.21	**99.80**	**78.83**	**93.02**	**85.24**	**94.49**
ViT-B-16	EBO	**90.91**	93.9	**88.81**	87.23	**69.72**	**83.49**	62.93	78.71
	BHL	79.38	**96.14**	81.38	**97.98**	62.19	81.40	**74.06**	**88.43**

Table 5. EBO and BHL compared on different models.

strongly improves the OoD detection capabilities of the model on both cases. However, this comes at the cost of a slight decrease in ID accuracy, due to the additional $\mathcal{L}_{\text{R-EBO}}$ loss.

Covariate shift. Table 4 presents the detailed OoD results of CIFAR-10 against every corruption type present in CIFAR-10-C. As with the semantic shift, we observe that covariate shift is better identified by the hidden layers rather than by the final logits. Table 4 also presents R-EBO results under covariate shift conditions, evaluated using the same checkpoints from Table 3. The findings suggest that regularizing layers with a semantically distinct $\mathcal{D}_{\text{out}}^{seen}$ does not consistently enhance the identification of covariate shift.

5.3. Analysis on different architectures

Table 5 presents the complete results for EBO and BHL, averaged over multiple runs, using ResNet18/50, EfficientNet-B7, and ViT-B-16 as backbones. The findings are consistent with earlier observations: BHL improves performance in most setups, except for certain Near OoD cases.

6. Discussion and Limitations

Our findings show that intermediate representations are capable of discriminating out-of-distribution samples better than the logits. Both semantic, in the form of unseen classes, and covariate shift, in the form of image corruptions, are strongly captured by intermediate layers. However, a robust selection criterion for which layer to use is still an open question, since the proposed aggregation method underperforms compared to simpler logit-based alternatives.

Regularization of the intermediate layer's energies improves the results even further, albeit with a trade-off in ID accuracy. We suspect that the influence of $\mathcal{D}_{\text{out}}^{\text{seen}}$ leads to sub-optimal filters for the discrimination of ID classes, thus motivating further research involving regularization which exploits $\mathcal{D}_{\text{in}}$ only. Additionally, regularization using synthetic generated data [45] applied to intermediate layers could also be a promising direction, as it would reduce dependence on specific datasets, promote privacy-preservation, and enhance the generalization.

Finally, the findings on this paper pave the way for real-time optimized out-of-distribution detection, enabling the identification of OoD samples in earlier layers during network propagation. By detecting such samples promptly, the system can flag them and halt further processing, reducing computational overhead and improving efficiency.

Acknowledgements

Gianluca Guglielmo acknowledges the support of KAI GmbH and Infineon Technologies Austria. Marc Masana acknowledges the support by the "University SAL Labs" initiative of Silicon Austria Labs (SAL).

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv e-prints*, pages arXiv–1606, 2016. 1

[2] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023. 5

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[5] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 5

[6] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv e-prints*, pages arXiv–1802, 2018. 2

[7] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023. 1

[8] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely Simple Activation Shaping for Out-of-Distribution Detection. *arXiv e-prints*, art. arXiv:2209.09858, 2022. 2

[9] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[11] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, pages 6009–6033. PMLR, 2022. 1, 2

[12] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org. 1

[13] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2019. 3

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 2, 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 2, 4, 5

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv e-prints*, pages arXiv–1610, 2016. 2

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 2

[19] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist challenge 2017 dataset. *CoRR*, abs/1707.06642, 2017. 5

[20] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 4

[22] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 3, 5

[23] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 3

[24] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1, 2, 4, 5, 6

[25] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv e-prints*, pages arXiv–2108, 2021. 1, 2

[26] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 3, 5, 6

[27] Shuo Lu, YingSheng Wang, LuJun Sheng, AiHua Zheng, LinXiao He, and Jian Liang. Recent advances in ood detection: Problems and approaches. *arXiv e-prints*, pages arXiv–2409, 2024. 1

[28] Marc Masana, Idoia Ruiz, Joan Serrat, Joost van de Weijer, and Antonio M Lopez. Metric learning for novelty and anomaly detection. In *British Machine Vision Conference (BMVC)*, 2018. 2, 3

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011, 2011. 5

[30] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 1, 5

[31] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 1, 3

[32] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 2

[33] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 4

[34] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020. 2

[35] Junjiao Tian, Yen-Change Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Exploring covariate and concept shift for detection and calibration of out-of-distribution data. *arXiv e-prints*, pages arXiv–2110, 2021. 2

[36] Aishwarya Venkataramanan, Assia Benbihi, Martin Laviale, and Cédric Pradalier. Gaussian latent representations for uncertainty estimation using mahalanobis distance in deep classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4488–4497, 2023. 2

[37] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4911–4920, 2022. 5

[38] Hongjun Wang, Sagar Vaze, and Kai Han. Dissecting out-of-distribution detection and open-set recognition: A crit-

ical analysis of methods and benchmarks. *International Journal of Computer Vision*, pages 1–26, 2024. 2

[39] Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. *Advances in neural information processing systems*, 36:73274–73286, 2023. 2

[40] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 6

[41] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022. 2

[42] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 1, 2

[43] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv e-prints*, pages arXiv–2306, 2023. 2, 5, 6

[44] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452–1464, 2018. 5

[45] Jianing Zhu, Yu Geng, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *Advances in Neural Information Processing Systems*, 36:22702–22734, 2023. 7