OODformer: Out-Of-Distribution Detection Transformer

Rajat Koner[1]
koner@dbs.ifi.lmu.com

Poulami Sinhamahapatra[2]
poulami.sinhamahapatra@iks.fraunhofer.de

Karsten Roscher[2]
karsten.roscher@iks.fraunhofer.de

Stephan Günnemann[3]
stephan.guennemann@tum.de

Volker Tresp[14]
volker.tresp@siemens.com

[1] Ludwig Maximilian University of Munich, Germany

[2] Fraunhofer-Institut für Kognitive Systeme IKS, Germany

[3] Technical University of Munich, Germany

[4] Siemens AG, Munich, Germany

Abstract

A serious problem in image classification is that a trained model might perform well for input data that originates from the same distribution as the data available for model training, but performs much worse for out-of-distribution (OOD) samples. In real-world safety-critical applications, in particular, it is important to be aware if a new data point is OOD. To date, OOD detection is typically addressed using either confidence scores, auto-encoder based reconstruction, or contrastive learning. However, the global image context has not yet been explored to discriminate the non-local objectness between in-distribution and OOD samples. This paper proposes a first-of-its-kind OOD detection architecture named OODformer that leverages the contextualization capabilities of the transformer. Incorporating the transformer as the principal feature extractor allows us to exploit the object concepts and their discriminatory attributes along with their co-occurrence via visual attention. Based on contextualised embedding, we demonstrate OOD detection using both class-conditioned latent space similarity and a network confidence score. Our approach shows improved generalizability across various datasets. We have achieved a new state-of-the-art result on CIFAR-10/-100 and ImageNet30. Code is available at : https://github.com/rajatkoner08/oodformer.
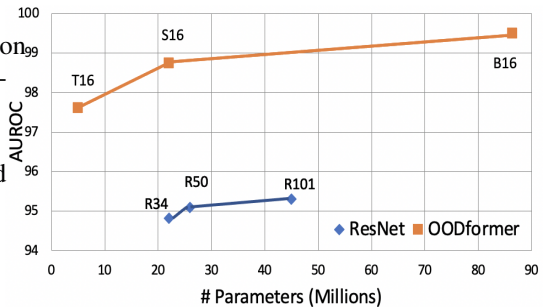
1 Introduction

Deep learning has been shown to give excellent results when the data in an application comes from the same distribution as the data that was available for model training, also called in-distribution (ID) data. Unfortunately, performance might deteriorate drastically for out-of-distribution (OOD) data. The reason why application data might be OOD can be manifold and is often attributed to complex distributional shifts, the appearance of an entirely new concept, or random noise coming from a faulty sensor. As deep learning becomes the core of

many safety-critical applications like autonomous driving, surveillance system, and medical applications, distinguishing ID from OOD data is of paramount importance.

The recent progress in generative modelling and contrastive learning has led to a significant advancement in various OOD detection methods. [25, 35] improved *softmax* distribution for outlier detection. [44, 47, 51] used contrastive learning for OOD detection. The common idea in these works is that in a contrastively trained network, similar objects will have similar embeddings while dissimilar objects will be repelled by the contrastive loss. However, these approaches often require fine-tuning with OOD data or rely on several negative samples and often suffer from inductive biases prevalent in convolution based architectures. Hence, it is difficult to train and deploy them out-of-the-box. This motivates us to go beyond the conventional practice of using negative samples or inductive biases in designing an OOD detector. Along this line, we argue that systematic exploitation of global image context presents a potential alternative to obtain a semantically compact representation. To systematically exploit the global image context, we leverage the multi-hop context accumulation of the vision-transformer (ViT) [13].

Transformer's visual attention significantly outperforms convolutional architectures on various image-classification [13, 48], object-detection [4], relationship-detection [29, 51] and other vision-oriented task [5, 23, 30, 37]. However, transformer's ability to act as a generalized OOD detector remains unexplored so far. As a suitable transformer candidate, we investigate the emerging transformer architecture in image-classification tasks, namely the Vision Transformer (ViT) [13], and its data-efficient variant DeiT [48]. ViT explores the global context of an image and its feature-wise correlations that are extracted from

Figure 1: Comparison of Transformer with traditional ResNet based variant, both trained with ID (CIFAR-10) samples, for distinguishing ID vs OOD (SVHN) samples .

small image patches using visual attention. Intuitively, the difference between an ID and OOD is a small-to-large perturbation in form of the incorrect ordering or corruption (incl. insertion or deletion). Thus, we argue that capturing class/object attribute interaction is important for OOD detection, which we have implemented via a transformer.

To the best of our knowledge, we are the first to propose and investigate the role of global context and feature correlation using vision transformer to generalise data in terms of OOD detection. Our key idea is to train ViT with an ID data set and use the similarity of its final representative embedding for outlier detection. Through the patch-wise attention on the attributes, we aim to reach a discriminatory embedding that is able to distinguish between ID and OOD. First, we follow a formal supervised training of ViT with an in-distribution data set using cross-entropy loss. In the second step, we extract the learned features from the classification head ([*class*] token) and calculate a class conditioned distance metric for OOD detection. Our experiment exhibits that *softmax* confidence prediction from the ViT is more generalisable in terms of OOD detection than the same from the convolutional counterpart. We also observe that the former overcomes multiple shortcomings of the latter, e.g., poor margin probabilities [14, 36]. Figure 1, illustrates our claim that, even with much fewer parameters, transformer-based architecture performs significantly better compared to

traditional convolution-based architecture. This shows the superiority of the transformer in the outlier detection task.

We evaluate the effectiveness of our model on several data sets such as CIFAR-10, CIFAR-100 [32], ImageNet30 [21] with multiple setting (e.g., class conditioned similarity, *softmax* score). Our model outperforms the convolution baseline and other state-of-the-art models in all settings by a large margin. Finally, we have conducted an extensive ablation study to strengthen the understanding of the generalizability of ViT and the impact of the shift in data. In summary, the key contributions of this paper are:

- We model the OOD task as an object-attribute-based compact semantic representation learning. In this context, we are the first to propose a vision-transformer-based OOD detection framework named OODformer and thoroughly investigate its efficiency.

- We probe extensively into the learned representation to verify the proposed method's efficacy using both class-conditioned embedding-based and *softmax* confidence score-based OOD detection.

- We provide an in-depth analysis of the learned attention map, softmax confidence, and embedding properties to intuitively interpret the OOD detection mechanism.

- We have achieved state-of-the-art results for numerous challenging datasets like CIFAR-10, CIFAR-100, and ImageNet30 with a significantly large gain.

2 Related Work

OOD detection approaches can be classified into a number of categories. The first and most intuitive approach is to classify the OOD sample using a confidence score derived from the network. [19] propose max *softmax* probability, consecutively improved by ODIN using temperature scaling [65]. [33] utilised Mahalanobis distance and [25] improved ODIN without using OOD data even further. Other density approximation-based generative modelling [68, 43, 45] also contributed in this approach. A second direction towards OOD detection is to train a generative model for likelihood estimation. [40, 54] focuses on learning representations of training samples, using a bottleneck layer for efficient reconstruction and generalisation of samples. Another advancement on OOD detection is based on self-supervised [6, 7, 8, 34] or supervised [9, 28] contrastive learning. It aims on learning effective discriminatory representations by forcing the network to learn similar representations for similar semantic classes while repelling the others. This property has been utilised by many recent works [44, 47, 51] for OOD detection. The common idea in these works is that in a contrastively trained network, semantically closer objects from ID samples will have similar representation, whereas OOD samples would be far apart in the embedding space.

3 Attention-based OOD detection

In this section, the OOD detection problem is presented followed by a brief background about vision transformer, and the feature similarity-based outlier detection. As OOD labels may not be available for most scenarios, our method primarily relies on similarity score-based detection.

3.1   Problem Decomposition : OOD Detection

Let $x_{in} \in X \subseteq \mathbb{R}^k$ a training sample with k^{th} dimensions and let $y_{in} \in Y = \{1,...,C\}$ be its class label, where C is number of classes. For a given neural feature extractor ($F_{feature}$), learned feature $X_{feat} \subseteq \mathbb{R}^d$ is obtained as $X_{feat} = F_{feature}(X)$. Finally we get posterior class probabilities as $P(y = c|x_{feat})$, where $x_{feat} \in X_{feat}$ and $y \in Y = F_{classifier}(X_{feat})$. $F_{feature}$ and $F_{classifier}$ are two learned functions that map an image from data space to feature space and then derive the posterior probability distribution. In a real world setting, data drawn from X in an application may not follow the same distribution as the training samples (x_{in}). We refer these data as OOD ($x_{ood} \in X, p(x_{ood}) \neq p(x_{in})$). A question now is to what extent OOD data diverges from an ID in their representations? A second question is how reliable the prediction of the posterior probability distributions is for OOD data. To quantify the shift in data from the ID samples, we compute the similarity of the embedding between the samples (x_{in} or x_{out}) and the nearest ID class mean. In an ideal scenario, this representational similarity should be much less for OOD (x_{out}) than ID (x_{in}). Also, its softmax confidence should be significantly lower than that of ID samples, allowing the use of a simple threshold to distinguish between ID and OOD samples.

3.2   Feature extraction: Vision Transformer

In our work we employ the Vision Transformer (ViT) [13] and its data efficient variant DeiT [48]; they use an identical transformer encoder [49] architecture with the same configuration. However, ViT uses ImageNet-21K for pre-training, whereas DeiT uses more robust data augmentation and is only trained with ImageNet. The encoder of ViT takes only a 1D sequence of tokens as input. However, to handle a 2D image with height H and width W, it divides the image into N number of small 2D patches and flattened into a 1D sequence $X \in \mathbb{R}^{N \times (P^2 \cdot C')}$, where C' is the number of channels, (P,P) is the resolution of each patch and N is the number of sequences obtained as $N = HW/P^2$. A [class] token (x_{cls}) similar to BERT [12] prepended (first position in sequence) to the sequence of patch embeddings, as expressed in Eq.1,

$$z_x = [x_{cls}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos}, E \in \mathbb{R}^{(P^2 \cdot C') \times d}, E_{pos} \in \mathbb{R}^{(N+1) \times d} \qquad (1)$$

where x_{cls}, E_{pos} are learnable embedding of dth dimension. E is the linear projection layer of d dimension for a patch embedding (e.g., $x_p^1 E$). The [class] token embedding ($x_{feat} \in \mathbb{R}^d$) from the final layer of the encoder is used for classification. This classification token serves as a representative feature from all patches accumulated using global attention.

At the heart of each encoder layer (for details, see supplementary) lies a multi-head self-attention (MSA) and a multi-layer perception (MLP) block. The MSA layers provide global attention to each image patch; thus, it has less inductive bias for local features or neighbourhood patterns than a CNN. One image patch or a combination thereof represents either a semantic or spatial attribute of an object. Therefore, we hypothesise that cumulative global attention on these patches would be helpful to encode discriminatory object features. Positional embedding (E_{pos}) of the encoder is also beneficial for learning the relative position of a feature or a patch. Subsequently, a local MLP block makes those image features translation invariant. This combination of MSA and MLP layers in an encoder jointly encodes the attributes' importance, associated correlation, and co-occurrence. In the end, [class] token being a representative of an image x, consolidate multiple attributes and their

related features via the global context, which is helpful to define and classify a specific object category.

The [*class*] token from the final layer is used for OOD detection in two ways; first, it is passed to $F_{classifier}(x_{feat})$ for *softmax* confidence score, and second it is used for latent space distance calculation as described in next section.

3.3   OOD detection

For an OOD sample, one or more attributes of the object are assumed to be different (e.g. corrupted or new attributes). Hence, an OOD sample should lie in a distribution that has significant distributional shift from training samples. We assume these shifts in data would be captured by the representational features extracted from the ViT. We quantify this variation of features or attributes compared to ID samples using 1) a distance metric on the latent space embedding and 2) *softmax* confidence score.

Algorithm 1: OOD detection using distance metric

Input: training samples $\{x_{in}^i\}$ and training labels $\{y_{in}^i\}$ for $i = 1 : n$, test sample x_{test}
Output: x_{test} is a outlier or not?
for *each class:* $c \leftarrow 1$ **to** C **do**
$\quad \{x_{feat}^i\}_{i=1:n_c} = F_{feature}(\{x_{in}^i | y_{in}^i = c\}_{i=1:n_c})$, where $n_c = |\{x_{in}^i | y_{in}^i = c\}|$
$\quad$ compute mean : $\mu_c = \frac{1}{n_c} \sum_{i=1}^{n_c} (x_{feat}^i)$
$\quad$ compute covarience : $\Sigma_c = \frac{1}{n_c-1} \sum_{i=1}^{n_c} (x_{feat}^i - \mu_c)(x_{feat}^i - \mu_c)^\top$
end
$x_{feat}^{test} = F_{feature}(x_{test})$
$distance = \min_c(\sum_{c=1}^{C} (x_{feat}^{test} - \mu_c)\Sigma_c^{-1}(x_{feat}^{test} - \mu_c)^\top)$
$conf = \max_c(softmax(F_{classifier}(x_{feat}^{test}))$
if $(distance > t_{distance})$ OR $(conf < t_{conf})$ **then**
$\quad$ | x_{test} is an outlier
else
$\quad$ | x_{test} is not an outlier
end

Distance in Latent Space: Multiple attributes of an object can be found in different spatial locations of an image. In vision transformer, image patches are ideal candidates for representing each of the individual attributes. Global information contextualisation from these attributes plays a crucial role in the classification of an object (see Sec. 4.2). Supervised learning (e.g.,cross-entropy) should benefit accumulation of object-specific semantic cues through [*class*] token for such global attributes and their context. This incentivises implicit clustering of object classes that have similar attributes and features, which are favourable for generalisation. To take advantage of attribute similarities, we compute class-wise distance over the activation of [*class*] token. First, we compute the mean (μ_c) of all class categories present in training samples. Second, for a test sample, we compute the distance between its embedding from the mean embedding of each class. Finally, the test sample is classified as OOD if its distance is more than the threshold ($t_{distance}$) to its nearest class. We have used the Mahalanobis distance metric for our experiment. The output from a representative [*class*] token is normalised with the transformer default layer normalisation [1] for every token. It makes the embedding distribution normal, thus Mahalanobis distance could utilise the normally distributed mean and co-variance unlike Euclidean distance which only uses mean. Mahalanobis distance

ID	CIAFR-10		CIFAR-100		IM-30	
(Out-of-Distribution)	SVHN	CIFAR-100	SVHN	CIFAR-10	CUB	Dogs
Baseline OOD[19]	95.9	89.8	78.9	78.0	-	-
ODIN[65]	96.4	89.6	60.9	77.9	-	-
Mahalanobis[33]	99.4	90.5	94.5	55.3	-	-
Residual Flows[52]	99.1	89.4	97.5	77.1	-	-
Outlier exposure[20]	98.4	93.3	86.9	75.7	-	-
Rotation pred[2]	98.9	90.9	-	-	-	-
Contrastive + Supervised[57]	99.5	92.9	95.6	78.3	86.3	95.6
CSI[1]	97.9	92.2	-	-	94.6	98.3
SSD+[13]	**99.9**	93.4	98.2	78.3	-	-
OODformer(Ours)	99.5	**98.6**	**98.3**	**96.1**	**99.7**	**99.9**

Table 1: Comparison of OODformer with state-of-the-art detectors trained with supervised loss.

from a sample to distribution of mean μ_c and covariance Σ_c can be defined as

$$D_c(x_{feat}^{test}) = (x_{feat}^{test} - \mu_c)\Sigma_c^{-1}(x_{feat}^{test} - \mu_c)^\top \qquad (2)$$

We have also examined Cosine and Euclidean distances as shown in Figure 3a. The complete algorithm to compute an outlier has been given to Algorithm 1.

Softmax confidence score: We get the final class probabilities from $F_{classifier}$ through *softmax*. Softmax based posterior probability has been reported in earlier studies to give erroneous high confidence score when exposed to outliers [53]. Prior works used this posterior probability for OOD detection task, using either a binary classifier [19] or temperature scaling [35]. We argue that for a good attribute cluster representation as in the case of a transformer no extra module for OOD detection is needed (see Table. 3). Thus in this work, we use only simple numerical thresholds (t_{conf}) from ID samples for the detection of outliers without using additional OOD data for fine-tuning.

4 Experiments

In this section, we evaluate the performance of our approach and compare it to state-of-the-art OOD detection methods. In Sec. 4.1, we report our results on labeled multi-class OOD detection and one class anomaly detection. It also contains a comparison with state-of-the-art methods and a ResNet [18] baseline. In Sec.4.2, we examine the influence of architectural variance and distance metric on the OODformer in the context of the OOD detection.

Area under PR and ROC : A single threshold score may not scale across all the data sets. In order to homogenise the performance across multiple data sets, a range of thresholds should be considered. Thus, we report the area under precision-recall (PR) and ROC curve (AUROC) for both the latent space embedding and the *softmax* score.

Setup: We use ViT-B-16 [13] as the base model of our all experiments and ablation study. Variants of ViT (Base-16, Large-16 with embedding size 768 and 1024 along with patch size 16) and DeiT (Tiny-16, Small-16 with corresponding embedding size of 192 and 384) used a similar architecture but differ on embedding and the number of attention head (full details provided in appendix). As transformers have inherent lesser inductive biases compared to CNNs, it can exploit similarities more efficiently when pre-trained with large datasets as prevalent with well-known architecture like BERT [12] or GPT2 [41]. Thus, pre-training with large datasets is a necessary precondition for most transformer based architectures and their use cases. Hence, all the models we use are pre-trained on ImageNet [11] as recommended in [13]. Supervised cross-entropy loss is used for training along SGD as

Methods	Airplane	Automobile	Bird	Cat	Dear	Dog	Frog	Horse	Ship	Truck	Average
GT[□]	74.7	95.7	78.1	72.4	87.8	87.8	83.4	95.5	93.3	91.3	86.0
Inv-AE[□]	78.5	89.8	86.1	77.4	90.5	84.5	89.2	92.9	92	85.5	86.6
Goad[□]	77.2	96.7	83.3	77.7	87.8	87.8	90	96.1	93.8	92	88.2
CSI[□]	89.9	**99.9**	93.1	86.4	93.9	**93.2**	95.1	98.7	97.9	95.5	94.3
SSD[□]	82.7	98.5	84.2	84.5	84.8	90.9	91.7	95.2	92.9	94.4	90.0
OODformer	**92.3**	99.4	**95.6**	**93.1**	**94.1**	92.9	**96.2**	**99.1**	**98.6**	**95.8**	**95.7**

Table 2: Comparison of OODformer with other methods on one-class OOD from CIFAR-10.

an optimizer and we followed the same data augmentation strategy like [28]. ResNet-50 [18] is our default convolution-based baseline architecture used across all result and ablation studies. Mahalanobis distance-based score on representational embedding is used as the default score for the AUROC calculation.

Datasets: We trained our networks on the following in-distribution data sets: CIFAR-10/-100 [62] and ImageNet-30 (IM-30) [21]. AS OOD data for CIFAR-10/-100 we have chosen resized Imagenet (ImageNet_r) [35] and LSUN [21] and SVHN [39] as specified in [47]. In case of IM-30 we flow the same setup of [47], and used Places-365 [62], DTD [10], Dogs [27], Food-101 [3], Caltech-256 [17], and CUB-200 [50]. Details about these data set are given in the appendix.

4.1 Results

Comparison with State-of-the-Art: Table 1, exhibits that even with a simple cross-entropy loss, OODformers achieve new state-of-the-art results on all data sets. Most importantly, OODformer supersedes its predecessor on the complex near OOD data sets, which strongly affirms substantial improvements in generalisation. In particular, detection of CIFAR-100 as OOD samples with a network trained in CIFAR-10 (ID) or vice-versa is the most challenging task since they share a significant amount of common attributes in similar classes. For example, a 'truck' is an ID sample in CIFAR-10, whereas a 'pickup-truck' from CIFAR-100 is an OOD sample, despite being semantically closer and having many similar attributes. These attributes vs class conflict leads to a significant drop in AUROC value for previous methods, especially when CIFAR-100 is ID and CIFAR-10 is OOD. We gained a notable 5.2% and 17.8% gain in AUROC when trained on CIFAR-10 and CIFAR-100, respectively. This can also be visualised in Figure.3b. These results show that the OODformer has superior generalizability of the representational embedding due to the global contextualization of object attributes from single or multiple patches, even with complex near OOD classification.

One class anomaly detection: Similar to OOD detection, anomaly detection is concerned with certain data types or specifically one class. Apart from that, the specific class presence of any class will be considered an anomaly. In this setting, we consider one of the classes from CIFAR-10 as an in-distribution, and the rest of the classes are anomalies or outliers. We train our network with one class (ID) and the rest of the nine classes become OOD and we repeat these experiments for every class similar to [47]. Table 2, shows our OODformer outperforms all existing methods and achieves new state-of-the-art results.

Comparison with ResNet: Table 3, shows the performance of OODformer in comparison with ResNet baseline. To show the generalisability and scalability, we trained our network with three in-distribution datasets and tested on nine different OOD datasets. Here, we use *softmax* confidence score in addition to our default embedding distance-based score for OOD detection. As discussed in Sec 1, *softmax* suffers from poor decision margin and lack of generalisation when used for ODD detection with CNNs. This can be addressed using our proposed OODformer. Table 3, shows that ours *softmax* based OOD detection significantly

ID	OOD	Emb-Distance		Softmax	
		ResNet	OODformer	ResNet	OODformer
CIFAR-10	CIFAR-100	87.8	**98.6**	87.4	**97.7**
	Imagenet_r	91.4	**98.8**	90.9	**96.0**
	LSUN	93.4	**99.2**	92.4	**97.6**
CIFAR-100	CIFAR-10	73.7	**96.1**	69.3	**88.9**
	Imagenet_r	79.9	**92.5**	72.4	**86.1**
	LSUN	79.0	**94.6**	72.9	**86.2**
IM-30	Places-365	82.9	**99.2**	91.8	**98.2**
	DTD	97.8	**99.3**	90.9	**98.2**
	Dogs	75.29	**99.9**	92.3	**99.0**
	Food-101	73.44	**99.2**	83.2	**97.2**
	Caltech256	86.37	**98.0**	91.4	**96.8**
	CUB-200	87.22	**99.7**	91.9	**99.4**

Table 3: Comparison of OODformer with ResNet baseline

and consistently outperforms ResNet and achieves an improvement as high as 19.6%. Our default AUROC score using embedding similarity also outperforms our baseline by a large margin. This result proves the objectness properties and exploitation of its related attributes through global attention play the most crucial role for outlier detection. It bolsters our hypothesis that, for outlier detection, a transformer can serve as the de-facto feature extractor without any bells and whistles.

Figure 2: Representative example of an ID sample of Truck (above), and an OOD sample of Mountain (below). Among the four attention maps of each image, top left represents the class token, the remaining correspond to the image position marked with green square block. The right bar plot shows the top 3 similar class embedding.

4.2 Dissection of OODformer

In this section, we extensively analyse the attention and impact of different settings via an ablation study. We have used CIFAR-10 as ID and ViT-B-16 as default model architecture for our experiments.

Global Context and Self-Attention: To understand how vision transformers discriminate between an ID and OOD, we analyse its self-attention and the embedding space. Figure 2 depicts examples of both ID and OOD samples with their attention maps collected from the

last layer. Self-attention accumulates information from all the related patches that define an object. For an ID sample ('truck'), [*class*] token attention focuses mostly on the object of interest, while other selected patches put their attention based on properties similar to them like colour or texture. However, in the case of an OOD sample, the object 'mountain' is unknown to the network, the [*class*] token attention mostly focused on the sky and water, while we draw similar observations for other patches as before. This misplaced attention and absence of known object attributes leads to a lower similarity score and predicts wrong classes just because of background similarity.

Figure 2, also shows objectness and its related attributes exploration in a hierarchical way is crucial for OOD distance score. This can also be inferred from Figure 3b, where we notice that the transformer not only reduces intra-class distance for ID samples, it also increases the distance of OOD samples from the ID class mean.

Model	Acc.	CIFAR100	Imagenet_r	LSUN
Resnet-34	95.6	87.2	89.7	91.4
Resnet-50	95.4	87.5	90.0	91.7
Resnet-101	95.7	87.8	91.4	93.4
Resnet-34(PT)	97.0	89.5	93.8	96.0
Resnet-50(PT)	97.1	89.9	94.1	96.1
Resnet-101(PT)	98.0	90.1	94.6	96.0
DeiT-T-16	95.4	94.4	95.2	97.3
DeiT-S-16	97.6	96.6	96.3	98.4
ViT-B-16	98.6	98.6	98.8	99.2
ViT-L-16	98.8	99.1	99.2	99.4

Table 4: Comparison of various ResNet baseline (both not pre-trained and pre-trained) and vision transformer architecture. Here, (PT) defines the ResNet has been pre-trained on large scale ImageNet dataset.

ResNets vs ViTs: This section examines the effect of model complexity or expressiveness of a model for OOD or outlier detection. We have conducted experiments on multiple variants of ViT in comparison with variants of ResNet baseline (top rows) as shown in Table 4. As all the variants of ViT and DeiT are pre-trained on ImageNet, for the sake of the comparison, we also evaluate ResNet variants that are pre-trained on ImageNet. As all the pre-trained ResNet models are trained with larger image size (224x224 pixel for ImageNet), thus for efficient utilisation of pre-training an up-scaling is needed. We fine-tune pre-trained ResNet on training in-distribution dataset using an up-scaling of 7x (from 32 to 224).

- **ResNet Baseline :** Table 4, shows that Deit-T-16 which is the lightest among all variants of ViT and much smaller in size than all ResNet baseline variants performs substantially better on OOD detection(94.4 compared to 87.8 of R-101 for CIFAR-100). Furthermore, despite being similar in accuracy DeiT-T-16 is significantly superior with respect to OOD detection like all other ViT variants.

- **ResNet Baseline with pre-trained on ImageNet :** Transformer based architecture relies on training with large-scale data in order to achieve superior performance. ViT also requires large-scale pre-training on ImageNet for an efficient performance similar to a few well-known transformer based architectures like BERT[12], GPT2[41]. Moreover, a few recent work [42], suggests pre-training of convolution offer better performance for outlier detection. Thus, to demonstrate the true performance of OODformer, we compare it with pre-trained convolutional models like ResNets (middle rows). Table 4, suggests fine-tuning with upscale images offers good accuracy, but limited gain over OOD detection. Pre-trained ResNet follows a similar trend as the baseline models like increasing model capacity has a marginal effect on OOD detection in contrast to ViT variants. One of the key observations we found, apart from pre-training, up-scaling of images helped to attain higher performance on accuracy and OOD detection score.

As per the above discussion based on Table 4, we can clearly observe irrespective of their accuracy, image size, pre-training features extracted from ResNet deliver sub-optimal performance compare to all variants of ViT. Furthermore, with an increase in model complexity, the performance of ResNet reaches a plateau, whereas the performance of ViTs consistently improves with an increase in expressive power. This distinctly demonstrates the importance of objectness context, the role of an object attribute, and their mutual correlation (e.g, spatial) as hypothesised in the Sec 1. Such observations contribute to our belief, that the transformer architecture is better suited for OOD detection tasks than classical CNNs primarily for less inductive bias, object attribute accumulation using attention.

(a)   (b)
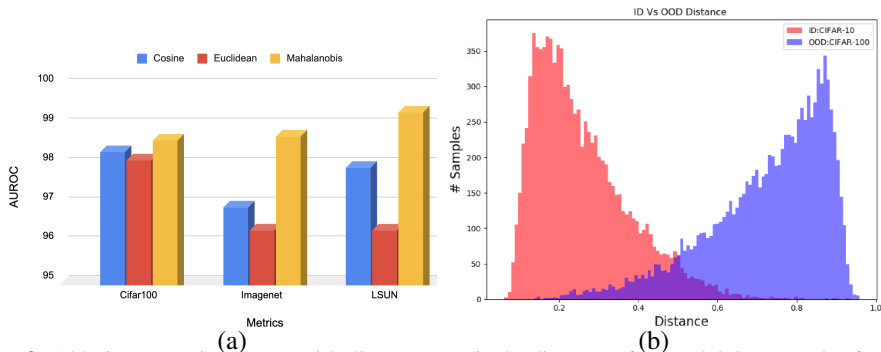
Figure 3: Ablation Experiment : a) with distance metric, b) distance of ID and OOD samples from its corresponding class mean.

Analyzing Distance Metrics: We examine the influence of various distance metrics like Cosine and Euclidean in Figure.3a. Euclidean utilises only the mean of the ID distribution thus it could be an intuitive reason for its underperformance. Furthermore, [44] shows that, in embedding space, Euclidean distance is dominated by higher eigenvalues that reduce the sensitivity towards outliers. Both Cosine and Mahalanobis perform very similarly. However, the slightly better performance of Mahalanobis, compared to Cosine, could be attributed to its less dependency on a higher norm between two features and its utilisation of both mean and covariance. Figure 3a, shows OODformer is quite robust ($\pm2\%$ variance) to various distances.

5   Conclusion

In this paper, we made an early attempt utilising a vision transformer namely the OODformer for OOD detection. Unlike prior approaches, which rely upon custom loss or negative sample mining, we alternatively formulate OOD detection as object-specific attribute accumulation problem using multi-hop context aggregation by a transformer. This simple approach is not only scalable and robust but also outperforms all existing methods by a large margin. OODformer is also more suitable for being deployed in the wild for safety-critical applications due to its simplicity and increased interpretability compare to other methods. Building on our work, emerging methods in self-supervision, pre-training, and contrastive learning will be of future interest to investigate in combination with OODformer.

Acknowledgement

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, November 2020.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv:2006.10029 [cs, stat]*, June 2020.

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297 [cs]*, March 2020. Comment: Tech report, 2 pages + references.

[9] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. *arXiv:2007.00224 [cs, stat]*, July 2020.

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

[11] 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*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Gamaleldin F Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *arXiv preprint arXiv:1803.05598*, 2018.

[15] Izhak Golan and Ran El-Yaniv. Deep Anomaly Detection Using Geometric Transformations. *arXiv:1805.10917 [cs, stat]*, November 2018.

[16] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[17] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[18] 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*, pages 770–778, 2016.

[19] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv:1610.02136 [cs]*, October 2018.

[20] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure. *arXiv:1812.04606 [cs, stat]*, January 2019. Comment: ICLR 2019; PyTorch code available at https://github.com/hendrycks/outlier-exposure.

[21] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

[22] 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. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 15663–15674. Curran Associates, Inc., October 2019. Comment: NeurIPS 2019; code and data available at https://github.com/hendrycks/ss-ood.

[23] Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*, 2020.

[24] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *arXiv preprint arXiv:1705.08741*, 2017.

[25] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.

[26] Chaoqing Huang, Jinkun Cao, Fei Ye, Maosen Li, Ya Zhang, and Cewu Lu. Inverse-transform autoencoder for anomaly detection. 2019.

[27] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.

[28] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. *arXiv:2004.11362 [cs, stat]*, April 2020.

[29] Rajat Koner, Poulami Sinhamahapatra, and Volker Tresp. Relation transformer network. *arXiv preprint arXiv:2004.06193*, 2020.

[30] Rajat Koner, Hang Li, Marcel Hildebrandt, Deepan Das, Volker Tresp, and Stephan Günnemann. Graphhopper: Multi-hop scene graph reasoning for visual question answering, 2021.

[31] Rajat Koner, Poulami Sinhamahapatra, and Volker Tresp. Scenes and surroundings: Scene graph generation using relation transformer. *arXiv preprint arXiv:2107.05448*, 2021.

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

[33] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *arXiv:1807.03888 [cs, stat]*, October 2018. Comment: Accepted in NIPS 2018.

[34] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical Contrastive Learning of Unsupervised Representations. *arXiv:2005.04966 [cs]*, July 2020.

[35] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[36] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016.

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[38] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? *arXiv:1810.09136 [cs, stat]*, February 2019. Comment: ICLR 2019.

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

[40] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative Probabilistic Novelty Detection with Adversarial Autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6822–6833. Curran Associates, Inc., 2018.

[41] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[42] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021.

[43] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14707–14718. Curran Associates, Inc., 2019. Comment: Accepted to NeurIPS 2019.

[44] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=v5gjXpMR8J.

[45] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, Jordi Luque, Olga Slizovskaia, Jos&#xE9, F. N&#xFA, &#xF1, ez, and Jordi Luque. Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models. In *Proc. ICLR 2020*, September 2019.

[46] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[47] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. *arXiv:2007.08176 [cs, stat]*, July 2020. Comment: 23 pages; Code is available at https://github.com/alinlab/CSI.

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[50] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[51] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive Training for Improved Out-of-Distribution Detection. *arXiv:2007.05566 [cs, stat]*, July 2020.

[52] Bolei Zhou, Agata 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(6):1452–1464, 2017.

[53] Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13994–14003, 2020.

[54] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.