XAI Saliency Maps Can Be Used for OOD Detection

Jonatan Hoffmann Hanssen
Department of Informatics
University of Oslo
jonatahh@uio.no

Hugo Lewi Hammer
Department of Computer Science
OsloMet
Department of Holistic systems
SimulaMet
hugo.hammer@oslomet.no

Abstract

When neural networks are used in high-impact settings such as cancer detection or autonomous driving, we must not only require that they make predictions with high accuracy, but also that they are aware of their shortcomings, and alert us when faced with unusual data. The need for models which "know what they do not know" has lead to the field of Out-of-Distribution (OOD) detection, which attempts to detect when models are exposed to data points that are far outside of their training data and thus unlikely to be classified correctly. OOD detection is a young and developing field, and there are, to date, no methods which achieve superior performance on all benchmarks. Thus, there is a need for novel techniques which push the field forward. In this paper, we show that simply aggregating the saliency values produced by Explainable Artificial Intelligence (XAI) methods such as Integrated Gradients, Local Interpretable Model-Agnostic Explanations (LIME) or Guided Backpropagation (GBP) is an effective way to detect OOD data points. Furthermore, we show that these aggregates are only weakly correlated with the model's confidence in the predicted class, which allows us to combine the Maximum Logit Score with saliency aggregates, achieving AUROC scores which are close to State-of-the-Art (SoTA) OOD detection methods.

1 Introduction

Machine Learning (ML) generally, and Deep Learning (DL) specifically, has seen a tremendous increase in performance in recent years, performing comparable to humans in tasks such as image classification, speech and handwriting recognition, as well as many others [3]. Consequently, DL methods have been deployed in a multitude of fields, and have become a part of our daily lives through their role in web search, text translation, computer vision, as well as in many other technologies which are taken for granted. In the medical field, DL has the potential to provide faster and more accurate detection of diseases by being trained on cases from thousands of previous patients [8]. Despite this, the adoption DL in high impact fields, such as medicine, has been slow, with Rajkomar et al. [9] stating that: "surprisingly little in health care is driven by machine learning".

To explain this discrepancy, we should consider that despite their impressive performance, the application of DL methods is not without drawbacks. Firstly, deep neural networks are inherently unexplainable due to the large number of parameters that any non-trivial network has. State-of-the-Art (SoTA) models will perform millions of operations to evaluate a single data point, and it is therefore impossible for humans to comprehend and explain the entire process which led the model to make a particular decision. In medicine, this is a major limitation of DL methods, as both doctors and patients expect to be able to understand why a decision was made [16]. In other high-impact fields, such as autonomous driving, this lack of transparency also has serious practical and legal ramifications.

Preprint. Under review.

Secondly, although neural networks may attain high accuracy on test data and appear to have learned great insights about the tasks they are employed in, they often lack robustness and can suffer large drops in performance on data points which are slightly different from the training data. As Szegedy et al. [13] has shown, it is possible to create data points which are imperceptibly different from normal data points, yet still fool otherwise high performing models. More problematically, unlike humans, who recognize when they are faced with a novel situation where their expertise might be lacking, DL methods will predict equally confidently on data points which are far outside the data they have been trained on [16].

These two problems lead to the fields of Explainable Artificial Intelligence (XAI) and Out-of-Distribution (OOD) detection. XAI attempts to explain the reasons why a model came to a decision, which helps to remedy the black-box nature of complicated DL models. In a healthcare setting, such explanations can be inspected by medical practitioners to confirm the diagnosis, and can be used to give patients information about why decisions regarding their health were made. In autonomous driving or other automated high impact fields, XAI can be used to detect failure modes or to understand and improve trained models. OOD detection attempts to uncover when a data point is too different from the training data to be classified reliably. These methods could alert medical practitioners when such data points occur, thus avoiding potentially fatal misclassifications. In autonomous driving, the system could detect novel situations and cede control back to the user, avoiding accidents.

Both of these fields have seen increased interest in recent years, and are vital parts of any integration of DL in high impact settings. As two vibrant fields of study, there is great potential to combining insights from one field to improve performance in the other; an area which is underexplored. This paper has investigated the possibility of using XAI methods to aid OOD detection. The overarching intuition is that when saliency maps are generated on semantically shifted data, they are creating an explanation for a wrong prediction, which leads to systematic differences when compared to explanations generated on In-Distribution (ID) data. This methodology is essentially non-existent in the literature: OpenOOD [17, 19], the standard OOD detection benchmarking framework which includes over 40 different methods, contains no method which uses XAI as part of its functioning.

In this paper, we show that the magnitude of values generated by XAI saliency mapping methods are significantly different between ID and OOD data. By aggregating all saliency values for a given input image, we get a single number which can be used for OOD detection, achieving Area Under Receiver Operating Characteristic (AUROC) scores which are comparable to baseline methods such as the Maximum Softmax Probability (MSP) or the Maximum Logit Score (MLS). Furthermore, we show that saliency aggregates are only somewhat correlated with these baselines. Inspired by the work of Shekhar et Al. [11], we combine saliency aggregations with the MLS, to exploit the information gained both by the explanation and the model prediction for OOD detection.

The key contributions of this work is as follows:

- We show that XAI saliency mapping methods such as Local Interpretable Model-Agnostic Explanations (LIME), Gradient Class Activation Mapping (GradCAM) and Guided Back-propagation (GBP) capture valuable magnitude information which can be used for OOD detection. By forgoing normalization, which is common when displaying saliency maps, we achieve higher AUROC scores than previous works, such as Martinez et Al. [7].
- Further illustrating the importance of the magnitude information in saliency maps, we show how aggregations such as the Coefficient of Variation (CV), Relative Mean Absolute Difference (RMD) and Quartile Coefficient of Determination (QCD), which are magnitude invariant, achieve far worse results than aggregations which capture the magnitude.
- We show that for many XAI saliency mapping methods, saliency aggregates are only weakly correlated with the confidence of the prediction (the MLS). This allows us to combine the discriminative power of XAI saliency maps with traditional OOD detection, achieving results which are close to SoTA OOD detection methods.

2 Related Work

While the combination of XAI and OOD detection has been explored in many previous works, the majority of them focus on explaining why a data point was marked as OOD, as opposed to using XAI to aid the detection itself. Delaney et al. [4], Sipple et al. [12] and Tallón-Ballesteros et al. [14] used XAI methods to explain OOD detection decisions. Within network security, XAI has been as part of anomaly detection systems to detect malicious or faulty network traffic. Here, it has been used to explain detections [1, 6], but also to aid in detection itself by inspecting the explanations of the detection system [15, 18]. These methods thus use XAI to aid OOD detection in a similar manner to our work, however they are strictly focused on sequential network traffic data as opposed to images, and are mostly concerned with detection "unnatural" data samples such as intentionally malicious traffic or that generated by faulty equipment, as opposed to natural OOD data caused by semantic or covariate shift occurring when a model is deployed.

Martinez et al. [7] is the most relevant previous work. Here, the authors explicitly aim to use XAI to improve OOD detection on images. They do this by looking at saliency maps produced by GradCAMPlusPlus [2] during inference, i.e the heatmaps that explain which parts of the image was most influential to classify the image as a specific class. Using these heatmaps, they perform distance-based OOD detection: By collecting all explanations for each image in the ID dataset, they are able to construct archetypical explanations, and can make clusters of explanations. To perform OOD detection, they simply compare the explanation of a new data point to the clusters of archetypical explanations, and mark it as OOD if it has a distance which is over a certain threshold. In contrast to our work, Martinez et al. consider normalized heatmaps, not raw saliency values.

This method performs decently on toy benchmarks, achieving scores similar to SoTA methods when using *Fashion MNIST* as ID and *MNIST* as OOD. However, this method fails in more complicated scenarios, achieving an AUROC score of only 52% on *CIFAR10* vs *SVHN*, which is no better than pure guessing and far below most other methods. The paper thus ends with the authors concluding that "OoD detection approaches that are specifically designed for the purpose achieve in general better detection scores at the cost of an additional computational burden in the model's construction" [7].

For more potential related work, we can look to OpenOOD [17, 19], which aims to provide a comprehensive benchmark of all relevant methods in the field of OOD detection. Out of all 41 OOD detection methods included in this benchmark, there are no methods which use XAI. However, as many XAI methods utilize the gradients of the network to generate saliency values, we could also consider OOD detection models which utilize gradients in some form as tangentially related to this thesis. In this regard GradNorm [5] is somewhat related, as it utilizes the norm of the gradients of the network with respect to the Kullback-Leibler distance between the outputs and a uniform distribution to perform OOD detection.

3 Background

Bare skriv her om problem statement, som de fleste andre OOD detection artikler gjør. F.eks som ReAct, ODIN osv

4 Saliency Aggregation plus Logit

As Shekhar et Al. [11] has shown, SoTA OOD detection performance can be achieved by combining different methods which extract information from the network in different ways. Inspired by this work, we present the Saliency Aggregation plus Logit framework, which combines XAI saliency aggregation and the MLS:

Under this framework, the OOD score is a sum of a saliency aggregate and the maximum logit score. However, due to the fact that both the logits and saliency aggregates can be of arbitrary magnitude, we must normalize them before summing if we want each part to contribute equally to the final score. Thus, we can sum the Z-scores of each metric instead. This ensures that the values of the maximum logit and the saliency aggregate are distributed in the same way. To calculate the Z-scores, we simply subtract the mean and divide by the standard deviation over an entire ID validation dataset, for each metric. Thus, we calculate the mean and standard deviations of the maximum logit over an ID validation set μ^{id}_{MLS} and σ^{id}_{MLS}, as well as the mean and standard deviation of the aggregate of saliencies μ^{id}_{Agg} and σ^{id}_{Agg}.

To align ourselves with convention in the field of OOD detection, we must define the OOD detection score as one which is higher for ID data than for OOD data. However, given that our framework does not place any limits on the choice of aggregation, we cannot know in advance whether a specific aggregation will have higher or lower values for ID data. To keep the framework as general as possible, we therefore include a *sign* factor, which multiplies the saliency aggregate by 1 or -1, depending on whether the ID aggregates are higher or lower, respectively. To do this, we calculate the mean value of the aggregation metric over a validation OOD dataset. We denote this value as μ_{Agg}^{ood}. The sign factor can then be calculated by $\text{sign}(\mu_{\text{Agg}}^{id} - \mu_{\text{Agg}}^{ood})$, which we denote as S.

We now have the necessary values required to define this framework mathematically. We assume we have a model $f : \boldsymbol{x} \to \mathbb{R}^C$, an XAI saliency mapping method $s : (f, \boldsymbol{x}) \to \mathbb{R}^{K\times N\times M}$, and an aggregation function $A : \mathbb{R}^{K\times N\times M} \to \mathbb{R}$. The saliency mapping method is defined as one which does not normalize or rectify its outputs, meaning that if we use methods such as GradCAM, we must modify them to remove the normalization step. An OOD detector under this framework then has the following form, given a threshold δ:

$$g(\boldsymbol{x}; s, A, \delta) = \begin{cases} \text{in} & S \cdot \frac{A(s(\boldsymbol{x},f)) - \mu_{\text{Agg}}^{id}}{\sigma_{\text{Agg}}^{id}} + \frac{\max_i S(\boldsymbol{x}) - \mu_{\text{MLS}}^{id}}{\sigma_{\text{MLS}}^{id}} \geq \delta \\ \text{out} & S \cdot \frac{A(s(\boldsymbol{x},f)) - \mu_{\text{Agg}}^{id}}{\sigma_{\text{Agg}}^{id}} + \frac{\max_i S(\boldsymbol{x}) - \mu_{\text{MLS}}^{id}}{\sigma_{\text{MLS}}^{id}} < \delta \end{cases} \tag{1}$$

In fact, this detector can be simplified somewhat. Consider the following:

$$S \cdot \frac{A(s(\boldsymbol{x}, f)) - \mu_{\text{Agg}}^{id}}{\sigma_{\text{Agg}}^{id}} + \frac{\max_i S(\boldsymbol{x}) - \mu_{\text{MLS}}^{id}}{\sigma_{\text{MLS}}^{id}} = \tag{2}$$

$$S \left(\cdot \frac{A(s(\boldsymbol{x}, f))}{\sigma_{\text{Agg}}^{id}} - \frac{\mu_{\text{Agg}}^{id}}{\sigma_{\text{Agg}}^{id}} \right) + \frac{\max_i S(\boldsymbol{x})}{\sigma_{\text{MLS}}^{id}} - \frac{\mu_{\text{MLS}}^{id}}{\sigma_{\text{MLS}}^{id}} = \tag{3}$$

$$S \cdot \frac{A(s(\boldsymbol{x}, f))}{\sigma_{\text{Agg}}^{id}} + \frac{\max_i S(\boldsymbol{x})}{\sigma_{\text{MLS}}^{id}} - \left(S \cdot \frac{\mu_{\text{Agg}}^{id}}{\sigma_{\text{Agg}}^{id}} + \frac{\mu_{\text{MLS}}^{id}}{\sigma_{\text{MLS}}^{id}} \right) \tag{4}$$

Notice how all the values in the third term of the above summation are constants; they do not depend on $\boldsymbol{x}$. Thus, we can disregard these terms, as all they do is shift all outputs by a constant value. The final OOD detector thus has the following form:

$$g(\boldsymbol{x}; s, A, \delta) = \begin{cases} \text{in} & S \cdot \frac{A(s(\boldsymbol{x},f))}{\sigma_{\text{Agg}}^{id}} + \frac{\max_i S(\boldsymbol{x})}{\sigma_{\text{MLS}}^{id}} \geq \delta \\ \text{out} & S \cdot \frac{A(s(\boldsymbol{x},f))}{\sigma_{\text{Agg}}^{id}} + \frac{\max_i S(\boldsymbol{x})}{\sigma_{\text{MLS}}^{id}} < \delta \end{cases} \tag{5}$$

To develop a method under this framework, one simply chooses an XAI saliency mapping method and an aggregate. In our testing, we have chosen LIME, Occlusion, GradCAM, Integrated Gradients and GBP as the XAI saliency mapping methods. For aggregate functions, six aggregates which capture the magnitude of the saliency values in different ways have been chosen. These are the mean, median, vector norm, range, maximum value and third quartile. In addition, three aggregate functions which are (more or less) scale invariant have also been chosen, which measure the statistical spread of the saliencies without considering the magnitude. These aggregates are the CV, the RMD and the QCD. By including these aggregate functions, we can compare the performance of magnitude variant and invariant functions and investigate the importance of saliency magnitudes in an OOD detection context.

5 Results

In this section, we report the results of our experiments. This section is divided into three parts: First, we show that the magnitudes of saliency values, by themselves, are discriminative in an OOD detection context, even without combining them with the MLS. We do this by using different saliency aggregates to perform OOD detection directly on the ImageNet200 and CIFAR10 OpenOOD benchmarks. In addition, we show that aggregates which are magnitude invariant exhibit very low discriminative power, showcasing the importance of using raw saliency values as opposed to normalized heatmaps. Next, we show that saliency aggregates are not strongly correlated with baseline OOD detection metrics such as the MLS, justifying their combination in the Saliency Aggregation plus Logit framework. Finally, we test the performance of a selection of methods under this framework, and show that XAI based OOD detection can be competitive with SoTA methods.

In all cases, the experiments are conducted on pretrained ResNet architectures with weights provided as part of OpenOOD. This ensures that the results are comparable to other methods tested under this framework.[1]

5.1 The discriminative performance of saliency aggregates

Figure 1 shows the combined Near- and Far-OOD detection performance on both ImageNet200 and CIFAR10 for all combinations of XAI methods and aggregate functions. As we can see, simply aggregating the saliency maps generated on ID and OOD data points allows for efficient detection of OOD data, with AUROCs of around 0.80 given a correct choice of aggregation. In addition, we see that while the first six aggregations (those who capture magnitude) perform well, the last three (which are magnitude invariant) perform very poorly. This strengthens the claim that it is the magnitude information of saliency values which is important for OOD detection, not their spread or relation to each other.

In most cases, ID saliencies have higher magnitudes than OOD saliencies, similarly to how ID data points more often have higher maximum logit scores than OOD data points. However, depending on the choice XAI saliency mapping method and aggregate function, this is not always the case.

5.2 Correlation between saliency aggregates and the maximum logit

Figure 2 shows the correlation between saliency aggregates and the MLS, averaged over both the ImageNet200 and CIFAR10 benchmarks. As we can see, GradCAM is very strongly correlated with the MLS. In fact, it can be shown that performing mean aggregation on GradCAM saliencies is equivalent to MLS OOD detection, given certain conditions.[2] However, the other saliency methods are less correlated, showing that the discriminatory performance of saliency aggregation is not just due to their connection with the logit of the predicted class. These results also hint at the potential for combining such aggregates with traditional OOD detection methods.

5.3 Performance of Saliency Aggregation plus Logit

To demonstrate the performance of the framework, we select five combinations of saliency mapping methods and aggregations to test and compare against the baseline MLS. These combinations are selected by choosing the best performing aggregation for each XAI method, as reported in figure 1. To ensure unbiased results, the OpenOOD benchmarks have been split into validation and testing benchmarks, with the choice of aggregation being done on the validation benchmarks while the final results are reported on the testing benchmarks. In addition, the testing benchmarks have been bootstrapped ten times, allowing for statistical comparisons between the baseline MLS and the methods which complement MLS by adding a saliency aggregate. As we can see from table 1, in all benchmarks except for CIFAR100 Near-OOD, there are statistically significant improvements over just using the MLS for multiple methods, showing that XAI saliency aggregations add relevant information to traditional OOD detection methods.

Finally, we compare the best performing method under the Saliency Aggregation plus Logit framework, GBPNorm, to SoTA OOD detection methods, in table 2. In this case, the testing is done on the entire benchmark as developed by OpenOOD, as opposed to a testing split. This is done to ensure accurate comparison with the results reported by OpenOOD [19], which use the entire dataset.

[1] `https://zjysteven.github.io/OpenOOD/`

[2] See the appendix.

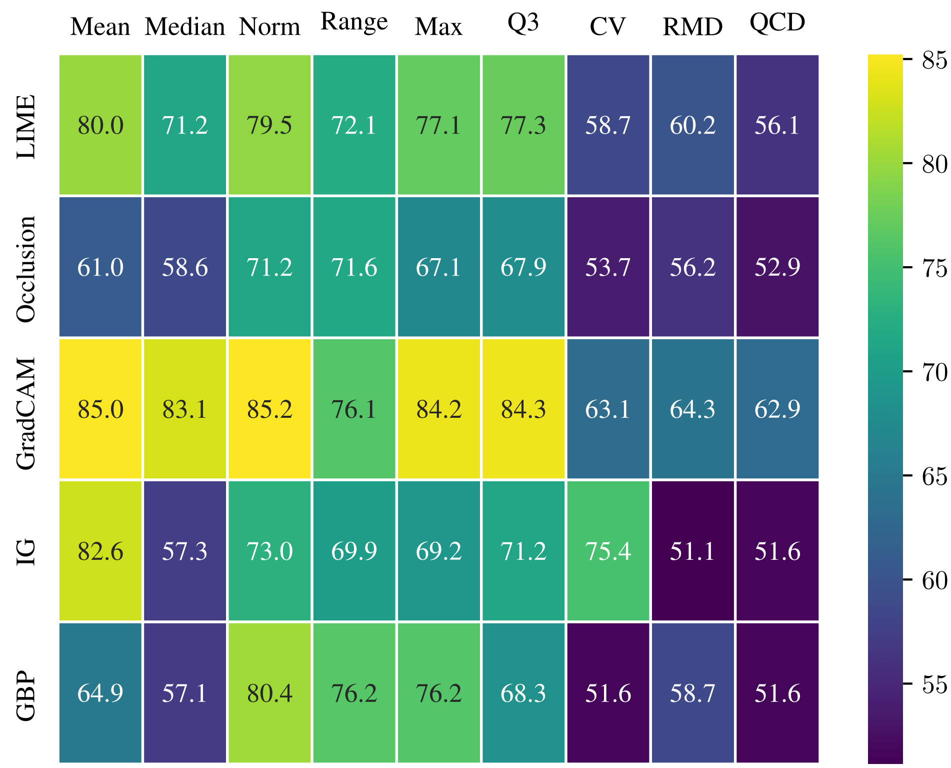

Figure 1: Heatmap of overall performance AUROC performance for all combinations of XAI methods and aggregations on ImageNet200 and CIFAR10

Dataset	MLS	LIMENorm	OccRange	GradCAMNorm	IGMean	GBPNorm
ImageNet200 Near-OOD	82.75	83.46**	80.21	83.15**	83.15**	**83.70****
ImageNet200 Far-OOD	91.28	93.12**	92.20**	92.09**	91.04	**94.03****
ImageNet1K Near-OOD	76.33	74.23	70.95	76.68**	75.37	**78.98****
ImageNet1K Far-OOD	89.51	92.17**	88.98	91.25**	88.85	**94.09****
CIFAR10 Near-OOD	86.97	87.80**	88.00**	86.92	87.21**	**89.85****
CIFAR10 Far-OOD	91.82	93.35**	90.54	91.81	92.47**	**94.79****
CIFAR100 Near-OOD	**81.24**	78.92	78.56	81.24	80.58	78.73
CIFAR100 Far-OOD	79.86	82.13**	81.39**	79.93**	81.95**	**84.18****

Table 1: Average AUROC scores over ten bootstraps for the five example methods under the Saliency Aggregation plus Logit framework, as well as the MLS. The asterisks denote the Bonferroni corrected statistical significance of a Wilcoxon signed-rank test done against the null hypothesis that the developed methods are no better than the MLS. For each benchmark, the best performing method is highlighted in bold.

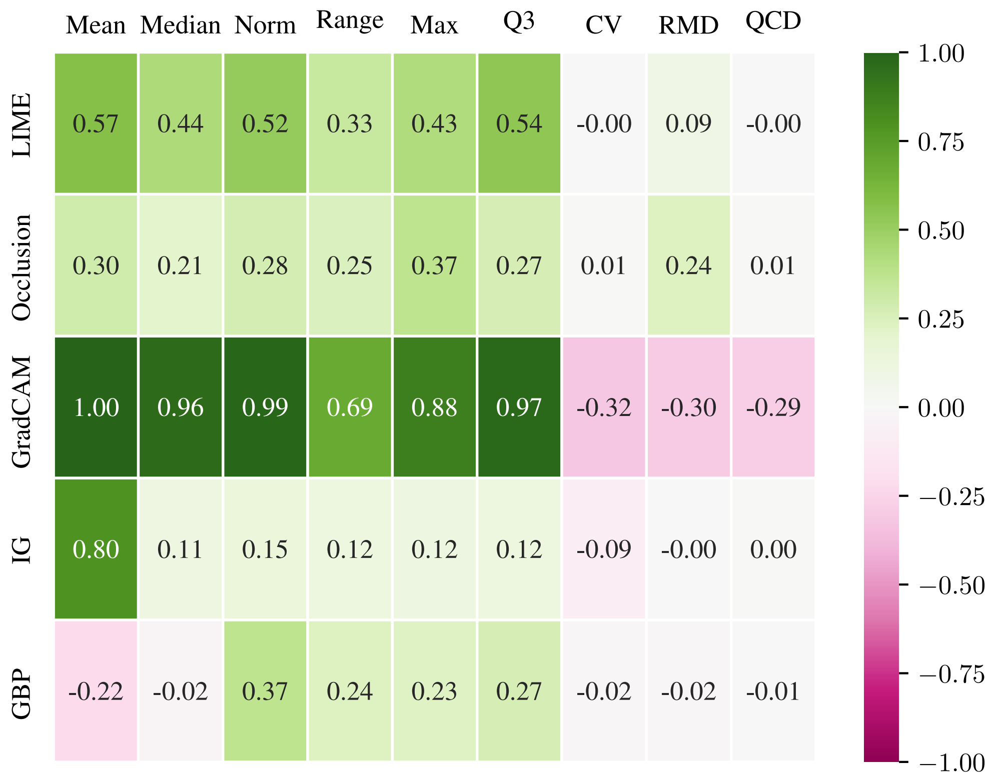

Figure 2: Heatmap of overall performance AUROC performance for all combinations of XAI methods and magnitude aggregations on ImageNet200 and CIFAR10

Dataset	RotPred	CombOOD	OE+MSP	AdaScale-L	GBPNorm
ImageNet200 Near-OOD	81.59	**95.74**	85.73	84.84	83.37
ImageNet200 Far-OOD	92.56	92.57	89.02	**94.86**	93.87
ImageNet1K Near-OOD	76.52	**95.22**	N/A	84.27	78.90
ImageNet1K Far-OOD	90.00	90.24	N/A	**97.28**	94.03
CIFAR10 Near-OOD	92.68	91.13	**94.82**	74.99	89.90
CIFAR10 Far-OOD	96.62	94.65	**96.00**	79.02	94.92
CIFAR100 Near-OOD	76.43	78.77	**88.30**	80.54	78.73
CIFAR100 Far-OOD	88.40	**85.87**	81.41	83.38	84.18

Table 2: The GBPNorm method under the Saliency Aggregation plus Logit compared to the performance of SoTA OOD detection methods as reported by Zhang et Al. [19]. The best performing method in each case is highlighted in bold.

From this table we see that although GBPNorm does not achieve the highest AUROC in any benchmark, its performance is quite close to the SoTA in many cases. This demonstrates the clear potential of using XAI saliency maps for OOD detection.

6 Conclusion

Appendix

6.1 Proof of Equivalence between GradCAM Saliency Aggregation and MLS

The following section proves that taking the mean value of the saliency values outputted by GradCAM is equivalent to MLS, up to a constant a.

In this case, we choose mean aggregation, and GradCAM as the saliency mapping method. In addition, we assume that the classification stage of the Convolutional Neural Network (CNN) is a simple Global Average Pooling (GAP) over the feature map followed by a single linear layer. Such a structure is the classification head of all ResNet models. Finally, we choose to perform GradCAM on the final layer of the network, which is recommended by Selvaraju et al. [10]. These conditions give rise to the following theorem:

Theorem 1. *Assume we have a network where $y^c = \sum_k mean(F_k) \cdot W_{ck}$, where F is a convolutional feature map and W is a linear layer of size channels $\times$ classes. Assume further that $A = mean$ and $s = GradCAM_F$, i.e GradCAM applied to the convolutional feature map F. Under these assumptions, the OOD score of Saliency Aggregation is equal to the MLS, up to a constant a:*

$$SalAgg(\boldsymbol{x}; s, A) = a \cdot MLS(\boldsymbol{x})$$

Proof. As defined in equation **??**, the OOD detection score of Saliency Aggregation has the following form:

$$\text{SalAgg}(\boldsymbol{x}; s, A) = \text{sign}(\mu_{id} - \mu_{ood}) \cdot A(s(\boldsymbol{x}, f)) \tag{6}$$

In this special case, A is equal to *mean* and s is equal to *GradCAM*. Following Selvaraju et al. [10], the saliency map generated by GradCAM has the following definition:

$$\text{GradCAM}_F(\boldsymbol{x}) = ReLU\left(\sum_k \left(\frac{1}{N \cdot M}\sum_i \sum_j \frac{\delta y^c}{\delta F^k_{ij}}\right) F^k\right). \tag{7}$$

Here, F^k is the k'th channel of the final convolutional feature map, while N and M are its dimensions.[3] The above equation simply describes averaging the gradients of the logit of class c for each channel, and using these values to perform a weighted sum of the channels in the feature map, as described in section **??**. While c can be any class index, in this case, we define $c = \max_i f_i(\boldsymbol{x})$, i.e we calculate the saliency map for the predicted class, as defined by the framework. From the formal definition for the Saliency Aggregation framework (section **??**), the saliency mapping method is specified as the unnormalized and unrectified saliency values. Thus, we remove the Rectified Linear Unit (ReLU) activation function from the GradCAM definition and use the raw saliency values instead. Using mean as the aggregation, the OOD score is then:

$$\text{SalAgg}(\boldsymbol{x}; s, A) = \text{sign}(\mu_{id} - \mu_{ood}) \cdot \text{mean}\left(\sum_k \left(\frac{1}{N \cdot M}\sum_i \sum_j \frac{\delta y^c}{\delta F^k_{ij}}\right) F^k\right). \tag{8}$$

Given the assumptions of the theorem, we have that the logit y^c for class c is calculated in the following manner:

[3]The attentive reader will notice that this is the same notation used for the dimensions of saliency maps throughout this thesis. This is intentional, as the saliency map generated by GradCAM has the same dimensions as the feature map on which the algorithm is performed.

$$y^c = \sum_k \text{mean}(F_k) \cdot W_{ck} \tag{9}$$

$$= \sum_k \left(\frac{\sum_i \sum_j F_{ij}^k}{N \cdot M} \cdot W_{ck} \right). \tag{10}$$

This equation simply describes GAP (all channels are averaged to a single number) followed by a single linear layer (each logit is a weighted sum of the average pooled channels, with the weights defined by a specific row/column in the weight matrix W). Given our definition of $c = \max_i f_i(\boldsymbol{x})$, $y^c = \text{MLS}(\boldsymbol{x})$. We return to the equation for $\text{SalAgg}(\boldsymbol{x})$:

$$\text{SalAgg}(\boldsymbol{x}; s, A) = \text{sign}(\mu_{id} - \mu_{ood}) \cdot \text{mean} \left(\sum_k \left(\frac{1}{N \cdot M} \sum_i \sum_j \frac{\delta y^c}{\delta F_{ij}^k} \right) F^k \right). \tag{11}$$

Given equation 10,

$$\frac{\delta y^c}{\delta F_{ij}^k} = \frac{W_{ck}}{N \cdot M}. \tag{12}$$

As we can see, the indices i and j have disappeared. This is to be expected, as global average pooling means that all values in each channel are multiplied by the same value when calculating the logit of a specific class. We may now substitute this derivative in equation 11:

$$\text{SalAgg}(\boldsymbol{x}; s, A) = \text{sign}(\mu_{id} - \mu_{ood}) \cdot \text{mean} \left(\sum_k \left(\frac{1}{N \cdot M} \sum_i \sum_j \frac{W_{ck}}{N \cdot M} \right) F^k \right). \tag{13}$$

We now perform some simple algebra, exploiting the fact that $\text{mean}(a \cdot \boldsymbol{x}) = a \cdot \text{mean}(\boldsymbol{x})$ and that $\sum_i c \cdot x_i = c \sum_i x_i$:

$$\text{SalAgg}(\boldsymbol{x}) = \text{sign}(\mu_{id} - \mu_{ood}) \cdot \text{mean} \left(\sum_k \left(\frac{1}{N \cdot M} \sum_i \sum_j \frac{W_{ck}}{N \cdot M} \right) F^k \right) \tag{14}$$

$$= \text{sign}(\mu_{id} - \mu_{ood}) \cdot \frac{1}{N \cdot M} \text{mean} \left(\sum_k \left(\sum_i \sum_j \frac{W_{ck}}{N \cdot M} \right) F^k \right) \tag{15}$$

$$= \text{sign}(\mu_{id} - \mu_{ood}) \cdot \frac{1}{N \cdot M} \text{mean} \left(\sum_k \left((N \cdot M) \frac{W_{ck}}{N \cdot M} \right) F^k \right) \tag{16}$$

$$= \text{sign}(\mu_{id} - \mu_{ood}) \cdot \frac{1}{N \cdot M} \text{mean} \left(\sum_k W_{ck} \cdot F^k \right) \tag{17}$$

$$= \left(\text{sign}(\mu_{id} - \mu_{ood}) \cdot \frac{1}{N \cdot M} \right) \cdot \left(\sum_k W_{ck} \cdot \text{mean}(F^k) \right). \tag{18}$$

The first factor above is a constant. It does not depend on $\boldsymbol{x}$, as all values are calculated before inference, or are themselves constants. As such, we can denote this factor as a. We recognize the second factor as $y^c = \text{MLS}(\boldsymbol{x})$ as described in equation 9. We then have

$$\text{SalAgg}(\boldsymbol{x}; s, A) = a \cdot \text{MLS}(\boldsymbol{x}), \tag{19}$$

which was what we wanted to prove.

$\square$

The constant factor has no effect on the OOD detection, as it just means that the thresholds δ will differ by this factor between the two detectors, with all predictions being the same for either method. Thus, theorem 1 states that Saliency Aggregation and MLS OOD detection are functionally equivalent given the conditions described above.

This theorem shows that XAI saliency mapping methods, although they have been developed for an entirely different purpose than OOD detection, also collect information from the network which can be used for OOD detection.

6.2 Further results

6.3 Hyperparameters used

References

[1] Osvaldo Arreche, Tanish R. Guntur, Jack W. Roberts, and Mustafa Abdallah. E-xai: Evaluating black-box explainable ai frameworks for network intrusion detection. *IEEE Access*, 12:23954–23988, 2024.

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.

[3] Shaveta Dargan, Munish Kumar, Maruthi Rohit Ayyagari, and Gulshan Kumar. A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*, 27(4):1071–1092, 09 2020.

[4] Eoin Delaney, Derek Greene, and Mark T. Keane. Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions, 2021.

[5] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild, 2021.

[6] Basim Mahbooba, Mohan Timilsina, Radhya Sahal, and Martin Serrano. Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021(1):6634811, 2021.

[7] Aitor Martinez-Seras, Javier Del Ser, and Pablo Garcia-Bringas. Can post-hoc explanations effectively detect out-of-distribution samples? In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–9, 2022.

[8] Sajid Nazir, Diane M. Dickson, and Muhammad Usman Akram. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in biology and medicine*, 156:106668, 2023.

[9] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[10] 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. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[11] Shashi Shekhar, Vagelis Papalexakis, Jing Gao, Zhe Jiang, and Matteo Riondato. *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2024.

[12] John Sipple and Abdou Youssef. A general-purpose method for applying explainable ai for anomaly detection. In Michelangelo Ceci, Sergio Flesca, Elio Masciari, Giuseppe Manco, and Zbigniew W. Raś, editors, *Foundations of Intelligent Systems*, pages 162–174, Cham, 2022. Springer International Publishing.

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[14] AJ Tallón-Ballesteros and C Chen. Explainable ai: Using shapley value to explain complex anomaly detection ml-based systems. *Machine Learning and Artificial Intelligence: Proceedings of MLIS 2020*, 332:152, 2020.

[15] Erzhena Tcydenova, Tae Woo Kim, Changhoon Lee, and Jong Hyuk Park. Detection of adversarial attacks in ai-based intrusion detection systems using explainable ai. *Human-Centric Comput Inform Sci*, 11, 2021.

[16] Jordan Zheng Ting Sim, Qi Wei Fong, Weimin Huang, and Cher Heng Tan. Machine learning in medicine: what clinicians should know. *Singapore Med J*, 64(2):91–97, May 2021.

[17] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WenXuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[18] Tahmina Zebin, Shahadate Rezvy, and Yuan Luo. An explainable ai-based intrusion detection system for dns over https (doh) attacks. *IEEE Transactions on Information Forensics and Security*, 17:2339–2349, 2022.

[19] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.