ENERGY-BASED OUT-OF-DISTRIBUTION DETECTION FOR MULTI-LABEL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection is essential to prevent anomalous inputs from causing a model to fail during deployment. Improved methods for OOD detection in multi-class classification have emerged, while OOD detection methods for multi-label classification remain limited and use rudimentary techniques. In this paper, we propose an energy-based method for OOD detection in the multi-label setting. Our energy scores are calculated from an aggregate of energy values across labels in a pre-trained multi-label classifier. We provide theoretical connections between the aggregated energy and the probability density of the training data, under certain distributional assumptions. Our results show an improvement in multi-label OOD detection over previous methods that are based on the maximum-valued scores. In particular, we evaluate ResNet pre-trained on common multi-label classification datasets MS-COCO, PASCAL-VOC, and NUS-WIDE. We show *SumEnergy* reduces the FPR95 by up to 7.8% compared to the previous best baseline, establishing state-of-the-art performance.

1 INTRODUCTION

Out-of-distribution (OOD) detection is central for reliably deploying machine learning models in open-world environments, where new forms of test-time data may appear that were nonexistent during the training time. The problem of OOD detection has gained significant research attention lately, given its importance for safety-critical applications such as unseen disease identification (Cao et al., 2020). However, recent studies have primarily focused on detecting OOD examples in multi-class classification, where each sample is assigned to one and *only one* label (Bevandić et al., 2018; Hein et al., 2019; Hendrycks & Gimpel, 2016; Lakshminarayanan et al., 2017; Lee et al., 2018; Liang et al., 2018; Mohseni et al., 2020; Chen et al., 2020; Hsu et al., 2020; Liu et al., 2020). This can be restrictive in many real-world settings where images often have multiple objects of interest. For example, self-driving cars must differentiate between the road, traffic signs, and obstacles within a frame. In the medical domain, multiple abnormalities may be present in a medical image (Wang et al., 2017). Multi-label classification is therefore better suited since there is no constraint on the number of classes an instance can be assigned to.

Currently, OOD detection in multi-label classification remains relatively underexplored. Out of the multi-label methods evaluated (Hendrycks et al., 2019), MaxLogit achieved the best performance. However, simply using the maximum-valued logit is limiting because it does not incorporate information available from other possible labels. Other baseline methods such as ODIN and Mahalanobis also derive scores based on the maximum score (*e.g.*, calibrated softmax score or Mahalanobis distance) among all labels, and fail to capture the joint information. While energy scores have recently demonstrated superior OOD detection performance in the multi-class setting (Liu et al., 2020), this method does not trivially generalize to a multi-label classification setting where labels are not mutually exclusive. Hitherto, a key challenge lies in how to leverage information across different labels.

In this paper, we propose an energy-based method for OOD detection in the multi-label setting, which estimates the OOD indicator scores *jointly from multiple labels*. We propose simple and effective aggregation mechanisms that seek to combine the OOD indicator scores derived from individually independent classes (see Figure 1). We show that the aggregated energy scores are mathematically meaningful under certain distributional assumptions. Our method is parameter-free and

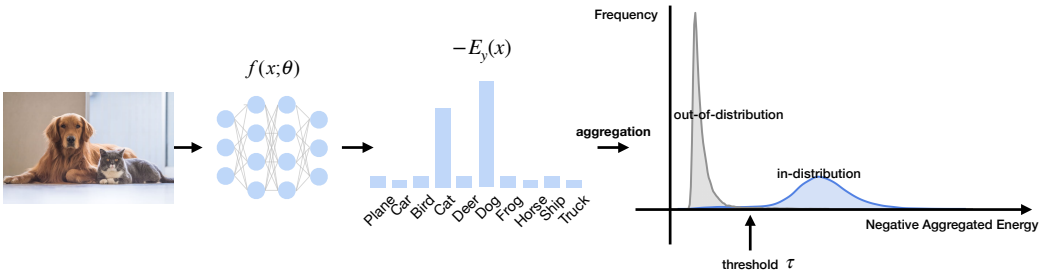

Figure 1: Energy-based out-of-distribution detection for multi-label classification. The energy can be used as a scoring function for any pre-trained neural network. During inference time, for a given input $\mathbf{x}$, the class-wise energy score $E_y(\mathbf{x})$ is calculated for each label, which will be aggregated across all labels. The OOD detector will classify the input as OOD if the aggregated energy score is smaller than the threshold value.

can be conveniently used for any pre-trained multi-label classification model. Below we describe our contributions in detail.

Contributions First, we propose energy-based scoring functions that are based on the aggregation over label-wise energy scores, and evaluate on common multi-label classification models. Extensive experiments show that *SumEnergy* outperforms the best-performing baseline across a wide range of multi-label OOD evaluation tasks. For example, on a network pre-trained with PASCAL-VOC (Everingham et al., 2015), our method reduces the false positive rate (at 95% TPR) by 4.36% when evaluated against ImageNet OOD data, compared to the best performing baselines. The performance improvement is more pronounced on NUS-WIDE, where *SumEnergy* reduces FPR95 by **7.8**%.

Additionally, we perform ablation studies and comparative analysis on how alternative aggregation methods affect OOD detection performance. In particular, we consider *MaxEnergy*, which takes the maximum energy score among the individual labels as the OOD indicator score. We show that *MaxEnergy* yields lower detection performance compared to *SumEnergy*, which underlines the importance of taking into account scores derived from multiple labels. We further considered *ExpSumEnergy*, which combines the exponential of negative label-wise energy scores and is theoretically connected to the input's probability density under certain distributional assumptions. Experiments show that *ExpSumEnergy* outperforms prior baselines, and compares favorably with *SumEnergy*.

Lastly, as an extension, we explore the effectiveness of applying aggregation methods to previous scoring functions such as logit (Hendrycks et al., 2019), MSP (Hendrycks & Gimpel, 2016), and ODIN (Liang et al., 2018), and Mahalanobis distance (Lee et al., 2018). This helps understand the extent to which previous approaches can generalize to the multi-label setting. Interestingly, we found that applying summation over individual logit/MSP/ODIN/Mahalanobis scores from each label does not yield competitive results, and in many cases worsens the performance. For example, simply summing over the logits across labels results in up to 64.4% degradation in FPR95 since the outputs are mixed with positive and negative numbers. In contrast, *SumEnergy* does not suffer from this issue since the energy scores for each label is non-negative, and is provably aligned with the probability density of the training data corresponding to the label. Our study therefore underlines the importance of properly choosing both the label-wise scoring function and the aggregation method. Our method demonstrates strong compatibility of between the label-wise energy function and aggregation functions, supported by both theoretical interpretations and empirical results.

2 METHODS

2.1 BACKGROUND: ENERGY FUNCTION IN MULTI-CLASS CLASSIFICATION

We consider a multi-class neural classifier $f(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}^K$, which maps an input $\mathbf{x} \in \mathbb{R}^D$ to K real-valued numbers as logits. A softmax function is used to derive a categorical distribution,

$$p(y \mid \mathbf{x}) = \frac{e^{f_y(\mathbf{x})}}{\sum_i^K e^{f_i(\mathbf{x})}}, \tag{1}$$

which indicates the probability for an input $\mathbf{x}$ to be of class y. Grathwohl et al. showed that a multi-class classifier can be interpreted from an energy-based perspective by viewing the logit $f_y(\mathbf{x})$ as an

energy function $E(\mathbf{x}, y) = -f_y(\mathbf{x})$. Therefore, Equation 1 can be rewritten as:

$$p(y \mid \mathbf{x}) = \frac{e^{-E(\mathbf{x},y)}}{\sum_i^K e^{-E(\mathbf{x},i)}} = \frac{e^{-E(\mathbf{x},y)}}{e^{-E(\mathbf{x})}} \tag{2}$$

By equalizing the two denominators above, we can express the *free energy* function $E(\mathbf{x})$ for any given input $\mathbf{x} \in \mathbb{R}^D$ in forms of:

$$E(\mathbf{x}) = -\log \sum_i^K e^{f_i(\mathbf{x})}. \tag{3}$$

Liu et al. propose to use free energy as a scoring function for OOD detection in the multi-class setting, with theoretical motivation that free energy is provably aligned with the likelihood of data, *i.e.,* $E(\mathbf{x}) \propto -\log p(\mathbf{x})$.

2.2 Energy Function for Multi-label Classification

In this work, we consider using energy function for OOD detection in multi-label classification networks, where an input can be associated with several labels (see Figure 1). In what follows, we first introduce label-wise energy function, and then propose aggregation methods that can leverage the joint information across labels in a theoretically meaningful way.

Label-wise Free Energy In multi-label classification, the prediction for each label $y \in \{1, 2, ..., K\}$ is independently made by a binary logistic classifier:

$$p(y = 1|\mathbf{x}) = \frac{e^{f_y(\mathbf{x})}}{1 + e^{f_y(\mathbf{x})}}.$$

This can be viewed as the softmax distribution with two logits—0 and $f_y(\mathbf{x})$, respectively. For each class y, we can define *label-wise free energy* as follows:

$$E_y(\mathbf{x}) = -\log(1 + e^{f_y(\mathbf{x})}). \tag{4}$$

Aggregation Methods We consider three aggregation mechanisms that seek to combine the label-wise energy scores derived above:

$$\textbf{Max}: \quad E_{\text{max}}(\mathbf{x}) = \max_y E_y(\mathbf{x}) \tag{5}$$

$$\textbf{Sum}: \quad E_{\text{sum}}(\mathbf{x}) = \sum_{y=1}^{K} E_y(\mathbf{x}) \tag{6}$$

$$\textbf{ExpSum}: \quad E_{\text{expsum}}(\mathbf{x}) = -\sum_{y=1}^{K} e^{-E_y(\mathbf{x})}, \tag{7}$$

where *MaxEnergy* finds the largest label-wise energy score among all labels; *SumEnergy* takes the summation of energy scores across all labels; *ExpSumEnergy* combines the exponential of negative energy scores. We provide mathematical interpretations of different aggregation methods below.

Mathematical Interpretation First, by resorting to the energy-based model (LeCun et al., 2006), we show that the label-wise energy score is provably aligned with the conditional likelihood function:

$$p(\mathbf{x}|y) = \frac{e^{-E_y(\mathbf{x})}}{\int_{\mathbf{x}|y} e^{-E_y(\mathbf{x})}}.$$

Since the normalized density $Z_y = \int_{\mathbf{x}|y} e^{-E_y(\mathbf{x})}$ is the same with respect to all $\mathbf{x}$ with label y, we can derive:

$$E_y(\mathbf{x}) \propto -\log p(\mathbf{x}|y).$$

Given this, *MaxEnergy* can be interpreted as taking the maximum of conditioned log-likelihood among all labels:

$$E_{\text{max}}(\mathbf{x}) \propto \max_y -\log p(\mathbf{x}|y), \tag{8}$$

which does not take into account the joint information across labels.

Next, we provide mathematical interpretation for *SumEnergy*, which utilize the energy scores across labels. Under the assumption that normalized densities $Z_y = \int_{\mathbf{x}|y} e^{-E_y(\mathbf{x})}$ are equivalent across labels, we can express *SumEnergy* as:

$$E_{\text{sum}}(\mathbf{x}) \propto \sum_y -\log p(\mathbf{x}|y) \tag{9}$$

$$\propto -\log \frac{p(\mathbf{x}, y_1)}{p(y_1)} \cdot \frac{p(\mathbf{x}, y_2)}{p(y_2)} \cdots \frac{p(\mathbf{x}, y_K)}{p(y_K)} \tag{10}$$

$$\propto -\log \prod_{y=1}^{K} p(\mathbf{x}, y). \tag{11}$$

In other words, *SumEnergy* can be interpreted as the proportional negative log probability for a data point $\mathbf{x}$ to be associated with all labels.

Lastly, we show that *ExpSumEnergy* can be provably aligned with the probability density of the multi-label training dataset under certain distributional assumptions. When the prior probability $p(y)$ is equal among all labels $\{1, 2, ..., K\}$, we can estimate the overall probability density over the entire input space $\mathbf{x} \in \mathbb{R}^D$:

$$p(\mathbf{x}) = \sum_{y=1}^{K} p(\mathbf{x}|y) \cdot p(y)$$

$$\propto \sum_{y=1}^{K} \frac{e^{-E_y(\mathbf{x})}}{\int_{\mathbf{x}|y} e^{-E_y(\mathbf{x})}}.$$

When the normalized densities $Z_y = \int_{\mathbf{x}|y} e^{-E_y(\mathbf{x})}$ (with respect to $\mathbf{x}$ with label y) are equivalent across labels, we have:

$$p(\mathbf{x}) \propto \sum_{y=1}^{K} e^{-E_y(\mathbf{x})} \tag{12}$$

$$\propto -E_{\text{expsum}}(\mathbf{x}), \tag{13}$$

which means a data point with a higher probability of occurrence is equivalent to having proportionally lower *ExpSumEnergy*.

2.3 Aggregated Energy for Multi-Label OOD Detection

We propose using the aggregated energy functions $E(\mathbf{x})$ defined in Section 2.2 for OOD detection:

$$G(\mathbf{x}; \tau) = \begin{cases} \text{out} & \text{if } -E(\mathbf{x}) \leq \tau, \\ \text{in} & \text{if } -E(\mathbf{x}) > \tau, \end{cases} \tag{14}$$

where τ is the energy threshold, and can be chosen so that a high fraction of in-distribution data is correctly classified by $G(\mathbf{x}; \tau)$. $E(\mathbf{x})$ could take on forms of *Max, Sum* or *ExpSum*. A data point with lower aggregated energy $E(\mathbf{x})$ is considered as in-distribution, and vice versa.

3 Experiments

In this section, we describe our experimental setup (Section 3.1) and demonstrate the effectiveness of our method on several OOD evaluation tasks (Section 3.2). We also conduct extensive ablation and comparative analysis that leads to an improved understanding of different aggregation methods.

3.1 Setup

In-distribution Datasets We consider three multi-label datasets: MS-COCO (Lin et al., 2014), PASCAL-VOC (Everingham et al., 2015), NUS-WIDE (Chua et al., 2009). MS-COCO consists

OOD Score	$\mathcal{D}_{\text{in}}$ Aggregation	MS-COCO	PASCAL-VOC	NUS-WIDE
		FPR95 / AUROC / AUPR		
			$\downarrow$ $\uparrow$ $\uparrow$	
Logit	Max	34.54 / 90.93 / 94.30	36.32 / 91.04 / 82.68	58.05 / 83.07 / 94.21
MSP	Max	34.54 / 90.93 / 94.30	36.32 / 91.04 / 82.68	58.05 / 83.07 / 94.21
ODIN	Max	34.58 / 90.26 / 93.69	36.32 / 91.04 / 82.68	58.05 / 83.07 / 94.21
Mahalanobis	Max	94.04 / 49.49 / 70.71	78.02 / 70.93 / 59.84	61.33 / 83.75 / 95.15
LOF	-	74.30 / 74.87 / 85.82	76.71 / 67.54 / 55.35	85.42 / 69.37 / 90.36
Isolation Forest	-	99.06 / 37.59 / 63.43	98.64 / 41.94 / 33.50	96.59 / 50.75 / 82.91
Energy (ours)	Max	34.54 / 90.93 / 94.30	36.32 / 91.04 / 82.68	58.05 / 83.07 / 94.21
	ExpSum	33.64 / 91.07 / 94.38	34.79 / 91.36 / 83.10	51.87 / 85.58 / 95.02
	Sum	**31.51 / 92.68 / 96.15**	**31.96 / 92.32 / 86.87**	**50.25 / 88.12 / 96.34**

Table 1: OOD detection performance comparison using energy-based approaches vs. competitive baselines. We use ResNet (He et al., 2016) to train on the in-distribution datasets. We use a subset of ImageNet classes as OOD test data, as described in Section 3.1. All values are percentages. $\uparrow$ indicates larger values are better, and $\downarrow$ indicates smaller values are better. **Bold** numbers are superior results.

of 82,783 training, 40,504 validation, and 40,775 testing images with 80 common object categories. PASCAL-VOC contains 22,531 images across 20 classes. NUS-WIDE includes 269,648 images across 81 concept labels. Since NUS-WIDE has invalid and untagged images, we follow the work (Zhu et al., 2017) and use 119,986 training images and 80,283 test images.

Out-of-distribution Datasets To evaluate the models trained on in-distribution datasets above, we follow the same set up as in Hendrycks et al., and use ImageNet (Deng et al., 2009) for its generality. In addition, we also evaluate on Textures dataset (Cimpoi et al., 2014b). For ImageNet, we use the same set of 20 classes chosen from ImageNet-22K as in Hendrycks et al.. These classes are chosen not to overlap with ImageNet-1k since the multi-label classifiers are pre-trained on ImageNet-1K. Specifically, we use the following classes for evaluations on MS-COCO and PASCAL-VOC pre-trained models: *dolphin, deer, bat, rhino, raccoon, octopus, giant clam, leech, venus flytrap, cherry tree, Japanese cherry blossoms, redwood, sunflower, croissant, stick cinnamon, cotton, rice, sugar cane, bamboo, and turmeric*. Since NUS-WIDE contains high-level concepts like animal, plants and flowers, we use a different collection of classes that are distinct from NUS-WIDE concepts: *asterism, battery, cave, cylinder, delta, fabric, filament, firebell, hornet nest, kazoo, lichen, naval equipment, newspaper, paperclip, pythium, satellite, thumb, x-ray tube, yeast, zither*.

Evaluation Metrics We measure the following metrics that are commonly used for OOD detection: (1) the false positive rate (FPR95) of OOD examples when the true positive rate of in-distribution examples is at 95%; (2) the area under the receiver operating characteristic curve (AUROC); and (3) the area under the precision-recall curve (AUPR).

Training Details We use a ResNet-101 backbone architecture that is pre-trained on ImageNet-1K. We replace the final layer of with 2 fully connected layers, and apply the logistic sigmoid function for multi-label prediction. We use the Adam optimizer (Kingma & Ba, 2014) with standard parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The initial learning rate is 10^{-4} for the fully connected layers and 10^{-5} for convolutional layers. We also augmented the data with random crops and random flips to obtain color images of size 256×256. After training, the mAP is 87.73% for PASCAL-VOC, 72.77% for MS-COCO, and 61.47% for NUS-WIDE. We provide details on the hyperparameter tuning for baseline detection methods such as ODIN and Mahalanobis in the Appendix.

3.2 Results

How does energy-based approach compare to common OOD detection methods? In Table 1, we compare energy-based approach against competitive OOD detection methods in literature. For fair comparisons, we consider approaches that rely on pre-trained models (without performing retraining or fine-tuning). Following the set up in Hendrycks et al., all the numbers reported are evaluated on ImageNet OOD test data, as described in Section 3.1. We provide additional results for the Texture OOD test dataset in Appendix A. Most baselines such as MaxLogit (Hendrycks et al., 2019), Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2018) and Mahalanobis (Lee et al., 2018) derive OOD indicator scores based on the maximum-valued statistics among all labels. LOF method uses K-nearest neighbors (KNN) to estimate the local density, where

	$\mathcal{D}_{\text{in}}$	MS-COCO	PASCAL	NUS-WIDE
		FPR95 / AUROC / AUPR		
OOD Score	**Aggregation**	$\downarrow$ $\uparrow$ $\uparrow$		
Logit	Sum	95.63 / 53.52 / 73.25	96.36 / 49.44 / 43.07	96.49 / 49.83 / 81.78
Prob	Sum	43.69 / 87.21 / 93.14	35.97 / 84.68 / 76.61	55.86 / 82.97 / 94.92
ODIN	Sum	43.69 / 87.21 / 93.14	53.77 / 74.50 / 67.15	55.24 / 81.84 / 94.59
Mahalanobis	Sum	94.47 / 46.82 / 67.06	78.56 / 70.84 / 59.34	62.79 / 83.19 / 94.96
LOF	Sum	-	-	-
Isolation Forest	Sum	-	-	-
Energy	Sum	**31.51 / 92.68 / 96.15**	**31.96 / 92.32 / 86.87**	**50.25 / 88.12 / 96.34**

Table 2: Ablation study on the effect of aggregation methods for prior approaches. We use ResNet (He et al., 2016) to train on the in-distribution datasets. We use ImageNet as OOD test data as described in Section 3.1. Note that *Sum* is not applicable to tree-based or KNN-based approaches (e.g., LOF and Isolation Forest).

OOD examples are detected from having lower density compared to their neighbors. IForest is a tree-based approach, which detects anomaly based on the path length from the root node to the terminating node.

Among different approaches, we show that *SumEnergy* outperforms the best-performing baseline across all three common multi-label classification datasets considered. In particular, on a network pre-trained with the PASCAL-VOC dataset, *SumEnergy* reduces FPR95 by 4.36%, compared to MaxLogit. The performance improvement is most pronounced on NUS-WIDE, where *SumEnergy* reduces FPR95 by **7.8**%. We additionally evaluate on a different architecture, DenseNet (Huang et al., 2017), for which we observe consistent improvement and provide details in Appendix.

We note here that existing approaches using a pre-trained model (such as ODIN (Liang et al., 2018) and Mahalanobis (Lee et al., 2018)) have hyperparameters that need to be tuned. In contrast, using an energy-based method on pre-trained models is parameter-free, easy to use, and deploy. In particular, Mahalanobis approach is based the assumption that the data is distributed as a mixture of Gaussian, and hence less well suited for multi-label setting.

How does different aggregation methods affect the OOD detection performance? In Table 1, we also perform a comparative analysis of the effect of different aggregation functions that combine label-wise energy scores. Among those, we observe that *MaxEnergy* does not outperform the other two aggregation methods *SumEnergy* or *ExpSumEnergy*—both of which utilize information jointly from across the labels. The performance of *MaxEnergy* is on par with MaxLogit since the negative of *MaxEnergy*, given by $\max_y \log(1 + e^{f_y(\mathbf{x})})$, is approximately close to the MaxLogit when $f_y(\mathbf{x})$ is large. The results underline the importance of taking into account information from other labels, not just the maximum-valued label. This is because, in multi-label classification, the model may assign high probabilities to several classes.

Interestingly, we also observe that *ExpSumEnergy* performs favorably compared to *SumEnergy*, though slightly suboptimal. As seen in Section 2.2, both aggregation methods are theoretically meaningful, under certain distributional assumptions. Since the normalized densities Z_y can be intractable to compute in practice, our aggregation method is based on the simplified scenario where all normalized densities are equal across labels. We note here that this might not strictly hold true on the real-world datasets evaluated. Despite this, we notice that our aggregation methods are overall effective across several multi-label evaluation tasks, which suggest its practical utility.

What is the effect of applying the aggregation method to prior methods? As an extension, we explore the effectiveness of applying the aggregation method to previous scoring functions such as logit (Hendrycks et al., 2019), MSP (Hendrycks & Gimpel, 2016) and ODIN (Liang et al., 2018). The results are summarized in Table 2. We calculate scores based on the logit $f_y(\mathbf{x})$, sigmoid of the logit $\frac{1}{1+e^{-f_y(\mathbf{x})}}$, ODIN score, as well as Mahalanobis distance score $M_y(\mathbf{x})$. We then perform summation across the label-wise scores as the overall OOD score. This ablation essentially replaces the *Max* aggregation with *Sum*, which helps understand the extent to which previous approaches are amenable in the multi-label setting. Note that the summation aggregation method is not applicable to tree-based or KNN-based approaches such as LOF and Isolation Forest.

Interestingly, we found that applying summation over individual logit/MSP/ODIN/Mahalanobis scores from each label does not yield competitive results, and in many cases worsens the performance. For example, simply summing over the logits across the labels leads to severe degradation in performance since the outputs are mixed with positive and negative numbers. On MS-COCO, the FPR degrades from 34.54% using MaxLogit to 95.63% (using SumLogit). In contrast, *SumEnergy* does not suffer from this issue since the energy scores for each individual label is non-negative. More importantly, the label-wise scores derived from energy is theoretically more meaningful than logit/MSP/ODIN/Mahalanobis, since it is provably aligned with the probability density of the training data corresponding to the label. This underlines the importance of properly choosing the label-wise scoring function in order to be compatible with the aggregation method.

Qualitative case study Lastly, to provide further insights on our method, we qualitatively examine examples from multi-label classification datasets. Figure **??** shows OOD inputs from the ImageNet dataset that are correctly classified by *SumEnergy* but not MaxLogit. We see that...

4 Related Work

Multi-label classification The task of identifying multiple classes within an input example is of significant interest in many applications (Tsoumakas & Katakis, 2007), with deep neural networks being commonly used as the classifier. Natural images usually contain several objects and may have many associated tags (Wang et al., 2016). Gong et al. use convolutional neural networks (CNN) to annotate images with 3 or 5 tags on the NUS-WIDE dataset. Chen et al. use CNNs to tag images of road scenes from 52 possible labels. In the medical domain, Wang et al. present a chest X-ray dataset in which one image may contain multiple abnormalities. Multi-label classification is also prominent in natural language processing (Nam et al., 2014). The OOD detection method proposed in our work is therefore relevant to a wide range of applications in the real world.

Out-of-distribution uncertainty for pre-trained models The softmax confidence score has become a common baseline for OOD detection (Hendrycks & Gimpel, 2016). A theoretical investigation (Hein et al., 2019) shows that neural networks with ReLU activation can produce arbitrarily high softmax confidence for OOD inputs. Several works attempt to improve the OOD uncertainty estimation by using deep ensembles (Lakshminarayanan et al., 2017), ODIN score (Liang et al., 2018), Mahalanobis distance-based confidence score (Lee et al., 2018), and generalized ODIN score (Hsu et al., 2020). DeVries & Taylor propose to learn the confidence score by using an auxiliary branch to derive the OOD score. Recent work using energy score demonstrated state-of-the-art performance on OOD detection tasks (Liu et al., 2020). However, previous methods primarily focused on multi-class classification networks. In contrast, in our work, we propose a parameter-free measurement that allows effective OOD detection in the underexplored *multi-label* setting, where the information from across various labels are taken into account in a theoretically meaningful manner.

Out-of-distribution detection with model fine-tuning While our work primarily focused on OOD detection for pre-trained neural networks, a parallel line of research has also explored using separate outlier training set to help the OOD detector generalize better. Auxiliary data allows the model to be explicitly regularized through fine-tuning, producing lower confidence on anomalous examples (Bevandić et al., 2018; Geifman & El-Yaniv, 2019; Malinin & Gales, 2018; Mohseni et al., 2020; Subramanya et al., 2017). A loss function is used to force the predictive distribution of OOD samples toward uniform distribution Lee et al. (2017). Recently, Mohseni et al. explore training by adding background class for an OOD score. Chen et al. propose using hard outlier mining which improves the OOD detection performance on both clean and perturbed natural images. In contrast, our approach does not alter the semantic class space and can be used without auxiliary OOD data. We leave the fine-tuning aspect with auxiliary data as future exploration.

Generative Modeling Based Out-of-distribution Detection. Generative models (Dinh et al., 2016; Kingma & Welling, 2013; Rezende et al., 2014; Van den Oord et al., 2016; Tabak & Turner, 2013) can be alternative approaches for detecting OOD examples, as they directly estimate the in-distribution density and can declare a test sample to be out-of-distribution if it lies in the low-density regions. However, as shown by Nalisnick et al. (2018), deep generative models can assign a high likelihood to out-of-distribution data. Deep generative models can be more effective for out-of-distribution detection using alternative metrics (Choi & Jang, 2018), likelihood ratio (Ren et al., 2019; Serrà et al., 2019). Though our work is based on discriminative classification models,

we show that energy scores can be theoretically interpreted from a data density perspective under certain data assumptions. Our method therefore inherits the merits of generative-based approaches, while circumventing the difficult optimization process in training generative models, especially on large, complex input data.

Energy-based learning Energy-based machine learning models date back to Boltzmann machines (Ackley et al., 1985; Salakhutdinov & Larochelle, 2010), networks of units with energy defined for the overall network. Energy-based learning (LeCun et al., 2006; Ranzato et al., 2007a;b) provides a unified framework for many probabilistic and non-probabilistic approaches to learning. Recent work (Zhao et al., 2019) also demonstrated using energy functions to train GANs (Goodfellow et al., 2014), where the discriminator uses energy values to differentiate between real and generated images. Grathwohl et al. showed that a discriminative classifier can be interpreted from an energy-based perspective. The proposed JEM optimization objective estimates the joint distribution $p(\mathbf{x}, y)$ from a generative perspective, which requires estimating the normalized densities and can be intractable and unstable to compute. Liu et al. propose to use energy score for OOD detection that is derived from a pure discriminatively trained classifier, which demonstrated superior performance for multi-class classification networks. In contrast, our work focuses on a multi-label setting, where we contribute aggregation methods that utilize information jointly from across all labels.

5 Conclusion

In this work, we propose energy scores for OOD detection in the multi-label classification setting. We show that aggregating energies over all labels into *SumEnergy* results in better discrimination between in-distribution and OOD inputs compared to using information from only one label's information (i.e. MaxLogit, MSP, ODIN, or Mahalanobis). Additionally, we justify the relationship between an input's probability distribution and its energy score through a mathematical derivation. Energy scores obtain better OOD detection performance compared to competitive baseline methods, establishing new state-of-the-art on this task. Applications of multi-label classification can benefit from our methods, and we anticipate further research in OOD detection to extend this work. We hope our work will increase the attention toward a broader view of OOD uncertainty estimation from an energy-based perspective.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection via informative outlier mining. *arXiv preprint arXiv:2006.15207*, 2020.

Long Chen, Wujing Zhan, Wei Tian, Yuhang He, and Qin Zou. Deep integration: A multi-label architecture for road scene recognition. *IEEE Transactions on Image Processing*, 28(10):4883–4898, 2019.

Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. 2018.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., 2009.

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)*, 2014a.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014b.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. *arXiv preprint arXiv:1901.09192*, 2019.

Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*, 2019.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification—revisiting neural networks. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 437–452. Springer, 2014.

Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann L Cun. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pp. 1137–1144, 2007a.

Marc'Aurelio Ranzato, Y-Lan Boureau, Sumit Chopra, and Yann LeCun. A unified energy-based framework for unsupervised learning. In *Artificial Intelligence and Statistics*, pp. 371–379, 2007b.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pp. 14680–14691, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 693–700, 2010.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

Akshayvarun Subramanya, Suraj Srinivas, and R Venkatesh Babu. Confidence estimation in deep neural networks via density modelling. *arXiv preprint arXiv:1707.07013*, 2017.

Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pp. 4790–4798, 2016.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2285–2294, 2016.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2019.

Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5513–5522, 2017.