Sunnybrook ML Journal Club

Open Set Recognition
May 10, 2019

David Burns

Orthopaedic Surgery Resident
PhD(c) with Orthopaedic Biomechanics Lab, IBBME
University of Toronto

Supervisor: C. Whyne

Paper

Published as a conference paper at ICLR 2018

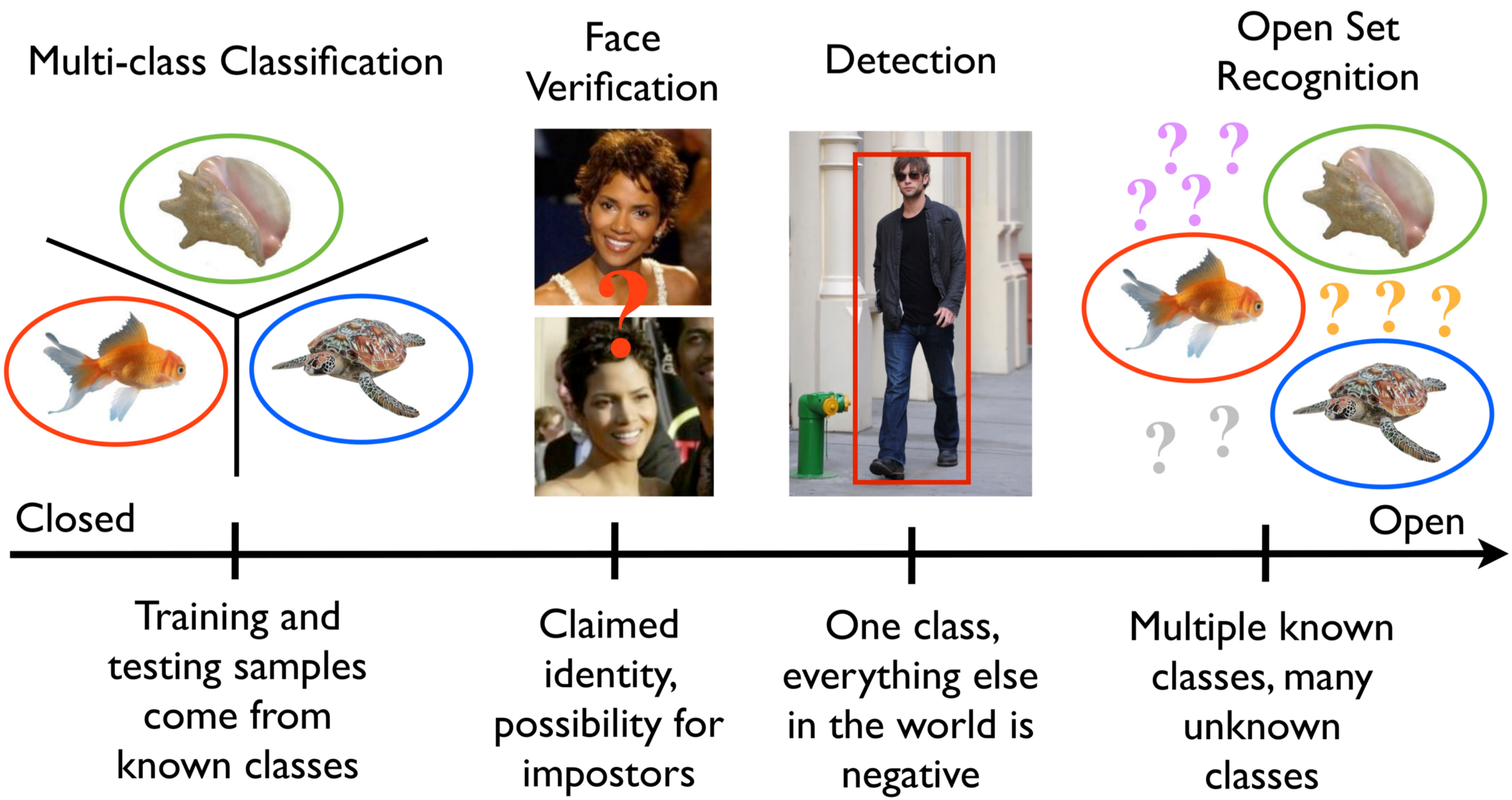
ENHANCING THE RELIABILITY OF OUT-OF-DISTRIBUTION IMAGE DETECTION IN NEURAL NETWORKS

Shiyu Liang
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
sliang26@illinois.edu

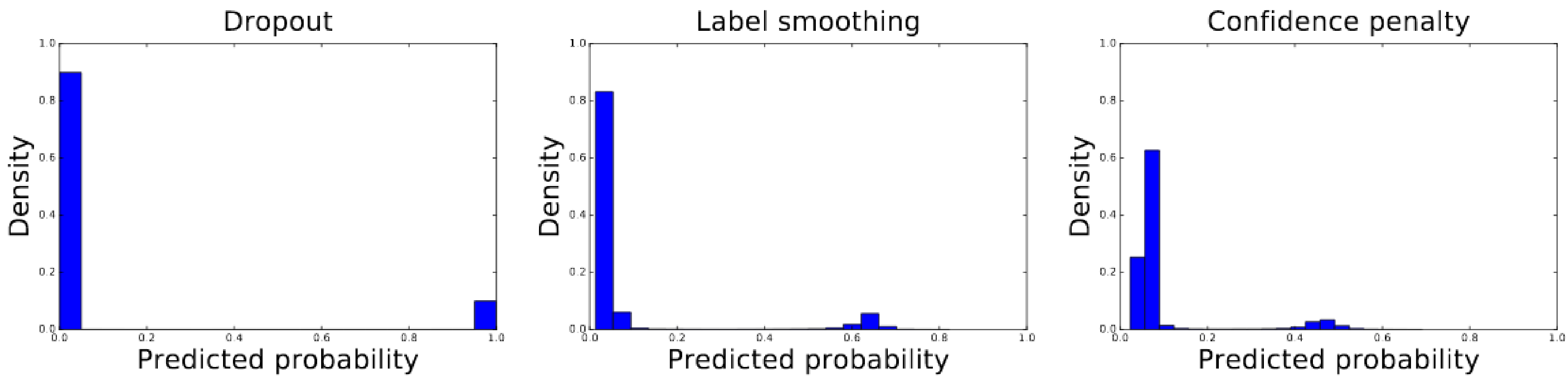
Yixuan Li
Facebook Research
yixuanl@fb.com

R. Srikant
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
rsrikant@illinois.edu

Open Set Recognition



Overconfidence



- MNIST, 2 layer FC network

Pereyra (Google Brain). ICRL 2017

Modern / Deep ANNs are Worse

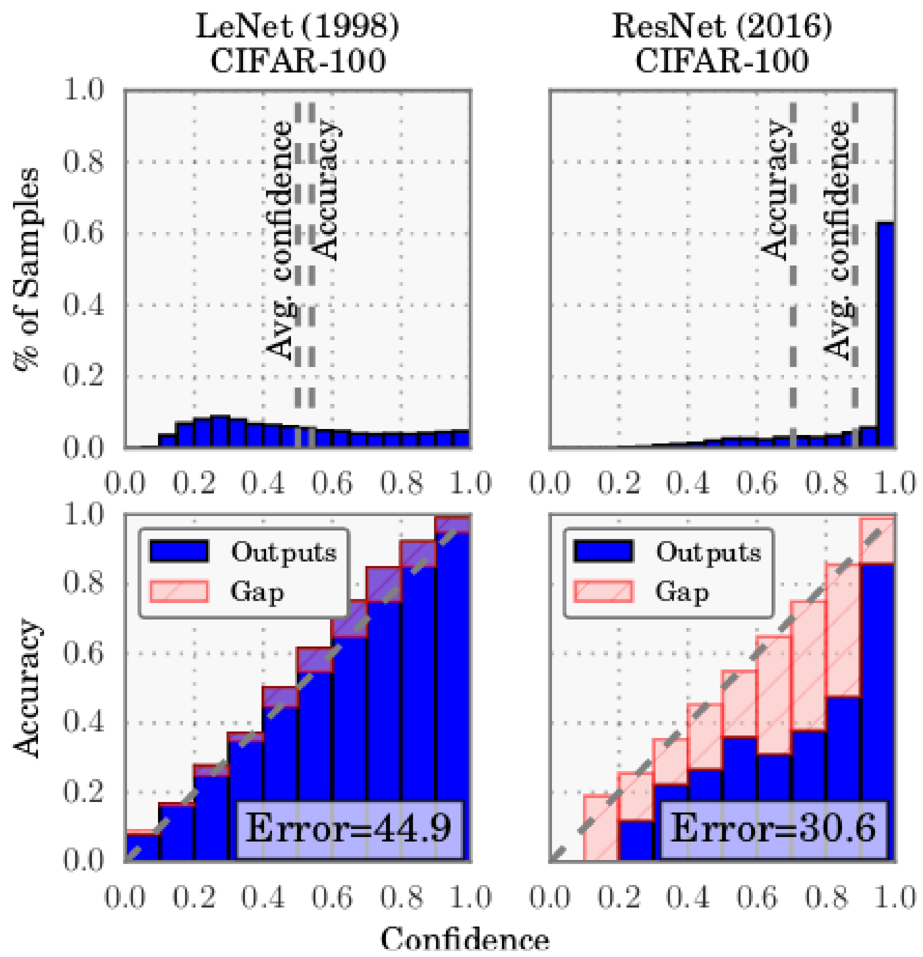


Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

Guo et al. ICRL 2017.

Open Set Recognition - Techniques

- Distance
- Domain
- Reconstruction
- Generative
- *Information-theoretic*
- *Adversarial*
- **Confidence**

Distance OSR

- Example distance metrics
  - Mahalanobis
  - Standardized Euclidean
  - Euclidean
- Semantic gap
  - low-level features vs high-level concepts
  - feature similarity != semantic similarity

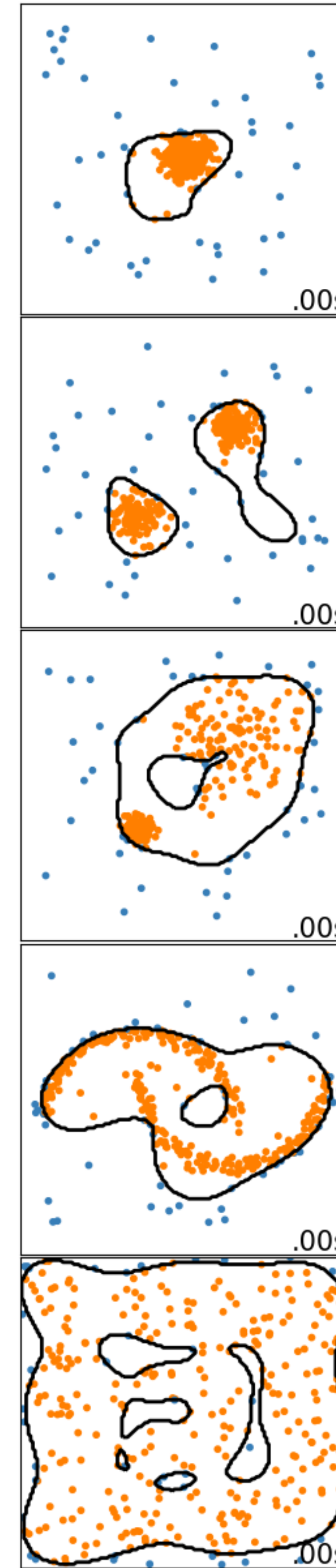
Domain OSR

- One-class SVM – Scholkopf. NeurIPS 2000.

$$\min_{w\in F,\boldsymbol{\xi}\in\mathbb{R}^{\ell},\rho\in\mathbb{R}} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho$$

$$\text{subject to} \quad (w\cdot\Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0.$$

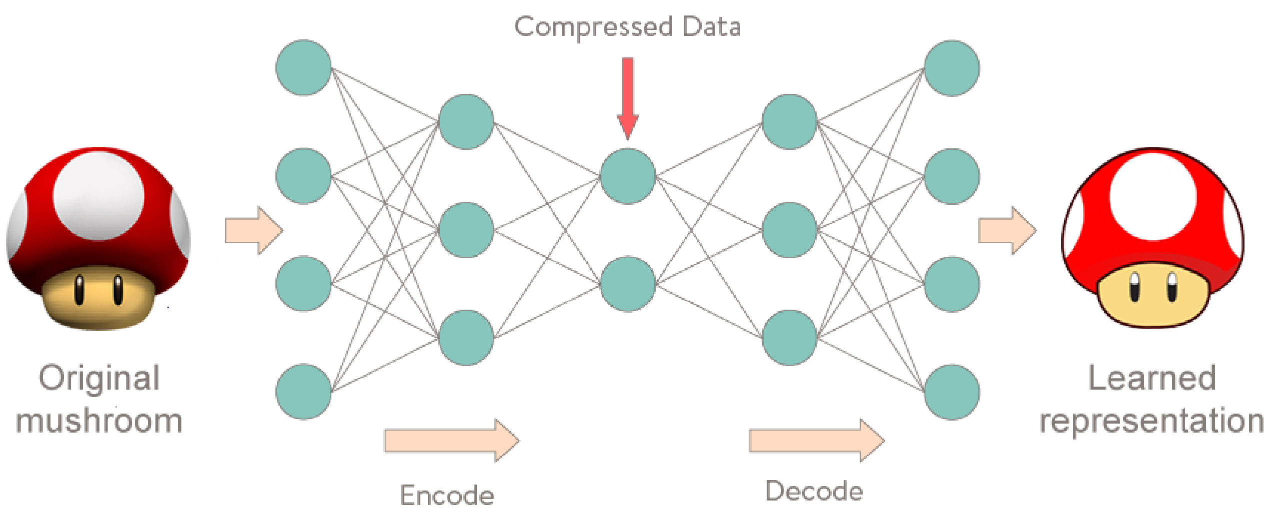
$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho\right)$$



Reconstruction OSR

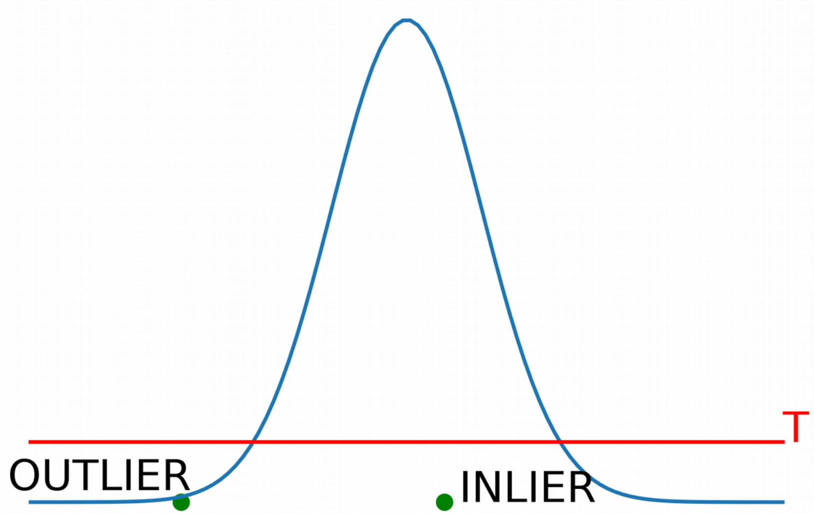
- Autocoders – learn compact data representation
- Greater reconstruction error for OD classes

$$L(x, x') = ||\,x - x'\,||^2$$



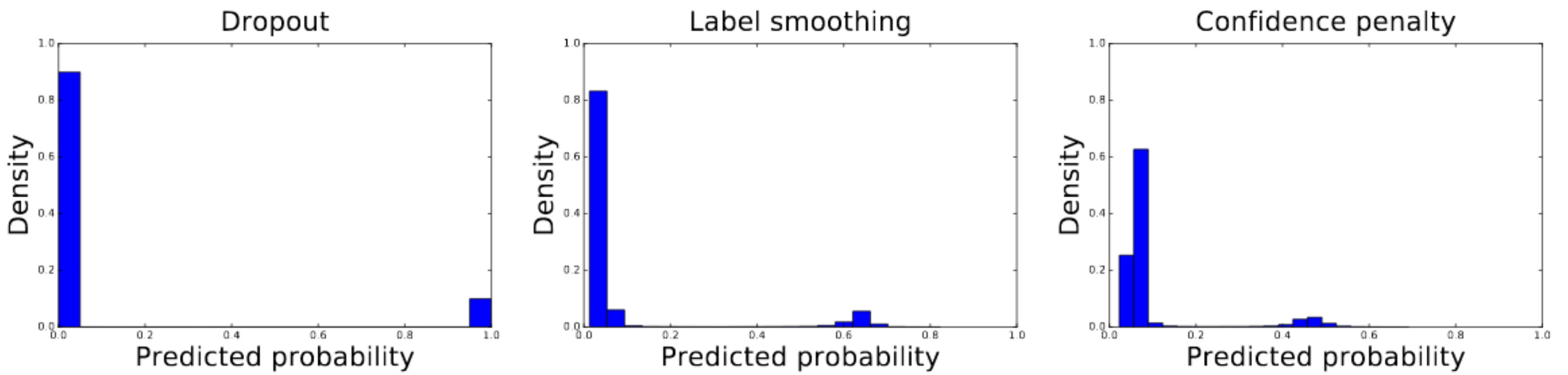
Probabilistic / Generative OSR

- Build a probabilistic model of the data $p_{model}(x)$
- Outliers are unlikely to have a high likelihood under the model
- Most generative models allow exact evaluation of $p_{model}(x)$



$$p_{model}(X|\theta) = \eta(X|\mu, \sigma)$$

Confidence Calibration



- Better generalization
- Separate P_{in} and P_{out} using thresholding
- This assumes a relationship between uncertain classes and unknown classes

Methods of ANN Calibration

- Post-hoc (recalibration) methods
  - Temperature scaling
  - Histogram binning / isotonic regression (binary classification)
  - Openmax
  - **ODIN**
- Calibrated training
  - Entropy regularization
  - Label smoothing regularization
  - Bayesian networks

Temperature Scaling

$$S_i(\boldsymbol{x};T) = \frac{\exp\left(f_i(\boldsymbol{x})/T\right)}{\sum_{j=1}^{N} \exp\left(f_j(\boldsymbol{x})/T\right)}$$

$$S_{\hat{y}}(\boldsymbol{x};T) = \max_i S_i(\boldsymbol{x};T)$$

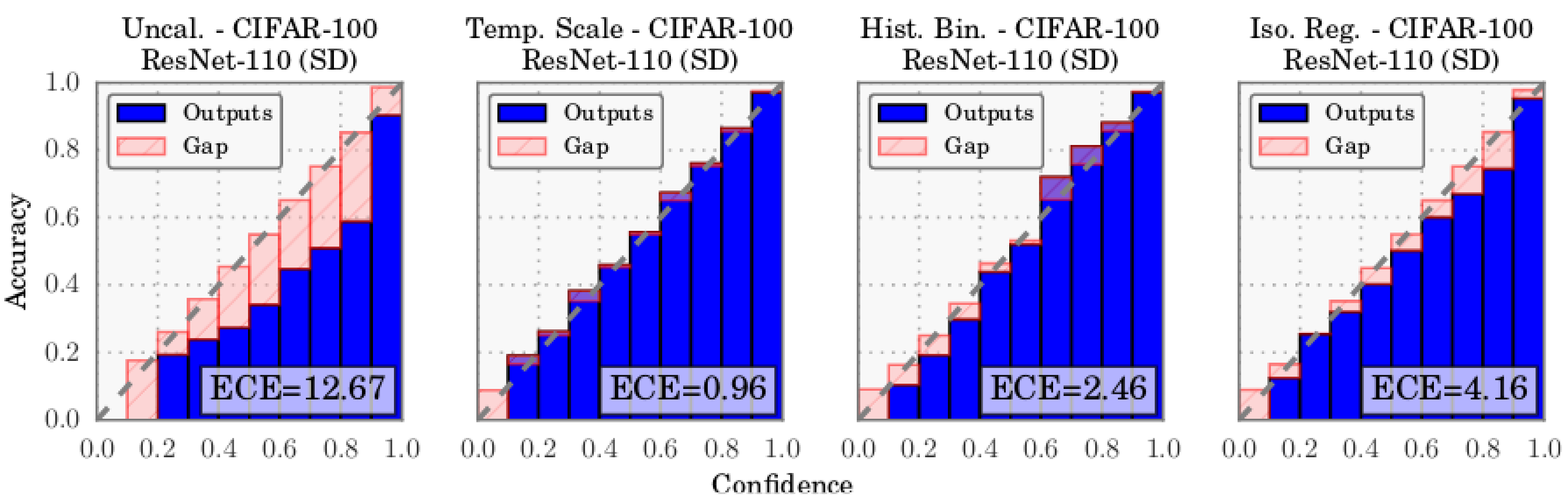
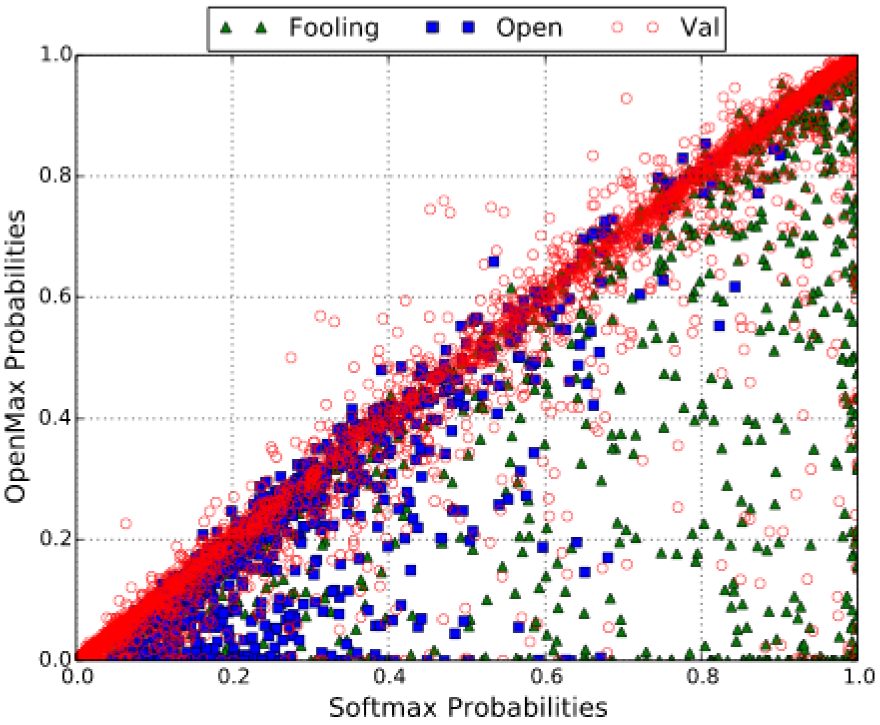


Figure 4. Reliability diagrams for CIFAR-100 before (far left) and after calibration (middle left, middle right, far right).

OpenMax

- Activation vector (AV): Penultimate ANN layer (prior to softmax)
- Classes represented by a mean activation vector fit by a Weibull distribution
- Openmax layer estimates probability for top few classes and an unknown unknown class



Bendale et al. Towards Open Set Deep Networks. CVPR 2015.

Regularizing Neural Networks by Penalizing Confident Output Distributions

Gabriel Pereyra *†
Google Brain
pereyra@google.com

George Tucker *†
Google Brain
gjt@google.com

Jan Chorowski
Google Brain
chorowski@google.com

Łukasz Kaiser
Google Brain
lukaszkaiser@google.com

Geoffrey Hinton
University of Toronto & Google Brain
geoffhinton@google.com

- Entropy regularization: neural network is trained to penalize confident output distributions

$$\mathcal{L}(\theta) = -\sum \log p_\theta(\boldsymbol{y}|\boldsymbol{x}) - \beta H(p_\theta(\boldsymbol{y}|\boldsymbol{x})),$$

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$

Rethinking the Inception Architecture for Computer Vision CVPR 2016

Christian Szegedy
Google Inc.
szegedy@google.com

Vincent Vanhoucke
vanhoucke@google.com

Sergey Ioffe
sioffe@google.com

Jonathon Shlens
shlens@google.com

Zbigniew Wojna
University College London
zbigniewwojna@gmail.com

- Label smoothing regularization

$$q(k|x) = \delta_{k,y}$$

$$q'(k) = (1 - \epsilon)\delta_{k,y} + \frac{\epsilon}{K}.$$

OSR Evaluation

- Datasets split into
  - ID
    - Train
    - Test
  - OD
  - Fooling / adversarial
- Metrics
  - Binary OD performance
  - ID performance

ENHANCING THE RELIABILITY OF OUT-OF-DISTRIBUTION IMAGE DETECTION IN NEURAL NETWORKS

Shiyu Liang
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
sliang26@illinois.edu

Yixuan Li
Facebook Research
yixuanl@fb.com

R. Srikant
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
rsrikant@illinois.edu

ODIN Method

- **O**ut-of-**DI**stribution detector for **N**eural networks
- Combines temperature scaling with input perturbations to scale the predictive distribution from a **pre-trained classifier**

Thor’s dad

ODIN Method

1) Input (x) fed through classifier with temperature scaled softmax output

$$S_i(\boldsymbol{x};T) = \frac{\exp\left(f_i(\boldsymbol{x})/T\right)}{\sum_{j=1}^{N}\exp\left(f_j(\boldsymbol{x})/T\right)} \qquad S_{\hat{y}}(\boldsymbol{x};T) = \max_i S_i(\boldsymbol{x};T)$$

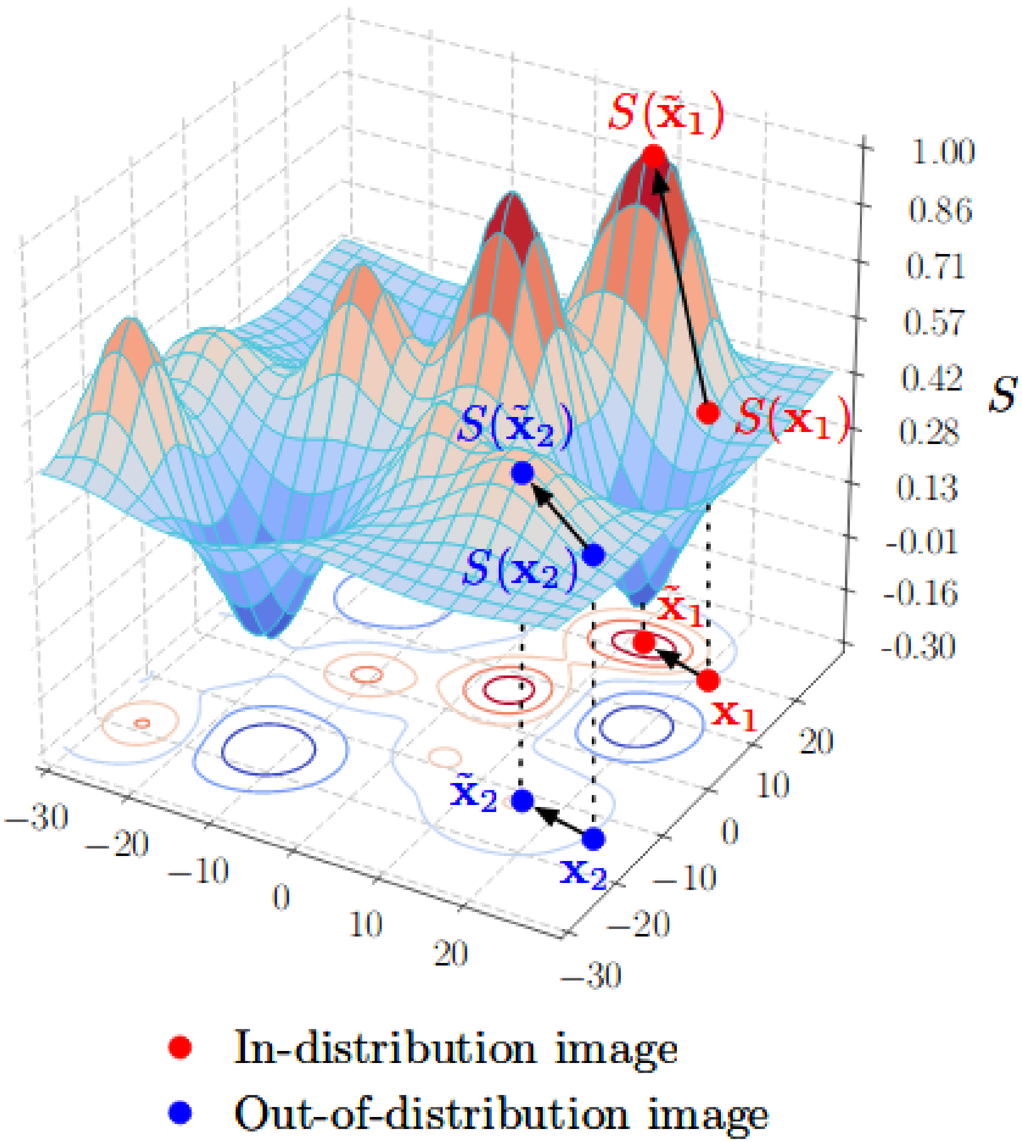
2) Pertubations generated using fast gradient sign method (Goodfellow 2015)

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \varepsilon \operatorname{sign}(-\nabla_{\boldsymbol{x}} \log S_{\hat{y}}(\boldsymbol{x};T))$$

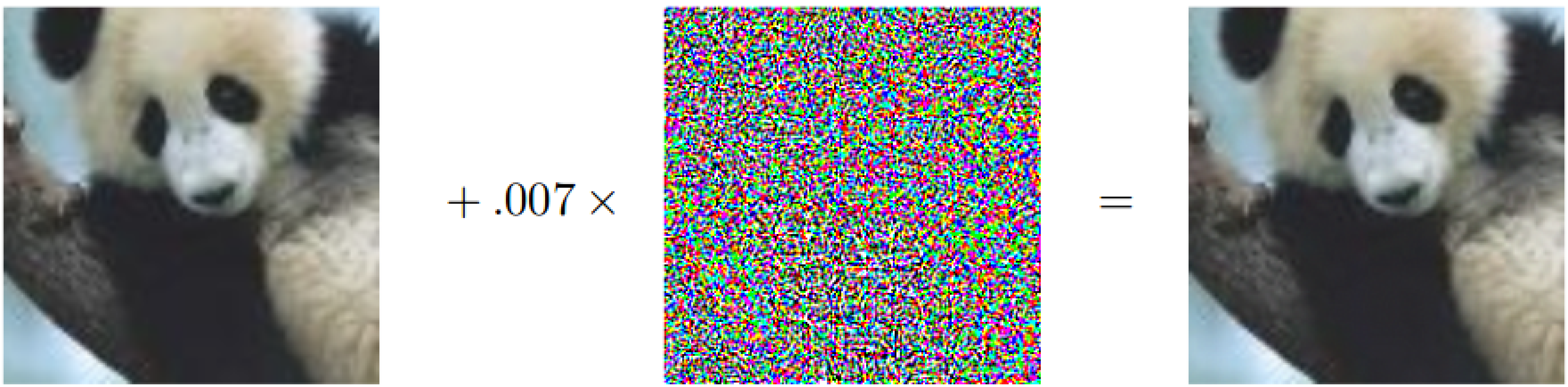
3) Outlier detection: softmax score on preturbed image compared to threshold

$$g(\boldsymbol{x};\delta,T,\varepsilon) = \begin{cases} 1 & \text{if} \max_i p(\tilde{\boldsymbol{x}};T) \le \delta, \\ 0 & \text{if} \max_i p(\tilde{\boldsymbol{x}};T) > \delta. \end{cases}$$

Input Perturbation



Input Perturbation



Images very close in pixel space, far in feature space

Experimental Setup

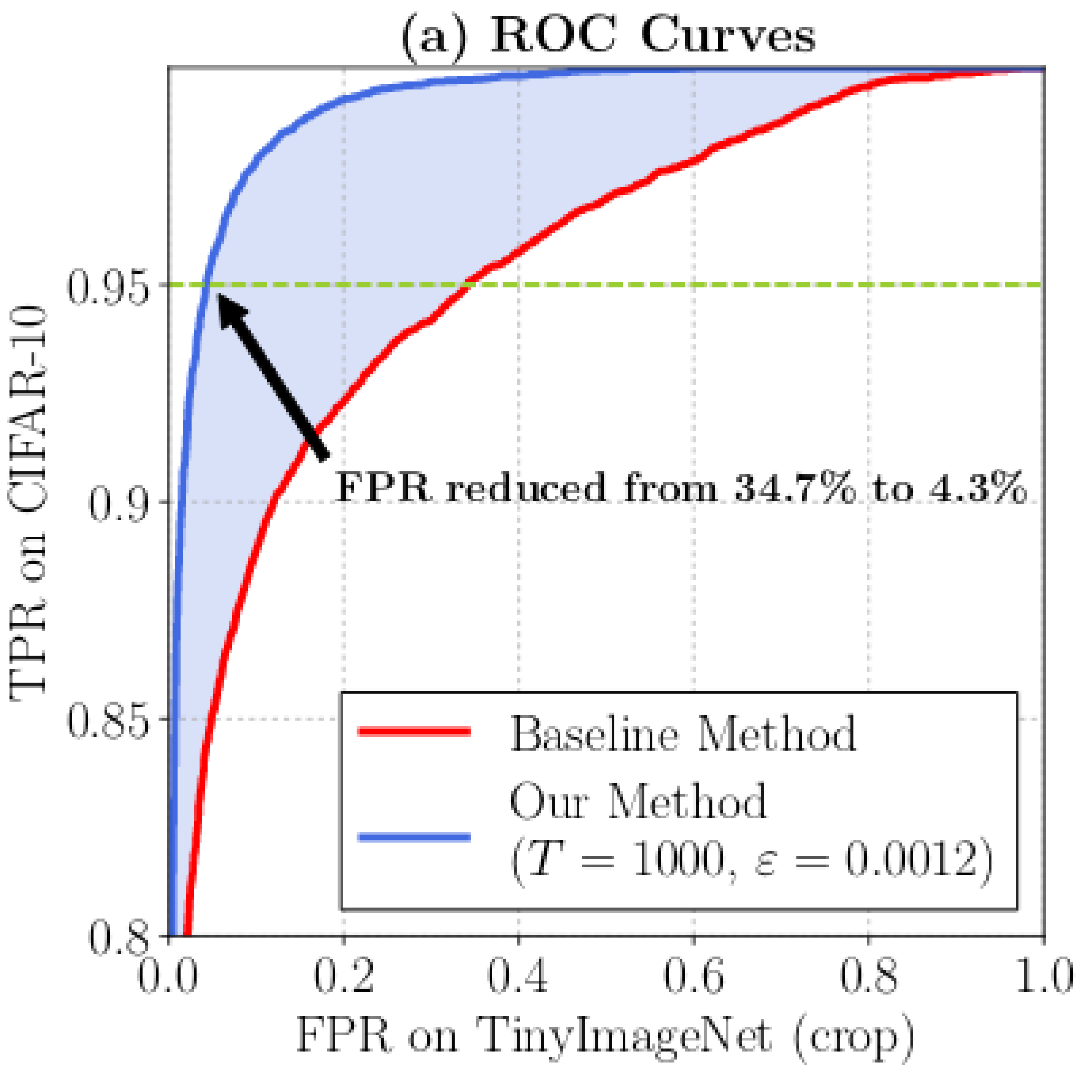
- Networks
  - DenseNet (2016)
  - Wide ResNet (2016)
- Datasets (ID)
  - CIFAR-10
  - CIFAR-100
- Datasets (OD)
  - TinyImageNet
  - LSUN
  - ISUN
  - Gaussian / Uniform Noise

- Metrics
  - FPR at 95% TPR
  - Detection Error
  - AUROC
  - AUPR
- Baseline
  - Threshold softmax score

Hyperparameter Optimization

- Randomly held out 1000 images from each test set for tuning T and ε
- Grid search over
  - T: 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000
  - ε: linspace([0, 0.004], 21)
- Free δ threshold parameter

Results


(a) ROC Curves
0.95
0.9
0.85
0.8
TPR on CIFAR-10
FPR reduced from 34.7% to 4.3%
Baseline Method
Our Method
(T = 1000, ε = 0.0012)
0.0
0.2
0.4
0.6
0.8
1.0
FPR on TinyImageNet (crop)

Dataset Difficulty

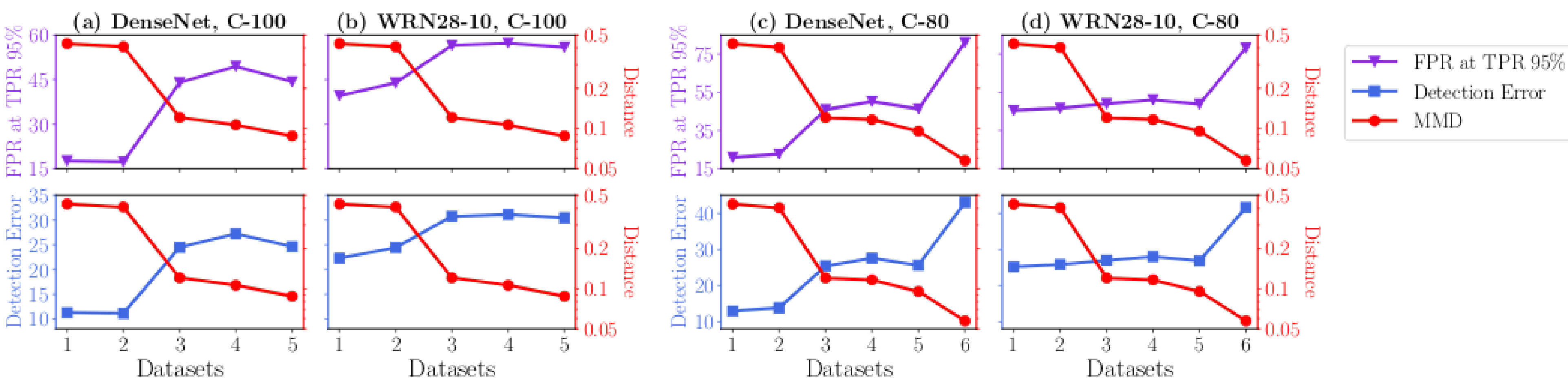


Figure 2: (a)-(d) Performance of our method vs. MMD between in- and out-of-distribution datasets. Neural networks are trained on CIFAR-100 and CIFAR-80, respectively. The out-of-distribution datasets are 1: LSUN (cop), 2: TinyImageNet (crop), 3: LSUN (resize), 4: is iSUN (resize), 5: TinyImageNet (resize) and 6: CIFAR-20.

ODIN Criticisms

- Strengths
  - Simple implementation
  - Works with pre-trained networks
  - Does not affect the prediction accuracy for ID classification or change predictions
- Weaknesses
  - Introduces 3 hyperparameters
  - Optimize the hyperparameters on the test set
  - Very weak baseline comparison
  - Used different datasets for OD data

Key Papers

- Guo et al. On Calibration of Modern Neural Networks. ICML 2017.
- Pereyra et al. Regularizing Neural Networks by Penalizing Confident Output Distributions. ICLR 2017.
- Liang et al. Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks. ICLR 2018.
- Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR 2015.
- Pementel et al. A Review of Novelty Detection. Signal Processing 2014.
- Bendale et al. Towards Open Set Deep Networks. CVPR 2016.
- Szegedy et al. Rethinking the Inception Architecture for Computer Vision. CVPR 2016.

Questions

David Burns MD PhD (c)
d.burns@utoronto.ca
github: dmbee