Learning Confidence for Out-of-Distribution Detection in Neural Networks

Terrance DeVries [1 2] Graham W. Taylor [1 2]

[1]School of Engineering, University of Guelph, Guelph, Ontario, Canada [2]Vector Institute, Toronto, Ontario, Canada. Correspondence to: Terrance DeVries <terrance@uoguelph.ca>, Graham W. Taylor <gwtaylor@uoguelph.ca>.

arxiv-2018

Outline

- Introduction
- Confidence Estimation
- Implementation Details
- Experiment

Introduction

Out of distribution(OOD) detection

- detect whether a data is from a different distribution from that of training data.
- Application: Anomaly detection. For instance, Fraud Detection and cancer detection.

How to detect OOD?

- Idea of this work: Learn a confidence of prediction. Very confidence-in the distribution; Not so confidence-Out of distribution.

Challenge: we don’t have confidence label.

How to learn a good confidence estimation without label?

Confidence Estimation

- Imagine a scenario when a student is taking a test. He can ask for hints, but for each hint he receive, he will also receive a penalty.
- The optimal strategy is only ask for hints when the question is uncertain to answer. In other words, the student should have a good estimation for the confidence of his answer.

Confidence Estimation

Prediction loss

Confidence loss

$$L = L_t + \lambda L_c$$

Cross entropy

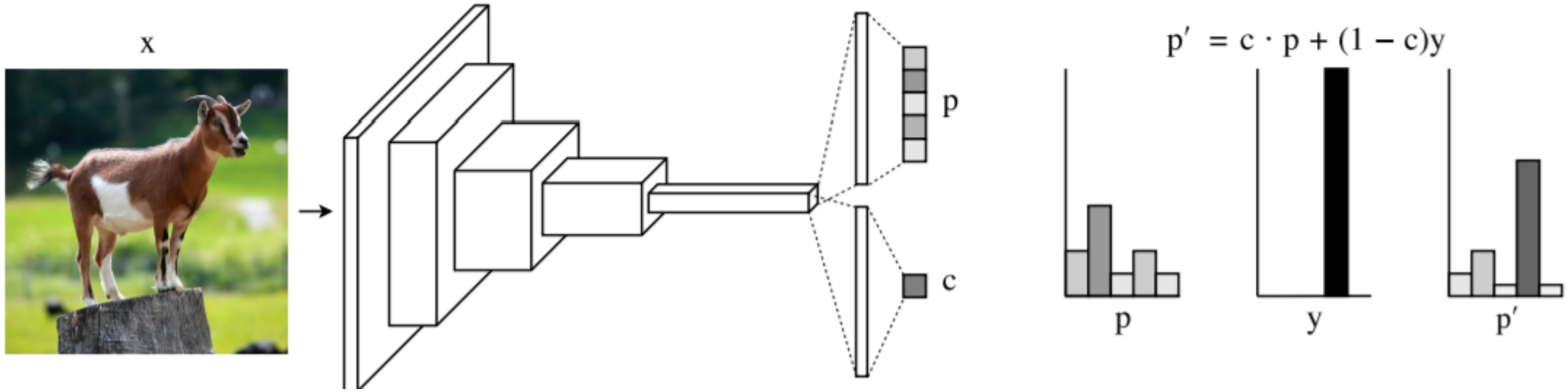
Cross entropy

$$\mathrm{L} = -\sum_{i=1}^{M} \log(p_i')\, y_i + \lambda[-\log(c)]$$

p_i'

$$L = -\sum_{i=1}^{M} \log(c * p_i + (1-c)y_i)\, y_i + \lambda[-\log(c)]$$

NN architecture



Implementation Details

First **challenge**: c often converges to unity for all samples.

Solution: introduce budget parameter β. $L_c > \beta$, increase λ. $L_c < \beta$, decrease λ.

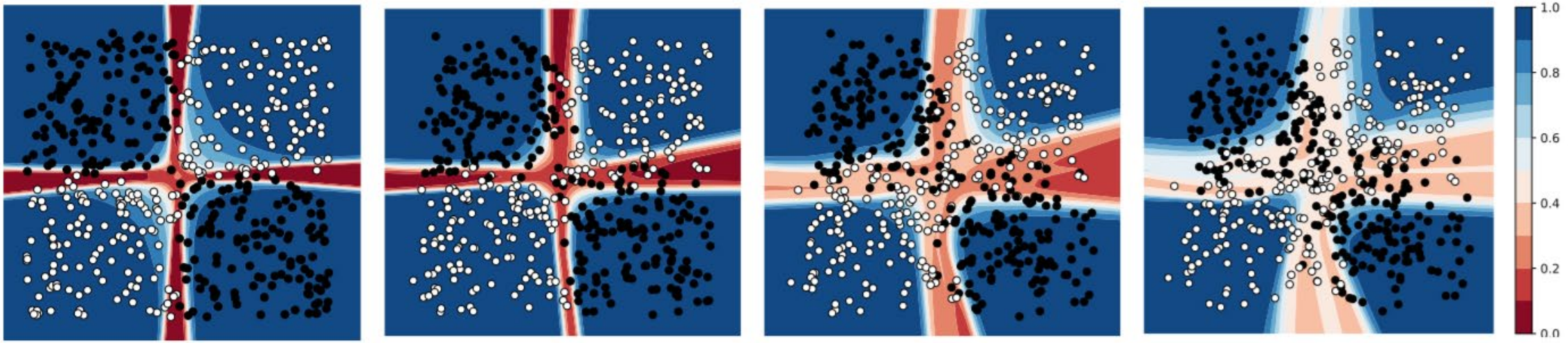
Second **challenge**: network automatically ignore outliers and noisy regions.

Solution: apply $\mathrm{p}_\mathrm{i}' = c * p_i + (1 - c)y_i$ to only half of the batch.

Third **challenge**: when training on small datasets, the model overfits and can't learn the concept of confidence.

Solution: use aggressive data augmentations to some of the data.

Experiment



(a) Noise = 0%

(b) Noise = 10%

(c) Noise = 20%

(d) Noise = 30%

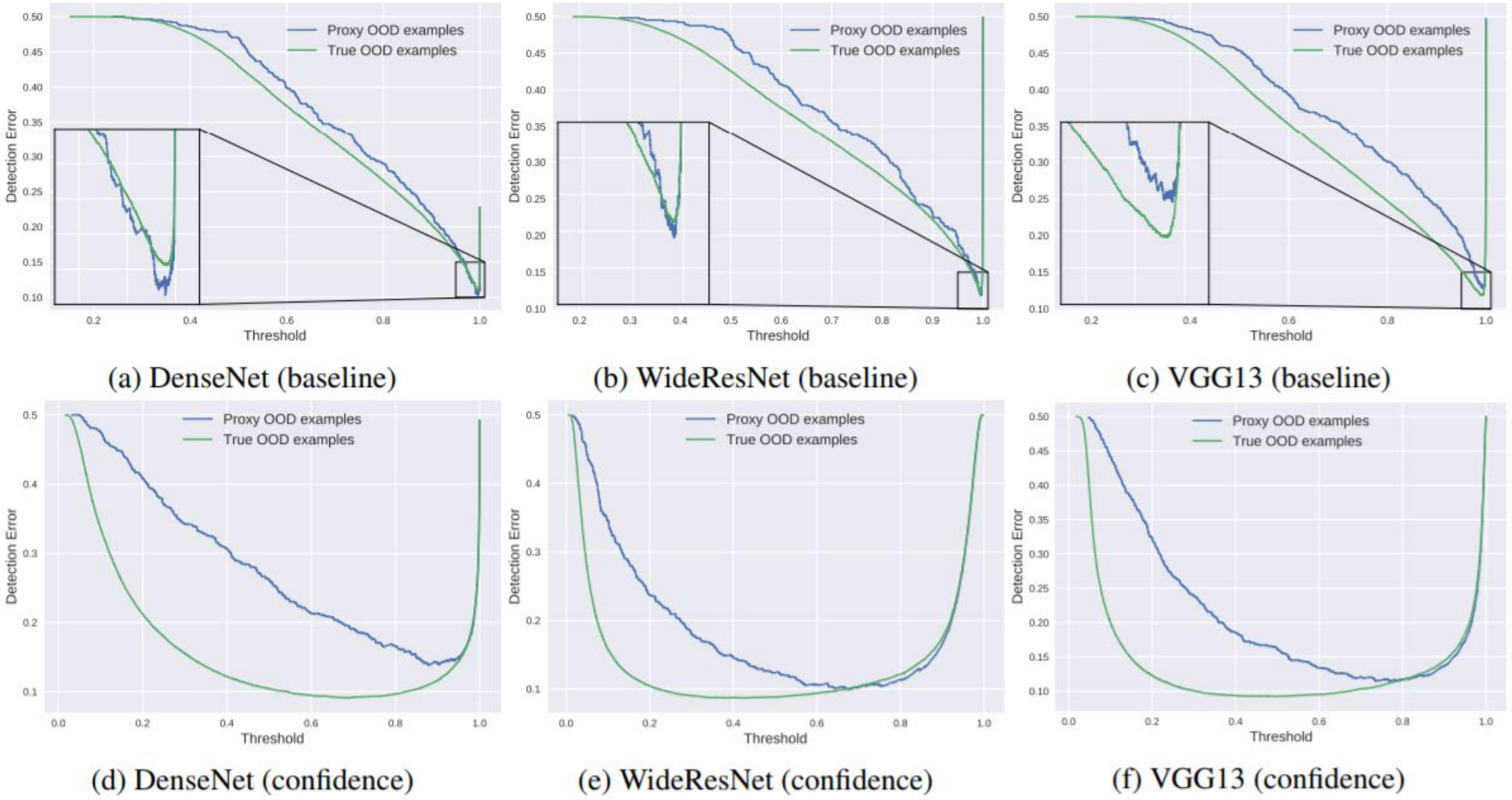
Experiment

Model In-distribution Dataset	**Out-of-distribution Dataset**	**Classification Error** ↓	**FPR (95% TPR)** ↓	**Detection Error** ↓	**AUROC** ↑	**AUPR In** ↑	**AUPR Out** ↑
		Baseline (Hendrycks & Gimpel, 2017)/Confidence Thresholding					
DenseNet-BC SVHN	TinyImageNet (resize)	2.89/**2.77**	7.2/**1.5**	5.3/**2.8**	98.4/**99.5**	99.4/**99.8**	95.6/**98.7**
	LSUN (resize)		6.0/**1.0**	4.9/**2.3**	98.6/**99.7**	99.5/**99.9**	96.0/**99.0**
	iSUN		6.0/**0.9**	4.9/**2.3**	98.6/**99.7**	99.5/**99.9**	95.7/**98.8**
	All Images		12.2/**4.2**	7.2/**4.5**	97.3/**98.9**	95.1/**97.4**	98.4/**99.4**
WRN-16-8 SVHN	TinyImageNet (resize)	2.77/**2.66**	10.6/**1.5**	6.1/**2.7**	97.8/**99.6**	99.2/**99.8**	93.6/**99.2**
	LSUN (resize)		9.5/**0.6**	5.8/**1.8**	98.0/**99.8**	99.3/**99.9**	94.0/**99.5**
	iSUN		9.6/**0.8**	5.9/**2.1**	98.0/**99.8**	99.3/**99.9**	93.4/**99.4**
	All Images		15.7/**5.3**	7.9/**5.0**	96.7/**98.7**	94.1/**96.8**	97.9/**99.4**
VGG13 SVHN	TinyImageNet (resize)	3.05/**2.98**	11.4/**1.8**	6.2/**3.1**	97.8/**99.6**	99.2/**99.8**	93.7/**99.1**
	LSUN (resize)		9.4/**0.8**	5.7/**2.0**	98.1/**99.8**	99.3/**99.9**	94.3/**99.6**
	iSUN		10.0/**1.0**	6.0/**2.2**	98.0/**99.8**	99.3/**99.9**	93.7/**99.5**
	All Images		14.2/**4.3**	7.1/**4.6**	97.3/**99.2**	95.9/**98.5**	98.2/**99.6**
DenseNet-BC CIFAR-10	TinyImageNet (resize)	**4.17**/4.39	44.9/**33.8**	12.8/**12.3**	93.2/**94.2**	94.6/**95.0**	91.2/**93.0**
	LSUN (resize)		38.6/**30.7**	10.8/**10.3**	94.6/**95.4**	95.9/**96.4**	92.8/**93.9**
	iSUN		41.4/**31.6**	11.6/**11.0**	94.1/**95.0**	95.8/**96.3**	91.3/**93.0**
	All Images		40.9/**28.9**	11.6/**10.9**	94.1/**95.3**	87.6/**88.1**	98.3/**98.7**
WRN-28-10 CIFAR-10	TinyImageNet (resize)	**3.25**/3.46	41.0/**26.6**	14.3/**11.6**	91.0/**94.5**	88.9/**94.1**	90.5/**94.0**
	LSUN (resize)		34.7/**24.0**	11.7/**9.1**	93.7/**96.0**	93.4/**96.6**	92.7/**94.5**
	iSUN		36.7/**24.9**	12.6/**9.8**	92.8/**95.7**	92.6/**96.5**	91.1/**94.0**
	All Images		36.1/**23.3**	12.4/**9.7**	92.9/**95.7**	73.3/**86.7**	98.1/**98.8**
VGG13 CIFAR-10	TinyImageNet (resize)	**5.28**/5.44	43.8/**18.4**	12.0/**9.4**	93.5/**97.0**	94.6/**97.3**	91.7/**96.9**
	LSUN (resize)		41.9/**16.4**	11.5/**8.3**	94.0/**97.5**	95.1/**97.8**	92.2/**97.2**
	iSUN		41.2/**16.3**	11.4/**8.5**	94.0/**97.5**	95.5/**98.0**	91.5/**96.9**
	All Images		41.6/**19.2**	11.7/**9.1**	93.9/**97.1**	85.5/**92.0**	98.2/**99.3**

Experiment

	In-distribution dataset	FPR (95% TPR) ↓	Detection Error ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
		ODIN (Liang et al., 2018)/Confidence + Input Preprocessing				
DenseNet-BC	SVHN	8.6/**4.2**	6.8/**4.5**	97.2/**98.9**	92.5/**97.5**	98.6/**99.4**
	CIFAR-10	**7.8**/16.2	**6.0**/8.6	**98.4**/97.0	**95.3**/91.4	**99.6**/99.2
WideResNet	SVHN	13.4/**5.2**	8.4/**5.0**	96.5/**98.7**	92.2/**97.1**	97.9/**99.3**
	CIFAR-10	25.0/**18.9**	12.0/**9.4**	93.4/**96.2**	71.6/**87.2**	98.3/**99.0**
VGG13	SVHN	7.3/**4.1**	6.0/**4.5**	98.2/**99.2**	96.8/**98.6**	98.9/**99.5**
	CIFAR-10	20.2/**11.2**	10.2/**6.9**	95.8/**98.0**	85.9/**94.5**	98.9/**99.5**

Experiment



(a) DenseNet (baseline) (b) WideResNet (baseline) (c) VGG13 (baseline)

(d) DenseNet (confidence) (e) WideResNet (confidence) (f) VGG13 (confidence)