RankFeat: Rank-1 Feature Removal for Out-of-distribution Detection

Yue Song¹, Nicu Sebe¹, and Wei Wang²

¹Department of Information Engineering and Computer Science, University of Trento, Italy
²Beijing Jiaotong University, China
yue.song@unitn.it

Abstract

The task of out-of-distribution (OOD) detection is crucial for deploying machine learning models in real-world settings. In this paper, we observe that the singular value distributions of the in-distribution (ID) and OOD features are quite different: the OOD feature matrix tends to have a larger dominant singular value than the ID feature, and the class predictions of OOD samples are largely determined by it. This observation motivates us to propose RankFeat, a simple yet effective post hoc approach for OOD detection by removing the rank-1 matrix composed of the largest singular value and the associated singular vectors from the high-level feature. RankFeat achieves state-of-the-art performance and reduces the average false positive rate (FPR95) by 17.90% compared with the previous best method. Extensive ablation studies and comprehensive theoretical analyses are presented to support the empirical results.

1 Introduction

In the real-world applications of deep learning, understanding whether a test sample belongs to the same distribution of training data is critical to the safe deployment of machine learning models. The main challenge stems from the fact that current deep learning models can easily give over-confident predictions for out-of-distribution (OOD) data [43]. Recently a rich line of literature has emerged to address the challenge of OOD detection [60, 27, 3, 7, 52, 14, 11, 54, 66, 37, 13, 68, 10, 15, 18].

Previous OOD detection approaches either rely on the feature distance [38], activation abnormality [52], or gradient norm [27]. In this paper, we tackle the problem of OOD detection from another perspective: by analyzing the spectrum of the high-level feature matrices (e.g., the output of Block 3 or Block 4 of a typical ResNet [19] model), we observe that the feature matrices have quite different singular value distributions for the in-distribution (ID) and OOD data (see Fig. 1(a)): the OOD feature tends to have a much larger dominant singular value than the ID feature, whereas the magnitudes of the rest singular values are very similar. This peculiar behavior motivates us to remove the rank-1 matrix composed of the dominant singular value and singular vectors from the feature. As displayed in Fig. 1(b), removing the rank-1 feature drastically perturbs the class prediction of OOD samples; a majority of predictions have been changed. On the contrary, most ID samples have consistent classification results before and after removing the subspace. This phenomenon indicates that the over-confident prediction of OOD samples might be largely determined by the dominant singular value and the corresponding singular vectors.

Based on this observation, we assume that the first singular value of OOD feature tends to be much larger than that of ID feature. The intuition behind is that the OOD feature corresponds to a larger
The distribution of top-5 singular values for the ID and OOD features on ImageNet-1k and SUN. The OOD feature matrix tends to have a significantly larger dominant singular value. After removing the rank-1 matrix composed by the dominant singular value and singular vectors, the class predictions of OOD data are severely perturbed, while those of ID data are moderately influenced. This observation indicates that the decisions of OOD data heavily depend on the dominant singular value and the corresponding singular vectors of the feature matrix. In light of this finding, we get motivated to propose RankFeat for OOD detection by removing the rank-1 matrix from the high-level feature. Both observations also hold for other OOD datasets.

Figure 2: The score distributions of Energy [40] (top row) and our proposed RankFeat (bottom row) on four OOD datasets. Our method can better separate the ID and OOD data.

PCAs explained variance ratio (being less informative), and the well-trained network weights might cause and amplify the difference (see Sec. E of the supplementary for the detailed illustration). Hence, we conjecture that leveraging this gap might help to better distinguish ID and OOD samples. To this end, we propose RankFeat, a simple but effective post hoc approach for OOD detection. RankFeat perturbs the high-level feature by removing its rank-1 matrix composed of the dominant singular value and vectors. Then the logits derived from the perturbed features are used to compute the OOD score function. By removing the rank-1 feature, the over-confidence of OOD samples is mitigated, and consequently the ID and OOD data can be better distinguished (see Fig. 2). Our RankFeat establishes the state-of-the-art performance on the large-scale ImageNet benchmark and a suite of widely used OOD datasets across different network depths and architectures. In particular, RankFeat outperforms the previous best method by 17.90% in the average false positive rate (FPR95) and by 5.44% in the area under curve (AUROC). Extensive ablation studies are performed to reveal important insights of RankFeat, and comprehensive theoretical analyses are conducted to explain the working mechanism. Code is publicly available via [https://github.com/KingJamesSong/RankFeat](https://github.com/KingJamesSong/RankFeat).

The key results and main contributions are threefold:

- We propose RankFeat, a simple yet effective post hoc approach for OOD detection by removing the rank-1 matrix from the high-level feature. RankFeat achieves the state-of-the-art performance across benchmarks and models, reducing the average FPR95 by 17.90% and improving the average AUROC by 5.44% compared to the previous best method.
• We perform extensive ablation studies to illustrate the impact of (1) removing or keeping the rank-1 matrix, (2) removing the rank-n matrix (n>1), (3) applying our RankFeat at various network depths, (4) the number of iterations to iteratively derive the approximate rank-1 matrix for acceleration but without performance degradation, and (5) different fusion strategies to combine multi-scale features for further performance improvements.

• Comprehensive theoretical analyses are conducted to explain the working mechanism and to underpin the superior empirical results. We show that (1) removing the rank-1 matrix reduces the upper bound of OOD score more, (2) removing the rank-1 matrix makes the statistics of OOD feature closer to random matrices, and (3) both RankFeat and ReAct work by optimizing the upper bound containing the largest singular value. ReAct indirectly and manually clips the underlying term, while RankFeat directly subtracts it.

2 RankFeat: Rank-1 Feature Removal for OOD Detection

In this section, we introduce the background of OOD detection task and our proposed RankFeat that performs the OOD detection by removing the rank-1 matrix from the high-level feature.

Preliminary: OOD detection. The OOD detection is often formulated as a binary classification problem with the goal to distinguish between ID and OOD data. Let \( f \) denote a model trained on samples from the ID data \( D_{in} \). For the unseen OOD data \( D_{out} \) at test time, OOD detection aims to define a decision function \( \mathcal{G}(\cdot) \):

\[
\mathcal{G}(x) = \begin{cases} \text{in} & S(x) > \gamma, \\ \text{out} & S(x) < \gamma. \end{cases}
\]

where \( x \) denotes the data encountered at the inference stage, \( S(\cdot) \) is the seeking scoring function, and \( \gamma \) is a chosen threshold to make a large portion of ID data correctly classified (e.g., 95%). The difficulty of OOD detection lies in designing an appropriate scoring function \( S(\cdot) \) such that the score distributions of ID and OOD data overlap as little as possible.

RankFeat: rank-1 feature removal. Consider the reshaped high-level feature map \( X \in \mathbb{R}^{C \times HW} \) of a deep network (the batch size is omitted for simplicity). Here 'high-level feature' denotes the feature map that carries rich semantics in the later layers of a network (e.g., the output of Block 3 or Block 4 of a typical deep model like ResNet). Our RankFeat first performs the Singular Value Decomposition (SVD) on each individual feature matrix in the mini-batch to decompose the feature:

\[
X = USV^T
\]

where \( S \in \mathbb{R}^{C \times HW} \) is the rectangular diagonal singular value matrix, and \( U \in \mathbb{R}^{C \times C} \) and \( V \in \mathbb{R}^{HW \times HW} \) are left and right orthogonal singular vector matrices, respectively. Then RankFeat removes the rank-1 matrix from the feature as:

\[
X' = X - s_1 u_1 v_1^T
\]

where \( s_1 \) is the largest singular value, and \( u_1 \) and \( v_1 \) are the corresponding left and right singular vectors, respectively. The perturbed feature is fed into the rest of the network to generate the logit predictions \( y' \). Finally, RankFeat computes the energy score of the logits for the input \( x \) as:

\[
\text{RankFeat}(x) = \log \sum \exp(y')
\]

By removing the rank-1 matrix composed by the dominant singular value \( s_1 \), the over-confident predictions of OOD data are largely perturbed. In contrast, the decisions of ID data are mildly influenced. This could help to separate the ID and OOD data better in the logit space.

Acceleration by Power Iteration. Since RankFeat only involves the dominant singular value and vectors, there is no need to compute the full SVD of the feature matrix. Hence our method can be potentially accelerated by Power Iteration (PI). The PI algorithm is originally used to approximate the dominant eigenvector of a Hermitian matrix. With a slight modification, it can also be applied to general rectangular matrices. Given the feature \( X \), the modified PI takes the coupled iterative update:

\[
v_k = \frac{Xu_k}{||Xu_k||}, \quad u_{k+1} = \left( \frac{v_k^T X}{||v_k||} \right)^T
\]
where $u_0$ and $v_0$ are initialized with random orthogonal vectors and converge to the left and right singular vectors, respectively. After certain iterations, the dominant singular value is computed as $s_1 = v_1^T X u_k$. As will be illustrated in Sec. 4.5, the approximate solution yielded by PI achieves very competitive performance against the SVD but with much less time overhead.

**Combination of multi-scale features.** Our RankFeat works at various later depths of a model, i.e., Block 3 and Block 4. Since intermediate features might focus on different semantic information, their decision cues are very likely to be different. It is thus natural to consider fusing the scores to leverage the distinguishable information of both features for further performance improvements. Let $y'$ and $y''$ denote the logit predictions of Block 3 and Block 4 features, respectively. RankFeat performs the fusion at the logit space and computes the score function as $\log \sum \exp((y'+y'')/2)$. Different fusion strategies are explored and discussed in Sec. 4.3.

### 3 Theoretical Analysis

In this section, we perform some theoretical analyses on RankFeat to support the empirical results. We start by proving that removing the rank-1 feature with a larger $s_1$ would reduce the upper bound of RankFeat score more. Then based on Random Matrix Theory (RMT), we show that removing the rank-1 matrix makes the statistics of OOD features closer to random matrices. Finally, the theoretical connection of ReAct and our RankFeat is analyzed and discussed: both approaches work by optimizing the score upper bound determined by $s_1$. ReAct manually uses a pre-defined threshold to clip the term with $s_1$, whereas our RankFeat directly optimizes the bound by subtracting this term.

**Removing the rank-1 matrix with a larger $s_1$ would reduce the upper bound of RankFeat more.**

For our RankFeat score function, we can express its upper bound in an analytical form. Moreover, the upper bound analysis explicitly indicates that removing the rank-1 matrix with a larger first singular value would reduce the upper bound more. Specifically, we have the following proposition.

**Proposition 1.** The upper bound of RankFeat score is defined as $\text{RankFeat}(x) < \frac{1}{HW} \left( \sum_{i=1}^{N} s_i - s_1 \right) ||W||_{\infty} + ||b||_{\infty} + \log(Q)$ where $Q$ denotes the number of categories, and $W$ and $b$ are the weight and bias of the last layer, respectively. A larger $s_1$ would reduce the upper bound more.

**Proof.** For the feature $X \in \mathbb{R}^{C \times HW}$, its SVD $USV^T = X$ can be expressed as the summation of rank-1 matrices $X = \sum_{i=1}^{N} s_i u_i v_i^T$. The feature perturbed by RankFeat can be computed as:

$$X' = X - s_1 u_1 v_1^T = \sum_{i=2}^{N} s_i u_i v_i^T$$  \hspace{1cm} (6)

where $N$ denotes the shorter side of the matrix (usually $N=HW$). In most deep models [19, 20], usually the last feature map needs to pass a Global Average Pooling (GAP) layer to collapse the width and height dimensions. The GAP layer can be represented by a vector

$$m = \frac{1}{HW} [1, 1, \ldots, 1]^T$$  \hspace{1cm} (7)

The pooled feature map is calculated as $X'm$. Then the output logits are computed by the matrix-vector product with the classification head as:

$$y' = WX'm + b = \sum_{i=2}^{N} (s_i Wu_i v_i^T m) + b$$  \hspace{1cm} (8)

where $W \in \mathbb{R}^{W \times C}$ denotes the weight matrix, $b \in \mathbb{R}^{Q \times 1}$ represents the bias vector, and $y' \in \mathbb{R}^{Q \times 1}$ is the output logits that correspond to the perturbed feature $X'$. Our RankFeat score is computed as:

$$\text{RankFeat}(x) = \log \sum_{i=1}^{Q} \exp(y'_i)$$  \hspace{1cm} (9)

where $x$ is the input image, and $Q$ denotes the number of categories. Here we choose Energy [40] as the base function due to its theoretical alignment with the input probability density and its strong
empirical performance. Eq. (9) can be re-formulated by the Log-Sum-Exp trick
\[
\log \sum_{i=1}^{Q} \exp(y'_i) = \log \sum_{i=1}^{Q} \exp(y'_i - \max(y')) + \max(y') \tag{10}
\]
The above equation directly yields the tight bound as:
\[
\max(y') < \log \sum_{i=1}^{Q} \exp(y') < \max(y') + \log(Q) \tag{11}
\]
Since \(\max(y') \leq \|y'\|_\infty\), we have
\[
\text{RankFeat}(x) = \log \sum_{i=1}^{Q} \exp(y') < \max(y') + \log(Q) \leq \|y'\|_\infty + \log(Q) \tag{12}
\]
The vector norm has the property of triangular inequality, i.e., \(\|a + c\| \leq \|a\| + \|c\|\) holds for any vectors \(a\) and \(c\). Moreover, since both \(u\) and \(v\) are orthogonal vectors, we have the relation \(\|u\|_\infty \leq 1\) and \(\|v\|_\infty \leq 1\). Relying on these two properties, injecting eq. (8) into eq. (12) leads to
\[
\text{RankFeat}(x) < \sum_{i=2}^{N} s_i ||W_i v_i^T m||_\infty + ||b||_\infty + \log(Q) \leq \sum_{i=2}^{N} s_i ||W_i m||_\infty + ||b||_\infty + \log(Q) \tag{13}
\]
Since \(m\) is a scaled all-ones vector, we have \(||W m||_\infty = ||W||_\infty / HW\). The bound is simplified as:
\[
\text{RankFeat}(x) < \frac{1}{HW} \left( \sum_{i=1}^{N} s_i - s_1 \right) ||W||_\infty + ||b||_\infty + \log(Q) \tag{14}
\]
As indicated above, removing a larger \(s_1\) would reduce the upper bound of \(\text{RankFeat}\) score more. □

**Remark:** Considering that OOD feature usually has a much larger \(s_1\) (see Fig. 1(a)), \(\text{RankFeat}\) would reduce the upper bound of OOD samples more.

*Notice that our bound analysis strives to improve the understanding of OOD methods from new perspectives instead of giving a strict guarantee of the score.* For example, the upper bound can be used to explain the shrinkage and skew of score distributions in Fig. 2. Subtracting \(s_1\) would largely reduce the numerical range of both ID and OOD scores, which could squeeze score distributions. Since the dominant singular value \(s_1\) contributes most to the score, removing \(s_1\) is likely to make many samples have similar scores. This would concentrate samples in a smaller region and further skew the distribution. Given that the OOD feature tends to have a much larger \(s_1\), this would have a greater impact on OOD data and skew the OOD score distribution more.

Figure 3: The exemplary eigenvalue distribution of ID/OOD feature and the fitted MP distribution. After the rank-1 matrix is removed, the lowest bin of OOD feature has a larger reduction and the middle bins gain some growth, making the ODD feature statistics closer to the MP distribution.

**Removing the rank-1 matrix is likely to make the statistics of OOD features closer to random matrices.** Now we turn to use RMT to analyze the statistics of OOD and ID feature matrices. For a random matrix of a given shape, the density of its eigenvalue asymptotically converges to the Manchenko-Pastur (MP) distribution [41, 50]. Formally, we have:

**Theorem 1** (Manchenko-Pastur Law [41, 50]). Let \(X\) be a random matrix of shape \(t \times n\) whose entries are random variables with \(E(X_{ij}) = 0\) and \(E(X_{ij}^2) = 1\). Then the eigenvalues of the sample covariance \(Y = \frac{1}{n}XX^T\) converges to the probability density function: \(\rho(\lambda) = \frac{t}{n} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi \lambda \sigma^2} \) for \(\lambda \in [\lambda_-, \lambda_+]\) where \(\lambda_- = \sigma^2(1 - \sqrt{\frac{\mu}{t}})^2\) and \(\lambda_+ = \sigma^2(1 + \sqrt{\frac{\mu}{t}})^2\).

5
This theorem implies the possibility to measure the statistical distance between ID/OOD features and random matrices. To this end, we randomly sample 1,000 ID and OOD feature matrices and compute the KL divergence between the actual eigenvalue distribution and the fitted MP distribution.

Table 1: The KL divergence between ID/OOD feature and the fitted MP distribution. When the rank-1 feature is removed, the statistics of OOD matrix are closer to random matrices.

<table>
<thead>
<tr>
<th>Matrix Type</th>
<th>Block 4</th>
<th>Block 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ID</td>
<td>OOD</td>
</tr>
<tr>
<td>Original feature</td>
<td>18.36</td>
<td>18.24</td>
</tr>
<tr>
<td>Removing rank-1 matrix</td>
<td>17.07</td>
<td><strong>15.79</strong></td>
</tr>
</tbody>
</table>

Fig. [3] and Table [1] present the exemplary eigenvalue distribution and the average evaluation results of Block 4 and Block 3 features, respectively. For the original feature, the OOD and ID feature matrices exhibit similar behaviors: the distances to the fitted MP distribution are roughly the same (dif f. ≈ 0.1). However, when the rank-1 matrix is removed, the OOD feature matrix has a much larger drop in the KL divergence. This indicates that removing the rank-1 matrix makes the statistics of OOD feature closer to random matrices, i.e., the OOD feature is very likely to become less informative than the ID feature. The result partly explains the working mechanism of RankFeat: by removing the feature matrix where OOD data might convey more information than ID data, the two types of distributions have a larger discrepancy and therefore can be better separated.

Connection with ReAct [52]. ReAct clips the activations at the penultimate layer of a model to distinguish ID and OOD samples. Given the feature $X$ and the pooling layer $m$, the perturbation can be defined as:

$$\min(Xm, \tau) = Xm - \max(Xm - \tau, 0)$$

where $\tau$ is a pre-defined threshold. Their method shares some similarity with RankFeat formulation $Xm - s_1 u_1 v_1^T m$. Both approaches subtract from the feature a portion of information that is most likely to cause the over-confidence of OOD prediction. ReAct selects the manually-defined threshold $\tau$ based on statistics of the whole ID set, while RankFeat generates the structured rank-1 matrix from the feature itself. Taking a step further, ReAct has the score inequality following eq. (12):

$$\text{ReAct}(x) < ||WXm - W\max(Xm - \tau, 0)||_\infty + ||b||_\infty + \log(Q)$$

Since $X$ is non-negative (output of ReLU), we have $\max(Xm) \geq \max(x)/HW$. Exploiting the vector norm inequality $||X||_\infty \geq ||X||_2$ leads to the relation $\max(X) \geq s_1 / \sqrt{CHW}$. Relying on this property, the above inequality can be re-formulated as:

$$\text{ReAct}(x) < \frac{1}{HW} \sum_{i=1}^{N} s_i ||W||_\infty - \frac{1}{HW} \max(s_1 \sqrt{CHW} - \tau, 0) ||W||_\infty + ||b||_\infty + \log(Q)$$

As indicated above, the upper bound of ReAct is also determined by the largest singular value $s_1$. In contrast, the upper bound of our RankFeat can be expressed as:

$$\text{RankFeat}(x) < \frac{1}{HW} \sum_{i=1}^{N} s_i ||W||_\infty - \frac{1}{HW} s_1 ||W||_\infty + ||b||_\infty + \log(Q)$$

The upper bounds of both methods resemble each other with the only different term boxed. From this point of view, both methods distinguish the ID and OOD data by eliminating the impact of the term containing $s_1$ in the upper bound. ReAct optimizes it by clipping the term with a manually-defined threshold, which is indirect and might be sub-optimal. Moreover, the threshold selection requires statistics of the whole ID set. In contrast, our RankFeat does not require any extra data and directly subtracts this underlying term which is likely to cause the over-confidence of OOD samples.

4 Experimental Results

In this section, we first discuss the setup in Sec. 4.1 and then present the main experimental results on ImageNet-1k in Sec. 4.2, followed by the extensive ablation studies in Sec. 4.3.
### 4.1 Setup

**Datasets.** In line with [26, 52, 27], we mainly evaluate our method on the large-scale ImageNet-1k benchmark [6]. The large-scale dataset is more challenging than the traditional CIFAR benchmark [56] because the images are more realistic and diverse (i.e., 1.28M images of 1,000 classes). For the OOD datasets, we select four testsets from subsets of iNaturalist [58], SUN [63], Places [70], andTextures [5]. These datasets are crafted by [26] with non-overlapping categories from ImageNet-1k. Besides the experiment on the large-scale benchmark, we also validate the effectiveness of our approach on Species [24] and CIFAR [56] benchmark. (see Supplementary Material).

**Baselines.** We compare our method with 6 recent post hoc OOD detection methods, namely MSP [22], ODIN [39], Energy [40], Mahalanobis [38], GradNorm [27], and ReAct [52]. The detailed illustration and settings of these methods are kindly referred to Supplementary Material.

**Architectures.** In line with [27], the main evaluation is done using Google BiT-S model [35] pretrained on ImageNet-1k with ResNetV2-101 [20]. We also evaluate the performance on SqueezeNet [29], an alternative tiny architecture suitable for mobile devices and on T2T-ViT-24 [67], a tokens-to-tokens vision transformer that has impressive performance when trained from scratch.

*For the implementation details and evaluation metrics, please refer to Supplementary Material.*

#### Table 2: Main evaluation results on ResNetV2-101 [20]. All values are reported in percentages, and these post hoc methods are directly applied to the model pre-trained on ImageNet-1k [6]. The best three results are highlighted with red, blue, and cyan.

<table>
<thead>
<tr>
<th>Methods</th>
<th>iNaturalist</th>
<th>SUN</th>
<th>Places</th>
<th>Textures</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPR95 (↑)</td>
<td>AUROC (↑)</td>
<td>FPR95 (↑)</td>
<td>AUROC (↑)</td>
<td>FPR95 (↑)</td>
</tr>
<tr>
<td>MSP [22]</td>
<td>63.69</td>
<td>87.59</td>
<td>79.89</td>
<td>78.34</td>
<td>81.44</td>
</tr>
<tr>
<td>ODIN [39]</td>
<td>62.69</td>
<td>89.36</td>
<td>71.67</td>
<td>83.92</td>
<td>76.27</td>
</tr>
<tr>
<td>Energy [40]</td>
<td>64.91</td>
<td>88.48</td>
<td>65.33</td>
<td>85.32</td>
<td>73.02</td>
</tr>
<tr>
<td>Mahalanobis [38]</td>
<td>46.48</td>
<td>84.64</td>
<td>89.75</td>
<td>65.20</td>
<td>82.23</td>
</tr>
<tr>
<td>GradNorm [27]</td>
<td>50.03</td>
<td>90.33</td>
<td>46.48</td>
<td>89.03</td>
<td>60.86</td>
</tr>
<tr>
<td>ReAct [52]</td>
<td>44.52</td>
<td>91.81</td>
<td>52.71</td>
<td>90.16</td>
<td>62.66</td>
</tr>
<tr>
<td>RankFeat (Block 4)</td>
<td>46.54</td>
<td>81.49</td>
<td>27.88</td>
<td>92.18</td>
<td>38.26</td>
</tr>
<tr>
<td>RankFeat (Block 3)</td>
<td>49.61</td>
<td>91.42</td>
<td>29.91</td>
<td>92.01</td>
<td>51.82</td>
</tr>
<tr>
<td>RankFeat (Block 3 + 4)</td>
<td>41.31</td>
<td>91.91</td>
<td>29.27</td>
<td>94.07</td>
<td>39.34</td>
</tr>
</tbody>
</table>

### 4.2 Results

**Main results.** Following [27], the main evaluation is conducted using Google BiT-S model [35], pretrained on ImageNet-1k with ResNetV2-101 architecture [20]. Table 2 compares the performance of all the post hoc methods. For both Block 3 and Block 4 features, our RankFeat achieves the best evaluation results across datasets and metrics. More specifically, RankFeat based on the Block 4 feature outperforms the second-best baseline by 15.01% in the average FPR95, while the Block 3 feature-based RankFeat beats the second-best method by 4.09% in the average AUROC. Their combination further surpasses other methods by 17.90% in the average FPR95 and by 5.44% in the average AUROC. The superior performances at various depths demonstrate the effectiveness and general applicability of RankFeat. The Block 3 feature has a higher AUROC but slightly falls behind the Block 4 feature in the FPR95, which can be considered a compromise between the two metrics.

**Our RankFeat is also effective on alternative CNN architectures.** Besides the experiment on ResNetV2 [20], we also evaluate our method on SqueezeNet [29], an alternative tiny network suitable for mobile devices and on-chip applications. This network is more challenging because the tiny network size makes the model prone to overfit the training data, which could increase the difficulty to distinguish between ID and OOD samples. Table 3 top presents the performance of all the methods. Collectively, the performances of RankFeat are very competitive at both depths, as well as the score fusion. Our RankFeat achieves the state-of-the-art performances, outperforming the second-best baseline by 14.22% in FPR95 and by 7.48% in AUROC.

**Our RankFeat also suits transformer-based architectures.** To further demonstrate the applicability of our method, we evaluate RankFeat on Tokens-to-Tokens Vision Transformer (T2T-ViT) [67], a popular transformer architecture that can achieve competitive performance against CNNs when trained from scratch. Similar to the CNN, RankFeat removes the rank-1 matrix from the final token
Table 3: The results on SqueezeNet [29] and T2T-ViT-24 [67]. All values are reported in percentages, and these post hoc methods are directly applied to the model pre-trained on ImageNet-1k [6]. For results on SqueezeNet [29], the best three results are highlighted with red, blue, and cyan.

<table>
<thead>
<tr>
<th>Model</th>
<th>Methods</th>
<th>Naturalist</th>
<th>SUN</th>
<th>Places</th>
<th>Textures</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
<td>AUROC</td>
</tr>
<tr>
<td>SqueezeNet [29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSP [23]</td>
<td>89.83</td>
<td>65.41</td>
<td>83.03</td>
<td>72.25</td>
<td>87.27</td>
<td>70.00</td>
</tr>
<tr>
<td>ODIN [39]</td>
<td>90.79</td>
<td>65.75</td>
<td>78.32</td>
<td>78.37</td>
<td>83.23</td>
<td>73.31</td>
</tr>
<tr>
<td>Energy [28]</td>
<td>79.27</td>
<td>73.30</td>
<td>56.41</td>
<td>87.88</td>
<td>67.74</td>
<td>82.73</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>91.50</td>
<td>51.79</td>
<td>90.33</td>
<td>62.18</td>
<td>92.26</td>
<td>56.63</td>
</tr>
<tr>
<td>GradNorm [33]</td>
<td>76.31</td>
<td>73.92</td>
<td>51.63</td>
<td>87.55</td>
<td>65.99</td>
<td>83.28</td>
</tr>
<tr>
<td>ReAct [52]</td>
<td>76.78</td>
<td>68.56</td>
<td>87.57</td>
<td>66.37</td>
<td>88.80</td>
<td>66.20</td>
</tr>
</tbody>
</table>

| T2T-ViT-24 [67]|         |           |         |        |         |        |         |        |         |        |
| MSP [23]       | 48.92  | 88.95     | 61.77   | 81.37  | 69.54   | 80.03  | 62.91   | 82.31  | 60.79   | 83.17  |
| ODIN [39]      | 44.07  | 88.17     | 63.83   | 78.46  | 68.19   | 75.33  | 54.27   | 83.63  | 57.59   | 81.40  |
| Energy [28]    | 52.95  | 82.93     | 68.55   | 73.06  | 74.24   | 68.17  | 51.05   | 83.25  | 61.70   | 78.85  |
| Mahalanobis     | 90.50  | 58.13     | 91.71   | 50.52  | 93.32   | 49.60  | 80.67   | 64.06  | 89.05   | 55.58  |
| GradNorm [33]  | 99.30  | 25.86     | 98.37   | 28.06  | 99.01   | 25.71  | 92.68   | 38.80  | 97.34   | 29.61  |
| ReAct [52]     | 52.17  | 89.51     | 65.23   | 81.05  | 68.93   | 78.20  | 52.54   | 85.46  | 59.72   | 83.55  |

| RankFeat (Block 4) | 61.67  | 83.09     | 46.72   | 88.31  | 61.31   | 80.52  | 38.04   | 88.82  | 51.94   | 85.19  |
| RankFeat (Block 3) | 77.84  | 81.59     | 49.18   | 90.43  | 62.94   | 85.82  | 50.14   | 79.52  | 58.53   | 85.28  |
| RankFeat (Block 3 + 4) | 65.81  | 83.06     | 46.64   | 90.17  | 61.56   | 84.51  | 42.54   | 85.00  | 54.14   | 85.69  |

4.3 Ablation Studies

In this subsection, we conduct several ablation studies based on Google-BiT-S ResNetv2-101 model. Unless explicitly specified, we apply RankFeat on the Block 4 feature by default.

Table 5: Ablation studies on keeping only the rank-1 matrix and removing the rank-n matrix.

<table>
<thead>
<tr>
<th>Baselines</th>
<th>ANaturalist</th>
<th>SUN</th>
<th>Places</th>
<th>Textures</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
</tr>
<tr>
<td></td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>GradNorm [33]</td>
<td>50.03</td>
<td>90.33</td>
<td>46.48</td>
<td>89.03</td>
<td>60.86</td>
</tr>
<tr>
<td>ReAct [52]</td>
<td>44.52</td>
<td>91.81</td>
<td>52.71</td>
<td>90.16</td>
<td>62.66</td>
</tr>
<tr>
<td>Keeping Only Rank-1</td>
<td>48.97</td>
<td>91.93</td>
<td>62.63</td>
<td>84.62</td>
<td>72.42</td>
</tr>
<tr>
<td>Removing Rank-3</td>
<td>55.19</td>
<td>90.03</td>
<td>48.97</td>
<td>91.26</td>
<td>56.63</td>
</tr>
<tr>
<td>Removing Rank-2</td>
<td>50.04</td>
<td>89.30</td>
<td>48.55</td>
<td>90.99</td>
<td>56.23</td>
</tr>
<tr>
<td></td>
<td>46.54</td>
<td>81.49</td>
<td>27.88</td>
<td>92.18</td>
<td>38.26</td>
</tr>
</tbody>
</table>

Removing the rank-1 matrix outperforms keeping only it. Instead of removing the rank-1 matrix, another seemingly promising approach is keeping only the rank-1 matrix and abandoning the rest of the matrix. Table 5 presents the evaluation results of keeping only the rank-1 matrix. The performance falls behind that of removing the rank-1 feature by 18.8% in FPR95, which indicates that keeping only the rank-1 feature is inferior to removing it in distinguishing the two distributions. Nonetheless,
it is worth noting that even keeping only the rank-1 matrix achieves very competitive performance against previous best methods, such as GradNorm [27] and ReAct [52].

Removing the rank-1 matrix outperforms removing the rank-n matrix (n>1). We evaluate the impact of removing the matrix of a higher rank, i.e., performing $X - \sum_{i=1}^{n} s_i u_i v_i^T$ where $n>1$ for the high-level feature $X$. Table 5 compares the performance of removing the rank-2 matrix and rank-3 matrix. When the rank of the removed matrix is increased, the average performance degrades accordingly. This demonstrates that removing the rank-1 matrix is the most effective approach to separate ID and OOD data. This result is coherent with the finding in Fig. 1(a): only the largest singular value of OOD data is significantly different from that of ID data. Therefore, removing the rank-1 matrix achieves the best performance.

**Table 7:** The approximate solution by PI yields competitive performance and costs much less time.

<table>
<thead>
<tr>
<th>Layer</th>
<th>iNaturalist</th>
<th>SUN</th>
<th>Places</th>
<th>Textures</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
</tr>
<tr>
<td>Block 1</td>
<td>87.81</td>
<td>77.00</td>
<td>59.15</td>
<td>87.29</td>
<td>65.50</td>
</tr>
<tr>
<td>Block 2</td>
<td>71.84</td>
<td>85.80</td>
<td>61.44</td>
<td>86.46</td>
<td>71.68</td>
</tr>
<tr>
<td>Block 3</td>
<td>49.61</td>
<td>91.42</td>
<td>39.91</td>
<td>92.01</td>
<td>51.82</td>
</tr>
<tr>
<td>Block 4</td>
<td>46.54</td>
<td>91.49</td>
<td>27.88</td>
<td>92.18</td>
<td>38.26</td>
</tr>
</tbody>
</table>

**Block 3 and Block 4 features are the most informative.** In addition to exploring the high-level features at Block 3 and Block 4, we also investigate the possibility of applying RankFeat to features at shallow network layers. As shown in Table 6, the performances of RankFeat at the Block 1 and Block 2 features are not comparable to those at deeper layers. This is mainly because the shallow low-level features do not embed as rich semantic information as the deep features. Consequently, removing the rank-1 matrix of shallow features would not help to separate the ID and OOD data.

**Table 7:** The approximate solution by PI yields competitive performance and costs much less time consumption. The test batch size is set as 16.

<table>
<thead>
<tr>
<th>Computation Technique</th>
<th>Processing Time Per Image (ms)</th>
<th>iNaturalist</th>
<th>SUN</th>
<th>Places</th>
<th>Textures</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
<td>AUROC</td>
<td>FPR95</td>
<td>AUROC</td>
</tr>
<tr>
<td>GradNorm [27]</td>
<td>80.01</td>
<td>50.03</td>
<td>90.33</td>
<td>46.48</td>
<td>89.03</td>
<td>60.86</td>
</tr>
<tr>
<td>ReAct [52]</td>
<td>8.79</td>
<td>44.52</td>
<td>91.81</td>
<td>52.71</td>
<td>90.16</td>
<td>62.66</td>
</tr>
<tr>
<td>SVD</td>
<td>18.01</td>
<td>46.54</td>
<td>81.49</td>
<td>27.88</td>
<td>92.18</td>
<td>38.26</td>
</tr>
<tr>
<td>PI (#100 iter)</td>
<td>9.97</td>
<td>46.59</td>
<td>81.49</td>
<td>27.93</td>
<td>92.18</td>
<td>38.26</td>
</tr>
<tr>
<td>PI (#50 iter)</td>
<td>9.47</td>
<td>46.58</td>
<td>81.49</td>
<td>27.93</td>
<td>92.15</td>
<td>38.24</td>
</tr>
<tr>
<td>PI (#20 iter)</td>
<td>9.22</td>
<td>46.58</td>
<td>81.48</td>
<td>27.93</td>
<td>92.18</td>
<td>38.28</td>
</tr>
<tr>
<td>PI (#10 iter)</td>
<td>9.03</td>
<td>46.77</td>
<td>81.29</td>
<td>28.21</td>
<td>91.84</td>
<td>38.44</td>
</tr>
<tr>
<td>PI (#5 iter)</td>
<td>9.00</td>
<td>48.34</td>
<td>79.81</td>
<td>30.44</td>
<td>89.71</td>
<td>41.33</td>
</tr>
</tbody>
</table>

The approximate solution by PI yields competitive performances. Table 7 compares the time consumption and performance of SVD and PI, as well as two recent state-of-the-art OOD methods ReAct and GradNorm. The performance of PI starts to become competitive against that of SVD (<0.1%) from 20 iterations on with 48.41% time reduction. Compared with ReAct, the PI-based RankFeat only requires marginally 4.89% more time consumption. GradNorm is not comparable against other baselines in terms of time cost because it does not support the batch mode.

**Figure 4:** The impact of different fusion strategies on FPR95 and AUROC.

**Fusion at the logit space achieves the best performance.** Fig. 4 displays the performance of different fusion strategies in combining RankFeat at the Block 3 and Block 4 features. As can be
observed, averaging logits outperforms other fusion strategies in most datasets and metrics. This indicates that the fusing the logits can best coordinate the benefit of both features.

5 Related Work

Distribution shifts. Distribution shifts have been a long-standing problem in the machine learning research community [17, 45, 34, 62]. The problem of distributions shifts can be generally categorized as shifts in the input space and shifts in the label space. Shifts only in the input space are often deemed as covariate shifts [21, 44]. In this setting, the inputs are corrupted by perturbations or shifted by domains, but the label space stays the same [25, 53]. The aim is mainly to improve the robustness and generalization of a model [23]. For OOD detection, the labels are disjoint and the main concern is to determine whether a test sample should be predicted by the pre-trained model [39, 25].

Some related sub-fields also tackle the problem of distribution shifts in the label space, such as novel class discovery [16, 69], open-set recognition [48, 59], and novelty detection [1, 56]. These sub-fields target specific distribution shifts (e.g., semantic novelty), while OOD encompasses all forms of shifts.

OOD detection with discriminative models. The early work on discriminative OOD detection dates back to the classification model with rejection option [4, 12]. The OOD detection methods can be generally divided into training-need methods and post hoc approaches. Compared with training-needed approaches, post hoc methods do not require any extra training processes and could be directly applied to any pre-trained models. For the wide body of research on OOD detection, please refer to [65] for the comprehensive survey. Here we only highlight the representative post hoc methods. Nguyen et al. [43] first observed the phenomenon that neural networks easily give over-confident predictions for OOD samples. The following researches attempted to improve the OOD uncertainty estimation by proposing ODIN score [39], OpenMax score [2], Mahalanobis distance [38], and Energy score [40]. Huang et al. [26] pointed out that the traditional CIFAR benchmark does not extrapolate to real-world settings and proposed a large-scale ImageNet benchmark. More recently, Sun et al. [52] and Huang et al. [27] proposed to tackle the challenge of OOD detection from the lens of activation abnormality and gradient norm, respectively. In contrast, based on the empirical observation of singular value distributions, we propose a simple yet effective post hoc solution by removing the rank-1 subspace from the high-level features.

OOD detection with generative models. Different from discriminative models, generative models detect the OOD samples by estimating the probability density function [32, 55, 47, 57, 8, 28, 5, 50]. A sample with a low likelihood is deemed as OOD data. Recently, a multitude of methods have utilized generative models for OOD detection [46, 51, 61, 64, 33, 49, 31]. However, as pointed out in [42], generative models could assign a high likelihood to OOD data. Furthermore, generative models can be prohibitively harder to train and optimize than their discriminative counterparts, and the performance is often inferior. This might limit their practical usage.

6 Conclusion

In this paper, we present RankFeat, a simple yet effective approach for post hoc OOD detection by removing the rank-1 matrix composed by the largest singular value from the high-level feature. We demonstrate its superior empirical results and the general applicability across architectures, network depths, and benchmarks. Extensive ablation studies and comprehensive theoretical analyses are conducted to reveal the important insights and to explain the working mechanism of our method.

Acknowledgments and Disclosure of Funding

This research was supported by the EU H2020 projects AI4Media (No. 951911) and SPRING (No. 871245). We thank our colleague Zhun Zhong for the fruitful discussion and valuable suggestions.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We briefly discuss the slightly higher time consumption of RankFeat in Sec. 4.3.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not anticipate any negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Since our method is post hoc, it will not be influenced by random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The used data do not contain this kind of information.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]