Igeood: An Information Geometry Approach to Out-of-Distribution Detection
Eduardo Dadalto Câmara Gomes, Florence Alberge, Pierre Duhamel, Pablo Piantanida

To cite this version:
Eduardo Dadalto Câmara Gomes, Florence Alberge, Pierre Duhamel, Pablo Piantanida. Igeood: An Information Geometry Approach to Out-of-Distribution Detection. NeurIPS DistShift Workshop 2021, Dec 2021, Virtual, France. hal-03649034

HAL Id: hal-03649034
https://centralesupelec.hal.science/hal-03649034
Submitted on 22 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
IGEOD: An Information Geometry Approach to Out-of-Distribution Detection

Eduardo D. C. Gomes¹, Florence Alberge¹, Pierre Duhamel¹ and Pablo Piantanida¹

Abstract

In this paper, we introduce Icsoft, an effective method for detecting Out-of-Distribution (OOD) samples. Icsoft applies to any pre-trained neural network, works under different degrees of access to the ML model, does not require OOD samples or assumptions on the OOD data but can also benefit (if available) from OOD samples. By building on the geodesic (Fisher-Rao) distance between the underlying data distributions, our discriminator combines confidence scores from the logits outputs and the learned features of a deep neural network.

Background

Let \( \mathcal{X} \subseteq \mathbb{R}^d \) be the feature space and \( \mathcal{Y} \) a label space and let \( p_{\text{std}} \) be the underlying unknown probability density function (pdf) over \( \mathcal{X} \times \mathcal{Y} \).

In order to model the underlying problem, we introduce an artificial binary random variable \( Z \in \{0, 1\} \) indicating with \( z = 1 \) that the test sample \( x \) is OOD and \( z = 0 \) otherwise.

The open-world data can then be modeled as a mixture distribution \( p_{\text{mix}} \) defined by

\[
p_{\text{mix}}(x|z=0) = p_{\text{std}}(x), \quad p_{\text{mix}}(x|z=1) = q_{x}(x).
\]

The intrinsic difficulty arises from the fact that very little can be assumed about the unknown distributions \( p_{\text{std}} \) and \( q_{\text{std}} \) in particular for out-of-distribution.

Alternative: distance based criteria w.r.t an in-distribution probability reference.

Statistical Model

![Image](image.png)

Figure: We model the hidden layers’ outputs as class conditional Gaussian distributions and the DNN’s outputs as softmax probability distributions.

OOD Detection using the Fisher-Rao Distance

- Fisher-Rao distance: Let \( q_{\theta} (\cdot|f) \) be a probability distribution with parameters \( \theta \).

\[
d_{\text{FR}}(x) = \min_{\theta} \text{dist-gauss} \left( \mathcal{x}, \sigma^{(i)}, (\mu^{(i)}, \sigma^{(i)}) \right)
\]

- Feature ensemble: we combine the confidence scores of the logits and low-level features through a linear combination. If OOD data is available, we can also calculate \( \text{FR} \) with OOD statistics, obtaining Icsoft +.

\[
\text{FR}(x) = a_0 \text{FR}(x) + a_1 \text{FR}(x) + a_2 \text{FR}(x)
\]

Therefore, we have derived a unified OOD detection framework that combines a single distance for both the softmax outputs and the latent features of a neural network.

Experimental Results

- The Icsoft score increases the separation between in- and out-of-distribution data.

| Model | In-dist. Mahalanobis | / OOD+ (ours) | Mahalanobis: 63% | Mahalanobis: 87% | Mahalanobis: 48% |
|-------|----------------------|--------------|------------------|------------------|------------------|
| C-10  | 76.6 ±8/92.1/92.1 ±8 | 91.7 ±12/98.4 ±8 | 79.9 ±30/79.9 ±30 | 78.9 ±30/79.9 ±30 | 91.9 ±72/94.0 ±7 | 57.4 ±36/65.1 ±33 | 86.7 ±19/91.5 ±15 |
| DenseNet-C100 | 67.2±24/90.2 ±24 | 90.2 ±13/97.0 ±13 | 75.9 ±30/79.7 ±30 | 60.4 ±34/70.9 ±34 | 85.3 ±19/90.8 ±19 | 94.9 ±36/99.6 ±36 | 93.7 ±72/99.2 ±72 |
| SVHN | 93.8 ±8/99.0 ±8 | 98.6 ±8/99.9 ±8 | 79.2 ±24/79.2 ±24 | 81.0 ±24/81.0 ±24 | 98.6 ±24/98.4 ±24 | 97.6 ±24/97.6 ±24 | 93.8 ±72/97.0 ±72 |
| LSUN | 98.8 ±8/99.2 ±8 | 99.9 ±13/99.9 ±13 | 91.7 ±24/91.7 ±24 | 99.9 ±24/99.9 ±24 | 99.9 ±24/99.9 ±24 | 97.6 ±24/97.6 ±24 | 93.8 ±72/97.0 ±72 |
| Tiny ImgNet | 97.1 ±8/98.7 ±8 | 97.8 ±8/96.3 ±8 | 78.6 ±24/78.6 ±24 | 81.0 ±24/81.0 ±24 | 98.6 ±24/98.4 ±24 | 99.9 ±24/99.9 ±24 | 93.8 ±72/97.0 ±72 |
| iSUN | 97.8 ±8/99.3 ±8 | 99.9 ±13/99.9 ±13 | 91.7 ±24/91.7 ±24 | 99.9 ±24/99.9 ±24 | 99.9 ±24/99.9 ±24 | 97.6 ±24/97.6 ±24 | 93.8 ±72/97.0 ±72 |

| Model | SVHN/C-10 | iSUN | Tiny ImgNet | LSUN | Tiny ImgNet | iSUN | LSUN |
|-------|----------|------|-------------|------|-------------|------|------|
| C-10  | 87.8/97.6/ -/96.5/98.8/ | 97.8/99.3/ -/ -/97.2/ | 97.8/99.3/ -/ -/97.2/ | 98.8/99.2/ -/ -/97.2/ | 97.8/99.3/ -/ -/97.2/ | 98.8/99.2/ -/ -/97.2/ | 98.8/99.2/ -/ -/97.2/ |

Average and standard deviation of OOD detection performance for the outputs of each hidden block of a DenseNet model on CIFAR-10 (in-distribution) and SVHN (out-of-distribution) dataset.

- We increase the average TNR-95% by 11.8% and 25% with validation on OOD and adversarial data, respectively.

Table: Average and standard deviation of OOD detection performance for the Wurts-Box settings. The abbreviation TNR-95%, C-10 and C-100 stands for TNR at TPR-95%, CIFAR-10 and CIFAR-100, respectively.

![Image](image.png)

Figure: Histograms of the Mahalanobis and Icsoft scores for the outputs of each hidden block of a DenseNet model on CIFAR-10 (in-distribution) and SVHN (out-of-distribution) datasets.

\[ ^1 \text{Laboratoire des signaux et systèmes (L2S), Université Paris-Saclay CNRS CentraleSupélec, 91190 Gif-sur-Yvette, France.} \]