Heatmap-based Out-of-Distribution Detection

Julia Hornauer
Ulm University, Germany
julia.hornauer@uni-ulm.de

Vasileios Belagiannis
Friedrich-Alexander-University Erlangen-Nürnberg, Germany
vasileios.belagiannis@fau.de

Abstract

Our work investigates out-of-distribution (OOD) detection as a neural network output explanation problem. We learn a heatmap representation for detecting OOD images while visualizing in- and out-of-distribution image regions at the same time. Given a trained and fixed classifier, we train a decoder neural network to produce heatmaps with zero response for in-distribution samples and high response heatmaps for OOD samples, based on the classifier features and the class prediction. Our main innovation lies in the heatmap definition for an OOD sample, as the normalized difference from the closest in-distribution sample. The heatmap serves as a margin to distinguish between in- and out-of-distribution samples. Our approach generates the heatmaps not only for OOD detection, but also to indicates in- and out-of-distribution regions of the input image. In our evaluations, our approach mostly outperforms the prior work on fixed classifiers, trained on CIFAR-10, CIFAR-100 and Tiny ImageNet. The code is publicly available at: `https://github.com/jhornauer/heatmap_ood`.

1. Introduction

Despite the astonishing performance of deep neural networks on standard recognition datasets [12, 25], they cannot be trusted yet for safety-critical problems mainly because of two reasons. First, they do not necessarily generalize well to data that was not covered by the training distribution. When deep neural networks are exposed to such out-of-distribution (OOD) samples, they often make wrong predictions with high confidence. Second, they mostly lack providing an explanation on their decision that is understood by humans. If the deep neural network is a black-box model and does not have the necessary tools to assess whether the prediction is meaningful, it cannot be used in applications such as automated driving or medical imaging. Therefore, it is of utmost importance not only to detect OOD samples but also to highlight out-of-distribution regions of the input.

Out-of-distribution detection methods can be divided into the following two categories: approaches that define a score for a fixed model [5, 15] and approaches that further train a model to distinguish in- from out-of-distribution samples [6, 14]. In our work, we focus on OOD detection for a fixed classifier, because it suits well to real-world applications. In this context, the maximum softmax probability [5] of a neural network often serves as a built-in baseline for the detection of OOD samples. Nevertheless, overconfident deep neural networks usually origin from using the predictive class probability as a confidence measure. The probability distribution represented by the softmax function often gives a high response for unknown inputs [21]. This limitation can be mitigated by calibrating the confidence [16] or alternative scoring functions [15, 18]. We set ourselves apart by relying on a second model to generate heatmaps, which, in turn, are used to define a scoring function. Thus, our model does not only detect OOD samples but also indicates in- and out-of-distribution image regions.

(a) (b) (c)

Figure 1: Out-of-distribution training image examples (a) with the corresponding closest in-distribution image (b) and the defined heatmap (c) as a result of the normalized distance between the two images. Blue colors mark similar regions, whereas the red/yellow colors highlight regions that differ from the in-distribution image.

Our work addresses the detection of OOD samples as an output explanation problem. Approaches towards explaining the model's decision highlight the features that contribute [27, 3, 26] to the decision or select prototypi-

cal images [1]. Saliency [27] or attention [3] maps indicate where the trained neural network looks in the image for making a prediction. Prototypes on the other hand, demonstrate a similar training image [1] for interpreting the neural network prediction. Similar to post hoc explanation approaches [1, 27, 26], we also assume a model that is already trained and will not be further adapted. Motivated by the visual explanation idea, we propose the heatmap generation for OOD detection.

We phrase the problem of OOD detection as binary classification in the classifier's feature space assuming that the features extracted by the classifier are representative to learn the deviation of out-of-distribution from in-distribution features. Given a trained and fixed neural network classifier, we aim to generate heatmaps based on the classifier features and the class prediction. Therefore, we define the expected heatmaps such that in-distribution samples should have zero response. In contrast, OOD samples should have a high response for image regions that differ from the in-distribution samples. We create the heatmap of an OOD image by looking for the closest in-distribution image in the classifier feature space and forming the heatmap as the normalized image difference (see Fig. 1). In this way, the heatmap response acts as a margin between the in- and out-of-distribution image samples. To develop our idea, we introduce a decoder network to generate the heatmaps with input the features extracted by the classifier and the class prediction, while the in- and out-of-distribution defined heatmaps compose the expected output. Finally, based on the heatmaps, we define our scoring function, which outputs whether a sample is in- or out-of-distribution. Our approach generates the heatmaps not only for OOD detection but also as visualization of in- and out-of-distribution regions based on the classifier features and the class prediction. We summarize our contribution as follows: First, we propose the heatmaps for OOD detection. Learning to generate the heatmaps is based on our introduced decoder, while our main novelty lies in how to create heatmaps for the OOD image samples. Second, the proposed heatmaps visualize in- and out-of-distribution image regions based on the classifier features and the class prediction. In particular, for OOD samples the heatmaps represent the difference to the closest in-distribution sample as illustrated in Fig 1. Third, we present the OOD scoring function using the heatmap as input, which results in state-of-the-art OOD detection results for different classifiers trained on CIFAR-10 [12], CIFAR-100 [12] or Tiny ImageNet [25], compared to approaches that do not modify the classifier.

2. Related Work

Out-of-Distribution Detection Out-of-distribution detection approaches aim to identify samples that are not covered by the training distribution [5]. The prior work on image classification is mainly divided into two categories, based on whether the classifier parameters are fixed or modifiable. For fixed deep neural classifiers, the maximum softmax probability (MSP) serves as a common function to detect OOD samples. Nevertheless, the softmax probability is not sufficient due to the major drawback of deep neural networks being overconfident for unseen samples [21]. ODIN [16] improves the softmax score with temperature scaling and input perturbations validated on OOD samples. Hsu et al. [9] extend the temperature scaling to be independent of the OOD validation data. Liu et al. [18] introduce the energy score to discriminate between in- and out-of-distribution samples of a pre-trained classifier, but it also can be used as a cost function to optimize the classifier for the detection of OOD samples. Lin et al. [17], on the other hand, implement the energy score with an early exit strategy to address faster OOD detection. Sun et al. [28] use the observation that OOD data causes positively skewed activation units and improve the detection of OOD data by clipping the activations of the penultimate layer to an upper limit based on in-distribution activation values. We also assume a fixed classifier that is not further modified. Unlike the prior work [5, 15, 16], we learn to separate in-distribution from out-of-distribution samples with the heatmaps, generated by our proposed decoder. Importantly, our decoder not only performs OOD detection but also delivers a visualization of out-of-distribution regions through heatmap responses. To our knowledge, this is the first approach to perform OOD detection and at the same time illustrate in- and out-of-distribution image regions based on the features and the class prediction of an already trained and fixed classifier. In the second category, where the classifier is further optimized to not only classify the network inputs but also determine OOD samples, self-supervised methods such as deep autoencoders [22] or rotation prediction [7] can be leveraged to learn the in-distribution representation. Zaeemzadeh et al. [34], for instance, rely on feature compression by learning to embed in-distribution samples to a 1-dimensional subspace for OOD detection. Zisselman et al. [37] design a network architecture based on deep residual flow networks, customized for OOD detection, while Huang et al. [10] target large-scale OOD detection with a group-based softmax. In addition, there are approaches that utilize OOD training samples during the classifier's training stage. For example, random [6] or generated [14] OOD samples can be mapped to the Uniform distribution. In contrast, Yang et al. [32], apply clustering in the semantic space to detect in-distribution data within the OOD training samples. Based on the same principle, we also use OOD samples for the training of the proposed decoder, but basically without modifying the original classifier. We rely on the features learned by the already trained classifier to distinguish out-of-distribution from in-distribution samples.

Prediction Explainability There are different approaches to explain neural network predictions. Post hoc methods give explanations for trained models by local explanations [24] or global approximations with an interpretable surrogate model [13]. Another line of research visualizes prototypes [1] or highlights features that contribute to the classifier decision through activation [26], attention [3], or saliency [27] maps. Moreover, Liznerski et al. [19] make use of visual explanation to highlight conspicuous regions in images for anomaly detection. In this work, we phrase OOD detection as anomalies in the feature representation of an already trained and fixed model. In this context, we propose to generate heatmaps from the features extracted by a classifier in order to detect out-of-distribution inputs and visualize the corresponding regions of the input image.

3. Method

Let $P_{in}(\mathbf{x}, \mathbf{y})$ be the joint distribution of the image $\mathbf{x} \in \mathbb{R}^{w \times h \times 3}$ with width w and height h, as well as the one-hot vector label $\mathbf{y} \in \{0,1\}^C$, such that $\sum_{c=1}^{C} \mathbf{y}(c) = 1$ for a C-category classification problem. In our context, P_{in} denotes the training distribution from which the in-distribution dataset $\mathcal{D}_{in} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}_{in}|}$ is generated. Moreover, the multi-class deep neural network classifier $f(\cdot)$, parameterized by $\mathbf{w_f}$, is trained with the $\mathcal{D}_{in}$ dataset. Importantly, we consider the parameters of $f(\cdot)$ to be fixed and not modifiable anymore. During deployment, the classifier $f(\cdot)$ can be exposed to the data distribution $P_{out}(\mathbf{x}, \mathbf{y})$ that is different from the training distribution. We also consider the OOD training set $\mathcal{D}_{out} = \{(\mathbf{x}_o, \mathbf{y}_o)\}_{o=1}^{|D_{out}|}$ that is formed by sampling from P_{out}. In general, images $\mathbf{x}$ drawn from P_{out} have either a non-semantic shift or belong to a different object category [9].

Our goal is to differentiate in-distribution $P_{in}(\mathbf{x}, \mathbf{y})$ from out-of-distribution $P_{out}(\mathbf{x}, \mathbf{y})$ image samples based on the prediction and the features of the classifier $f(\cdot)$. We formulate the problem as binary classification where we assume that we have access to the feature layers of the fixed classifier $f(\cdot)$. In this context, we introduce the OOD heatmaps, which are utilized for spotting OOD samples and simultaneously illustrate in- and out-of-distribution image regions (3.1). Our main innovation lies in how to form expected heatmaps for the out-of-distribution samples. Also, we present the decoder neural network $d(\cdot)$, which is trained to generate heatmaps (3.2) from the classifier's features and class prediction. Finally, we rely on the decoder's heatmap generation for each image sample to compute the OOD detection score, as discussed in Sec. 3.3. Essentially, the generated heatmap indicates the corresponding image regions for the in- or out-of-distribution predictions (see Fig. 2). Note that for the evaluation, we consider the OOD test set $\mathcal{D}_{test} = \{(\mathbf{x}_l, \mathbf{y}_l)\}_{l=1}^{|\mathcal{D}_{test}|}$ from a different distribution $P_{test}(\mathbf{x}, \mathbf{y})$ that is not covered by P_{in} or P_{out}.

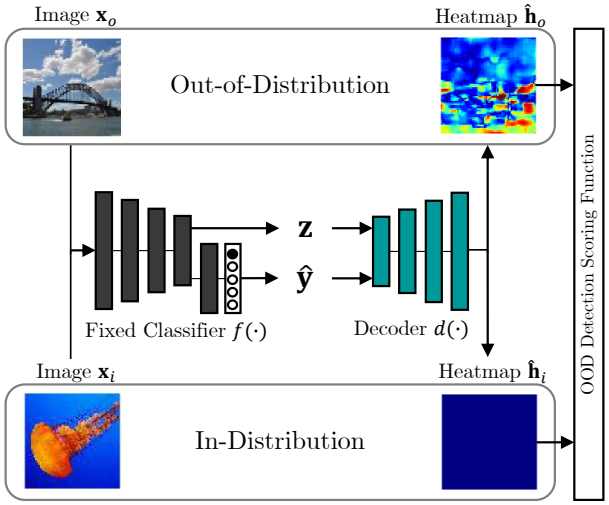


Figure 2: In our approach, we aim to detect OOD samples and at the same time visualize in- and out-of-distribution regions. Therefore, we generate heatmaps from the features $\mathbf{z}$ and the prediction $\hat{\mathbf{y}}$ of a fixed classifier. The generated heatmaps $\hat{\mathbf{h}}_i$ should show no response for in-distribution images $\mathbf{x}_i$, while the heatmaps $\hat{\mathbf{h}}_o$ should have a high response for OOD images $\mathbf{x}_o$.

3.1. Heatmaps

We define the expected heatmap $\mathbf{h} \in \mathbb{R}^{w \times h \times 3}$ with the same dimensions as the input image $\mathbf{x}$. For an in-distribution image, the heatmap should have zero response since it does not contain out-of-distribution information. In contrast, the heatmap of an out-of-distribution sample should have a high response for the image regions that differ from the in-distribution data. We provide an illustration of the heatmaps in Fig. 2. Consequently, the heatmap response acts as a margin between the in- and out-of-distribution image samples. For the in-distribution images, it is trivial to specify the expected heatmap since we aim to have a zero response. However, it is not clear how to define the heatmap for out-of-distribution samples. Below, we present an approach to make use of the in-distribution dataset $\mathcal{D}_{in}$ for defining the expected heatmaps of the images of the out-of-distribution dataset $\mathcal{D}_{out}$.

Out-of-Distribution Heatmap. For each image sample $\mathbf{x}_o$ of the out-of-distribution dataset $\mathcal{D}_{out}$, our idea is to form the heatmap based on the closest image sample of the in-distribution dataset $\mathcal{D}_{in}$. For that reason, we rely on the class prediction $\hat{\mathbf{y}}_o$ and the feature representation $\mathbf{z}_o$ of the fixed classifier with $\hat{\mathbf{y}}_o, \mathbf{z}_o = f(\mathbf{x}_o; \mathbf{w}_f)$. The selection is illustrated in Fig. 3. For the feature representation $\mathbf{z}_o$, we use the penultimate layer of the classifier. Based on these

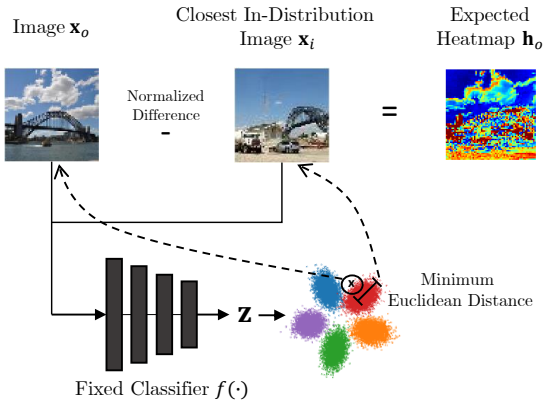


Figure 3: The definition of the expected OOD heatmap $\mathbf{h}_o$ is based on the distance to the closest in-distribution sample from $\mathcal{D}_{in}$. To find the closest in-distribution sample, we rely on the class prediction $\hat{\mathbf{y}}$ and the feature representation $\mathbf{z}$ of the fixed classifier. More precisely, we look for the sample $\mathbf{x}_i$ from $\mathcal{D}_{in}$ with the same class category, i.e. $\hat{\mathbf{y}}_o = \mathbf{y}_i$ and minimum Euclidean distance between $\mathbf{z}_o$ and $\mathbf{z}_i$.

two attributes, we look for the sample $\mathbf{x}_i$ from $\mathcal{D}_{in}$ with the same class category, i.e. $\hat{\mathbf{y}}_o = \mathbf{y}_i$, and similar feature representation, i.e. $\mathbf{z}_o \approx \mathbf{z}_i$. Note that even if $\hat{\mathbf{y}}_o$ is a wrong prediction, it does not have an impact on our approach. The goal is to find the closest in-distribution sample $\mathbf{x}_i$ based on the Euclidean distance between $\mathbf{z}_o$ and $\mathbf{z}_i$ given the class category. After obtaining the corresponding in-distribution sample $\mathbf{x}_i$ for the out-of-distribution sample $\mathbf{x}_o$, we define the heatmap $\mathbf{h}_o$ as their difference, where the images are normalized beforehand. The operation is described as:

$$\mathbf{h}_o = norm(\mathbf{x}_{i^*}) - norm(\mathbf{x}_o), \tag{1}$$

$$\text{where} \quad i^* = \underset{i \in \{1,\dots,|\mathcal{D}_{in}|\}}{\operatorname{argmin}} (\mathbf{z}_i - \mathbf{z}_o)^2, \quad \text{s.t. } \hat{\mathbf{y}}_o = \mathbf{y}_i. \tag{2}$$

The $norm(\cdot)$ operation corresponds to the normalization operation within the range -1 and 1. By considering the normalized image difference as the heatmap response, we define a margin between the in- and out-of-distribution image samples. Moreover, the margin is deliberated because both images have similar feature representations. In this way, we create the corresponding heatmap $\mathbf{h}_o$ for each image $\mathbf{x}_o$ of $\mathcal{D}_{out}$, resulting in our out-of-distribution training set. The complete image and heatmap set is defined as:

$$\mathcal{H}_{out} = \{(\mathbf{x}_o, \mathbf{h}_o)\}_{o=1}^{|\mathcal{D}_{out}|}. \tag{3}$$

We consider $\mathcal{H}_{out}$ as the out-of-distribution training set for training the proposed decoder.

In-Distribution Heatmap. In contrast to out-of-distribution samples, the zero response heatmap $\mathbf{h}_i = [0]^{w \times h \times 3}$ is representative for in-distribution images. These samples do not contain any out-of-distribution image regions to be indicated in the heatmap. Given the in-distribution images $\mathbf{x}_i$ and their heatmaps $\mathbf{h}_i$, we define our in-distribution training set as:

$$\mathcal{H}_{in} = \{(\mathbf{x}_i, \mathbf{h}_i)\}_{i=1}^{|\mathcal{D}_{in}|}. \tag{4}$$

Based on $\mathcal{H}_{in}$ and $\mathcal{H}_{out}$, we can train the proposed decoder to generate heatmaps for both types of image samples.

Heatmap Interpretation. The visual comparison between in- and out-of-distribution images is not directly possible in the image space due to the pixel value ambiguity and generally complex space. In our approach, the decoder learns a mapping from feature space back to pixel space, where the heatmap represents the difference to the closest in-distribution image. Therefore, OOD detection is based not only on the information in feature space, but also on the mapping back to pixel space learned by the decoder, which improves the OOD detection performance. In Fig. 1, features that differ from the closest in-distribution training sample are highlighted by red regions, which in turn are OOD features. In contrast, the blue color indicates similar features that are in-distribution regions.

3.2. Heatmap Decoder

The proposed decoder $d(\cdot)$, parameterized by $\mathbf{w_d}$, aims to estimate the heatmaps $\hat{\mathbf{h}}$. Similar to the heatmap definition, we rely on the feature representation $\mathbf{z}$ and the predictive probability distribution $\hat{\mathbf{y}}$ as input to the decoder. Fig. 2 shows an overview of our approach. At first, the classifier receives the image $\mathbf{x}$ either drawn from P_{in} or drawn from P_{out} as input and outputs the corresponding feature representations $\mathbf{z}$ and the predictive probability distribution $\hat{\mathbf{y}}$. Then, the features and the one-hot encoded class prediction are concatenated before being passed to the decoder to estimate the heatmaps $\hat{\mathbf{h}}$. Again, we rely only on a single feature representation of $\mathbf{z}$, which is the penultimate classifier layer. In the supplementary material we add an ablation study to validate this choice. We use thereby the $\mathcal{H}_{in}$ and $\mathcal{H}_{out}$ sets to train the decoder producing heatmaps. Below, the learning objective of the decoder is summarized as:

$$\underset{\mathbf{w_d}}{\operatorname{argmin}} \mathbb{E}_{\mathbf{x}_o, \mathbf{h}_o \sim \mathcal{H}_{out}} (1 + \alpha |\mathbf{h}_o|) \parallel \hat{\mathbf{h}}_o - \mathbf{h}_o \parallel^2 + \mathbb{E}_{\mathbf{x}_i, \mathbf{h}_i \sim \mathcal{H}_{in}} \parallel \hat{\mathbf{h}}_i - \mathbf{h}_i \parallel^2, \tag{5}$$

$$\text{where} \quad \hat{\mathbf{h}}_{\{i,o\}} = d(f(\mathbf{x}_{\{i,o\}}; \mathbf{w_f}); \mathbf{w_d}). \tag{6}$$

The OOD entries in the heatmap are weighted higher with a scaling factor $1 + \alpha|\mathbf{h}_o|$ depending on the defined heatmap

$\mathbf{h}_o$. Since in-distribution regions occur more frequently, the scaling factor has the effect that out-of-distribution regions are weighted higher in the loss calculation. For the in-distribution samples, we minimize effectively the term $\| \hat{\mathbf{h}}_\mathbf{i} \|^2$ since the $\mathbf{h}_i$ is always zero for in-distribution samples. During inference, where we have no knowledge of the distribution, the trained decoder estimates the heatmaps $\hat{\mathbf{h}}$ for illustrating in- and out-of-distribution image regions and for defining our OOD scoring function.

3.3. Out-of-Distribution Scoring Function

We leverage the generated heatmap $\hat{\mathbf{h}}$ to define the OOD scoring function for differentiating in- from out-of-distribution images. Since OOD detection is a binary classification problem, the OOD scoring function is described as:

$$OOD(\hat{\mathbf{h}}) = \begin{cases} 1 & \frac{1}{w \cdot h} \sum_{j=1}^{w \cdot h} \max(|\hat{\mathbf{h}}_j|) \leq \delta \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where j is used to iterate over the heatmap pixels and δ is the threshold to categorize a sample as in- or out-of-distribution.

4. Experiments

We conduct a detailed evaluation on standard OOD detection benchmarks using different network architectures. Therefore, we compare our method to state-of-the-art OOD detection approaches that operate on a pre-trained model (4.2). We demonstrate the heatmaps that serve as illustration of in- and out-of-distribution image regions (4.3) based on the distance to the closest in-distribution training sample. To further explain the heatmaps, we evaluate the effect of lighting conditions by influencing the brightness of images (4.4). Lastly, we show the importance of the OOD training data size in an ablation study (4.6).

4.1. Experimental Setting

Datasets and Models. We follow the prior work [6, 18] to evaluate our method on CIAFR-10 [12] and CIFAR-100 [12] as in-distribution datasets. Both in-distribution datasets have a resolution of 32×32. To cover a wide variety of OOD samples, we rely on the following OOD test sets: iSUN [31], LSUN-Crop [33], LSUN-Resize [33], SVHN [20], Textures [2], and Places365 [36]. All OOD images are downsampled to the in-distribution image resolution. As in the prior work [6], we use the 80 Million Tiny Images [29] database as the OOD training set. In this context, we train ResNet18 [4] and WideResNet [35] with depth 40 and width 2 on the in-distribution datasets and fix the classifier weights afterwards. In addition, we evaluate our method on the complex setting with ResNet50 [4] trained on Tiny ImageNet [25] as an in-distribution database. Here, the image resolution is 64×64. For the OOD test set, we rely on the iNaturalist [8], SUN [30], and Textures [2] databases, which are again downsampled to the same resolution as the in-distribution images. For the complex setup, we leverage Places365 as an OOD training set. As in prior work [6, 18], we evaluate with the entire in-distribution test set. As in literature [6, 18], the number of OOD samples per OOD test set is fixed to $\frac{1}{5}$ of the in-distribution dataset size.

Decoder Architecture and Training. The decoder network architecture is based on the DCGAN generator [23]. The heatmaps are normalized to $[-1, 1]$ with hyperbolic tangent as the last activation function in the decoder. As mentioned in 3.2, we rely on a single feature representation of the classifier, namely the penultimate classifier layer, as input to the decoder. Then, the features are normalized to the range of $[0, 1]$ and afterward concatenated with the one-hot encoded class prediction. Furthermore, the scaling factor α of the loss function is empirically chosen to be 5. Furthermore, the decoder is trained for 150 epochs with the Adam[11] solver, where the learning rate is set to 0.0002, β_1 to 0.5, and β_2 to 0.999. Finally, the input images are processed in a batch of 200 samples. The ratio from OOD samples to in-distribution samples is empirically chosen to be $\frac{1}{5}$. The CIFAR-10 and CIFAR-100 databases both consist of 50000 training images. This means the number of OOD training samples is set to 10000. For Tiny ImageNet, which contains 100000 training images, the OOD training data size then is 20000. Similar to prior work [6], the OOD training images are randomly chosen.

Evaluation Metrics. We evaluate our method based on the standard metrics [6, 18, 34], namely the AUROC, AUPR-Success (AUPR-S), AUPR-Error (AUPR-E), and FPR at 95% TPR (FPR-95). All metrics are independent of the OOD detection threshold δ. The AUROC integrates over the area under the receiver operating curve (ROC). The AUPR-Success indicates the area under the precision-recall curve with in-distribution samples as positive, while the AUPR-Error treats the OOD samples as positive. Compared to the AUROC metrics, the AUPR accounts for class imbalance. Lastly, FPR at 95% TPR computes the false positive rate (FPR) at 95% true positive rate (TPR). We directly apply the metric to the respective OOD detection score.

Comparison to Related Work. We compare with post hoc methods that operate on a fixed classifier, similar to our approach. First, we employ the maximum softmax probability (MSP) [5]. In addition, we use ODIN [16] and Mahalanobis [15] for comparison with our approach. For ODIN, we set the temperature to 1000, as proposed in the paper, and select the noise magnitude to be 0.0014 based on the

best outcome. In the case of Mahalanobis, we obtain the best results with a noise magnitude of 0.0028 when relying only on the features of the penultimate classifier layer. Furthermore, we compare our method to the recent Energy score [18] and ReAct [28]. In case of the Energy score, we choose the version that does not further optimize the classifier. Instead, the score is also determined based on the fixed classifier, since we train a second model but do not adjust the classifier. The temperature is set to 1 as proposed by the authors. For the related approaches, we make use of the official implementations. All approaches are evaluated on the same network architecture. Since the authors of the Energy score also conduct their evaluation with a pre-trained WideResNet, we additionally report the original results from their paper marked as Energy*. Since our approach is based on a trainable model, unlike the related approaches, we conduct each experiment five times and report the mean value over all runs.

4.2. Out-of-Distribution Detection Results

The OOD detection results are presented in Tab. 1. Similar to the related work [6, 18], we report the performance averaged over the respective OOD test sets. A detailed evaluation of the individual OOD datasets is provided in the supplementary material. In general, the degree of difficulty increases with a larger class number and higher resolution. Thus, the OOD detection for CIFAR-10 is less complex, while for CIFAR-100 it is more difficult and Tiny ImageNet is the most complex setup in our experiments. In the case of CIFAR-10, the built-in softmax score already achieves a good OOD detection performance in terms of AUROC and AUPR-S. The other related approaches besides ReAct improve the MSP especially in the detection of OOD samples (AUPR-E) and the FPR-95 metric. Nevertheless, we obtain better results in all cases, especially for the FPR-95 metric. For CIFAR-100, the OOD detection performance of all methods decreases with 100 instead of 10 in-distribution classes. Here, we outperform the other approaches in all metrics. For both architectures, the AUPR-E metric highlights that the OOD detection, in particular, is significantly improved compared to the other methods. The last experiment, with Tiny ImageNet as an in-distribution dataset, emphasizes that the more classes that need to be categorized, the harder it becomes to detect OOD samples. In this setup, we achieve comparable or even better results (AUPR-E, FPR-95) than ReAct and outperform the other methods in all metrics. Overall, the metrics indicate that all methods have a tendency towards the detection of in-distribution samples. This is shown by the higher value for AUPR-S in contrast to AUPR-E. However, our approach surpasses prior work in the detection of OOD samples with an improvement of almost 10% in terms of AUPR-E for the majority of the cases and works consistently for all cases.

4.3. Visual Illustrations

In Fig. 1, we demonstrate example heatmap definitions used for the decoder training. Fig. 1a shows OOD images from Places365, Fig. 1b visualizes their closest in-distribution training sample from Tiny ImageNet, while in Fig. 1c the resulting heatmap is illustrated. In Fig. 1, the blue colors mark similar regions, whereas the red/yellow colors highlight features that differ from the in-distribution image. The first row represents a far-distribution image, while the second row represents a near-distribution image. Here, the near-distribution heatmap shows a milder response compared to the far-distribution heatmap.

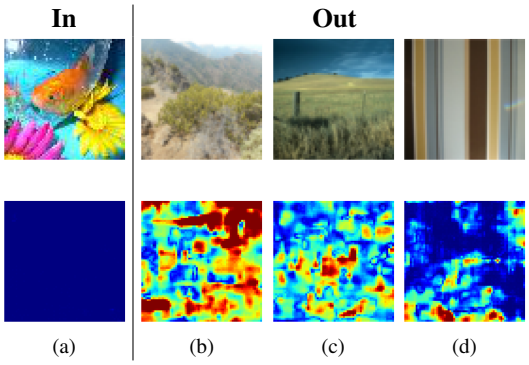


Figure 4: Visual results with ResNet trained on TinyImageNet (a) as an in-distribution database (**In**). Examples from the OOD databases (**Out**) iNaturalist (b), SUN (c) and Textures (d) are shown. The original images are displayed in the top row, whereas the estimated heatmaps are in the bottom row. Blue colors mark in-distribution regions, whereas the red/yellow colors highlight OOD regions.

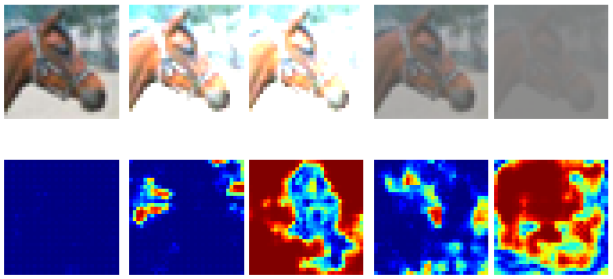
Fig. 4 shows qualitative results for ResNet trained on Tiny ImageNet. The first row presents the original input images, while the estimated heatmaps are in the second row. Fig. 4a visualizes an in-distribution example of Tiny ImageNet. The heatmap shows no response compared to the original image. Thus, the heatmap entries are close to zero, which in turn means that features extracted with the classifier are representative of in-distribution samples. From Fig. 4b to Fig. 4d, we illustrate three different OOD examples as input to the classifier trained on Tiny ImageNet. In all cases, regions over the entire images are highlighted by out-of-distribution responses indicating the differences from the closest in-distribution sample. The OOD examples cover different OOD types with landscapes (Fig. 4b, Fig. 4c) and textures (Fig. 4d) demonstrating the generalization capability of our approach. Further visual results are provided in the supplementary material.

D_{in} (model)	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
CIFAR-10 (ResNet)	MSP [5]	90.72	97.89	63.48	55.21
	ODIN [16]	88.33	96.67	71.49	38.35
	Mahalanobis [15]	92.33	98.29	71.30	39.52
	Energy [18]	91.72	97.90	72.12	37.97
	ReAct [28]	91.71	97.89	72.55	36.52
	Ours	**96.47**	**99.17**	**83.73**	**15.37**
CIFAR-10 (WideResNet)	MSP [5]	91.48	98.18	63.47	56.77
	ODIN [16]	95.01	98.68	84.39	21.09
	Mahalanobis [15]	92.03	98.09	75.44	32.73
	Energy [18]	94.91	98.75	80.89	24.26
	Energy* [18]	91.88	97.83	-	33.01
	ReAct [28]	51.92	85.46	17.53	97.12
	Ours	**96.36**	**99.07**	**86.73**	**14.06**
CIFAR-100 (ResNet)	MSP [5]	79.29	95.04	40.34	76.58
	ODIN [16]	83.28	95.96	48.74	67.96
	Mahalanobis [15]	73.46	93.00	35.90	79.46
	Energy [18]	82.07	95.71	43.92	74.45
	ReAct [28]	84.22	96.27	49.08	67.78
	Ours	**86.74**	**96.49**	**58.78**	**52.73**
CIFAR-100 (WideResNet)	MSP [5]	65.31	90.38	26.21	88.45
	ODIN [16]	79.43	94.60	43.98	73.19
	Mahalanobis [15]	73.99	92.58	43.80	68.45
	Energy [18]	77.11	93.95	39.07	78.03
	Energy* [18]	79.56	94.87	-	73.60
	ReAct [28]	80.74	95.24	48.04	67.47
	Ours	**85.98**	**95.96**	**61.14**	**49.86**
Tiny ImageNet (ResNet)	MSP [5]	72.16	93.12	29.06	87.93
	ODIN [16]	75.25	94.01	32.59	85.67
	Mahalanobis [15]	74.99	93.01	44.03	68.97
	Energy [18]	75.99	94.20	33.74	84.00
	ReAct [28]	**85.53**	**96.50**	54.52	61.10
	Ours	85.28	96.25	**56.14**	**54.66**

Table 1: Comparison of the OOD detection performance in terms of AUROC, AUPR-Success, AUPR-Error, and FPR at 95% TPR. The results are averaged over the number of OOD test sets. We compare our approach to methods that do not further optimize the classifier but operate on the pre-trained model. The results marked with * are taken from the original paper. ↑ indicates that larger values are better, whereas ↓ marks that lower values are better.

4.4. Lighting Effect

The performance of neural networks can be affected by external factors such as lighting conditions. Therefore, we evaluate the influence of brightness and contrast changes on our approach. We augment in-distribution test data with increased brightness (B) selected from B $= \{2.0, 2.5\}$ and reduced contrast (C) selected from C $= \{0.1, 0.5\}$. The higher the brightness and the lower the contrast, the more augmented in-distribution samples should be detected as OOD, since the relevant features are no longer recognizable. In Fig. 5, examples of a horse with augmented brightness and augmented contrast are visualized with the corresponding heatmap predictions. The heatmaps clearly show a higher response for images with higher brightness augmentation as well as for images with lower contrast augmentation. With increasing brightness the in-distribution features of the images are less recognizable and therefore the images should be labelled as OOD. The same applies to images with reduced contrast. In general, more image regions are visualized as OOD for increasing brightness values. For B $= 2.5$ (Fig. 5c), the pixels covering the horse's head are still marked as in-distribution, while most other pixels are highlighted as OOD. By contrast, when the brightness and contrast are not changed (Fig. 5a), the heatmap entries are



(a) (b) B=2.0 (c) B=2.5 (d) C=0.5 (e) C=0.1

Figure 5: Example images of a horse from the CIFAR-10 testset with corresponding heatmap predictions augmented with different brightness (B) or contrast (C) values. (a) is the original image without augmentation, whereas in (b) and (c) the brightness value is increased. In (d) and (e), the contrast is decreased. Blue colors mark in-distribution regions, whereas the red/yellow colors highlight out-of-distribution regions.

zero. For the image augmented with C $= 0.1$ (Fig. 5e), the corresponding heatmap also shows larger OOD regions compared to the original image. In the supplementary material, we report the OOD detection performance evaluation where the augmented in-distribution samples are labelled as OOD. As already evident from the qualitative heatmap examples, the OOD detection performance increases with higher brightness and reduced contrast.

4.5. Discussion

Overall, we implement a trainable model to generate the heatmaps. Since we optimize the parametrized model, we naturally have additional computational effort in comparison to prior work [5, 16]. Nevertheless, the heatmaps can not only be leveraged to detect OOD samples but also as a possibility to illustrate in- and out-of-distribution image regions based on the classifier features and the class prediction. At the moment, our approach is specifically designed for image classification models. Since the reference images are associated with the class prediction, the approach is limited to classification tasks. The adaptation to other domains, such as object detection, could be addressed in future work.

4.6. Ablation Study

We study the influence of the OOD training data size. Since AUPR-S and AUPR-E set the focus on the positive class, we report the AUROC and provide the other metrics in the supplementary material. We conduct the ablation studies with both network architectures pre-trained on CIFAR-10 and CIFAR-100 as in-distribution dataset and 80 Million Tiny Images as OOD training set. As mentioned in Sec. 4.1, the OOD training set size is set to $\frac{1}{5}$ of the in-distribution dataset size. We keep the number of

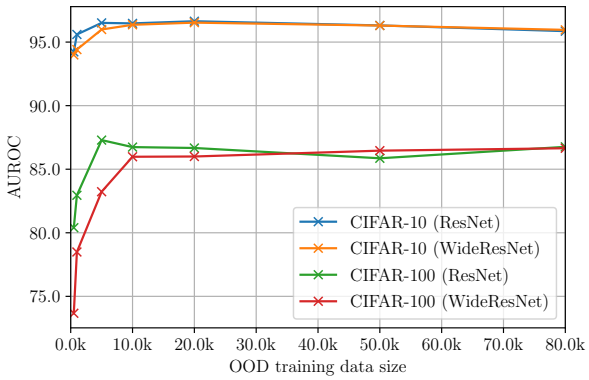


Figure 6: AUROC results when alternating the OOD training set. The in-distribution dataset size is fixed.

in-distribution samples fixed at $50K$ and vary the number of OOD samples. In Fig. 6, we present the AUROC results when alternating the OOD training set size with sets of $\{500, 1000, 5000, 10000, 20000, 50000, 80000\}$ samples. Especially for the WideResNet architecture, the performance considerably drops when using less than $10K$ samples. Between $10K$ to $80K$ OOD samples, the AUROC slightly decreases for both datasets and both architectures. The minor deterioration in performance can be explained by the increasing focus on OOD samples. Since the performance degradation is negligible, the OOD training set size should be at least $\frac{1}{5}$ of the in-distribution dataset size.

5. Conclusion

We presented an approach to learn heatmaps for OOD detection, which also serve as an illustration of in- and out-of-distribution image regions. We introduced the decoder that is trained to produce heatmaps zero response for an in-distribution sample and high response for OOD samples, based on the classifier features and the class prediction of a fixed classifier. Our main contribution consists of the OOD sample heatmap definition that is based on the normalized difference from the closest in-distribution sample. The heatmap eventually acts as a margin to distinguish between in- and out-of-distribution images. In our evaluations, we show that our OOD score function, based on the heatmap response, achieves state-of-the-art OOD detection performance compared to fixed classifiers approaches, trained on CIFAR-10, CIFAR-100, and Tiny ImageNet.

Acknowledgements. The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action" within the project "KI Delta Learning" (Förderkennzeichen 19A19013A). The authors would like to thank the consortium for the successful cooperation.

References

[1] Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, and Marius Popescu. A generic and model-agnostic exemplar synthetization framework for explainable ai. In Frank Hutter, Kristian Kersting, Jefrey Lijffijt, and Isabel Valera, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 190–205, Cham, 2021. Springer International Publishing.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[3] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural network. In *ICLR*, 2017.

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

[7] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[8] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[9] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10948–10957, 2020.

[10] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[13] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[14] 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*, 2018.

[15] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *NIPS*, volume 31. Curran Associates, Inc., 2018.

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

[17] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *CVPR*, 2021.

[18] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *NIPS*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

[19] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[20] Yuval Netzer, Tiejie Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS*, 2011.

[21] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[22] Stanislav Pidhorskyi, Ranya Almohsen, Donald A. Adjeroh, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *NeurIPS*, 2018.

[23] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[26] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE*

International Conference on Computer Vision (ICCV), Oct 2017.

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

[28] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

[29] 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*, 30:1958–1970, 2008.

[30] Jian xiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[31] Pingmei Xu, Krista A. Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *ArXiv*, abs/1504.06755, 2015.

[32] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[33] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *ArXiv*, abs/1506.03365, 2015.

[34] Alireza Zaeemzadeh, Niccoló Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *CVPR*, 2021.

[35] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

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

[37] E. Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13991–14000, 2020.