Out-of-distribution Detection with Deep Nearest Neighbors

Yiyou Sun
UW-Madison

Yifei Ming
UW-Madison

Xiaojin Zhu
UW-Madison

Yixuan Li
UW-Madison
Problem Introduction
**Closed-world:** Training and testing distributions **match**

**Open-world:** Training and testing distributions **differ**
Training examples:
traffic signs

$\begin{align*}
  p_i(x) &= \frac{\exp(f_i(x))}{\sum_{j=1}^{N} \exp(f_j(x))},
\end{align*}$

confidence: $\max_i p_i$

High confidence in classifying traffic signs.

Why is OOD Detection hard?
Test time: **out-of-distribution example**

*Ideally: Low confidence in predicting as traffic sign*
Neural Networks Can Be Over-confident to Out-of-distribution Examples

[Nguyen et al. 2015]

In-Distribution (ID) Samples

Out-of-distribution (OOD) Samples
Prior literature on distance-based OOD detection

- Mahalanobis
  [Lee et al. NeurIPS 2018]
- CSI
  [Tack et al. NeurIPS 2020]
- SSD
  [Sehwag et al. ICLR 2021]

In-Distribution (ID) Embeddings

Out-of-distribution (OOD) Embeddings
Can we leverage the non-parametric nearest neighbor approach for OOD detection?
Methodology
Problem Statement

• Supervised learning setting
  • input space: $X = \mathbb{R}^d$
  • label space: $Y = \{1, 2, \ldots, K\}$
  • a training dataset drawn i.i.d. from a joint distribution: $P = X \times Y$
  • training a neural network: $f(x; \theta)$

• In test time, test data can come from a different distribution
  • whose label set has not interaction with $Y$
  • should not be predicted by model

• OOD detection can be made via a level set estimation

$$G_\lambda(x; f) = \begin{cases} 
    \text{ID} & S(x; f) \geq \lambda \\
    \text{OOD} & S(x; f) < \lambda 
  \end{cases}$$
Set up

For all $x_i$ in the training data, collect embeddings: $Z_n = (z_1, z_2, \ldots, z_n)$
OOD Detection with k-NN Distance

1. Given a test sample $x^*$, we calculate feature vector $z^* = \phi(x^*) / \|\phi(x^*)\|_2$

2. Calculate the Euclidean distances $\|z_i - z^*\|_2$ with respect to embedding vectors $z_i \in \mathbb{Z}_n$

3. Reorder $\mathbb{Z}_n$ according to the increasing value of $\|z_i - z^*\|_2$, Denote the reordered data sequence as $\mathbb{Z}'_n = (z(1), z(2), \ldots, z(n))$

4. Calculate distance to the k-th nearest neighbor (k-NN): $r_k(z^*) = \|z^* - z(k)\|_2$

5. The decision function is given by $G_\lambda(x^*; \lambda) = \begin{cases} 
\text{ID} & -r_k(z^*) \geq \lambda \\
\text{OOD} & -r_k(z^*) < \lambda 
\end{cases}$

(The threshold $\lambda$ is typically chosen so that a high fraction of ID data (e.g., 95%) is correctly classified.)
Experiments
(1) KNN:
Cross-entropy Loss

$$L_{\text{ce}} = - \sum_{i=1}^{C} y_i \log(p_i), \text{ for } C \text{ classes}$$

(2) KNN+:
Supervised Contrastive Loss

[Khosla, et al. 2020]
Experiments Setting: Backbone and Datasets

Model

ID
CIFAR-10
ImageNet-1k

OOD
SVHN
SUN
Places
Textures

ID
iNaturalist
SUN
Places
Textures
### Table 1. Results on CIFAR-10. Comparison with competitive OOD detection methods. All methods are based on a discriminative model trained on ID data only, without using outlier data. ↑ indicates larger values are better and vice versa.

<table>
<thead>
<tr>
<th>Method</th>
<th>SVHN</th>
<th>LSUN</th>
<th>iSUN</th>
<th>Texture</th>
<th>Places365</th>
<th>Average</th>
<th>ID ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPR↓</td>
<td>AUROC↑</td>
<td>FPR↓</td>
<td>AUROC↑</td>
<td>FPR↓</td>
<td>AUROC↑</td>
<td></td>
</tr>
<tr>
<td>MSP</td>
<td>59.66</td>
<td>91.25</td>
<td>45.21</td>
<td>93.80</td>
<td>54.57</td>
<td>92.12</td>
<td>62.46</td>
</tr>
<tr>
<td>ODIN</td>
<td>20.93</td>
<td>95.55</td>
<td>7.26</td>
<td>98.53</td>
<td>33.17</td>
<td>94.65</td>
<td>56.40</td>
</tr>
<tr>
<td>Energy</td>
<td>54.41</td>
<td>91.22</td>
<td>10.19</td>
<td>98.05</td>
<td>27.52</td>
<td>95.59</td>
<td>55.23</td>
</tr>
<tr>
<td>GODIN</td>
<td>15.51</td>
<td>96.60</td>
<td>4.90</td>
<td>99.07</td>
<td>34.03</td>
<td>94.94</td>
<td>46.91</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>9.24</td>
<td>97.80</td>
<td>67.73</td>
<td>73.61</td>
<td>6.02</td>
<td>98.63</td>
<td>23.21</td>
</tr>
<tr>
<td>KNN (ours)</td>
<td>24.53</td>
<td>95.96</td>
<td>25.29</td>
<td>95.69</td>
<td>25.55</td>
<td>95.26</td>
<td>27.57</td>
</tr>
</tbody>
</table>

#### Without Self-supervised Learning

#### With Self-supervised Learning

- KNN+ (ours) 2.42 99.52 1.78 99.48 20.06 96.74 8.09 98.56 23.02 95.36

---

- **Bold values indicate superior performance.**
Contrastively Learned Representation Helps More Distinguishable
K-NN Distance Achieves Superior Performance (ImageNet)

Table 4. Results on ImageNet. All methods are based on a model trained on ID data only (ImageNet-1k (Deng et al., 2009)). We report the OOD detection performance, along with the per-image inference time.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Inference time (ms)</th>
<th>iNaturalist</th>
<th>SUN</th>
<th>Places</th>
<th>Textures</th>
<th>Average</th>
<th>ID ACC</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>
<td>FPR95</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↑</td>
<td>↓</td>
<td>↑</td>
<td>↓</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>MSP</td>
<td>7.04</td>
<td>54.99</td>
<td>87.74</td>
<td>70.83</td>
<td>80.86</td>
<td>73.99</td>
<td>79.76</td>
</tr>
<tr>
<td>ODIN</td>
<td>7.05</td>
<td>47.66</td>
<td>89.66</td>
<td>60.15</td>
<td>84.59</td>
<td>67.89</td>
<td>81.78</td>
</tr>
<tr>
<td>Energy</td>
<td>7.04</td>
<td>55.72</td>
<td>89.95</td>
<td>59.26</td>
<td>85.89</td>
<td>64.92</td>
<td>82.86</td>
</tr>
<tr>
<td>GODIN</td>
<td>7.04</td>
<td>61.91</td>
<td>85.40</td>
<td>60.83</td>
<td>85.60</td>
<td>63.70</td>
<td>83.81</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>35.83</td>
<td>97.00</td>
<td>52.65</td>
<td>98.50</td>
<td>42.41</td>
<td>98.40</td>
<td>41.79</td>
</tr>
<tr>
<td>KNN (α = 100%)</td>
<td>10.31</td>
<td>59.00</td>
<td>86.47</td>
<td>68.82</td>
<td>80.72</td>
<td>76.28</td>
<td>75.76</td>
</tr>
<tr>
<td>KNN (α = 1%)</td>
<td>7.04</td>
<td>59.08</td>
<td>86.20</td>
<td>69.53</td>
<td>80.10</td>
<td>77.09</td>
<td>74.87</td>
</tr>
<tr>
<td>Rotation</td>
<td>7.04</td>
<td>76.65</td>
<td>82.63</td>
<td>88.54</td>
<td>54.01</td>
<td>88.14</td>
<td>53.63</td>
</tr>
<tr>
<td>SSD+</td>
<td>28.31</td>
<td>57.16</td>
<td>87.77</td>
<td>78.23</td>
<td>73.10</td>
<td>81.19</td>
<td>70.97</td>
</tr>
<tr>
<td>KNN+ (α = 100%)</td>
<td>10.47</td>
<td>30.18</td>
<td>94.89</td>
<td>48.99</td>
<td>88.63</td>
<td>59.15</td>
<td>84.71</td>
</tr>
<tr>
<td>KNN+ (α = 1%)</td>
<td>7.04</td>
<td>30.83</td>
<td>94.72</td>
<td>48.91</td>
<td>88.40</td>
<td>60.02</td>
<td>84.62</td>
</tr>
</tbody>
</table>

CLNN outperforms the competitive baseline by **24.77%** in false positive rate (FPR95)
# K-NN Distance Has Similar Inference Speed

Table 4. **Results on ImageNet.** All methods are based on a model trained on ID data only (ImageNet-1k (Deng et al., 2009)). We report the OOD detection performance, along with the per-image inference time.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Inference time (ms)</th>
<th>iNaturalist FPR95</th>
<th>AUROC</th>
<th>SUN FPR95</th>
<th>AUROC</th>
<th>Places FPR95</th>
<th>AUROC</th>
<th>Textures FPR95</th>
<th>AUROC</th>
<th>Average FPR95</th>
<th>AUROC</th>
<th>ID ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSP</td>
<td>7.04</td>
<td>54.99</td>
<td>87.74</td>
<td>70.83</td>
<td>80.86</td>
<td>73.99</td>
<td>79.76</td>
<td>68.00</td>
<td>79.61</td>
<td>66.95</td>
<td>81.99</td>
<td>76.65</td>
</tr>
<tr>
<td>ODIN</td>
<td>7.05</td>
<td>47.66</td>
<td>89.66</td>
<td>60.15</td>
<td>84.59</td>
<td>67.89</td>
<td>81.78</td>
<td>50.23</td>
<td>85.62</td>
<td>56.48</td>
<td>85.41</td>
<td>76.65</td>
</tr>
<tr>
<td>Energy</td>
<td>7.04</td>
<td>55.72</td>
<td>89.95</td>
<td>59.26</td>
<td>85.89</td>
<td>64.92</td>
<td>82.86</td>
<td>53.72</td>
<td>85.99</td>
<td>58.41</td>
<td>86.17</td>
<td>76.65</td>
</tr>
<tr>
<td>GODIN</td>
<td>7.04</td>
<td>61.91</td>
<td>85.40</td>
<td>60.83</td>
<td>85.60</td>
<td>63.70</td>
<td>83.81</td>
<td>77.85</td>
<td>73.27</td>
<td>66.07</td>
<td>82.02</td>
<td>70.43</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>35.83</td>
<td>97.00</td>
<td>52.65</td>
<td>98.50</td>
<td>42.41</td>
<td>98.40</td>
<td>41.79</td>
<td>55.80</td>
<td>85.01</td>
<td>87.43</td>
<td>55.47</td>
<td>76.65</td>
</tr>
<tr>
<td>KNN (α = 100%)</td>
<td>10.31</td>
<td>59.00</td>
<td>86.47</td>
<td>68.82</td>
<td>80.72</td>
<td>76.28</td>
<td>75.76</td>
<td>11.77</td>
<td>97.07</td>
<td>53.97</td>
<td>85.01</td>
<td>76.65</td>
</tr>
<tr>
<td>KNN (α = 1%)</td>
<td>7.04</td>
<td>59.08</td>
<td>86.20</td>
<td>69.53</td>
<td>80.10</td>
<td>77.09</td>
<td>74.87</td>
<td>11.56</td>
<td>97.18</td>
<td>54.32</td>
<td>84.59</td>
<td>76.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Without Self-supervised Learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rotation</td>
<td>7.04</td>
<td>76.65</td>
<td>82.63</td>
<td>88.54</td>
<td>54.01</td>
<td>88.14</td>
<td>53.63</td>
<td>70.71</td>
<td>82.29</td>
<td>81.01</td>
<td>68.14</td>
<td>73.10</td>
</tr>
<tr>
<td>SSD+</td>
<td>28.31</td>
<td>57.16</td>
<td>87.77</td>
<td>78.23</td>
<td>73.10</td>
<td>81.19</td>
<td>70.97</td>
<td>36.37</td>
<td>88.52</td>
<td>63.24</td>
<td>80.09</td>
<td>79.10</td>
</tr>
<tr>
<td>KNN+ (α = 100%)</td>
<td>10.47</td>
<td>30.18</td>
<td>94.89</td>
<td>48.99</td>
<td>88.63</td>
<td>59.15</td>
<td>84.71</td>
<td>15.55</td>
<td>95.40</td>
<td><strong>38.47</strong></td>
<td><strong>90.91</strong></td>
<td><strong>79.10</strong></td>
</tr>
<tr>
<td>KNN+ (α = 1%)</td>
<td>7.04</td>
<td>30.83</td>
<td>94.72</td>
<td>48.91</td>
<td>88.40</td>
<td>60.02</td>
<td>84.62</td>
<td>16.97</td>
<td>94.45</td>
<td>39.18</td>
<td>90.55</td>
<td><strong>79.10</strong></td>
</tr>
<tr>
<td></td>
<td><strong>With Self-supervised Learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Other Experiments

1. Ablation on k and random sampling ratio
2. Nearest Neighbor Approach is competitive on hard OOD datasets
3. Comparison with other non-parametric methods
4. Nearest Neighbor Approach is competitive on ViT
5. KNN can be further boosted by activation rectification
6. Ablation on the penultimate layer and the projection layer
7. Ablation on the k-th and the averaged k-NN distance
8. Results on different architectures

(See more details in the paper)
Discussion and Theoretical Insight
Feature normalization is critical to KNN’s success
Theoretical Insight

• **Set Up:** We consider OOD detection task as a special binary classification task, where the negative samples (OOD) are only available in the testing stage.

• **Main Result:** KNN-based OOD detector can reject inputs equivalent to the estimated Bayesian binary decision function.

**Theorem 6.1.** If \( \hat{p}_{out}(z_i) = \hat{c}_0 \mathbf{1} \left\{ \hat{p}_{in}(z_i; k, n) < \frac{\beta \epsilon \hat{c}_0}{(1-\beta)(1-\epsilon)} \right\} \), and \( \lambda = -m^{-1} \sqrt{\frac{(1-\beta)(1-\epsilon)k}{\beta \epsilon c_b n \hat{c}_0}} \), we have

\[
\mathbf{1} \left\{ -r_k(z_i) \geq \lambda \right\} = \mathbf{1} \left\{ \hat{p}(g_i = 1 \mid z_i) \geq \beta \right\}
\]
Summary

1. We present the first study exploring and demonstrating the efficacy of non-parametric density estimation with nearest neighbors for OOD detection.

2. We demonstrate the superior performance of the KNN-based method on several OOD detection benchmarks, different model architectures.

3. We offer new insights on the key components to make KNN effective in practice, including feature normalization and a compact representation space.

4. We provide theoretical analysis, showing that KNN-based OOD detection can reject inputs equivalent to the Bayes optimal estimator.
Thank you!

https://github.com/deeplearning-wisc/knn-ood