SNEAKOSCOPE: REVISITING UNSUPERVISED OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The problem of detecting out-of-distribution (OOD) examples in neural networks has been widely studied in the literature, with state-of-the-art techniques being supervised in that they require fine-tuning on OOD data to achieve high-quality OOD detection. But supervised OOD detection methods also have a disadvantage in that they require expensive training on OOD data, curating the OOD dataset so that it is distinguishable from the in-distribution data, and significant hyper-parameter tuning. In this work, we propose a unified evaluation suite, *Sneakoscope*, to revisit the problem with in-depth exploration of unsupervised OOD detection. Our surprising discovery shows that (1) model architectures play a significant role in unsupervised OOD detection performance; (2) unsupervised approaches applied on large-scale pre-trained models can achieve competitive performance compared to their supervised counterparts; and (3) unsupervised OOD detection based on Mahalanobis Distance with the support of a pre-trained model consistently outperforms other unsupervised methods by a large margin and compares favorably with results from state-of-the-art supervised OOD detection methods reported in the literature. We thus provide new baselines for unsupervised OOD detection methods.

1 INTRODUCTION

Deep neural networks (DNNs) have attained unprecedented success in various tasks across domains such as object recognition in computer vision (Krizhevsky et al., 2012), machine translation in natural language processing (Bahdanau et al., 2014), and protein structure prediction in biology (Jumper et al., 2021). However, such breakthrough hinges on the assumption that unseen test data distribute identically as their training counterparts. In fact, DNNs have been found to make over-confident predictions for semantically meaningless inputs. For example, (Hendrycks & Gimpel, 2017) shows an MNIST image classifier gives a predicted class probability of 91% to an input of random Gaussian noise. The problem of DNNs being overly assertive to semantic data distribution shifts has motivated a rich literature of detecting anomalous inputs like out-of-distribution (OOD) examples (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018a;b; Hendrycks et al., 2019a;b; Grathwohl et al., 2020; Liu et al., 2020; Fort et al., 2021; Lin et al., 2021).

Initial OOD detection methods were *unsupervised* (e.g., Maximum Softmax Probability-based detector (Hendrycks & Gimpel, 2017) or unsupervised version of Mahanalobis Distance-based detector (Lee et al., 2018b)), in that they only required access to a classifier trained on in-distribution examples, but not any OOD data itself. But, state-of-the-art OOD detection methods tend to be *supervised*, in which the detector is further fine-tuned (trained) on OOD data (Lee et al., 2018b; Hendrycks et al., 2019b; Liu et al., 2020). This fine-tuning on OOD data is considered essential for boosting detector performance. For example, Hendrycks et al. (2019b) cuts down false positive rate from 42.80% (unsupervised) to 12.20% (supervised) on one OOD dataset by fine-tuning the MSP-based CIFAR-10 detector on a very broad set of OOD examples. Unfortunately, this fine-tuning also has its own limitations. Training on wide range of ODD data involves considerable additional effort, making it harder to do quick development of OOD detectors. Such detectors (e.g., ODIN (Liang et al., 2018)) also require an auxiliary OOD dataset to fine-tune the hyper-parameters of the trained classifier or the pre-processing pipeline, but the OOD dataset may not be representative of incoming OOD examples. Some supervised detectors, such as Outlier Exposure (Hendrycks et al., 2019b),

heavily rely on training the detector against massive amounts of OOD data; such data may not always be available. In Outlier Exposure, it requires careful curation relative to the training classes.

In this work, we revisit unsupervised approaches to the OOD detection problem and ask the question if they can achieve state-of-the-art results on OOD detection without needing supervised fine-tuning on OOD data, overcoming some of the limitations of supervised methods. To do an in-depth exploration of this issue, we propose *Sneakoscope*, a unified evaluation suite that systematically investigates the performance of different unsupervised OOD detection methods on classifiers with different model architectures and with and without large-scale pre-training.

We show that an unsupervised approach based on Mahalanobis Distance (Lee et al., 2018b) consistently stands out after it is augmented with support of a classifier that is pretrained on a large-scale dataset (e.g., the Vision Transformer or ResNet-based Big Transfer). Mahalanobis Distance was originally proposed by Lee et al. (2018b) for OOD detection, but the original approach did not work well in unsupervised mode (false positive rate of 45.60%); their best supervised version had a false positive rate of 3.58% on CIFAR-10 vs. SVHN but required fine-tuning of input pre-processing and feature ensemble on OOD data. However, our experiment results indicate that Mahalanobis Distance without any fine-tuning is still a better OOD detection method when compared to other unsupervised approaches based on MSP and energy scores. We also demonstrate that using an underlying classifier that uses large-scale pre-training uplifts the performance of all evaluated unsupervised detection methods while Mahalanobis Distance gains the most from pre-training. Surprisingly, the results compare favorably with reported results from state-of-the-art supervised detectors.

In summary, *Sneakoscope*'s analysis suggests unsupervised OOD detection methods that are based on existing classifiers that rely on large-scale pre-trained datasets (e.g., the Vision Transformer and Big Transfer) can compare favorably with state-of-the-art supervised OOD detection methods that require fine-tuning a detector on OOD data. Furthermore, Mahalanobis Distance is a good underlying measure for unsupervised OOD detection and outperforms MSP and Energy Score. Our work provides new baselines for unsupervised OOD detection methods. For instance, the combination of pre-trained ViT and Mahalanobis Distance reduces the false positive rate (at 95% true positive rate) on CIFAR-10 vs. SVHN from 67.81% (ResNet with MSP) to 0.43%. The rest of paper includes *Sneakoscope* overview, experiments results, and detailed analysis and discussion that provides insights into the findings.

2 *Sneakoscope*: A Unified Evaluation Suite

In this section, we start by describing the problem setup, and then give an overview of key components in *Sneakoscope*. Unlike prior works, we focus on studying the effects of model architectures and large-scale pre-training on unsupervised OOD detection. *Sneakoscope* incorporates both factors into the evaluation, and provide visual and quantitative analysis to justify the results.

2.1 Preliminaries

We consider the problem of predicting whether an input at inference time comes from a distribution different from training data. Given a trained classifier $\mathcal{F}$ on an in-distribution $\mathcal{X}$ with a fixed label space $\mathcal{Y} = \{1, \ldots, K\}$, a detector $\mathcal{G} : \mathcal{X}' \to \{0, 1\}$ (where $\mathcal{X}'$ is the sample space) assigns label 0 to an out-of-distribution sample (negative) and label 1 to an in-distribution sample (positive).

We also describe here the prominent issues of supervised detection approaches.

- Hyper-parameter fine-tuning or training a separate detector with a modified training objective requires additional OOD data that are not always available;
- There is no guarantee that available OOD data are representative, thus the detector may overfit to OOD samples seen in the fine-tuning;
- Depending on the downstream classifiers, available OOD data need to be curated to avoid class overlapping, which is laborious and inefficient.

So in this work, we are primarily interested in evaluating unsupervised OOD detection and, if possible, improving them.

2.2 UNSUPERVISED OOD DETECTION TECHNIQUES

Sneakoscope currently includes three popular unsupervised detection techniques as described below.

Maximum Softmax Probability (MSP). Hendrycks & Gimpel (2017) proposes a simple yet effective method for OOD detection. For a given input x, it takes the negative prediction probability, i.e., $-max_y\ p(y|x)$, as the anomaly score. x is considered an OOD sample if the anomaly score is greater than a specified threshold.

Mahalanobis Distance. Lee et al. (2018b) proposes to estimate the probability density of intermediate features from a trained classifier by fitting class-conditional Gaussian distributions to training examples. For each class c, the empirical mean $\hat{\mu}_c$ is computed as $\frac{1}{N_c}\sum_{i:y_i=c} f(x_i)$ where N_c is the number of training examples in class c and $f(\cdot)$ denotes the output of an intermediate layer. A shared covariance $\hat{\Sigma}$ is given by $\frac{1}{N}\sum_c \sum_{i:y_i=c}(f(x_i)-\hat{\mu}_c)(f(x_i)-\hat{\mu}_c)^T$. The anomaly score of a test sample x is defined as the closest Mahalanobis distance $min_c(f(x)-\hat{\mu}_c)^T\hat{\Sigma}^{-1}(f(x)-\hat{\mu}_c)$.

Energy Score. Liu et al. (2020) proposes to use energy scores directly derived from a discriminative model as an alternative to MSP-based methods that are more susceptible to the overconfidence issue. Energy scores are non-probabilistic scalars that claim to align with the probability density of the inputs (Liu et al., 2020). In other words, samples with higher energies are less likely to occur. Without changes to a trained classifier f, the energy score is defined as $E(x;f) = -\log\sum_i^K e^{f_i(x)}$ where x is an input, K is the number of classes, and $f_i(x)$ is the logit of the i-th class.

To summarize, MSP has served as a long-standing baseline of OOD detection due to its simplicity and effectiveness. Energy-based detection also becomes popular in recent works (Grathwohl et al., 2020; Liu et al., 2020; Lin et al., 2021). Despite having less recognition, we find Mahalanobis Distance is a potent approach with strong detection performance as shown later.

2.3 DETECTION BACKBONE MODELS

Although unsupervised OOD detection techniques are applicable to any DNN-based classifier, few works assess the sensitivity of those techniques to different model architectures. So *Sneakoscope* sources a collection of widely-used model architectures for evaluation.

ResNets. Residual Networks (He et al., 2016) are one of the most successful families of DNNs. The novel idea of having skip connections to learn the residual mapping solves the problems of unstable training and degrading performance with very deep neural networks.

WideResNets. Wide Residual Networks (Zagoruyko & Komodakis, 2016) retain the overall architecture of ResNets while trading the depth with the width of the networks.

DenseNets. Densely Connected Convolutional Networks (Huang et al., 2017) are another type of CNN architecture where all layers are directly connected, and each layer takes in feature maps from all preceding layers while passing on its own feature maps to all subsequent layers.

Big Transfer. Big Transfer (Kolesnikov et al., 2020) introduces a set of large-scale pre-trained ResNets that can effectively adapt to downstream tasks with few-shot finetuning.

ViT. Unlike CNN architectures above, the Vision Transformer (ViT) (Vaswani et al., 2017) carries on self-attention and large-scale pre-training from its NLP predecessor. With minor modifications of the original transformer architecture, ViT attains compelling results in image classification.

2.4 UNIFIED ANALYSIS

Sneakoscope provides both visual and quantitative analysis to understand the results through the same scope. In particular, *Sneakoscope* examines model confidence calibration and hidden representations to illustrate the following research questions.

- **RQ1**: Why does the detection performance of the same method on the same dataset vary significantly across different model architectures?
- **RQ2**: How does large-scale pre-training reshape the landscape of OOD detection?
- **RQ3**: Why does Mahalanobis Distance have the biggest win from pre-training?

Confidence Calibration. Confidence calibration refers to the problem whether the model confidence reflects the true correctness likelihood. Since MSP employs the predicting probability (as the model confidence measure) for OOD detection, we can expect the effectiveness of MSP to rely on the degree of confidence calibration of a classifier.

We follow Guo et al. (2017) and measure model confidence calibration with two empirical approximations: Reliability Diagram and Expected Calibration Error (ECE). The reliability diagram assigns predictions into evenly-spaced confidence bins, and plots both expected accuracy and empirical accuracy against each confidence interval. By showing the gap between two accuracy, the reliability diagram visually captures confidence calibration of a classifier. Different from the reliability diagram, ECE is a scalar statistical indicator of confidence calibration. It calculates a weighted sum of the absolute difference between the accuracy and confidence of each bin, thus taking into account the percentage of examples in each bin.

Note that the predicting probability is not the sole proxy of model confidence. With slight transformations, non-probabilistic energy-based score $S_E(x; f)$ and Mahalanobis Distance-based score $S_M(x; f)$ (both are non-negative, and the higher the better) can also be converted to model confidence. We apply function $g(z) = 1 \ / \ (1+e^{-z}) \times 2 - 1$ where $z \geq 0$ to squash unbounded values into the interval $[0, 1]$ while maintaining the relative order of original values. Thereby, we can derive the following confidence measures: (1) *Energy Score-based confidence* $= 1 \ / \ (1 + e^{-S_E(x;f)}) \times 2 - 1$; and (2) *Mahalanobis Distance-based confidence* $= 1 \ / \ (1 + e^{-S_M(x;f)}) \times 2 - 1$.

With probabilistic confidence measures, we are able to utilize Reliability Diagram and Expected Calibration Error to inspect whether a model is calibrated with respect to Energy Score and Mahalanobis Distance. It is worth pointing out that the absolute values of Expected Calibration Error are not comparable across confidence measures. Nevertheless, a good confidence measure should assign high values to in-distribution samples and low values to OOD samples.So we can anticipate a more calibrated model to have better detection performance, which helps explain the varying results of the same detection method over different model architectures.

Learned Representation. Confidence calibration illuminates the aforementioned questions from the output aspect. We next turn to the input perspective for more insights. Both MSP and Energy Score obtain detection scores from the model logits — MSP takes the predicting probability from the softmax-normalized logits and Energy Score simply applies *logsumexp* operator to logits. In contrast, Mahalanobis Distance was motivated to characterize OOD samples in the feature space rather than "label-overfitted" output space (Lee et al., 2018b). So we take a particular look at features of the penultimate layer (which go through fully connected layer(s) and become logits).

To analyze hidden features, we leverage Central Kernel Alignment (CKA) (Kornblith et al., 2019) that can quantitatively compare learned representations across model architectures. CKA is essentially a normalized version of Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2007) that measures the statistical independence between two distributions. With CKA[1], we are able to examine the effect of model architectures and pre-training on hidden representations and the downstream task of OOD detection.

3 EXPERIMENTS

We start with the experiment setup in this section. We then demonstrate (1) model architectures have a significant effect on OOD detection performance; (2) large-scale pre-training can dramatically improve the results of all three detection methods while Mahalanobis Distance benefits most from pre-training. Finally, we present analysis that helps explain and understand the experiment results.

3.1 DATASETS & MODELS

We consider CIFAR-10/100 (Krizhevsky et al., 2009) as in-distribution datasets along with SVHN (Netzer et al., 2011), Textures (Cimpoi et al., 2014), Fashion-MNIST (Xiao et al., 2017), Gaussian noise, Rademacher Noise, and Blobs as out-of-distribution datasets. The last three are synthetic images. We also use CIFAR-100 as OOD for CIFAR-10 and vice versa.

[1] `https://github.com/google-research/google-research/tree/master/representation_similarity`

To circumvent the issues associated with training practices and biases towards less effective approaches identified in this work, we use trained models released from prior works: ResNet and DenseNet from Lee et al. (2018b); WideResNet from Liu et al. (2020). To study the effect of pre-training, We took a ViT CIFAR-10 model from Hugging Face Model Zoo[2]. Since we didn't find the rest of models publicly available, we fine-tuned Big Transfer[3] on CIFAR-10 and CIFAR-100, and ViT[4] on CIFAR-100 with the recommended hyper-parameters from their sources. Both Big Transfer and ViT were pre-trained on ImageNet-21K. Table 1 gives test accuracy of all models.

Table 1: Test accuracy in percentage of each model on CIFAR-10 and CIFAR-100.

Test Accuracy	ResNet	DenseNet	WideResNet	Big Transfer	ViT
CIFAR-10	93.67	95.19	94.84	97.46	98.52
CIFAR-100	78.34	77.63	75.96	86.93	92.90

3.2 METRICS

We evaluate the unsupervised detection methods with the following metrics: (1) **AUROC**, the area under the receiver operating characteristic curve; (2) **AUPR**, the area under the precision-recall curve; (3) **FPR95**, the fraction of out-of-distribution examples (negative) misclassified as in-distribution data (positive) when the true positive rate is at 95%.

Note that AUROC and AUPR are independent of thresholds whereas FPR95 sets a specific threshold such that 95% in-distribution data are correctly detected. Unlike AUROC, AUPR also adjusts for different positive and negative base rates.

3.3 RESULTS

Table 2: OOD detection results in percentage with CIFAR-10 as the in-distribution dataset. The overall best numbers are in bold while the best numbers of each detection technique is in italics.

OOD: SVHN	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	89.89	85.95	67.81	93.92	95.48	45.60	*91.20*	*85.84*	52.09
DenseNet	89.91	84.61	59.61	96.72	92.08	14.26	90.93	84.61	47.73
Wide-ResNet	*91.91*	*86.50*	*48.43*	**97.66**	**99.49**	**13.20**	91.08	79.22	*35.36*
OOD: Textures	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	*89.28*	*94.06*	60.90	95.14	97.47	30.92	*89.54*	*93.54*	*49.66*
DenseNet	88.53	93.07	*59.50*	95.17	96.62	17.82	84.98	88.21	58.51
WideResNet	88.42	92.78	59.66	**97.28**	**98.38**	**15.21**	85.35	88.44	52.48
OOD: CIFAR-100	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	86.77	87.87	66.39	*88.81*	*90.08*	*59.05*	87.52	87.66	55.38
DenseNet	*89.74*	*91.28*	*58.77*	68.05	68.07	81.31	**90.95**	**91.72**	**45.72**
Wide-ResNet	88.18	89.09	62.47	84.11	85.70	69.20	87.57	86.62	49.66
OOD: Gaussian	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	89.47	92.78	76.62	97.24	97.92	17.07	89.69	92.70	72.18
DenseNet	*97.69*	*98.40*	*11.94*	**100.00**	**100.00**	**0.00**	*98.10*	*98.93*	*0.06*
Wide-ResNet	94.79	96.26	42.17	**100.00**	**100.00**	**0.00**	72.14	79.98	100.00

The Effect of Architectures on OOD. Table 2 presents CIFAR-10 OOD detection results without considering pre-training. We can see that even for the same detection technique and the same dataset, there is a large variation of the detection performance across model architectures, especially regarding FPR95. For instance, in the setting of CIFAR-10 vs. SVHN, ResNet, DenseNet, and WideResNet share similar AUROC and AUPR under the MSP column; however, as regard to FPR95, ResNet is more than 8% worse than DenseNet and nearly 20% worse than WideResNet; Under the Mahalanobis column, ResNet is more than 30% worse than other models; Energy Score also exhibits similar discrepancy among the three models.

[2] https://huggingface.co/nateraw/vit-base-patch16-224-cifar10

[3] https://github.com/google-research/big_transfer

[4] https://github.com/jeonsworld/ViT-pytorch

Consistent with CIFAR-10 vs. SVHN, the numbers in CIFAR-10 vs. Textures/CIFAR-100/Gaussian show the considerable variations of all three metrics over architectures as well. We also observe that Mahalanobis Distance applied on WideResNet obtains the best results in three settings (except CIFAR-100) where the OOD samples are semantically dissimilar to objects in CIFAR-10. But in the case of CIFAR-100 as the OOD dataset, Mahalanobis Distance suffers from high FPR95 regardless of architectures, which we will explain in the analysis section.

We have the same observation of the fluctuating results over architectures with CIFAR-100 as the in-distribution dataset. Nevertheless, as there are much more classes in CIFAR-100, MSP and Energy Score are struggling with all three metrics, again especially with FPR95. AUROC and AUPR of MSP and Energy Score are mainly around or below 80% while FPR95 is generally above 75%, which is very concerning and suggests both techniques are broken. Due to the limit of space, we show the rest of CIFAR-10 results in table 5 and table 6, and the full CIFAR-100 results in table 7 and table 8 in appendix.

Table 3: Consider pre-training in OOD detection with CIFAR-10 as the in-distribution dataset. The overall best numbers are in bold while the best numbers of each detection technique is in italics.

OOD: SVHN	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	89.89	85.95	67.81	93.92	95.48	45.60	*91.20*	*85.84*	52.09
Big Transfer	95.64	87.22	16.20	99.74	99.51	**0.36**	94.10	81.93	20.84
ViT	*99.07*	*98.28*	*3.54*	**99.81**	**99.60**	0.43	*99.44*	*98.79*	*2.10*
OOD: Textures	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	89.28	94.06	60.90	95.14	97.47	30.92	89.54	93.54	49.66
Big Transfer	97.66	98.39	8.51	**100.00**	**100.00**	**0.00**	93.91	94.95	22.71
ViT	*99.94*	*99.97*	*0.14*	99.98	99.99	**0.00**	*99.97*	*99.98*	*0.11*
OOD: CIFAR-100	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	86.77	87.87	66.39	*88.81*	*90.08*	*59.05*	87.52	87.66	55.38
Big Transfer	89.52	85.21	31.11	96.27	96.36	18.85	86.08	80.21	38.20
ViT	*97.53*	*97.57*	*11.36*	**98.71**	**98.74**	**7.02**	*97.52*	*97.30*	*10.07*
OOD: Gaussian	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	89.47	92.78	76.62	97.24	97.92	17.07	89.69	92.70	72.18
Big Transfer	96.57	95.23	11.28	**100.00**	**100.00**	**0.00**	91.03	87.11	29.13
ViT	*99.99*	*99.99*	**0.00**	99.98	99.99	**0.00**	*99.99*	**100.00**	**0.00**

OOD with Pre-training. Table 3 shows the CIFAR-10 OOD detection results considering large-scale pre-training. We assess the impact of pre-training on OOD detection by comparing ResNet trained from scratch, Big Transfer (i.e., pre-trained ResNet), and ViT [5].

All three techniques have a remarkable gain from pre-training. For MSP and Energy Score, the improvements with respect to FPR95 range from 30% to 70%; the improvements of Mahalanobis Distance with respect to FPR95 have a relatively lower range from 17% to 52% while Mahalanobis Distance applied on pre-trained models achieves the lowest FPR95. We see CIFAR-10 vs. CIFAR-100 remains a challenging setting where even Big Transfer has a relatively high FPR95 compared to ViT, which we also find explanation in our analysis section.

Comparison with Supervised Results. Due to the magnificent uplift by large-scale pre-training, Mahalanobis Distance obtains nearly perfect detection performance, which is even better than dedicated fine-tuned results reported in Hendrycks et al. (2019b) using MSP and Liu et al. (2020) using Energy Score. This indicates supervision is not the only way to reach the best results in OOD detection. The combination of large-scale pre-trained models and a competitive detection technique such as Mahalanobis Distance overcomes the aforementioned drawbacks of supervised detection.

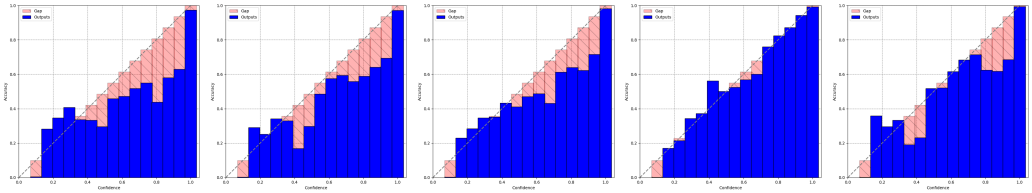
3.4 Analysis

Confidence Calibration. Figure 1 shows the reliability diagram of each model architecture on CIFAR-10 with the predicting probability as confidence measure. It is visually clear that pre-trained models Big Transfer and ViT are better calibrated than models without pre-training, which explains

[5]There are no available large-scale pre-trained DenseNets and WideResNets in Google's Big Transfer project, and we are not able to do large-scale pre-training due to limited computing resources.

why pre-trained models yield superb MSP detection results. One pitfall of reliability diagrams is that they treat all of the bins equally without considering the proportion of examples in each bin. While Big Transfer appears to be an almost perfectly calibrated model, Expected Calibration Error (ECE) gives a precise measure of confidence calibration. Table 4 shows ECE relative to each model architecture and confidence measure on CIFAR-10 and CIFAR-100. ECE in general reflects the trend of OOD detection performance of each method over model architectures. In CIFAR-10, more calibrated models (without pre-training) such as DenseNet and WideResNet have comparably lower FPR95 as opposed to less calibrated ResNet with the worst FPR95. On the other hand, in CIFAR-100, WideResNet instead is the most poorly calibrated model and indeed FPR95 of WideResNet is often falling behind. It is also clear from table 4 that large-scale pre-training notably reduces the calibration error for all confidence measures. Thus, we see in table 3 that Big Transfer and ViT outperform ResNet without pre-training and unsurprisingly sweep through the OOD detection task.

Figure 1: Reliability diagrams with MSP as the confidence measure on CIFAR-10 where red bars reflect the gap between the expected accuracy and the actual accuracy of each bin.



(a) ResNet with ECE = 3.96 (b) DenseNet with ECE = 2.89 (c) WideResNet with ECE = 2.82 (d) Big Transfer with ECE = 1.41 (e) ViT with ECE = 0.34

Table 4: Expected Calibration Error in percentage (the lower the better) on CIFAR-10/100 with respect to each model architecture and confidence measure. Best numbers are highlighted in bold.

CIFAR-10	**MSP**	**Mahalanobis**	**Energy**
ResNet	3.96	6.33	6.19
DenseNet	2.89	4.81	4.79
WideResNet	2.82	5.16	5.12
Big Transfer	1.41	2.54	1.58
ViT	**0.34**	**1.48**	**1.15**

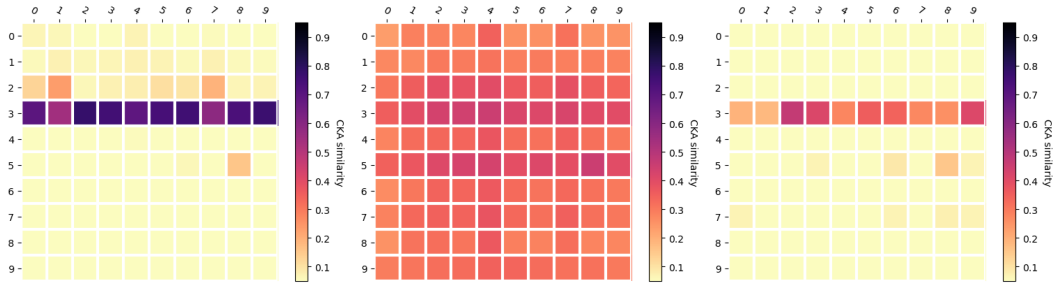
CIFAR-100	**MSP**	**Mahalanobis**	**Energy**
ResNet	7.52	21.66	21.54
DenseNet	8.51	22.37	22.37
WideResNet	10.02	24.04	24.03
Big Transfer	2.65	13.07	13.00
ViT	**1.70**	**7.10**	**7.09**

Learned Representation. The analysis of confidence calibration uncovers aforementioned research questions RQ1 and partly RQ2 from the output perspective. We now consider RQ2 and RQ3 from the input aspect of learned hidden representations. In particular, we examine the most successful detection approach — Mahalanobis Distance with support of pre-training. In principle, Mahalanobis Distance fits a class-conditional Gaussian distribution to intermediate features (in our case, the hidden representations of the penultimate layer) of training samples of each class. Due to the linear discriminant analysis assumption, all Gaussian distributions share the same covariance matrix. In other words, each class c is exactly characterized by the empirical mean $\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f(x_i)$ where N_c is the number of training examples in class c and $f(\cdot)$ denotes the output of the penultimate layer. It then becomes evident that Mahalanobis Distance-based detection will deteriorate if there is a hypothetical class of OOD inputs that also follows the Gaussian distribution with the mean close to the empirical mean of any training class.

Following the qualitative analysis, we visualize the "closeness" between in-distribution class means and out-of-distribution class means with CKA similarity heatmaps. Figure 2 indicates the ResNet empirical mean of class 3 of CIFAR-10 (cat) is close to several classes of SVHN (digit 2, 3, 6, 9) with high CKA similarity around/over 0.8. In contrast, CKA similarity between in- and out-of-distribution means of pre-trained models Big Transfer and ViT are generally below 0.6. Due to the "restrictive" representations learned by non-pre-trained models like ResNet, there could be multiple classes of OOD examples that fall close to in-distribution examples. With more generalizable representations, pre-training dramatically elevates the OOD performance. We also notice that Big Transfer and ViT seem to reach the best results through different paths. In the heatmap of Big

Transfer, each OOD class is very weakly similar to all in-distribution classes while ViT learns sharp representations such that the similarity concentrates on in-distribution class 3.

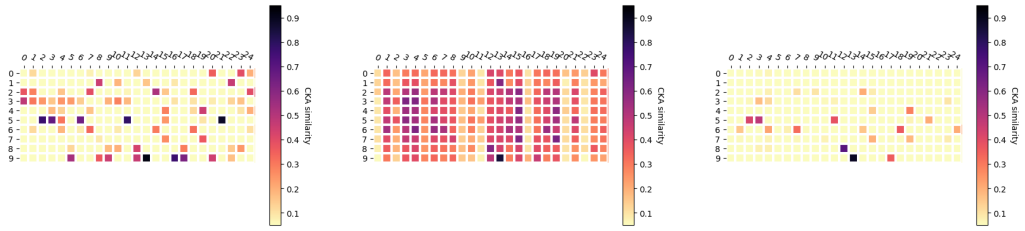
Figure 2: CKA similarity heatmaps of CIFAR-10 vs. SVHN where the y-axis marks the indices of CIFAR-10 classes and the x-axis marks the indices of SVHN classes.



(a) ResNet (b) Big Transfer (c) ViT

Figure 3 presents CKA similarity heatmaps of CIFAR-10 vs. CIFAR-100. There are a notable number of OOD classes close to class 5 (dog) and class 9 (truck) of CIFAR-10 that corrupt Mahalanobis Distance-based detection, which explains the poor performance on CIFAR-10 vs. CIFAR-100 we see in table 2. Nevertheless, pre-training mitigates the issue and Mahalanobis Distance resumes to top the OOD detection. We also observe from CKA similarity heatmaps that there are apparently much more confusing OOD classes in Big Transfer than in ViT, which explains why FPR95 of ViT is over 10% lower than Big Transfer even though both are pre-trained models.

Figure 3: CKA similarity heatmaps of CIFAR-10 vs. CIFAR-100 where the y-axis marks the indices of CIFAR-10 classes and the x-axis marks the indices of the first 25 classes of CIFAR-100.



(a) ResNet (b) Big Transfer (c) ViT

4 Related Work and Discussions

OOD Detection in Neural Networks. Due to the rich literature of OOD detection, we give an overview of OOD detection techniques that are more related to those evaluated in this work.

Following the baseline work (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018) improves on the baseline by enlarging the softmax score gap between in- and out-of-distribution data with temperature scaling and perturbations to the input images. However, the fine-tuning of hyper-parameters requires an additional validation dataset, which is not guaranteed to be representative of incoming OOD samples. So the hyper-parameters including the temperature scalar and the amount of perturbations added to the inputs could potentially overfit to the picked validation dataset. The extreme of fine-tuning is Outlier Exposure (Hendrycks et al., 2019b) that trains a MSP-based detector with large "outlier" datasets such as 80 Million Tiny Images (Torralba et al., 2008) and ImageNet-22K (Deng et al., 2009). We argue Outlier Exposure is impractical for the reason that depending on the specific classifier, one needs to carefully and laboriously curate the outlier dataset so that there is no overlapping between in- and out-of-distribution data. A more efficient way as shown is to hand-off the "exposure" workload to pre-training of the classifier that learns good enough representations for unsupervised OOD detection.

As MSP is criticized for giving arbitrarily high predicting score, recent works exploit energy-based models for OOD detection (Grathwohl et al., 2020; Liu et al., 2020; Lin et al., 2021). Grathwohl et al. (2020) proposes to reinterpret a classifier as a Joint Energy-based Model (JEM) and apply JEM to detect OOD samples. Although JEM shows promising results in applications not limited to OOD detection, there remain concerning issues in reliably training energe-based models. Liu et al. (2020) directly derives the energy score from the classifier model and avoids the unstable training problem of JEM. By further fine-tuning the detector with an energy-bounded learning objective, they obtain the state-of-the-art results on a collection of six OOD datasets. Nevertheless, they also report Energy Score without fine-tuning underperforms Mahalanobis Distance. We also show in this work that Mahalanobis Distance with pre-training achieves even better results than Energy Score with fine-tuning/pre-training. A more recent work (Lin et al., 2021) employs adjusted Energy Score and intermediate classifier outputs to efficiently detect OOD examples.

Anomaly Detection. Deep learning models are well known to be susceptible to adversarial examples (Goodfellow et al., 2014) and out-of-distribution examples. Beyond MSP and Energy Score, the Mahalanobis Distance-based approach can also detect adversarial examples (Lee et al., 2018b), but detection of adversarial examples is out of scope of our work.

Due to the promising progress in OOD detection, recent works (Ahmed & Courville, 2020; Winkens et al., 2020) further divide OOD detection to near-OOD tasks and far-OOD tasks depending on how semantically similar the outliers are to in-distribution examples. In this work, we consider both near- and far-OOD detection. For example, CIFAR-10 vs. SVHN is a pair of far-OOD datasets because the digit images in SVHN are semantically dissimilar to object images in CIFAR-10. In contrast, the pair of CIFAR-10 vs. CIFAR-100 is considered for near-OOD detection since they are both sourced from the 80 Million Tiny Images dataset and there are semantically similar categories like (big) *truck* in CIFAR-10 and *pickup truck* in CIFAR-100.

Pre-training for OOD Detection. Hendrycks et al. (2019a) demonstrates that pre-training can improve model robustness and uncertainty estimates. Our work distinguishes from them mainly in (1) the pre-training scale in Hendrycks et al. (2019a) is limited to a down-sampled version of ImageNet-1K (Chrabaszcz et al., 2017), which is considerably smaller than the ImageNet-21K pre-training scale evaluated in this work; (2) they focus on the improvement of AUROC and AUPR with only MSP-based detection while similar improvement can be reached by a stronger detection method like Mahalanobis distance. On the contrary, we study more detection methods with emphasis on FPR95 that is a more practical indicator than AUROC and AUPR; (3) We investigate the impact of model architectures on OOD detection that is mostly ignored in the literature. We also notice another more recent work (Fort et al., 2021) that studies pre-trained transformers for OOD detection. However, they consider far-OOD detection is nearly solved and primarily explore the improvement of only AUROC from pre-training on near-OOD tasks. Our work with a different goal studies unsupervised OOD detection more extensively.

We also call on the community to review and refine the definition of OOD detection in the context of pre-training. It remains unclear whether samples from classes seen in pre-training should be considered OOD for the downstream classifiers fine-tuned on a different set of classes.

5 Conclusion

We revisit the problem of unsupervised OOD detection and propose *Sneakoscope*, a unified evaluation suite that systematically investigates the performance of different unsupervised OOD detection methods on classifiers with different model architectures and with and without large-scale pre-training. Based on the investigation, we propose ways to significantly improve unsupervised OOD detection methods over prior reported results. For instance, a simple combination of pre-trained ViT and Mahalanobis Distance in unsupervised mode (no futher fine-tuning of the detector on OOD data is required) has a low false positive rate (at 95% true positive rate) on CIFAR-10 vs. SVHN of 0.43% versus 67.81% for an unsupervised OOD detector based on Resnet with MSP. These results compare even favorably with state-of-the-art OOD detectors that require substantial fine-tuning on OOD data. We thus provide new insights and baselines for unsupervised OOD detection methods.

REFERENCES

Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3154–3162, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

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.

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*, pp. 248–255. IEEE, 2009.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *arXiv preprint arXiv:2106.03004*, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 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*, 2020.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J Smola. A kernel statistical test of independence. In *NIPS*, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

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*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019a.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019b.

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.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018a.

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, 2018b.

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

Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15323, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

A Appendix

A.1 OOD Results not Shown in the Main Body

Table 5 presents the rest of OOD detection results with CIFAR-10 as the in-distribution dataset.

Table 6 presents the rest of OOD detection results considering large-scale pre-training with CIFAR-10 as the in-distribution dataset.

Table 7 presents the full OOD detection results with CIFAR-100 as the in-distribution dataset.

Table 8 presents the full OOD detection results considering large-scale pre-training with CIFAR-100 as the in-distribution dataset.

Table 5: The rest of OOD detection results in percentage with CIFAR-10 as the in-distribution dataset. The overall best numbers are in bold.

OOD: F-MNIST	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	93.00	94.37	45.81	94.62	96.04	39.96	95.61	96.09	24.88
DenseNet	97.26	97.87	19.56	95.50	96.65	29.44	**99.66**	**99.70**	**0.89**
Wide-ResNet	95.56	96.58	34.42	96.06	96.92	26.31	99.12	99.23	3.66
OOD: Rademacher Noise	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	89.47	92.78	76.62	97.24	97.92	17.07	89.69	92.70	72.18
DenseNet	97.69	98.40	11.94	**100.00**	**100.00**	**0.00**	98.10	98.93	0.06
Wide-ResNet	94.79	96.26	42.17	**100.00**	**100.00**	**0.00**	72.14	79.98	100.00
OOD: Blobs	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	93.59	95.23	49.44	99.39	99.47	2.61	94.29	95.46	36.49
DenseNet	64.05	66.11	96.88	99.71	99.80	**0.02**	48.17	53.00	99.39
Wide-ResNet	94.60	95.80	39.66	**99.79**	**99.83**	0.29	96.91	97.40	15.68

Table 6: Consider large-scale pre-training. The rest of OOD detection results in percentage with CIFAR-10 as the in-distribution dataset. The overall best numbers are in bold.

OOD: F-MNIST	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	93.00	94.37	45.81	94.62	96.04	39.96	95.61	96.09	24.88
Big Transfer	89.98	81.14	20.48	**99.66**	**99.70**	**1.11**	86.26	75.09	25.60
ViT	98.03	98.05	9.53	99.38	99.44	3.42	98.16	98.08	8.15
OOD: Rademacher Noise	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	93.60	95.79	57.10	99.29	99.48	0.99	93.21	95.66	60.15
Big Transfer	97.10	96.15	9.30	**100.00**	**100.00**	**0.00**	92.48	89.14	24.86
ViT	99.98	99.99	**0.00**	99.99	99.99	**0.00**	99.99	99.99	**0.00**
OOD: Blobs	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	93.59	95.23	49.44	99.39	99.47	2.61	94.29	95.46	36.49
Big Transfer	99.45	99.52	0.83	**100.00**	**100.00**	**0.00**	99.09	99.15	2.45
ViT	98.99	99.22	2.44	99.90	99.92	0.00	99.58	99.64	0.22

Table 7: The full OOD detection results in percentage with CIFAR-100 as the in-distribution dataset. The overall best numbers are in bold.

OOD: SVHN	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	79.34	65.86	80.12	89.41	**83.59**	56.80	79.19	64.04	82.18
DenseNet	82.64	75.09	73.76	85.96	71.72	54.70	87.91	81.86	65.52
WideResNet	71.38	57.79	84.35	**90.45**	82.18	**44.01**	73.87	60.50	85.61
OOD: Textures	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	77.89	86.43	78.69	81.24	88.50	72.66	79.13	86.85	76.76
DenseNet	72.62	80.84	81.47	90.04	92.53	**31.88**	74.03	78.85	78.78
WideResNet	73.59	83.12	83.28	**90.76**	**94.36**	40.12	76.35	84.47	79.65
OOD: CIFAR-10	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	78.24	81.40	79.91	78.28	79.24	79.94	78.35	81.19	81.52
DenseNet	76.67	80.67	80.20	50.24	56.80	98.81	77.09	78.98	81.50
WideResNet	76.36	80.49	80.58	69.40	72.51	92.11	**79.11**	**82.10**	**78.09**
OOD: F-MNIST	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	86.97	88.99	62.97	91.68	93.21	49.57	88.28	89.86	60.61
DenseNet	92.67	93.27	35.09	90.12	92.69	63.59	99.00	99.02	5.26
WideResNet	93.68	94.32	33.00	70.93	80.15	98.68	**99.28**	**99.31**	**3.54**
OOD: Gaussian	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	66.75	78.30	100.00	65.05	79.28	100.00	67.21	78.38	100.00
DenseNet	86.41	90.65	82.74	**100.00**	**100.00**	**0.00**	95.18	97.03	41.77
WideResNet	64.42	76.01	99.87	**100.00**	**100.00**	**0.00**	43.33	61.56	100.00
OOD: Rademacher Noise	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	69.71	80.25	99.51	88.08	92.30	85.68	74.03	82.38	99.19
DenseNet	51.92	69.34	100.00	**100.00**	**100.00**	**0.00**	67.29	79.65	100.00
WideResNet	79.21	85.91	97.64	**100.00**	**100.00**	**0.00**	42.80	60.54	100.00
OOD: Blobs	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	90.11	92.28	59.06	90.59	93.55	69.77	90.40	92.16	54.94
DenseNet	88.01	90.46	61.32	98.19	98.90	1.50	88.06	91.97	88.00
WideResNet	75.35	80.03	89.30	**99.92**	**99.92**	**0.09**	51.18	61.15	100.00

Table 8: Consider large-scale pre-training. The full OOD detection results in percentage with CIFAR-100 as the in-distribution dataset. The overall best numbers are in bold.

OOD: SVHN	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	79.34	65.86	80.12	89.41	83.59	56.80	79.19	64.04	82.18
Big Transfer	92.68	86.64	39.88	**99.35**	**98.81**	**1.84**	93.51	86.01	31.85
ViT	92.55	84.89	36.37	98.02	96.00	8.12	96.58	92.03	14.63
OOD: Texture	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	77.89	86.43	78.69	81.24	88.50	72.66	79.13	86.85	76.76
Big Transfer	90.52	94.04	47.50	**100.00**	**100.00**	**0.00**	86.76	90.80	59.77
ViT	96.42	97.91	19.91	99.31	99.58	2.45	99.02	99.43	4.91
OOD: CIFAR-10	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	78.24	81.40	79.91	78.28	79.24	79.94	78.35	81.19	81.52
Big Transfer	84.00	84.46	61.37	77.16	78.95	80.31	82.23	82.01	62.36
ViT	93.28	93.89	32.98	**97.27**	**97.57**	**14.04**	96.23	96.22	16.90
OOD: F-MNIST	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	86.97	88.99	62.97	91.68	93.21	49.57	88.28	89.86	60.61
Big Transfer	89.86	88.01	43.91	93.50	94.97	40.14	84.99	81.23	55.66
ViT	97.68	97.83	13.19	**99.39**	**99.49**	**2.03**	99.21	99.28	3.14
OOD: Gaussian	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	66.75	78.30	100.00	65.05	79.28	100.00	67.21	78.38	100.00
Big Transfer	95.07	96.00	30.66	**100.00**	**100.00**	**0.00**	94.56	95.50	31.98
ViT	99.79	99.84	0.05	99.94	99.97	**0.00**	99.97	99.98	**0.00**
OOD: Rademacher Noise	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	69.71	80.25	99.51	88.08	92.30	85.68	74.03	82.38	99.19
Big Transfer	94.80	95.78	32.86	**100.00**	**100.00**	**0.00**	94.38	95.29	32.70
ViT	99.99	99.99	**0.00**	99.94	99.97	**0.00**	99.99	99.99	**0.00**
OOD: Blobs	**MSP**			**Mahalanobis**			**Energy**		
	AUROC ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑	**FPR95** ↓
ResNet	90.11	92.28	59.06	90.59	93.55	69.77	90.40	92.16	54.94
Big Transfer	94.76	95.76	32.21	**100.00**	**100.00**	**0.00**	91.97	93.30	50.57
ViT	99.96	99.97	0.01	99.97	99.98	**0.00**	99.98	99.99	**0.00**

A.2 OTHER OBSERVATIONS

Synthetic OOD Examples. Beyond realistic OOD inputs, we also evaluate the robustness of OOD detection against synthetic noise. MSP and Energy Score are highly susceptible to noise regardless of model architectures. On the contrary, Mahalanobis Distance is immune to synthetic noise even without pre-training except in the case of ResNet.

Unstable Energy Score. When comparing results of Energy Score with those reported in Liu et al. (2020), we notice a more than 15% discrepancy of AUPR in the setting of CIFAR-10 vs. Textures using WideResNet. We use their released code and the same pre-trained model (the one without energy-based fine-tuning)[6], but different base ratios of in-distribution data and OOD data. In fact, Liu et al. (2020) followed the setup in Hendrycks et al. (2019b) and set the number of OOD examples as $\frac{1}{5}$ of in-distribution examples. As Energy Score can unexpectedly fluctuate and AUPR is sensitive to base ratios of positive and negative examples, AUPR drops from 97.67% to 79.22% once we use all of the OOD samples in evaluation. In contrast, MSP and Mahalanobis Distance are not affected by the change of base ratios. The issue of Energy Score being unstable is also observed by Lin et al. (2021) but in a different scenario.

[6] https://github.com/wetliu/energy_ood