ViM: Out-Of-Distribution with Virtual-logit Matching

Haoqi Wang†* Zhizhong Li†* Litong Feng† Wayne Zhang‡‡

1SenseTime Research 2Qing Yuan Research Institute, Shanghai Jiao Tong University

{wanghaoqi,lizz,fenglitong,wayne.zhang}@sensetime.com

Abstract

Most of the existing Out-Of-Distribution (OOD) detection algorithms depend on single input source: the feature, the logit, or the softmax probability. However, the immense diversity of the OOD examples makes such methods fragile. There are OOD samples that are easy to identify in the feature space while hard to distinguish in the logit space and vice versa. Motivated by this observation, we propose a novel OOD scoring method named Virtual-logit Matching (ViM), which combines the class-agnostic score from feature space and the In-Distribution (ID) class-dependent logits. Specifically, an additional logit representing the virtual OOD class is generated from the residual of the feature against the principal space, and then matched with the original logits by a constant scaling. The probability of this virtual logit after softmax is the indicator of OOD-ness.

1. Introduction

Considering most deep image classification models are trained in the closed-world setting, the out-of-distribution (OOD) issue arises and deteriorates customer experience when the models are deployed in production, facing inputs coming from the open world [9]. For instance, a model may wrongly but confidently classify an image of crab coming from the open world [9]. For instance, a model may wrongly but confidently classify an image of crab into the clapping class, even though no crab-related concepts appear in the training set. OOD detection is to decide whether an input belongs to the training distribution. OOD detection complements classification and finds its application in fields such as autonomous driving [19], medical analysis [30] and industrial inspection [1]. A comprehensive review of OOD and related topics including open set recognition, novelty detection and anomaly detection can be found in [38].

The core of an OOD detector is a scoring function \( \phi \) that maps an input feature \( \mathbf{x} \) to a scalar in \( \mathbb{R} \), indicating to what extent the sample is likely to be OOD. In testing, a threshold \( \tau \) is decided, ensuring that the validation set retains at least a given true-positive rate (TPR), \( e.g. \) the typical value of 0.95. The input example is regarded as OOD if \( \phi(\mathbf{x}) > \tau \) and as ID (i.e., in-distribution) otherwise. In cases where a score indicating the ID-ness is convenient, we can mentally use the negative of OOD score as the ID score.

Researchers have designed quite a few scoring functions by seeking properties that are naturally held by ID examples and easily violated by OOD examples, or vice versa. Scores are mainly derived from three sources: (1) the probability, such as the maximum softmax probabilities [13], the minimum KL-divergence between the softmax and the mean class-conditional distributions [12]; (2) the logit, such as the maximum logits [12], the logsumexp function over log-

---

† These two authors contribute equally to the work.

‡ Corresponding author: Wayne Zhang.
its [25]; and (3) the feature, such as the norm of the residual between feature and the pre-image of its low-dimensional embedding [27], the minimum Mahalanobis distance between the feature and the class centroids [23], etc. In these methods, OOD scores can be directly computed from existing models without re-training, making the deployment effortless. However, as illustrated in Fig. 1, their performances are limited by the singleness of their information source: using features exclusively disregards the classification weights with class-dependent information; using the logit or the softmax solely misses feature variations in the null space [3], which carries class-agnostic information; and the softmax further discards the norm of logits. To cope with the immense diversity that manifests in OOD samples, we ask the question, is it helpful to design an OOD score that utilizes multiple sources?

Built upon the success of prior arts, we design a novel scoring function termed the Virtual-logit Matching (ViM) score, which is the softmax score of a constructed virtual OOD class whose logit is jointly determined by the feature and the existing logits. To be specific, the scoring function first extracts the residual of the feature against a principal subspace, and then converts it to a valid logit by matching its mean over training samples to the average maximum logit. Finally, the softmax probability of the devised OOD class is the OOD score. From the construction of ViM, we can see intuitively that the smaller the original logits and the greater the residual, the more likely it is to be OOD.

Different from the aforementioned methods, another line of works tailors the features learned by the network to better identify ID and OOD by imposing dedicated regularization losses [5, 16, 18, 40] or by exposing generated or real collected OOD samples [22, 37]. As they all require the re-training of the network, we briefly mention them here and will not delve into the details.

Recently, OOD detection in large-scale semantic space has attracted increasing attention [12, 15, 18, 29], advancing OOD detection methods toward real-world applications. However, the current shortage of clean and realistic OOD datasets for large-scale ID datasets becomes an impediment to the field. Previous OOD datasets were curated from public datasets which were collected with a predefined tag list, such as ImageNet-O pre-trained with the latest ViT-B16 [8], RepVGG [7], DeiT [33] and Swin Transformer [26]. From the results on four OOD datasets, including OpenImage-O, ImageNet-O, Texture, and iNaturalist, we found that model selection affected the performance of many baseline methods, while our method performs stably well. Specially, our method achieved an average AUROC of 90.91% using the BiT model, which greatly surpasses the best baseline whose average AUROC is 86.62%.

Our contributions are threefold. (1) We proposed a novel OOD detection method ViM, that works well for a large range of models and datasets, owing to the effective fusion of information from both features and logits. The method is lightweight and fast, requiring neither extra OOD data nor re-training. (2) We conducted comprehensive experiments and ablation studies on the ImageNet-1K dataset, including CNNs and vision transformers. (3) We curated a new OOD dataset for ImageNet-1K called OpenImage-O, which is very diverse and contains complex scenes. We believe it will facilitate research on large-scale OOD detection.

2. Related Work

OOD/ID Score Design Hendrycks et al. [13] presented a baseline method using the maximum predicted softmax probability (MSP) as the ID score. ODIN [24] enhances MSP by perturbing the inputs and rescaling the logits. Hendrycks et al. [12] also experimented with the MaxLogit and the KL matching method on the ImageNet dataset. The energy score [25] computes the $\log\sum_{x}\exp$ on logits, and ReAct [32] strengthens the energy score by feature clipping. In [27] the norm of the difference between the feature and the pre-image of its low-dimensional manifold embedding is used. Lee et al. [23] computes the minimum Mahalanobis distance between the feature and the class-wise centroids. NuSA [3] uses the ratio of the norm of feature projected onto the column space of the classification weight matrix to the original norm as the ID score. The gradients are also used as evidence for ID and OOD distinction in [17]. For methods using logits/probabilities, feature variations on the null space of the weight matrix are completely ignored; while for methods that operate on the features space, the class-dependent information on weight matrix is dropped.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image Distribution</th>
<th>#Image</th>
<th>Labeling Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenImage-O</td>
<td>natural class statistics</td>
<td>17,632</td>
<td>image-level manual</td>
</tr>
<tr>
<td>iNaturalist [18, 34]</td>
<td>predefined tag list</td>
<td>10,000</td>
<td>tag-level manual</td>
</tr>
<tr>
<td>ImageNet-O [18]</td>
<td>hard adversarial OOD</td>
<td>2,000</td>
<td>image-level manual</td>
</tr>
</tbody>
</table>

Table 1. OpenImage-O follows natural class statistics, while ImageNet-O is adversarially built to be hard. Both datasets have image-level OOD annotation. Texture and iNaturalist are selected by tags, and their OOD labels are annotated in tag-level.

4922
Our method combines the strengths of feature-based scores and logit-based scores by the novel mechanism of virtual logit, and gets substantial improvements.

Network/Loss Design Many works redesign the training loss to be OOD-aware [5] or add regularization terms [18, 40] to push part ID/OOD features. DeVries et al. [5] augment the network with a confidence estimation branch that uses misclassified in-distribution examples as a proxy for out-of-distribution examples. MOS [18] modifies the loss to use the pre-defined group structure so that the minimum group-wise “else” class probability can indicate the OOD-ness. Zaeemzadeh et al. [40] forces the ID samples to embed into a union of 1-dimensional subspaces during training and computes the minimum angular distance from the feature to the class-wise subspaces. Generalized ODIN [16] uses a dividend/divisor structure to encode the prior knowledge of decomposing the confidence of class probability. Different from these methods, our method does not require model retraining, thus not only is it easier to apply, but the ID classification accuracy is also preserved.

OOD Data Exposure Outlier Exposure [14] utilizes an auxiliary OOD dataset to improve OOD detection. Dhamija et al. [6] regularize samples from extra background classes to have uniform logits and to have small feature norms. Lee et al. [22] use GAN to generate OOD samples that lie near the ID samples and push the prediction of OOD samples to the uniform distribution. Several methods, including MCD [39], NGC [36] and UDG [37], can utilize external unlabeled noisy data to enhance the OOD detection performances. Different from these methods, our method does not require additional OOD data and thus avoids biases towards the introduced OOD samples [31].


For a series of OOD detection methods that are based on logits or softmax probabilities, we find that their performances are limited. In Fig. 1, feature-based OOD scores such as Mahalanobis and Residual are good at detecting OOD in ImageNet-O, while all methods that are based on logit/probability lag behind. This is not an accident, as is again shown in Fig. 2. The AUROC of the state-of-the-art probability-based method KL Matching is still lower than straightforwardly designed OOD scores in feature space on Texture dataset. This motivates us to study the influence of the lost information going from features to logits.

Consider a $C$-class classification model whose logit $l \in \mathbb{R}^C$ is transformed from the feature $x \in \mathbb{R}^N$ by a fully connected layer with weight $W \in \mathbb{R}^{N \times C}$ and bias $b \in \mathbb{R}^C$, i.e. $l = W^T x + b$. The predicted probability is $p(x) = \text{softmax}(l)$. For convenience, we set the point $o := - (W^T)^+ b$, where $(\cdot)^+$ is the Moore-Penrose inverse, as the origin of a new coordinate system of feature space,

$$l = W^T x' = W^T (x - o), \quad \forall x. \quad (1)$$

Geometrically, each logit $l_i$ is the inner product between the feature $x'$ and the class vector $w_i$ (the $i$-th column of $W$). Later when generalizing logits to virtual logits, we will replace $w_i$ with a subspace, and replace the inner product with a projection. The bias term is safely omitted in the new coordinate system. In the remaining part of the paper, we assume the feature space uses the new coordinate system.

Logits contain class-dependent information, yet there is class-agnostic information in feature space that is not recoverable from logits. We study two cases (null space and principal space) and discuss the two OOD scores (NuSA and Residual) that rely on them, respectively.

OOD Score Based on Null Space A feature $x$ can be decomposed into $x = x^{W^\perp} + x^W$, where $W$ is the column space of $W$, $x^{W^\perp}$ and $x^W$ are projections of $x$ to $W^\perp$ and $W$, respectively. $W^\perp$ is the null space of $W^T$, and we have
$W^T x^{W^+} = 0$. The component $x^{W^+}$ does not affect classification, but it influences OOD detection. It is demonstrated in [3] that one can perturb an image intensely yet constrain the difference between the features in $W^\perp$. The resulting outlier images are not like any of the ID images but retains high confidence in classification. Taking advantage of this, they define an ID score NuSA (null space analysis) as in [3] that one can perturb an image intensely yet constrain ordisification, but it influences OOD detection. It is demonstrated in [3] that one can perturb an image intensely yet constraint the difference between the features in $W^\perp$. The resulting outlier images are not like any of the ID images but retains high confidence in classification. Taking advantage of this, they define an ID score NuSA (null space analysis) as

$$\text{NuSA}(x) = \frac{\sqrt{\|x\|^2 - \|x^{W^+}\|^2}}{\|x\|}. \quad (2)$$

Intuitively, NuSA uses the angle ($= \arccos(\text{NuSA}(x))$) between $x$ and $W$ to indicate the OOD-ness. From Fig. 2 we can see that the simple angle information clearly distinguishes OOD examples in Texture with an AUROC 95.50%, surpassing methods based on logits and the competitive method KL Matching based on softmax probability.

**OOD Score Based on Principal Space** It is generally assumed that features lie in low-dimensional manifolds [27, 40]. For simplicity, we use linear subspace (in the new coordinate system) passing through the origin $o$ as the model. We define the principal space as the $D$-dimensional subspace $P$ spanned by eigenvectors of the largest $D$ eigenvalues of the matrix $X^T X$, where $X$ is the ID data matrix. Features that deviate from the principal space are likely to be OOD examples. We can define

$$\text{Residual}(x) = \|x^{P^\perp}\|, \quad (3)$$

to capture the deviation of features from the principal space. Here $x = x^P + x^{P^\perp}$ and $x^{P^\perp}$ is the projection of $x$ to $P^\perp$. The residual score is similar to the reconstruction error in [27] except that they employ nonlinear manifold learning for dimension reduction. Note that after the projection onto logits, this deviation is corrupted since the matrix $W^T$ projects to a lower dimensional space than the feature space. Fig. 2 shows that Residual score improves over the NuSA score on both datasets, making the performance contrast between feature-based methods with logit/probability-based methods more striking.

**Fusing Class-dependent and Class-agnostic Information** In contrast to methods on logit/probability, both the NuSA and the Residual do not consider information that is specific to individual ID classes, namely they are class-agnostic. As a consequence, these scores ignore the feature similarity to each ID class, and are ignorant about which class the input resembles most. This gives an explanation of their worse performance on the iNaturalist OOD benchmark, as iNaturalist samples need to distinguish subtle differences between fine-grained classes. We hypothesize that unifying the information from feature space and the logits could improve the detection performance on a broader type of OOD samples. Such a solution is presented in Sec. 4 using the concept of virtual logit.

**4. Virtual-logit Matching**

To unify the class-agnostic and class-dependent information for OOD detection, we propose an OOD score by Virtual-logit Matching, abbreviated as ViM. The pipeline is illustrated in Fig. 3, where there are three steps, operating at the feature, the logit, and the probability, respectively. To be specific, for feature $x$, (1) extract the residual $x^{P^\perp}$ of $x$ against the principal subspace $P$; (2) convert the norm $\|x^{P^\perp}\|$ to a virtual logit by rescaling; and (3) output the softmax probability of the virtual logit as the ViM score. Below we give more details. Recall the notations: $C$ is the number of classes, $N$ is the feature dimension, and $W$ and $b$ are the classification weight and bias, respectively.

**Principal Subspace and Residual** Firstly we offset the feature space by a vector $o = -(W^T)^+ b$ so that it is bias-free in the computation of logits as Eq. (1). The principal subspace $P$ is defined by the training set $X$, where rows are features in the new coordinate system with origin $o$. Suppose the eigendecomposition on the matrix $X^T X$ is

$$X^T X = Q\Lambda Q^{-1}, \quad (4)$$

where eigenvalues in $\Lambda$ are sorted decreasingly, then the span of the first $D$ columns is the $D$-dimensional principal subspace $P$. The residual $x^{P^\perp}$ is the projection of $x$ onto $P^\perp$. Let the $(D+1)$-th column to the last column of $Q$ in Eq. (4) be a new matrix $R \in \mathbb{R}^{N \times (N-D)}$, then $x^{P^\perp} = RR^T x$. The residual $x^{P^\perp}$ is sent to the next step.

**Virtual-logit Matching** The virtual logit

$$l_0 := \alpha \|x^{P^\perp}\| = \alpha \sqrt{\frac{x^T RR^T x}{\|x\|^2}} \quad (5)$$

is the norm of the residual rescaled by a per-model constant $\alpha$. The norm $\|x^{P^\perp}\|$ cannot be used as a new logit directly since the latter softmax will normalize over the exponential of logits and thus is very sensitive to the scale of logits. If the residual is very small compared to the largest logit, then after the softmax the residual will be buried in the noise of logits. To match the scales of the virtual logit, we compute the average norm of the virtual logit on the training set and also the mean of the maximum logit on the training set, then

$$\alpha := \frac{\sum_{i=1}^{K} \max_{j=1,...,C} \{l_i^j\}}{\sum_{i=1}^{K} \|x_i^{P^\perp}\|}, \quad (6)$$

where $x_1, x_2, \ldots, x_K$ are uniformly sampled $K$ training examples, and $l_i^j$ is the $j$-th logit of $x_i$. In this way, on average, the scale of the virtual logit is the same as the maximum of the original logits.
The ViM Score  We append the virtual logit to the original logits and compute the softmax. The probability corresponding to the virtual logit is defined as ViM. Mathematically, let the $i$-th logit of $x$ be $l_i$, and then the score is

$$\text{ViM}(x) = \frac{e^{\alpha \|x^p\|^2}}{\sum_{i=1}^C e^{l_i} + e^{\alpha \|x^p\|^2}}.$$  

(7)

This equation reveals that two factors affect the ViM score: if its original logits are larger, then it is less of an OOD example; while if the norm of residual is larger, it is more likely to be OOD. The computational overhead is comparable to the last fully-connected layer (mapping from feature to logit) in the classification network, which is small.

Connection to Existing Methods  Note that applying a strictly increasing function to the scores does not affect the OOD evaluation. Apply the function $t(x) = -\ln (\frac{1}{x} - 1)$ to the ViM score, then we have an equivalent expression

$$\alpha \|x^p\| - \ln \sum_{i=1}^C e^{l_i}.$$  

(8)

The first term is the virtual logit in Eq. (5) while the second term is the energy score [25]. ViM completes the energy method by feeding extra residual information from features. The performance is much superior to energy and residual.

5. OpenImage-O Dataset

We build a new OOD dataset called OpenImage-O for the ID dataset ImageNet-1K. It is manually annotated, comes with a naturally diverse distribution, and has a large scale with 17,632 images. It is built to overcome several shortcomings of existing OOD benchmarks. OpenImage-O is selected image-by-image from the test set of OpenImage-V3, including 125,436 images collected from Flickr without a predefined list of class names or tags, leading to natural class statistics and avoiding an initial design bias.
6. Experiment

In this section, we compare our algorithm with state-of-the-art OOD detection algorithms. Following the prior work on large-scale OOD detection, we choose ImageNet-1K as the ID dataset. We benchmark the algorithms using both the CNN-based and the transformer-based models. Detailed experimental settings are as follows.

OOD Datasets Four OOD datasets (Tab. 1) are used to comprehensively benchmark the algorithms. OpenImage-O is our newly collected large-scale OOD dataset. Texture [2] consists of natural textural images and we removed four categories (bubbly, honeycombed, cobwebbed, spiralled) that overlapped with ImageNet. iNaturalist [34] is a fine-grained species classification dataset. We use the subset from [18]. Images in ImageNet-O [15] are adversarially filtered so that they can fool OOD detectors.

Evaluation Metrics Two commonly used metrics are reported. The AUROC is a threshold-free metric that computes the area under the receiver operating characteristic curve. Higher value indicates better detection performance. FPR95 is short for FPR@TPR95, which is the false positive rate when the true positive rate is 95%. The smaller FPR95 the better. We report both their numbers in percentage.

Experiment Settings BiT (Big Transfer) [20] is a variant of ResNet-v2, which employs group normalization and weight standardization. The BiT-S model series is pre-trained on ImageNet-1K, and we take the officially released checkpoint of BiT-S-R101×1 for experiments. ViT (Vision Transformer) [8] is a transformer-based image classification model which treats images as sequences of patches. We use the officially released ViT-B/16 model, which is pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K. Since the compared algorithms do not require re-training, the ID accuracies are not affected. Results on more model architectures, including CNN-based RepVGG [7], ResNet-50d [11], and transformer based Swin [26] and DeiT [33], are listed in Sec. 6.3. Their pre-trained weights are obtained from the timm repo [35]. When estimating the principal space, \( K = 200,000 \) images are randomly sampled from the training set. For features spaces with dimension \( N > 1500 \), we set the dimension of principal space to \( D = 1000 \), and set \( D = 512 \) otherwise.

Baseline Methods We compare ViM with eight baselines that do not require fine-tuning. They are MSP [13], Energy [25], ODIN [24], MaxLogit [12], KL Matching [12], Residual, ReAct [32] and Mahalanobis [23]. For Mahalanobis, we followed the setting in [10], which uses only the final feature instead of an ensemble of multiple layers [18, 23]. For ReAct, we use the Energy+ReAct setting with rectification percentile \( p = 99 \). The Residual is defined in Eq. (3).

6.1. Results on BiT

We present the results of the Bit model at the first half of Tab. 2. The best AUROC is shown in bold and the second and third place ones are shown with underlines.

ViM vs. Baselines On three datasets, including OpenImage-O, Texture, and ImageNet-O, ViM achieves the largest AUROC and the smallest FPR95. On average ViM has 90.91% AUROC, which surpasses the second place by 4.29%. The average FPR95 is also the lowest among them. In particular, regarding Eq. (8), an interpretation of ViM in terms of the Residual score and the Energy score, the results show that ViM is significantly better than the two methods on all datasets. This indicates that ViM non-trivially combined the OOD information in Residual and in Energy. However, on iNaturalist, ViM is only on the third place. We hypothesize that its moderate performance on iNaturalist relates to how much information is contained in the residual, because iNaturalist has the smallest average residual norm among four OOD datasets (iNaturalist 4.65, OpenImage-O 5.04, ImageNet-O 5.16, and Texture 8.16).

Effect of Information Source For OOD detection performances on BiT model, Tab. 2 shows an interesting pattern regarding the information source. If feature variations in the null space are absent, such as in methods that rely on logits and softmax, performances on Texture and ImageNet-O are restricted. For example, on the Texture dataset, the best performing method that relies on logit and softmax is KL Matching, which has 86.92% AUROC and is far behind ViM, Mahalanobis, and Residual, which operate on the feature space. In contrast, if the class-dependent information is dropped, such as in the Residual method, performances in iNaturalist and OpenImage-O are also limited. The proposed ViM score, however, is competent regardless of dataset types.

6.2. Results on ViT

[10] has discussed the benefit of large-scale pre-trained transformers on OOD tasks. However, their experiments are conducted on CIFAR100/10 and only two baseline methods are compared. We provide a comprehensive OOD evaluation on ImageNet-1K over a wide range of methods in the second half of Tab. 2.

ViM vs. Baselines The two best-performing methods for the ViT model are ViM and Mahalanobis. Their AU-
is pre-trained on the ImageNet-21K dataset, the semantics
between ViT and BiT

Difference between ViT and BiT Since the ViT model
is pre-trained on the ImageNet-21K dataset, the semantics
it has been seen is much larger than the BiT model. The OOD performance is relatively saturated. Although on most OOD datasets ViT is significantly better than BiT, we observe that ViT performs less competitively on the Texture dataset. We hypothesize that it is related to the observation in [28] that higher layers of ViT maintain spatial location information more faithfully than ResNets. ViT has high responses for local patches. However, textural images with similar local patches but not revealing the whole object are regarded as OOD of ImageNet (see example images in Fig. 2).

### Table 2. OOD detection for ViM and baseline methods. The ID dataset is ImageNet-1K, and OOD datasets are OpenImage-O, Texture, iNaturalist and ImageNet-O. Both metrics AUROC and FPR95 are in percentage. A pre-trained BiT-S-R101×1 model and a pre-trained ViT-B/16 model is tested. The best method is emphasized in bold, and the 2nd and 3rd ones are underlined. ODIN needs backpropagation for producing input perturbations, so it is prob+grad. ReAct clips feature and uses Energy subsequently, so it is feat+logit. Mahalanobis need gt labels to compute the class-wise mean feature, so it is feat+label. ↑: Residual is defined in Eq. (3).

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Source</th>
<th>OpenImage-O AUROC</th>
<th>Texture AUROC</th>
<th>iNaturalist AUROC</th>
<th>ImageNet-O AUROC</th>
<th>Average AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSP [13]</td>
<td>prob</td>
<td></td>
<td>84.16 73.72</td>
<td>79.80 76.65</td>
<td>87.92 64.09</td>
<td>57.12 96.85</td>
<td>77.25 77.83</td>
</tr>
<tr>
<td>Energy [25]</td>
<td>logit</td>
<td></td>
<td>84.77 73.42</td>
<td>81.09 73.91</td>
<td>84.47 74.98</td>
<td>63.59 96.40</td>
<td>78.48 79.68</td>
</tr>
<tr>
<td>ODIN [24]</td>
<td>prob+grad</td>
<td></td>
<td>85.64 72.83</td>
<td>81.60 74.07</td>
<td>86.73 70.75</td>
<td>63.00 96.85</td>
<td>79.24 78.63</td>
</tr>
<tr>
<td>MaxLogit [12]</td>
<td>logit</td>
<td></td>
<td>85.67 72.68</td>
<td>81.66 73.72</td>
<td>86.76 70.59</td>
<td>63.01 96.85</td>
<td>79.27 78.46</td>
</tr>
<tr>
<td>BiT</td>
<td>KL Matching [12]</td>
<td>prob</td>
<td>88.96 51.51</td>
<td>86.92 51.05</td>
<td><strong>92.95 33.28</strong></td>
<td>65.68 86.65</td>
<td>83.63 55.62</td>
</tr>
<tr>
<td>Residual↑</td>
<td>feat</td>
<td></td>
<td>80.58 67.85</td>
<td>97.66 11.16</td>
<td>76.76 80.41</td>
<td>81.57 65.50</td>
<td>84.14 56.23</td>
</tr>
<tr>
<td>ReAct [32]</td>
<td>feat+logit</td>
<td></td>
<td>88.94 54.97</td>
<td>90.64 50.25</td>
<td>91.45 48.60</td>
<td>67.07 91.70</td>
<td>84.53 61.38</td>
</tr>
<tr>
<td>Mahalanobis [23]</td>
<td>feat+label</td>
<td></td>
<td>83.10 64.32</td>
<td>97.33 14.05</td>
<td>85.70 64.95</td>
<td>80.37 70.05</td>
<td>86.62 53.34</td>
</tr>
<tr>
<td><strong>ViM (Ours)</strong></td>
<td>feat+logit</td>
<td></td>
<td><strong>91.54 43.96</strong></td>
<td><strong>98.92 4.69</strong></td>
<td><strong>89.30 55.71</strong></td>
<td><strong>83.87 61.50</strong></td>
<td><strong>90.91 41.46</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MSP</td>
<td>78.10 70.55</td>
<td>77.99 67.96</td>
<td>87.57 43.44</td>
<td>79.48 66.43</td>
</tr>
<tr>
<td>Energy</td>
<td>76.38 78.99</td>
<td>71.08 78.39</td>
<td>87.77 35.08</td>
<td>72.80 70.14</td>
</tr>
<tr>
<td>ODIN</td>
<td>77.72 72.68</td>
<td>75.27 68.56</td>
<td>88.00 36.58</td>
<td>77.13 63.92</td>
</tr>
<tr>
<td>MaxLogit</td>
<td>77.56 73.50</td>
<td>75.39 69.34</td>
<td>88.40 35.28</td>
<td>76.79 64.49</td>
</tr>
<tr>
<td>KL Matching</td>
<td>81.35 61.65</td>
<td>82.72 64.81</td>
<td>88.87 46.99</td>
<td>83.49 64.80</td>
</tr>
<tr>
<td>Residual↑</td>
<td>84.19 59.00</td>
<td>87.01 58.55</td>
<td>92.87 37.38</td>
<td>84.15 74.13</td>
</tr>
<tr>
<td>ReAct</td>
<td>49.14 98.96</td>
<td>82.93 58.63</td>
<td>90.27 31.56</td>
<td>77.37 67.00</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>86.07 59.39</td>
<td>88.33 55.70</td>
<td>92.16 40.89</td>
<td>85.03 73.18</td>
</tr>
<tr>
<td><strong>ViM (Ours)</strong></td>
<td><strong>87.81 50.50</strong></td>
<td><strong>89.22 52.61</strong></td>
<td><strong>94.11 31.04</strong></td>
<td><strong>85.25 69.95</strong></td>
</tr>
</tbody>
</table>

Table 3. Results on RepVGG, ResNet50-d, Swin and DeiT. D due to space limitation, only their average AUROC (A↑) and average FPR95 (F↓) are reported. The numbers are in percentage. All models are using pre-trained weights taken from timm [35].

ROC are close on all four datasets. However, Mahalanobis needs to compute the class-wise Mahalanobis distance, which makes its computation costly. In contrast, our method is lightweight and fast. Four methods, ReAct, Energy, MaxLogit, and ODIN, are the second best ones, and the remaining three methods have relatively low AUROCs.

**6.3. Results on More Model Architectures**

We show more results on a variety of model architectures. In particular, we choose two CNN-based models RepVGG [7] and ResNet50d [11] and two transformer-based models Swin Transformer [26] and DeiT [33]. Their average AUROCs and average FPR95s over the four OOD datasets are listed in Tab. 3. It is shown that ViM is robust to model architecture changes. The detailed experiment setting and results are in the supplementary materials.

**6.4. The Effect of Hyperparameter**

The **Dimension D of Principal Space** In [40] the feature of each class is represented by a 1-dimensional subspace, so a natural choice for the dimension D of principal space

---

4927
is the number of classes $C$. For models like ViT whose feature dimension $N$ may be less than the number of classes $C$, we empirically suggest taking a number in the range $[N/3, 2N/3]$. We show in Fig. 4 that our method is robust to the selection of dimensions. However, if the application permits, one can adjust this parameter according to a hold-out OOD dataset. In our experiments, we set $D = 1000$ for BiT and $D = 512$ for ViT.

The Matching Parameter $\alpha$ The matching parameter controls the relative importance of the trade-off between different OOD features. Since OOD distribution is unknown, we suggest keeping them to be of equal importance. This is how $\alpha$ is defined in Eq. (6). It is easy to tune the parameter to fit some types of OOD datasets, but it is hard to improve all datasets at the same time. We show the result of perturbing the matching parameter by multiplying a factor in Fig. 5. If the multiple is larger, then information from the feature space is given more weight. Otherwise, information from logits is given more importance. Overall the best choice is no perturbation, suggesting that the defined $\alpha$ is a good choice.

6.5. The Effect of Grouping

In addition, we also compare with MOS [18], which exploits grouping structure in large-scale semantic spaces. Two methods are added to the comparison. (1) MaxGroup is the group version of MSP, which first obtains the group-wise probability by summing over the constituent classes, and then takes the maximum group probability as the ID score. (2) ViM+Group also takes the maximum group probability as the ID score, except that the probabilities are taken from the $(C + 1)$ dimensional vector, with an extra ViM virtual class participating in the softmax normalization.

MaxGroup and ViM+Group are evaluated on the pre-trained weights of BiT, while MOS needs to fine-tune the model using group-based learning. Results are shown in Tab. 4. We observe that (1) the average AUROC of MaxGroup improves over the vanilla MSP from 77.25% to 79.23%, showing the usefulness of group information; and (2) both our original ViM and the group version of ViM are better than MOS on three of four datasets by large margins.

6.6. Limitation of ViM

As we have noticed in Sec. 6.1, ViM shows less performance gains on OOD datasets that have small residuals, such as iNaturalist. Besides, the property that ViM does not need training is a double-edged sword. It means that ViM is limited by the feature quality of the original network.

7. Conclusion

In this paper, we present a novel OOD detection method: the Virtual-logit Matching (ViM) score. It combines the information from both the feature space and the logits, which provides the class-agnostic information and the class-dependent information, respectively. Extensive experiments on the large-scale OOD benchmarks show the effectiveness and robustness of the method. Especially, we tested ViM on both CNN-based models and transformer-based models, showing its robustness across model architectures. To facilitate the evaluation of large-scale OOD detection, we create the OpenImage-O dataset for ImageNet-1K, which is of high-quality and large-scale.

Acknowledgement

This work was supported in part by Innovation and Technology Commission of the Hong Kong Special Administrative Region, China (Enterprise Support Scheme under the Innovation and Technology Fund B/E030/18). Haoqi Wang was also supported by the Technology Leaders of Tomorrow (TLT) Programme of HKSTP InnoAcademy.
References


