# Feature Space Singularity for Out-of-Distribution Detection

Haiwen Huang, Zhihan Li, LuluWang, Sishuo Chen, Bin Dong, Xinyu Zhou
Corresponding: haiwen.huang2@cs.ox.ac.uk

NEURAL INFORMATION PROCESSING SYSTEMS

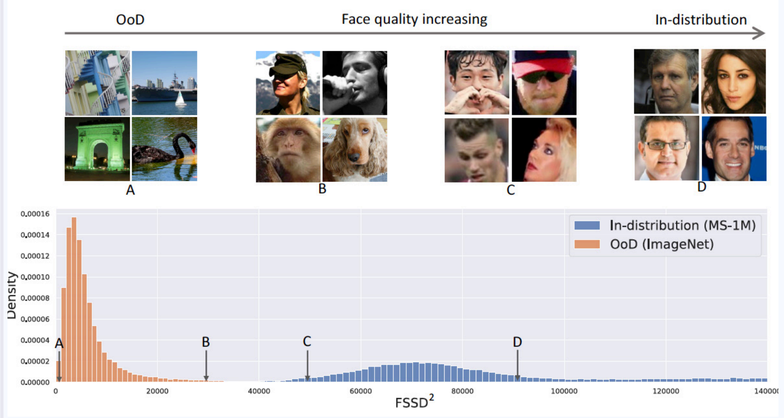
北京大学 PEKING UNIVERSITY

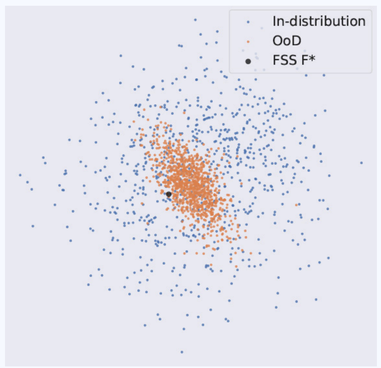
MEGVII 旷视

UNIVERSITY OF OXFORD

## Observation

In a trained NN, OoD samples concentrate in the feature space.

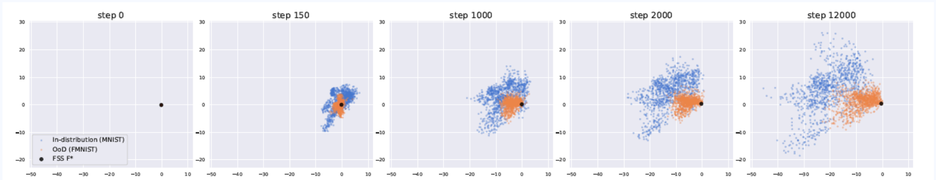


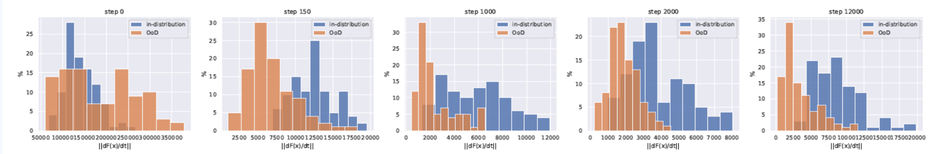

$$\mathrm{FSSD}\left(x\right) := \left\| F_\theta\left(x\right) - F^* \right\| \qquad (1)$$

## Understanding

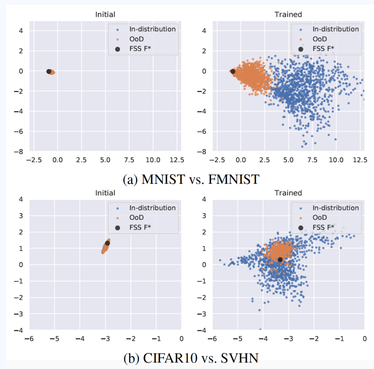
OoD features move slowly during training.




(a) The dynamics of features $F_{\theta_t}\left(x\right)$.




(b) The norm of the derivative, i.e., "moving speed", of last-layer feature vector at different time steps.




(a) MNIST vs. FMNIST

(b) CIFAR10 vs. SVHN

All features concentrate at initial stage.

"Moving speed" of the feature vector

$$\begin{aligned}\frac{\mathrm{d}F_{\theta_t}\left(x\right)}{\mathrm{d}t} &= \frac{\partial F_{\theta_t}\left(x\right)}{\partial \theta_t}\frac{\mathrm{d}\theta_t}{\mathrm{d}t} \\ &= -\sum_{m=1}^{M}\frac{\partial F_{\theta_t}\left(x\right)}{\partial \theta_t}\frac{\partial F_{\theta_t}\left(x_m\right)^{\mathsf{T}}}{\partial \theta_t}\partial_m \mathcal{L}_\phi .\end{aligned}$$

Empirical Neural Tangent Kernel

Integrate

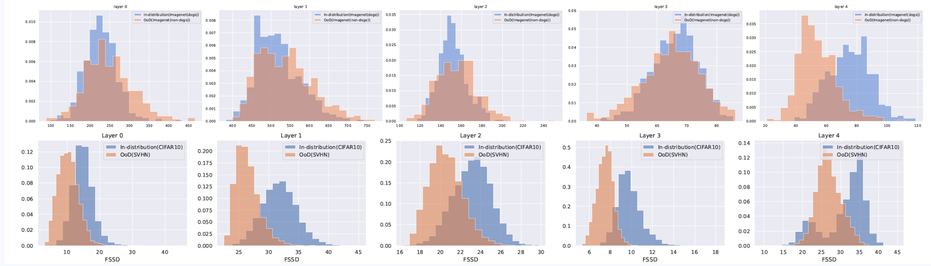
$$\begin{aligned}\mathrm{FSSD}\left(x\right) &\overset{\substack{F^*\approx F_{\theta_0}(x) \\ \text{Equation (1)}}}{\approx} \left\| F_{\theta_t}\left(x\right) - F_{\theta_0}\left(x\right) \right\| \\ &= \left\| \sum_{m=1}^{M}\int_0^T \Theta_t\left(x, x_m\right)\partial_m \mathcal{L}_\phi \,\mathrm{d}t \right\| \\ &\approx \left\| \sum_{m=1}^{M} \Theta\left(x, x_m\right)\nu_m \right\|,\end{aligned}$$

where $\nu_m = \int_0^T \partial_m \mathcal{L}_\phi \,\mathrm{d}t$.

## Experiments

Table 2: Main results. All values are in %.

| | Datasets (Architecture) | Metrics | *Base* | *ODIN* | *Maha* | *DE* | *MCD* | *OE* | *FSSD* |
|---|---|---|---|---|---|---|---|---|---|
| Small-scale benchmarks | FMNIST vs. MNIST (LeNet) | AUROC | 77.3 | 96.9 | **99.6** | 83.9 | 81.7 | **99.6** | **99.6** |
| | | AUPRC | 79.2 | 93.0 | **99.7** | 83.3 | 85.3 | 99.6 | **99.7** |
| | | FPR80 | 43.5 | 2.5 | **0.0** | 27.5 | 36.8 | **0.0** | **0.0** |
| | CIFAR10 vs. SVHN (ResNet34) | AUROC | 89.9 | 96.7 | 99.1 | 93.7 | 96.7 | 90.4 | **99.5** |
| | | AUPRC | 85.4 | 92.5 | 98.1 | 90.6 | 93.9 | 89.8 | **99.5** |
| | | FPR80 | 10.1 | 4.7 | **0.3** | 3.7 | 2.4 | 12.5 | 0.4 |
| | ImageNet dogs vs. non-dogs (ResNet34) | AUROC | 88.5 | 90.8 | 83.3 | 89.0 | 67.2 | 92.5 | **93.1** |
| | | AUPRC | 86.1 | 88.6 | 83.0 | 89.0 | 66.9 | **92.6** | 92.5 |
| | | FPR80 | 19.5 | 15.2 | 30.1 | 18.8 | 59.2 | **7.9** | 10.2 |
| Large-scale benchmarks | CelebA non-blurry vs. blurry (ResNeXt50) | AUROC | 71.7 | 73.3 | 73.9 | 74.5 | 69.8 | 71.5 | **78.3** |
| | | AUPRC | 89.9 | 91.4 | 90.9 | 91.4 | 88.7 | 90.7 | **92.8** |
| | | FPR80 | 52.0 | 50.3 | 46.0 | 47.1 | 53.2 | 54.2 | **39.2** |
| | MS-1M vs. IJB-C (ResNeXt50) | AUROC | 60.0 | 61.3 | 82.5 | 63.0 | 65.5 | 52.6 | **86.7** |
| | | AUPRC | 53.3 | 55.9 | 80.6 | 56.1 | 59.4 | 46.6 | **86.1** |
| | | FPR80 | 61.8 | 59.4 | 29.6 | 56.7 | 58.8 | 64.2 | **22.1** |
| Sequence benchmark | Bacteria Genome (LSTM) | AUROC | 69.6 | 70.6 | 70.4 | 70.0 | 69.3 | NA | **74.8** |
| | | AUPRC | 69.9 | 71.9 | 69.3 | 56.0 | 70.2 | NA | **75.8** |
| | | FPR80 | 57.4 | 55.9 | 53.7 | **30.0** | 58.3 | NA | 47.4 |




(a) ImageNet (dogs) vs. ImageNet (non-dogs)

(b) CIFAR10 vs. SVHN

FSSDs from different layers behave differently. Blue is in-distribution.

## Our Algorithm

Layer ensemble helps

**Algorithm 1:** Computation of FSSD-Ensem

**Input:** Test samples $\boldsymbol{x} = \{x_n^{\text{test}}\}_{n=1}^N$, ensemble weights $\alpha_k$, perturbation magnitude $\epsilon$, feature extractors $\{F_{(k)}\}_{k=1}^K$

**for** *each feature extractor* $\{F_{(k)}\}_{k=1}^K$ **do**

1. Estimate FSS $F^*_{(k)} = \sum_{s=1}^S F_{(k)}\left(x_s^{\text{noise}}\right)/S$, where $x_s^{\text{noise}} \sim \mathcal{U}[0,1]$, $s = 1, \cdots, S$
2. Add perturbation to test sample: $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon\,\mathrm{sign}(\nabla_{\boldsymbol{x}}\left\|F_{(k)}\left(\boldsymbol{x}\right) - F^*_{(k)}\right\|)$
3. Calculate $\mathrm{FSSD}^{(k)}\left(\boldsymbol{x}\right) = \left\|F_{(k)}\left(\tilde{\boldsymbol{x}}\right) - F^*_{(k)}\right\|$

**end**

**Return** $\text{FSSD-Ensem}\left(\boldsymbol{x}\right) = \sum_{k=1}^K \alpha_k\,\mathrm{FSSD}^{(k)}\left(\boldsymbol{x}\right)$

## Future Work

1. Study the phenomenon in different phases of training corresponding to the recent advances in NTK;

2.Explore the connection to the probabilistic models, e.g. Gaussian Processes.