Igeood: An Information Geometry Approach to Out-of-Distribution Detection

Eduardo Dadalto Câmara Gomes, Florence Alberge, Pierre Duhamel, Pablo Piantanida

IGEOOD: An Information Geometry Approach to Out-of-Distribution Detection

Eduardo D. C. Gomes[1], Florence Alberge[1], Pierre Duhamel[1] and Pablo Piantanida[1]

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

- In this paper, we introduce IGEOOD, an effective method for detecting Out-of-Distribution (OOD) samples.
- IGEOOD 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_{XY} 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 $\boldsymbol{x}$ is OOD and $z = 0$ otherwise.
- The open-world data can then be modeled as a *mixture* distribution $p_{X|Z}$ defined by

$$p_{X|Z}(\boldsymbol{x}|z=0) \triangleq p_X(\boldsymbol{x}), \quad p_{X|Z}(\boldsymbol{x}|z=1) \triangleq q_X(\boldsymbol{x}).$$

- The intrinsic difficulty arises from the fact that very little can be assumed about the unknown distributions p_X and q_X, in particular for out-of-distribution.
- **Alternative:** distance based criteria w.r.t an in-distribution probability reference.

Statistical Model

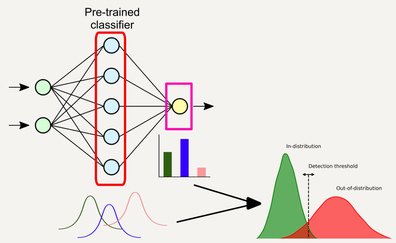


Figure: We model the hidden layers' outputs as class conditional Gaussian distributions and the DNN's outputs as softmax probability distributions.

OOD Detection using the Fisher-Rao Distance

- **Fisher-Rao distance**: Let $q_\theta(\cdot|\cdot)$ be a probability distribution with parameters θ.

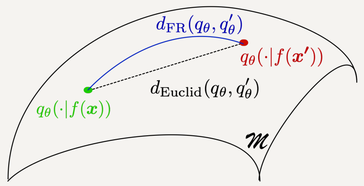


Figure: Illustration of the shortest path between probability distributions without leaving the underlying statistical manifold.

$$d_{\mathrm{FR}}(q_\theta, q'_\theta) \triangleq \inf_{\gamma} \int_0^1 \sqrt{\frac{d\gamma^\top}{dt} \boldsymbol{G}(\gamma) \frac{d\gamma}{dt}} dt$$

Where $\gamma(0) = \theta$, $\gamma(1) = \theta'$, $\boldsymbol{G}$ is the Fisher Information Matrix and dt is the infinitesimal length element.

- **IGEOOD score using the softmax probability:** Let $q_\theta(\cdot|f(\boldsymbol{x}))$ be the softmax probability distribution of the outputs. We can define the Fisher-Rao distance between softmax distributions as:

$$d_{\mathrm{FR-Logits}}\big(q_\theta(\cdot|f(\boldsymbol{x})), q_\theta(\cdot|f(\boldsymbol{x}'))\big) \triangleq 2 \arccos\left(\sum_{y\in\mathcal{Y}} \sqrt{q_\theta\big(y|f(\boldsymbol{x})\big) q_\theta\big(y|f(\boldsymbol{x}')\big)}\right)$$

  - From which we derive the IGEOOD score for the logits.

$$\mathrm{FR}_0(\boldsymbol{x}) \triangleq \sum_{y\in\mathcal{Y}} d_{\mathrm{FR-Logits}}\big(q_\theta(\cdot|f(\boldsymbol{x})), q_\theta(\cdot|\boldsymbol{\mu}_y)\big)$$

  - Where $\boldsymbol{\mu}_y$ are the class conditional centroids given by:

$$\boldsymbol{\mu}_y \triangleq \min_{\boldsymbol{\mu}\in\mathbb{R}^{|\mathcal{Y}|}} \frac{1}{N_y} \sum_{\forall\, i:\, y_i=y} d_{\mathrm{FR-Logits}}\big(q_\theta(\cdot|f(\boldsymbol{x}_i)), q_\theta(\cdot|\boldsymbol{\mu})\big).$$

- **IGEOOD score leveraging latent features:** For each layer, we model the features as a set of class-conditional Gaussian distributions with diagonal standard deviation matrix. The distribution parameters are calculated accordingly.

$$\boldsymbol{\mu}_y^{(\ell)} = \frac{1}{N_y} \sum_{\forall\, i:\, y_i=y} f^{(\ell)}(\boldsymbol{x}_i), \quad \text{and} \quad \boldsymbol{\sigma}^{(\ell)} = \mathrm{diag}\left(\sqrt{\frac{1}{N} \sum_{y\in\mathcal{Y}} \sum_{\forall\, i:\, y_i=y} \left(f_j^{(\ell)}(\boldsymbol{x}_i) - \mu_{y,j}^{(\ell)}\right)^2}\right)$$

  - We derive a confidence score by calculating the Fisher-Rao distance between the test sample $\boldsymbol{x}$ and the closest class-conditional diagonal Gaussian distribution.

$$\mathrm{FR}_\ell(\boldsymbol{x}) = \min_{y\in\mathcal{Y}} d_{\mathrm{FR-Gauss}}\left(\left(\boldsymbol{x}, \boldsymbol{\sigma}^{(\ell)}\right), \left(\boldsymbol{\mu}_y^{(\ell)}, \boldsymbol{\sigma}^{(\ell)}\right)\right)$$

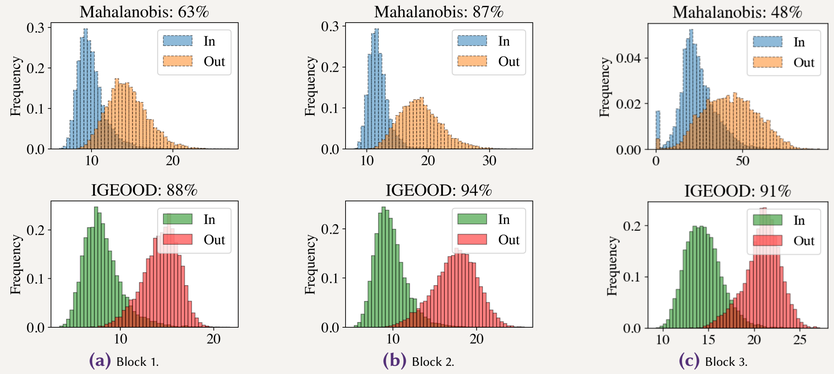
- **Feature ensemble:** we combine the confidence scores of the logits and low-level features through a linear combination. If OOD data is available, we can also calculate $\mathrm{FR}'_\ell\left(\boldsymbol{x}; \boldsymbol{\mu}^{(\ell)\prime}, \boldsymbol{\sigma}^{(\ell)\prime}\right)$ with OOD statistics, obtaining IGEOOD+.

$$\mathrm{FR}(\boldsymbol{x}) \triangleq \alpha_0 \mathrm{FR}_0(\boldsymbol{x}) + \sum_\ell \alpha_\ell \cdot \mathrm{FR}_\ell(\boldsymbol{x}) + \alpha'_\ell \cdot \mathrm{FR}'_\ell(\boldsymbol{x})$$

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



(a) Block 1. **(b)** Block 2. **(c)** Block 3.

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

- We increase the average TNR-95% by 11.8% and 2.5% with validation on OOD and adversarial data, respectively.

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

		Validation on OOD data		Validation on adversarial data	
		TNR-95%	AUROC	TNR-95%	AUROC
Model	In-dist.	Mahalanobis / IGEOOD+ (ours)		Mahalanobis / IGEOOD (ours)	
DenseNet	C-10	76.6±31/**92.6**±14	92.1±12/**98.4**±3.0	75.9±30/**77.9**±29	91.7±12/**94.0**±9.0
	C-100	67.2±28/**90.2**±21	90.2±13/**97.7**±5.0	60.4±34/**70.9**±35	85.3±19/**90.8**±13
	SVHN	93.3±8.0/**98.0**±2.0	98.6±1.0/**99.6**±0.1	**93.7**±10/92.2±9.0	**98.6**±2.0/98.4±1.0
ResNet	C-10	82.5±23/**91.6**±16	96.5±4.0/**98.4**±3.0	**78.6**±24/77.3±32	**95.3**±6.0/90.0±15
	C-100	70.4±30/**86.4**±23	91.9±10/**97.1**±5.0	57.4±36/**65.1**±33	86.9±13/**88.6**±15
	SVHN	96.8±6.0/**98.9**±2.0	99.2±1.0/**99.7**±0.1	**96.3**±8.0/93.6±14	**99.1**±1.0/98.4±3.0
Average and Std.		81.1±11/**92.9**±4.0	94.8±4.0/**98.5**±1.0	77.0±15/**79.5**±10	92.8±5.4/**93.4**±3.9

Table: TNR at TPR-95% (%) performance comparison in a WHITE-BOX setting considering the original results from [1,2,3,4]. Methods with a (*) were tuned without OOD data.

	OOD	CIFAR-10	CIFAR-100	SVHN
	dataset	Mahalanobis [1] / Gram Matrix* [2] / DeConf-C* [3] / Res-Flow [4] / IGEOOD / IGEOOD+		
DenseNet	iSUN	95.3/99.0/ - / - /97.7/**99.8**	87.0/95.9/ - / - /93.8/**99.7**	**99.9**/99.4/ - / - /98.3/**99.9**
	LSUN	97.2/99.5/99.4/98.2/98.5/**99.9**	91.4/97.2/98.7/96.3/95.2/**99.9**	**99.9**/99.5/ - /**100**/97.1/**99.9**
	TinyImgNet	95.0/98.8/99.1/96.4/95.7/**99.8**	86.6/95.7/98.6/93.0/94.5/**99.5**	**99.9**/99.1/ - /**100**/98.2/**99.9**
	SVHN/C-10	90.8/96.1/98.8/94.9/98.9/**99.9**	82.5/89.3/95.9/84.9/93.3/**99.6**	96.8/80.4/ - /**99.0**/91.6/98.3
ResNet	iSUN	97.8/99.3/ - / - /97.2/**99.9**	89.9/94.8/ - / - /93.4/**99.8**	99.7/99.4/ - / - /99.8/**100**
	LSUN	98.8/99.6/ - /99.0/98.4/**100**	90.9/96.6/ - /96.2/94.3/**100**	**99.9**/99.6/ - /**100**/99.7/**99.9**
	TinyImgNet	97.1/98.7/ - /97.8/96.3/**99.6**	90.9/94.8/ - /94.6/90.1/**99.6**	**99.9**/99.3/ - /**100**/99.7/**99.9**
	SVHN/C-10	87.8/97.6/ - /96.5/98.8/**99.8**	91.9/80.8/ - /93.0/91.6/**99.7**	98.4/85.8/ - /99.4/97.7/**99.7**

[1] Kimin L. et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018.
[2] Sastry & Oore. Detecting out-of-distribution examples with Gram matrices, 2020.
[3] Hsu et al. Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data, 2020.
[4] Zisselman & Tamar. Deep residual flow for novelty detection, 2020.

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