An Exploration of Near Out-of-Distribution Detection

Juan Rios [1]

Abstract

Out-of-distribution (OOD) detection is an active research area with many successfully models developed in the last few years. A forthcoming concern is the ability of a learning model to correctly flag OOD inputs when the input is considered near-OOD. Existing works test near-OOD performance when training on CIFAR-10 and testing on CIFAR-100 or vice-versa. We propose a more rigorous benchmark to test new frameworks in the near-OOD domain. First, we follow the intuition that OOD distance correlates with input object semantics similarity. Then, we use our benchmark to test this intuition using energy-based OOD detection. Furthermore, we propose a few adjustments to the m_{out} parameters in the energy loss function, and to the calculation of the energy score $E(x)$. Finally, we test these adjustments in an extreme near-OOD testing environment.

1. Introduction

Out-of-distribution (OOD) detection is crucial for the safe deployment of machine learning models such as in domains where erroneous predictions have substantial negative impact. One cause is the attempted classification of inputs whose distribution differs from the distribution of training samples. (Amodei et al., 2016) discuss in detail how, much like humans, machine learning models struggle in recognizing what the model does not "know". when the testing (real-world) distribution differs from training, these models often perform poorly, in addition to being highly-confident in their decisions. This behavior hinders the trust, and hence, the deployment of these models in critical areas such as medicine, law, autonomous vehicles, etc.

1.1. Out-of-Distribution Detection Frameworks

Much effort has been done in allowing models to detect OOD inputs. For example, (Hendrycks & Gimpel, 2018) establishes a baseline method for OOD detection using probabilities from softmax distributions. The proposed ODIN framework by (Liang et al., 2020) improves the baseline by adding temperature scaling and small perturbations to the input to further separate the softmax distributions of

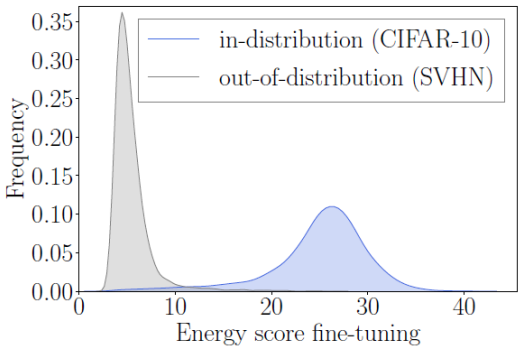

Figure 1. Energy scores for in-distribution and out-of-distribution samples from energy-based fine-tuning of a model.

in-distribution (ID) and OOD samples. Other OOD detection methods such as Outlier Exposure leverage data to train anomaly detectors against an auxiliary dataset of outliers (Hendrycks et al., 2019). (Liu et al., 2021) adapt the objective function to shape the energy surface explicitly for OOD detection. The model computes an energy score $E(x)$ for an input x where the distribution of energy scores for ID inputs is distinguished from the distribution of scores for OOD samples as shown in Figure 1. Another method by (Lee et al., 2018) derives a confidence score using the mahalanobis distance between class conditional Gaussian distributions with respect to deep model features. (Sastry & Oore, 2020) use gram matrices to characterize activity patterns and detecting anomalies by identifying abnormal gram matrix values to increase OOD detection rates. More recently, (Fort et al., 2021) explores the use of vision transformers in OOD detection.

1.2. The Problem of Near Out-of-Distribution

One burgeoning topic in OOD detection, is a distinction between *near* and *far* OOD. Many of the aforementioned frameworks perform well when the in-distribution (ID) and OOD samples are sufficiently different. For example, a model trained to discriminate cats and dogs will have an easier time flagging an image containing numerical digits versus an image containing coyotes. (Winkens et al., 2020) make such distinction, and evidence this issue by showing that a model trained on CIFAR-100 (containing images of

fish, trees, vehicles, etc.) performs better at flagging OOD samples from SVHN (containg street view house numbers) than from CIFAR-10, whose classes are semantically similar to CIFAR-100. (Sastry & Oore, 2020) achieve a 99% AUROC when the ID set is from CIFAR-100 and the OOD set from SVHN. But OOD detection rates for *near* OOD detection is much lower than *far* OOD such as in (Zhang et al., 2020), where the AUROC between CIFAR-100 (ID) and CIFAR-10 (OOD) is 85%.

Despite the difficulties, it is important for models to distinguish between *near*-OOD and ID inputs. For example a model trained to differentiate between lymphoma cancer cells against healthy tissue in lymph node images, where it is possible that abnormal tissue absent from the training samples (metastatic tissue) might be incorrectly classified, rather than flagged as OOD. Abnormal tissue might not appear in training since medical image databases tend to be small and hard to acquire, such as discussed by (Ling et al., 2021) due to the arduous methods of labeling images by medical professionals such as pathologists. A less speculative example is demonstrated by (Ren et al., 2019), where a model trained to detect bacteria through genomics sequences may be fooled by what can be considered *near*-OOD genomics sequences during real-world deployment.

2. Understanding Sample Similarity

The assumption has been that samples whose semantics or "similarity" are very different to training samples are considered *far*-OOD while similar samples are considered *near*-OOD. (Fort et al., 2021) and (Ren et al., 2021) apply this intuition classifying samples as *near* or *far* based on semantics. Such intuition is portrayed in Figure 2. Additionally, these works summarize near-OOD detection performance mainly by testing CIFAR-10 against CIFAR-100 or vice-versa.

The main goal of this paper is to design an experiment that (1) more rigorously tests this intuition of OOD distance corresponding to object semantic distance, and (2) design a dataset to test OOD detection frameworks in the realm of near-OOD.

Contributions:

- We build a dataset to gain insight better inisght regarding the intuition that object semantics distance follows OOD distance.
- We experiment with energy-based OOD detection (Liu et al., 2021) in this new environment.
- We further tweak energy-based OOD by adjusting the role of the m_{out} in the energy loss function, and adjust the energy score function $E(x)$ based on observations derived from this experimental setup.
- Lastly, we test these adjustments in an extremely near-OOD testing environment.

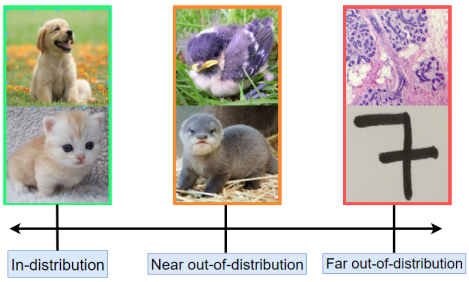

Figure 2. A diagram depicting the correlation between semantic similarities and the distribution distance of test samples. A model trained to differentiate cats and dogs will have a harder time flagging near but out-of-distribution samples such as otters or birds, compared to far out-of-distribution samples such as numeric digits or breast tissue.

2.1. Background: Object Semantics

Object semantic concepts from the domain of computational cognitive psychology are used to design the dataset and the experiment. These concepts are summarized as follows:

A proposed model for humans learning nouns relies on the *taxonomic assumption*, which states that words (class labels) refer to taxonomic classes in a tree-like structure of natural categories (Markam, 1991). The work of (Xu & Tenenbaum, 2007) builds on this to develop a Bayesian model of word inference to show how adult learners generalize new words to objects that are semantically similar to each other. These objects fall into different levels in the taxonomy tree, for example at a basic level, objetcs like cats, dogs, whales and other mammals share this level, but at a lower taxonomy level the dog category would consist of different dog breeds like dalmatians, terriers, huskies, etc.

The experiments in (Xu & Tenenbaum, 2007) are designed to explore how adult learners generalize examples of a novel word on the taxonomy tree. The experiments use every day objects with an intuitive taxonomy of varying levels. In one phase, participants are given a novel word in a new language and pair the word with images of the object the word applies to. Then the participants are shown a series of objects and decide which of these objects does the novel word also apply to. More relevant to this paper is the second phase, where participants are shown images of every day objects such as dalmatians, non-dalmatian dogs, pigs, trucks, vegetables, etc. The participants are tasked with giving similarity ratings between objects. The ratings were collected and a taxonomy tree of objects is built using a hierarchical

clustering algorithm. The closer a pair of objects appear in the tree, the higher the judged similarity between the pair, for example pairs of vegetables are very close. Objects judged less similar appear more distant on the tree, such as the dog-motorcycle pair. A partial view of the resulting tree is shown in Figure 3.

2.2. Background: Energy-Based Out-of-Distribution Detection

This paper uses the energy-based OOD detection framework from (Liu et al., 2021) for its good performance and straightforward implementation. A brief summary of this framework is given:

An energy-based model (LeCun et al., 2006) is a function that maps an input x in put space to a scalar called *energy*. Through the Gibbs distribution, a collection of energy values can be turned into a probability density such as:

$$p(y|x) = \frac{e^{-E(x,y)/T}}{e^{-E(x)/T}} \tag{1}$$

Consider x an image, and y the label. Equation (1) models the conditional probability of a label given an image x, where $-E(x, y)$ is the energy between x and y. The lower this energy, the higher is their joint probability. T is a temperature scaling parameter, and $E(x)$ is the *Helmholtz free energy* given by:

$$E(x) = -T \cdot log \int_{y'} e^{-E(x,y')/T} \, dx \tag{2}$$

Consider a classifier $f(x)$ that maps an image $x \in R^D \rightarrow R^K$ into K real-valued numbers called logits where K is the number of classes. These logits are passed to a softmax equation to derive a class distribution:

$$P(y|x) = \frac{e^{f_y(x)/T}}{\sum_i^K e^{f_i(x)/T}} \tag{3}$$

Here, $f_y(x)$ is the logit for the y^{th} class label computed by the classifier $f(x)$. The connection between equation (1) and equation (3) approximates $E(x, y) = -f_y(x)$. The free energy function $E(x)$ can be defined as:

$$E(x, f) = -T \cdot log \sum_i^k e^{f_i(x)/T} \tag{4}$$

a sample x is considered OOD if the negative of the free energy function is lower than a threshold. In other words, x is OOD if $-E(x, f) \leq \tau$ and ID otherwise. The classifier f calculates this free energy, while τ is chosen so that most of the samples in a set of ID data are classified as ID.

Lastly a modified loss function that promotes the separation of energy between ID and OOD data and is given by:

$$L_{\text{total}} = L_{\text{ce}} + \lambda \cdot L_{\text{energy}} \tag{5}$$

L_{ce} is the standard cross entropy loss between the ground-truth y and the prediction probability $p(\hat{y})$:

$$L_{\text{ce}} = -\sum_i^K y^i \cdot log(p(\hat{y}^i)) \tag{6}$$

λ is regularization parameter held constant at 0.1 and L_{energy} is a regularization loss given by:

$$\begin{aligned} L_{energy} = \max(0, E(x_{in}) - m_{in})^2 + \\ \max(0, m_{out} - E(x_{out}))^2 \end{aligned} \tag{7}$$

Here, $E(x_{in})$ is the energy for a training sample computed by equation (4), and $E(x_{out})$ is the energy from an OOD input acquired from an auxiliary unlabeled OOD dataset. Because we want low energy for ID and high energy for OOD images, this regularization loss is high when an ID image has high energy or an OOD image has low energy.

3. Experimental Setup

To understand the notion of OOD distance, object semantics, and model behavior, we design a series of experiments to see how energy-based OOD detection behave in the presence of data picked to resemble the taxonomy tree described in the background section and shown in Figure 3.

3.1. Data

Although we do not conduct any human experiments to construct our own tree and handcraft our own data, we approximate the taxonomy tree by using INaturalist data from the INat 2021 competition. The data consists of images of various animals and plants organized by species as in the biological taxonomy tree. For example a tiger belongs to the following categories with increasing granularity: The *animalia* kingdom, the *chordata* phylum, the *mammal* class, and the *carnivora* order.

The assumption is that this biological tree, much like the object tree in (Xu & Tenenbaum, 2007), is a clustering of animals based on their similarity. The farther the organisms are from *carnivora* class in the biological tree, the more dissimilar in both biological form and function. A second assumption is that more dissimilar organisms correspond

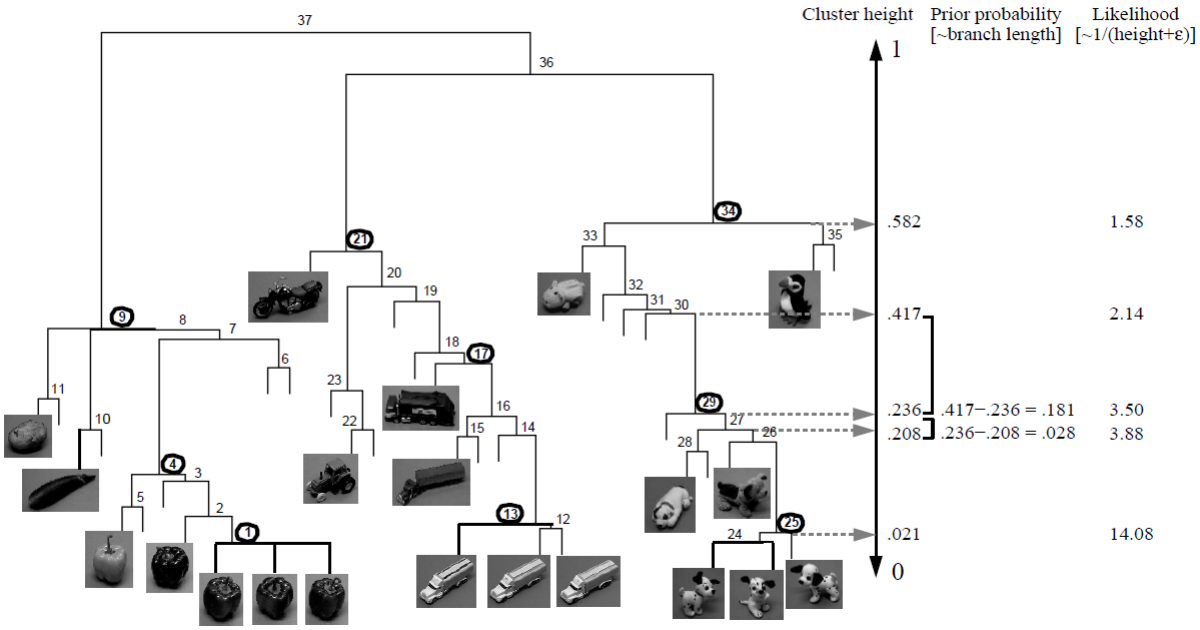

Figure 3. The resulting tree from hierarchical clustering of similarity ratings (Xu & Tenenbaum, 2007). This tree structure clusters objects that are semantically similar together, and inter-cluster distances increase as the semantics of the clusters deviate away from each other.

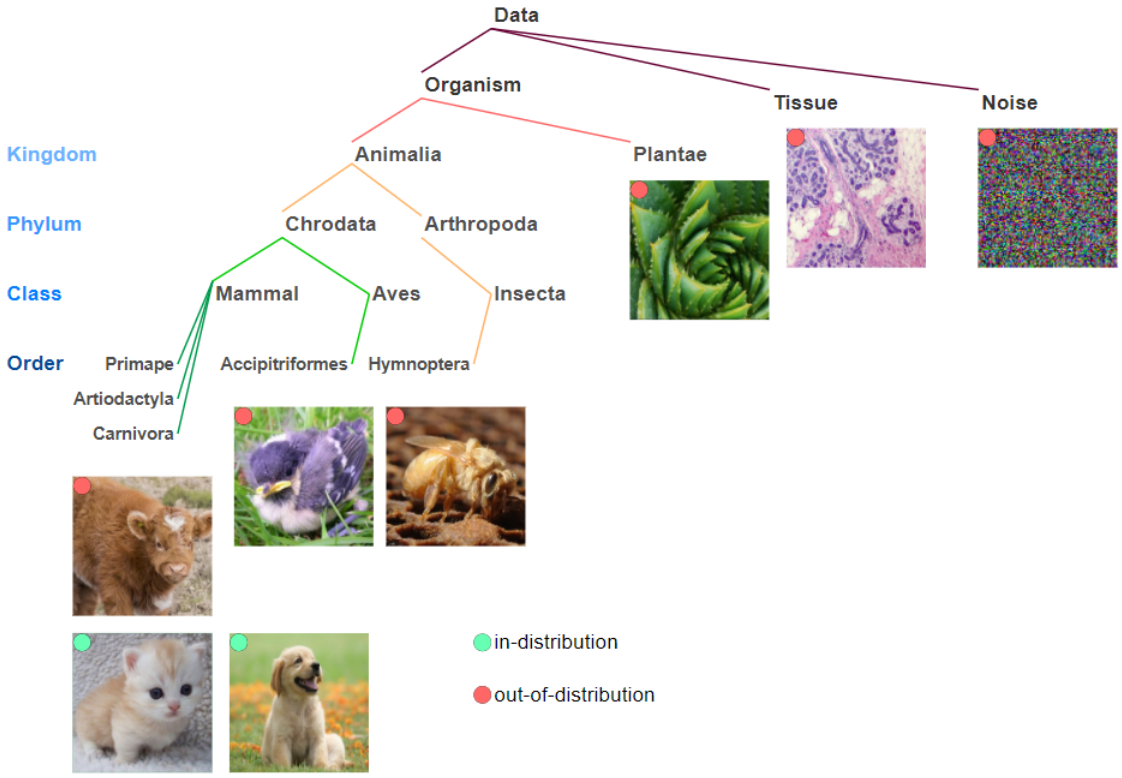

Figure 4. This tree structure describes the taxonomy of our data. This data shows the in-distribution classes as *carnivora* and out-of-distribution data of varying distances.

with samples that are farther OOD, and similar organisms are closer OOD. Not only does this apply to the organism depicted in the image, but also the background scenes. For example, close to *carnivora* are animals belonging to *artiodactyla* whose image backgrounds are shared by those of *carnivora* since they co-inhabit the same space. Birds and insects not only have dissimilar image protagonists, the background varies as well. Instead of the mammal's open plains, forests, and beach backgrounds, we now see treetops, tree barks, and zoomed-in scenery. Ideally, the model would see and predict objects invariant to the background, but this types of algorithms such as Invariant Risk Minimalization (Arjovsky et al., 2020) are not presented here but should be included in future work.

An example of the data organization is as follows: First, a model is trained to see animals in the *carnivora* order (lions, tigers, seals). *Near*-OOD organisms share the same *mammal* class but not the order (*artiodactyla*, *rodentia*, and *primape*). With increasing distance, birds belong to the same phylum but another class (*aves*). Insects belong to the same kingdom (*animalia*) but another phylum (*arthropoda*). Plants are also organisms but have different kingdoms (*plantae*). The most distant OOD images come from datasets collected from microscopic slides of breast and lymph node tissue, along with images of random noise. An overview of the taxonomy of the data is shown in Figure 4.

3.2. Training

The models used are pre-trained Resnet-18s using PyTorch. The ID training dataset contains 54 animals belonging to the *carnivora* order. the size of this data is approximately 15,000 images evenly distributed among 54 classes. The batch size for all training is 50, the initial learning rate is 0.005 and follows a cosine decay schedule. A momentum of 0.9 used, and the training loss is either standard cross entropy from equation (6) or the total loss from equation (5). The auxiliary OOD data is Imagenet (Deng et al., 2009) with images of organisms removed.

we use the energy-based model from (Liu et al., 2021) as described. The first experiment takes a pre-trained model and fine-tunes on the *carnivora* dataset with only the cross-entropy loss. Subsequent experiments train a model using the combined cross-entropy and energy loss. The default m_{in} is -27 as done by (Liu et al., 2021) and m_{out} is -5, although m_{out} will be adjusted in a later experiment.

3.3. Evaluation Metrics

The *carnivora* data is split $80/20$ for training and an additional test set is used for OOD evaluation. The OOD datasets evaluated on are in order of semantic distance from carnivora as described in the Data subsection: *artiodactyla*, *primape*, *rodentia*, *accipitriformes*, *hymnoptera*, *plantae*, *tissue*, and *noise* The performance of a model in terms of accuracy is measure through its classification accuracy on the validation set by its top-1 and top-5 accuracy, which are consistently around 53% and 74% respectively. The images collected from INaturalist are visibly noisier than images from manicured datasets such as CIFAR-10. The OOD evaluation is done using FPR95 which is the false positive rate of OOD samples when the true positive rate for ID samples is at 95%

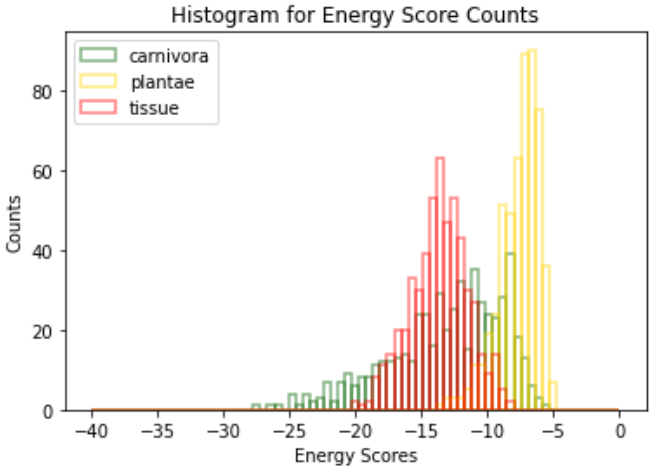

Figure 5. Energy scores for *carnivora* (ID), *plantae* (OOD), and *tissue* (OOD) datasets for a model trained with cross entropy loss.

4. Experimental Results

The first experiment takes model a trained with cross-entropy (at-inference) and evaluates the FPR95 of the OOD data. The results are shown in Table 1. The FPR95 shows that as the datasets become more distant in the taxonomy tree, the easier the model is able to detect the data as OOD. This provides the first evidence that the assumption of *near*- and *far*-OOD may correlate to semantic distance. This assumption is true for the most distant OOD datasets *tissue* and *noise*. Their mean energy scores are very close to the mean energy score for the ID data;. The model cannot distinguish between *carnivora*, *tissue*, and *noise*. A histogram of the energy scores for *carnivora*, *plantae*, *tissue* are shown in Figure 5. From this experiment, it can be concluded that using at-inference energy-based OOD detection does not necesarily follow object semantics in terms of the difficulty in OOD detection. For example, the nearest OOD animals in *artiodactyla* have a high FPR95 (0.92), but tissue and noise images have even higher FPR95s (1.00 for both).

A second experiment trains a model using L_{energy} described in equations (5) and (7). The results are shown in Table 2. The most drastic changes are for *tissue* and *noise* datasets,

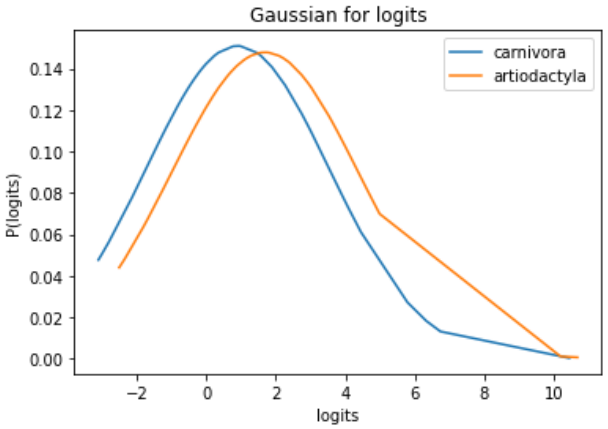

Figure 6. Distribution of logits for samples of *carnivora* and *artiodactyla* for class '0'(golden jackal) for model with cross entropy loss

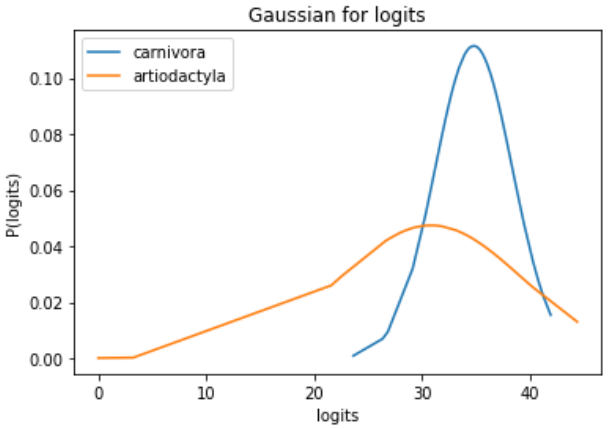

Figure 7. Distribution of logits for samples of *carnivora* and *artiodactyla* for class '0'(golden jackal) for model with cross entropy and energy loss.

with FPR95 of 0.13 and 0.00 respectively (down from 1.00). The datasets closer to *carnivora* experienced mixed results. Overall, the results show a trend in decreasing performance as the datasets get semantically closer to *carnivora*. This is not always the case. From the results, plants are harder to detect than insects (*hymnoptera*), even though insects are closer semantically to *carnivora*. One question is the effect of the background found in both images, which tend to be very similar, since these insects will often be pictured next to foliage. Additionally, plants and birds (*accipitriformes*) are on-par in terms of FPR95s, suggesting that object semantics is not always a clear indicator of OOD distance. However, there is a general trend in decreasing FPR95 with increasing semantic distance.

An interesting observation is that although the energy loss enables more separation in mean energy score between ID and OOD datasets, there is an increase in spread of these energy scores evidenced by the increasing standard deviation, reducing the effectiveness of OOD detection.

Since energy scores are non-deterministically computed from the logits, an investigation of these logits is shown in Figure 6 and 7. Figure 6 shows the distribution of logits for class '0' (golden jackal) for a series of samples from *carnivora* and *artiodactyla* regardless of predicted class. This distribution is for a model trained solely with cross-entropy loss (at-inference). Figure 7 shows the logits but with the additional energy loss. These distributions reveal the effect of the energy loss from equation (7) on the logits for class '0'. this trend is the same for every class.

There is a clear separation in energy means between *carnivora* and *artiodactyla* using energy loss. However, the distribution of logits from *artiodactyla* samples spread far apart, hindering the model from performing better. This is also evidenced by the increase in standard deviation of energy scores shown in Table 2 compared to Table 1

Initially, m_{out} is taken to be -5.0 with the intent of 'pulling up' OOD energy scores, and m_{in} is -27.0 with the intent of 'pulling down' ID scores. (Liu et al., 2021) recommends to take m_{out} as the mean of scores for OOD samples. The model punishes OOD scores below m_{out} and ID scores above m_{in}. This setup causes *near*-OOD samples to have such high variability in energy scores, and ID sample suffer similarly. An intuition is to train with $m_{out} = -20.0$. Initially, this will not punish most OOD samples, but acts as a barrier that slows down the increase in energy/logit spread. This proves to be true empirically; the mean of energy scores remaind separated, and the deviations narrow.

The effect of $m_{out} = -20.0$ gives energy scores of ID samples with higher spread compared to OOD samples across all OOD datasets increasing detection performance (decreasing FPR95s). This leads to a modification of equation 4. Minimizing $E(x, f)$ can be done by maximizing the sum $\sum_i^k e^{f_i(x)/T}$. Originally, the free energy $E(x) = -f_y(x)$. We add a small modification that results in $E(x) = -(f_y(x) + \sigma_{f(x)})$ where $\sigma_{f(x)}$ is the standard deviation of all K logits computed by passing a sample x into classifier f. Because the observed $\sigma_{f(x)}$ is higher for ID samples, this results in a greater decrease of energy scores for ID data. Substituting the modified $E(x)$ into equation (4) results in:

$$E(x, f) = -T \cdot log(\sum_i^k e^{f_i(x)/T}) - \sigma_{f(x)} \qquad (8)$$

The change in training to $m_{out} = -20.0$ and the use of equation (8) achieve a small overall improvement to FPR95

scores as shown in Table 3. The most considerable reduction of FPR95 is for *accipitriformes* from 0.57 to 0.43 and second is *rodentia* from 0.95 to 0.88. Most other OOD datasets experiences small reductions in FPR95 with the exception of *tissue* which increased by 0.1. Using $m_{out} = -20.0$ did not decrease the classification accuracy of the model. Although empirically implementing equation (8) reduces FPR95 by a small margin, this relies on the assumption that ID data has a larger varience in logit distribution that OOD data. This only holds true for the 7 OOD datasets observed, and may not apply for the countless possibilities of OOD samples. An illustration of the final results in Table 3 is shown in Figure 8.

4.1. Testing in Extreme Near-OOD

We additionally test the effect of $m_{out} = -20.0$ and equation (8) in a more extreme near-OOD environment. The ID data come from image patches of healthy breast and lymph node tissue. The OOD dataset include breast tissue patches with benign turmor, invasive ductal carcinoma, and in situ ductal carcinoma. The datasets are gathered from (Peikari et al., 2017), (et al., 2019), and (Campanella et al., 2019). Table 4 shows the FPR95 for benign, in situ, and invasive OOD sets using at-inference only, with energy loss, $m_{out} = -5.0$ and the original equation (4), and with energy loss at $m_{out} = -20.0$ and the proposed equation (8).

Curiously, the best results come from using inference only (no energy loss). And using energy loss vastly degrades performance. The explanation here is that energy loss does not separate mean energy scores between ID and OOD datasets. One intuition behind this is the use of the auxiliary dataset which is very far semantically from both the ID and OOD data (ImageNet). Future work should explore the effect of choice of auxiliary datasets. The use of $m_{out} = -20$ shows to reduce the standard deviation of OOD data, leaving ID data to have higher standard deviation in energy scores, but using $m_{out} = -20$ alone does not increase performance unless paired with the proposed equation (8).

5. Related work

The work of (Fort et al., 2021) explores the use of transformers for near-OOD detection. They note an improvement in AUROC scores for CIFAR-100 versus CIFAR-10 OOD detection from 85% to 96% by using vision transformers pre-trained on ImageNet-21k. Additionally, They report an improvement on a genomics benchmark from 66% to 77% using a vision transformer and unsupervised pre-training. Furthermore, they report that pairing vision transformers with few-shot outlier exposure improves the CIFAR-100 versus CIFAR-10 AUROC to 99.46%. One key reason for such an improvement is that vision transformers can naturally distinguish between classes better than Resnet. This is evidenced by projecting the image embeddings from the semi-last layer using principal component analysis, which shows a higher degree of class-separation for transformers comapred to ResNet.

(Winkens et al., 2020) explore the topic of near-OOD detection and develop a metric for measuring the difficulty of an OOD detection task termed the *Confusion Log Probability*. The next step in research would analyze how well this new metric aligns with the dataset proposed in this paper. If so, the *Confusion Log Probability* along with the data used here can be used to evaluate the performance of near-OOD detection frameworks. One top candidate for this future research is the vision transformers used by (Fort et al., 2021). The idea is to have a more rigorous testing compared to CIFAR-10/CIFAR-100 benchmarks.

6. Conclusion

We explore the domain of near out-of-distribution detection (OOD) by developing a benchmark for models using a dataset that follows an intuitive organization in a taxonomy tree. This proposed benchmark would be used to evaluate the effectiveness of OOD detection models in near-OOD environments. Additionally, we show the results of energy-based OOD detection (Liu et al., 2021) using this benchmark, and provide evidence for the intuition that OOD distance correlates with semantic distance, although not strictly. For example, detecting birds and plants had on-par performance, when the in-distribution data included carnivorous mammals (lions, tigers, coyotes).

Furthermore, we adjust the m_{out} value in the energy-loss function, and provide an additional term in the energy score equation to improve performance of the energy-based framework. Lastly, we show that these adjustments can help inprove performance in the extreme near-OOD task of flagging cancer tissue as OOD when the ID images contain healthy breast and lymph node tissue.

7. Contributions

Juan Rios constributed 100% of the effort.

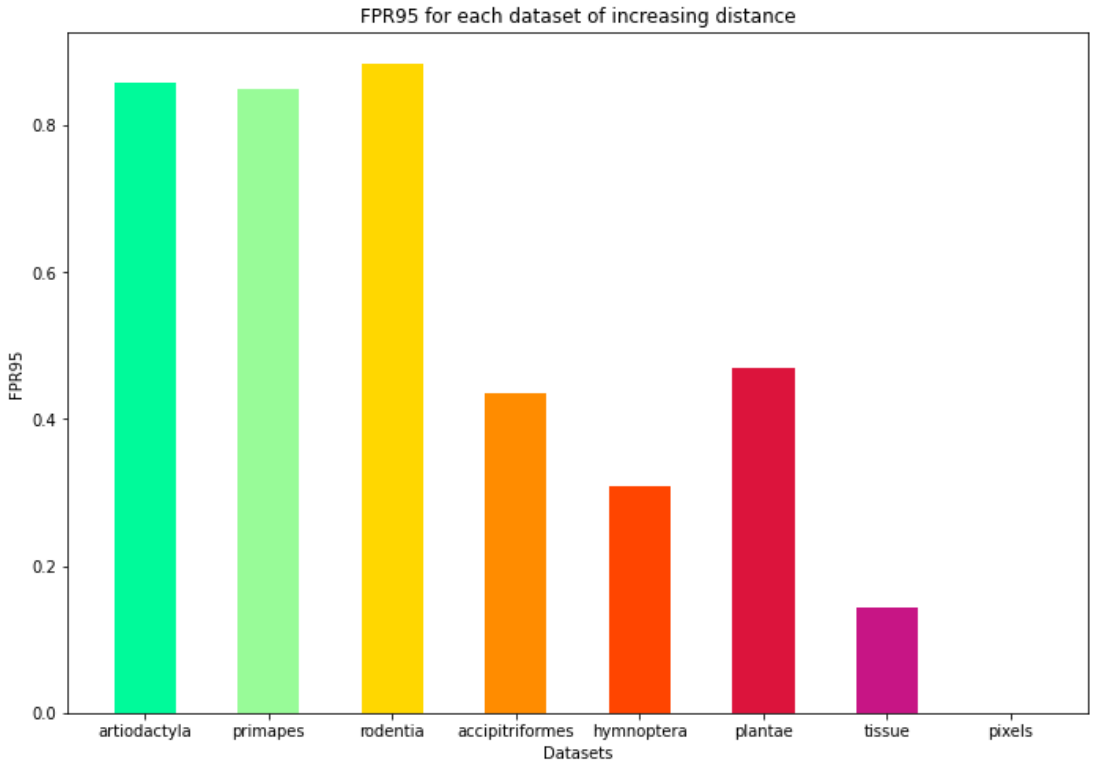

Figure 8. An illustration of FPR95 resultsfrom Table 3. The datasets are ordered with increasing semantic distance from left to right with respect to *carnivora* dataset.

Table 1. The FPR95 along with mean and standard deviation of energy scores for the in-distribution and out-of-distribution datasets using cross entropy loss only.

Cross-entropy loss only. OOD with increasing semantic distance $\rightarrow$									
Results	Carni. (in)	Artio.	Primape	Rodentia	Accipit.	Hymn.	Plantae	Tissue	Noise
FPR	0.05	0.92	0.90	0.78	0.58	0.70	0.45	1.00	1.00
$\mu_{E(x)}$	-13.11	-10.99	-9.28	-10.41	-8.24	-8.69	-7.71	-13.04	-14.22
$\sigma_{E(x)}$	4.29	2.75	2.13	2.53	1.81	1.85	1.54	2.24	0.67

Table 2. The FPR95 along with mean and standard deviation of energy scores for the in-distribution and out-of-distribution datasets using energy loss L_{energy}.

OOD with cross entropy loss									
Results	Carni. (in)	Artio.	Primape	Rodentia	Accipit.	Hymn.	Plantae	Tissue	Noise
FPR	0.05	0.89	0.87	0.95	0.57	0.34	0.57	0.13	0.00
$\mu_{E(x)}$	-43.87	-38.75	-38.31	-39.16	-27.44	-19.05	-26.10	-16.70	-4.91
$\sigma_{E(x)}$	10.91	9.70	11.11	8.18	13.73	14.04	11.77	8.89	0.11

Table 3. The FPR95 along with mean and standard deviation of energy scores for the in-distribution and out-of-distribution datasets using energy loss L_{energy} along with setting $m_{out} = -20.0$ and using equation (8)

OOD with cross entropy loss									
Results	Carni. (in)	Artio.	Primape	Rodentia	Accipit.	Hymn.	Plantae	Tissue	Noise
FPR	0.05	0.86	0.85	0.88	0.43	0.31	0.47	0.14	0.00
$\mu_{E(x)}$	-40.54	-35.53	-35.18	-35.83	-28.86	-25.99	-28.61	-24.99	-15.44
$\sigma_{E(x)}$	7.08	5.81	6.06	5.49	5.83	6.79	5.42	3.70	0.31

Table 4. The FPR95 for ductile carcinoma in situ, ductile carcinoma invasive, and benign tumors using energy-based out-of-distribution detection on breast tissue

Energy-Based FPR95			
Mode	Benign	In Situ	Invasive
At-inference	0.54	0.48	0.79
Energy Loss eq. 4 $m_{out} = -5$	0.97	0.98	1.00
Energy Loss eq. 4 $m_{out} = -20$	0.97	0.98	0.99
Energy Loss eq. 8, $m_{out} = -20$	0.78	0.75	0.93

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety, 2016.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2020.

Campanella, H., Andrea Brogi, M., and Fuchs, T. Breast metastases to axillary lymph nodes [data set]. 2019. doi: https://doi.org/10.7937/tcia.2019.3xbn2jcc.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

et al., G. A. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.05.010.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection, 2021.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure, 2019.

LeCun, Y., Chopra, S., Hadsell, R., Huang, F. J., and et al. A tutorial on energy-based learning. In *PREDICTING STRUCTURED DATA*. MIT Press, 2006.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks, 2020.

Ling, M., Lv, G., Wang, J., Hao, X., Shi, J., and An, H. Fast whole slide image analysis of cervical cancer using weak annotation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1037–1041, 2021. doi: 10.1109/ISBI48211.2021.9433964.

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based out-of-distribution detection, 2021.

Markam, E. Categorization and naming in children: Problems of induction. ellen markman. cambridge, ma: Mit press, 1989. pp. 250. *Applied Psycholinguistics*, 12(3): 385–392, 1991. doi: 10.1017/S0142716400009310.

Peikari, M., Salama, S., Nofech-Mozes, S., and Martel, A. L. Automatic cellularity assessment from post-treated breast surgical specimens. *Cytometry Part A*, 91(11): 1078–1087, 2017. doi: https://doi.org/10.7937/TCIA.2019.4YIBTJNO.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection, 2019.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. A simple fix to mahalanobis distance for improving near-ood detection, 2021.

Sastry, C. S. and Oore, S. Detecting out-of-distribution examples with in-distribution examples and gram matrices, 2020.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., Cemgil, T., Eslami, S. M. A., and Ronneberger, O. Contrastive training for improved out-of-distribution detection, 2020.

Xu, F. and Tenenbaum, J. Word learning as bayesian inference. *Psychological review*, 114:245–72, 05 2007. doi: 10.1037/0033-295X.114.2.245.

Zhang, H., Li, A., Guo, J., and Guo, Y. Hybrid models for open set recognition, 2020.