A ViT Architecture

Vision Transformer [13] uses transformer encoder [49] for patch based image classification. The core of ViT relies on multi-head self-attention (MSA) and multi-layer perception (MLP) for processing sequence of image patches.

Multi-head Self-Attention: The attention mechanism is formulated as a trainable weighted sum based approach. One can define self-attention as

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where Q,K,V are a set of learnable query, key and value and d is the embedding dimension. A query vector $q \in \mathbb{R}^d$ is multiplied with key $r \in \mathbb{R}^d$ using inner product obtained from the sequence of tokens as specified in Eq. 1. The important features from the query token is dynamically learned by taking a *softmax* on the product of query and key vectors. It is then multiplied with the value vector v that incorporates features from other tokens based on their learned importance.

Multi-Layer Perception: The transformer encoder uses a Feed-Forward Network (FFN) on top of each MSA layer. An FFN layer consists with two linear layer separated with GleU activation. The FFN processes the feature from the MSA block with a residual connection and normalizes with layer normalization [1]. Each of the FFN layer is local for every patch unlike the MSA (MSA act as a global layer), hence the FFN makes the encoder image translation invariant.

B Implementation Details

Our backbone ViT [13] and DeiT [48] are pretrained on ImageNet, and fine-tuned in an in-distribution dataset with SGD optimizer, a batch size of 256 and image size of 224×224. We use a learning rate of 0.01 with Cyclic learning rate scheduler [46], weight decay=0.0005 and train for 50 epochs. We follow the data augmentation scheme same as [28].

B.1 Model Detail

We use multiple variants of ViT and DeiT, primarily because DeiT offers lighter model, whereas ViT mainly focusses on havier model. The idea being an enhanced outlier detection performance with a lighter variant will bolster our assumption that exploring an object's attributes and their correlation using global attention plays a crucial role in OOD detection. In comparison, a heavier variant will offer increased model capacity to improve the performance of the OODformer. Table. 4 exhibits the performance of OODformer with multiple backbone variants in support of our hypothesis. Specially the significant performance gain with the smallest variant of DeiT (T-16) bolster our claim. Table 5 shows the variation of their parameter, number of layers, hidden or embedding size, MLP size, number of attention head.

Model	Prams	#Layers	Hidden Size	MLP Size	#Heads
DeiT-T-16	5	12	192	768	3
DeiT-S-16	22	12	384	1536	6
ViT-B-16	86.5	12	768	3072	12
ViT-L-16	307	24	1024	4096	16

Table 5: DeiT and ViT model architecture.

B.2 Dataset Details

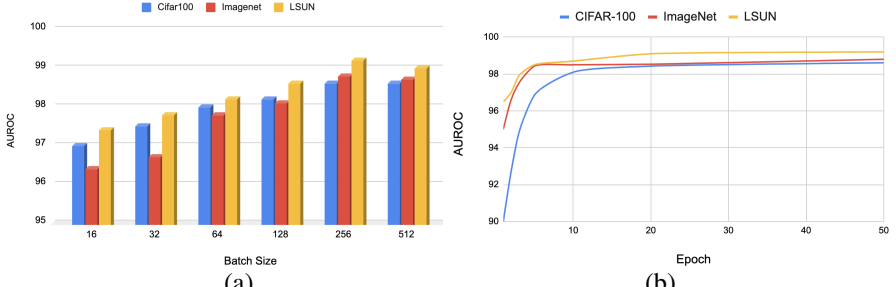
Among the in-distribution dataset, CIFAR-10/-100 [32] consists of 50K training and 10K test images with corresponding 10 and 100 classes. The CIFAR-100 dataset also contains twenty superclasses for all the hundred classes present in it. Even though CIFAR-10 and CIFAR-100 has no overlap for any class, some classes share similar attributes or concepts (e.g., 'truck' and 'pickup-truck') as discussed in Section.4.2. As a result of this close semantic similarity these two datasets poses the most challenging near OOD problem and the performance of OODformer in this context has shown in Table 1. Another in-distribution dataset, ImageNet-30 [21], is a subset of ImageNet[11] with 30 classes that contains 39K training and 3K test images.

Out-Of-Distribution dataset used for CIFAR-10/-100 are as follows : Street View Housing Number or SHVN [39] contains around 26K test images of ten digits, LSUN [21] consists of 10K test images of ten various scenes, ImageNet-resize [21] is also a subset of ImageNet with 10K images and two hundred classes. For multi-class ImageNet-30, we follow the same OOD datsets as specified in [47], they are : Places-365 [52], Describable Texture Dataset [10], Food-101 [3], Caltech-256 [17] and CUB-200 [50].

C Ablation and Interpretation

In addition to the analysis provided in Sec. 4.2, we ablate OODformer on various batch sizes, epochs and analyze the cluster in embedding space.

Figure. 4a, demonstrates large batch size helps in OOD detection, though we observe it



(a) (b)

Figure 4: Ablation Experiment : a) with various batch size, b) improvement of AUROC over the epochs.

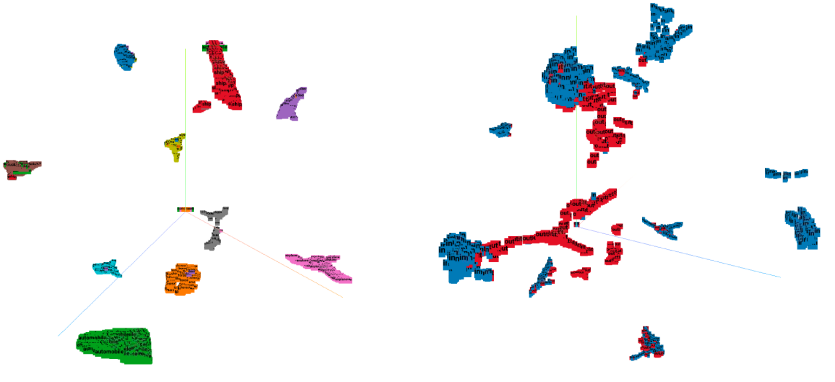
doesn't significantly impact accuracy on the in-distribution test set. An intuitive reason could be large batch size improves generalization [24], which enables the network to generalize object-specific properties that are helpful for outlier identification. Despite this gain, we observe OODformer remain relatively stable across all the batch sizes with OOD detection accuracy $\pm 1.5\%$. However, the gain in AUROC gradually becomes stagnant with an increase

of batch size suggest further scope of tuning learning rate is required using a linear scaling [16].

Figure. 4b, shows an increase of outlier detection accuracy with the number of epochs. One of the important observation is easier OOD dataset (e.g., LSUN, ImageNet) are distinguishable with fewer epochs whereas difficult OOD dataset like CIFAR-100 takes more time. In comparison with the state-of-the-art i.e. convolution [27] or contrastive [44], our proposed OODformer converges significantly faster, even with much less batch size. This promising result shows the efficacy of the OODfromer in a real-world scenario and directs to further scope of research of transformer in outlier detection.

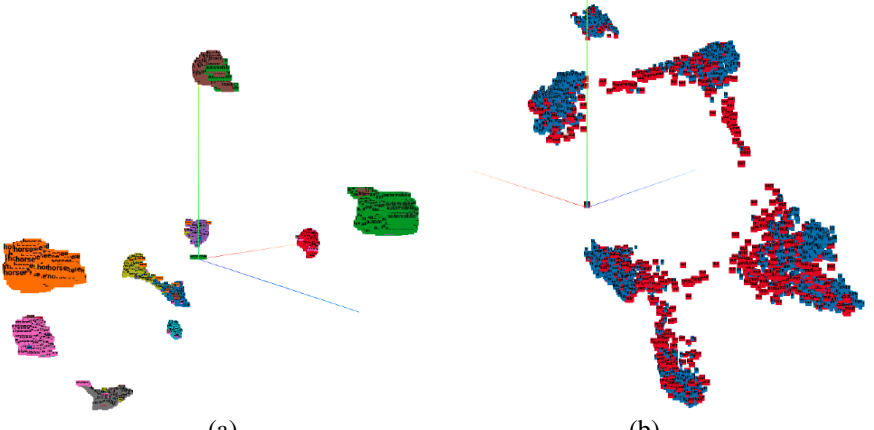
Manifold Analysis : Fig. 5a and 6a, shows both for OODformer and ResNet-50 baseline, all the classes in CIFAR-10 have formed a compact cluster as shown by their corresponding UMAP. As discussed in Sec. 3, we can observe supervise loss helps in the formation of the compact clustering, which can be exploited for class conditioned OOD detection provided there is a separability between ID and OOD data. Figure. 5b, shows that for OODformer, OOD samples in the embedding space lie far from any cluster center of an in-distribution sample due to its large distributional shift or lack of object-specific attributes. This variation of distance between an ID and OOD sample is effectively utilized by our distance metric. However, Fig. 6, suggests that despite being able to form a distinctive cluster for ID samples, our ResNet baseline has failed to maintain a clear separation between an ID and OOD samples.

This UMAP analysis supports our earlier assumption on results of Table 4, in spite of lower or similar accuracy for classification of ID samples, features extracted from transformer have more distinctive separable features for OOD detection.



(a) (b)

Figure 5: OODformer UMAP analysis: a) ID (CIFAR-10) samples and their corresponding cluster, b) ID (blue) and OOD (red) samples shown in UMAP clustering.



(a) (b)

Figure 6: ResNet-50 baseline UMAP analysis: a) ID (CIFAR-10) samples and their corresponding cluster, b) ID (blue) and OOD (red) samples shown in UMAP clustering.