2020.10.15. Open DMQA Seminar

Open Set Recognition with Background Data

김상훈

발표자 소개

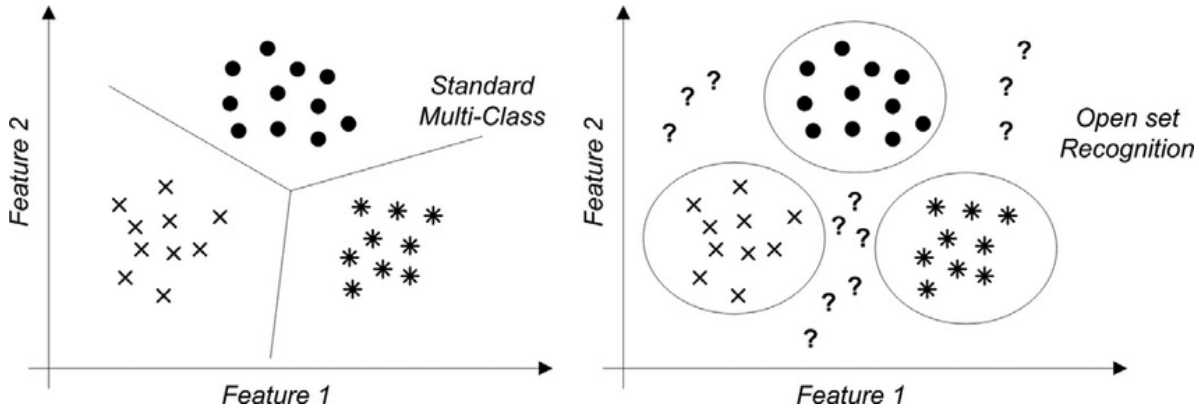
Sanghoon Kim (김상훈)

M.S. Student
(September 1, 2019 ~ Present)

Topic: Machine Learning Algorithms

Email: dawonksh@korea.ac.kr

- **김상훈 (Sanghoon Kim)**
  - ✓ Data Mining & Quality Analytics Lab.
  - ✓ M.S. Student (2019.09 ~ Present)
- **Research Interest**
  - ✓ Machine learning / Deep learning Algorithms
  - ✓ Open Set Recognition / Anomaly Detection



CONTENTS

Open Set Recognition

(1) Standard Multi-class Classification

Training images

Class 1: dog



Class 2: cat

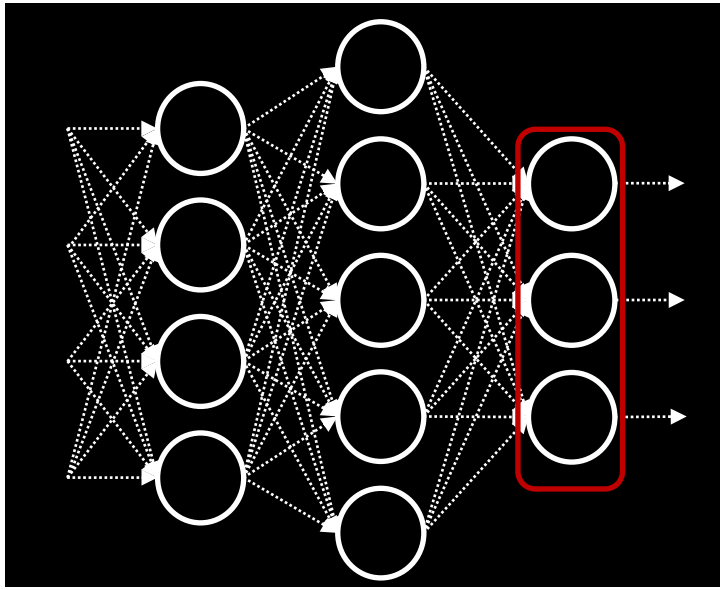


Class 3: rabbit



Classifier Model

$f(X)$



$P(Y = dog)$

$P(Y = cat)$

$P(Y = rabbit)$

Open Set Recognition

(1) Standard Multi-class Classification

Logit Vector = $\begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}$

값의 범위 : $(-\infty, \infty)$

Training images

Class 1: dog



Class 2: cat

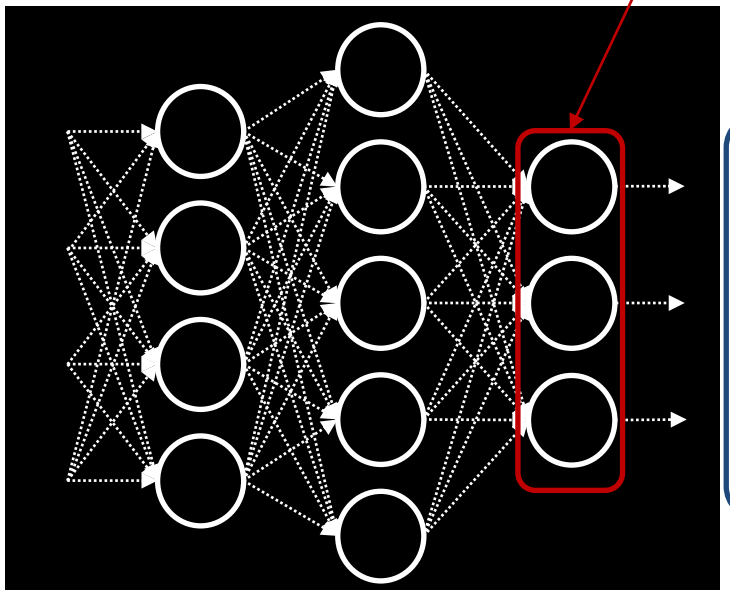


Class 3: rabbit



Classifier Model

$f(X)$



SoftMax Activation

값의 범위 : [0,1]

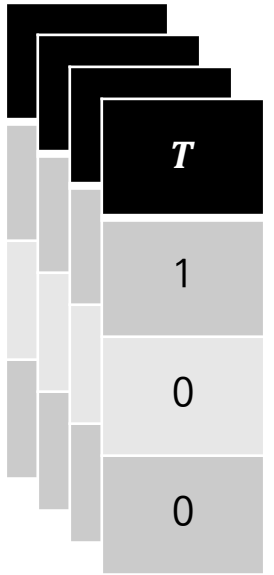
$S_i(x)$

$\frac{\exp(V_1)}{\sum_k \exp(V_k)} \rightarrow \boldsymbol{P(Y = dog)}$

$\frac{\exp(V_2)}{\sum_k \exp(V_k)} \rightarrow \boldsymbol{P(Y = cat)}$

$\frac{\exp(V_3)}{\sum_k \exp(V_k)} \rightarrow \boldsymbol{P(Y = rabbit)}$

True labels



Cross Entropy Loss

$$Loss = -\sum_{i=1}^{C} t_i \log S_i(x)$$

- C : $total\ number\ of\ classes$
- $S_i(x)$: $Softmax\ score\ of\ i-th\ class$
- t_i : $true\ label\ of\ i-th\ class$

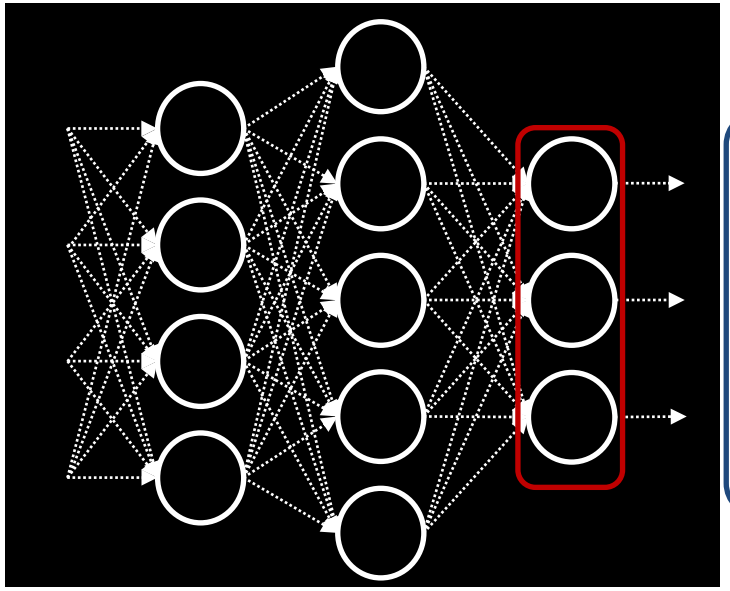
Open Set Recognition

(1) Standard Multi-class Classification

Testing image

Classifier Model

$f(X)$



$$\frac{\exp(V_1)}{\sum_k \exp(V_k)} \rightarrow \boldsymbol{P(Y = dog)} = 0.97$$

$$\frac{\exp(V_2)}{\sum_k \exp(V_k)} \rightarrow P(Y = cat) = 0.01$$

$$\frac{\exp(V_3)}{\sum_k \exp(V_k)} \rightarrow P(Y = rabbit) = 0.02$$

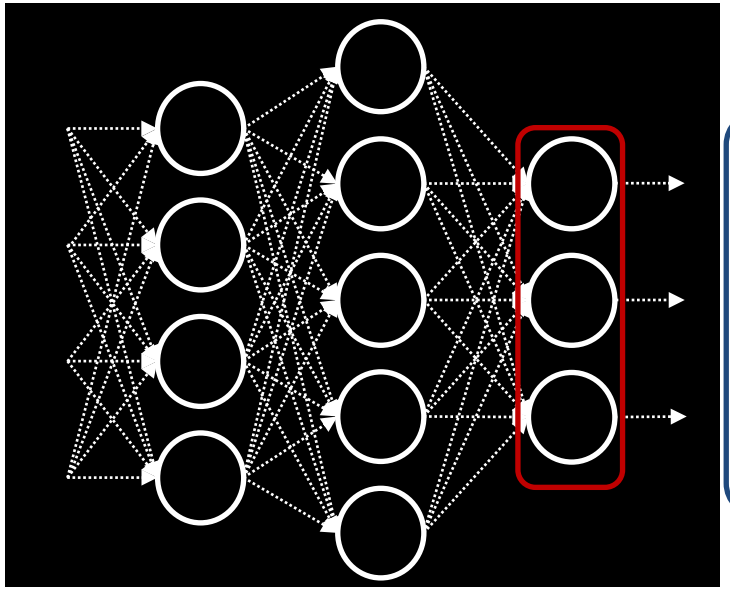
Open Set Recognition

(1) Standard Multi-class Classification

Testing image

Classifier Model

$f(X)$



$$\frac{\exp(V_1)}{\sum_k \exp(V_k)} \quad \frac{\exp(V_2)}{\sum_k \exp(V_k)} \quad \frac{\exp(V_3)}{\sum_k \exp(V_k)}$$

Predicted label

$P(Y = dog)$ = 0.97

$P(Y = cat)$ = 0.01

$P(Y = rabbit)$ = 0.02

T
1
0
0

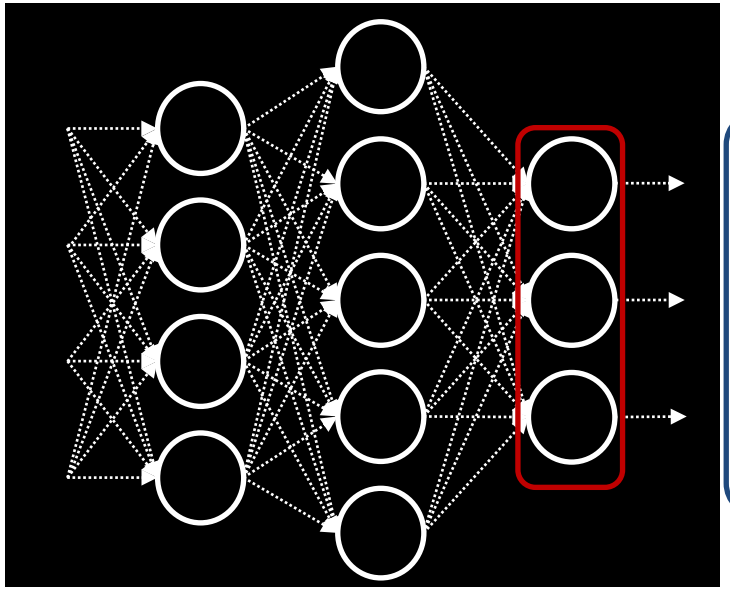
Open Set Recognition

(1) Standard Multi-class Classification

Unknown Image

Trained Classifier Model

$f(X)$



$$\frac{\exp(V_1)}{\sum_k \exp(V_k)} \rightarrow P(Y = dog)$$

$$\frac{\exp(V_2)}{\sum_k \exp(V_k)} \rightarrow P(Y = cat)$$

$$\frac{\exp(V_3)}{\sum_k \exp(V_k)} \rightarrow P(Y = rabbit)$$

Standard Multi-class Classifier는
학습한 Class에 대한 확률만 출력할 수 있다. (Closed Set Classification)

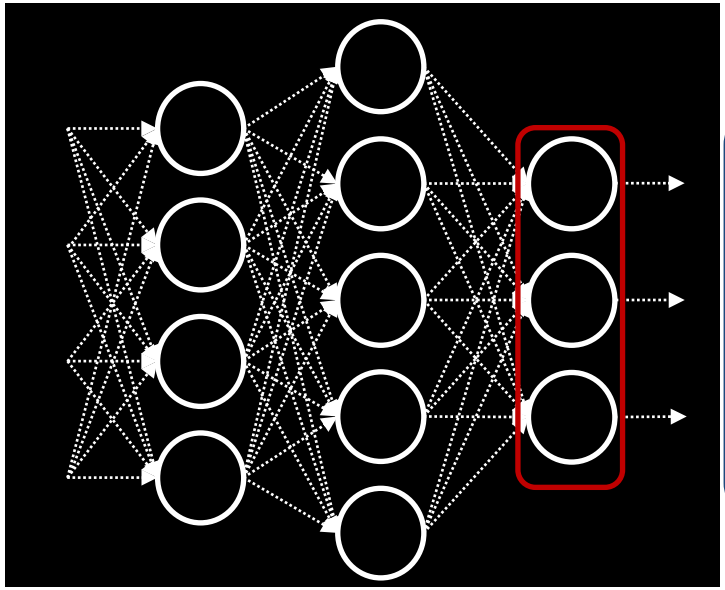
Open Set Recognition

(1) Standard Multi-class Classification

Unknown Image

Trained Classifier Model

$f(X)$



$$\frac{\exp(V_1)}{\sum_k \exp(V_k)}$$

$$\frac{\exp(V_2)}{\sum_k \exp(V_k)}$$

$$\frac{\exp(V_3)}{\sum_k \exp(V_k)}$$

Threshold = 0.7

$P(Y = dog)$ = 0.4 〈 0.7 reject

$P(Y = cat)$ = 0.3 〈 0.7 reject

$P(Y = rabbit)$ = 0.3 〈 0.7 reject

모든 Class에 대해 각 Class에 속할 확률이 특정 threshold보다 작다면?

Unknown Class

Open Set Recognition

(1) Standard Multi-class Classification

결정 경계 $y^* = \text{argmax}_j \, P(y = j|\mathbf{x})$

Open Set Recognition

(1) Standard Multi-class Classification

Class 1

결정 경계 $y^* = \text{argmax}_j \, P(y = j|\mathbf{x})$

Unknown Data

모든 Class에 대한 SoftMax 점수가
Threshold = 0.5보다 작은 구역

Class 2

Class 3

Unknown Classes의 Data들이 모든 Class에 대한 SoftMax 점수가 고르게 낮을 것이라는 가정

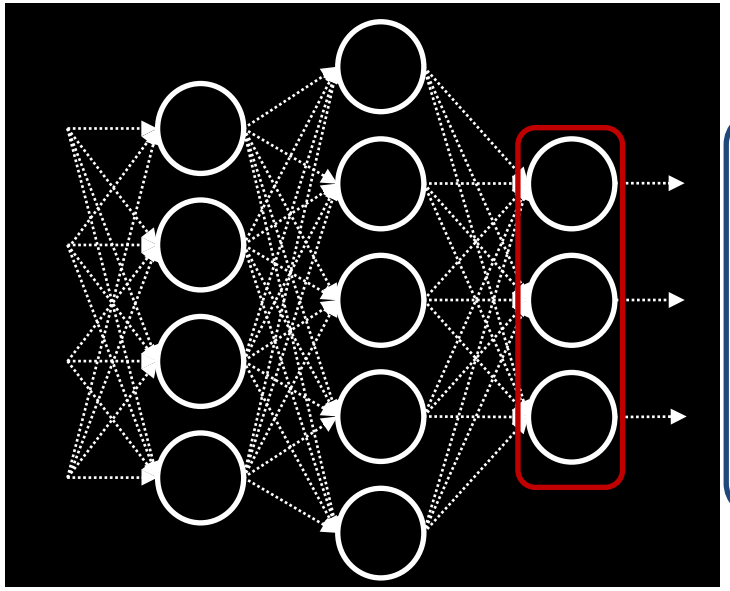
Open Set Recognition

(1) Standard Multi-class Classification

Unknown Image

Trained Classifier Model

$f(X)$



$\frac{\exp(V_1)}{\sum_k \exp(V_k)}$ → $P(Y = dog)$ = 0.02

$\frac{\exp(V_2)}{\sum_k \exp(V_k)}$ → $P(Y = cat)$ = 0.01

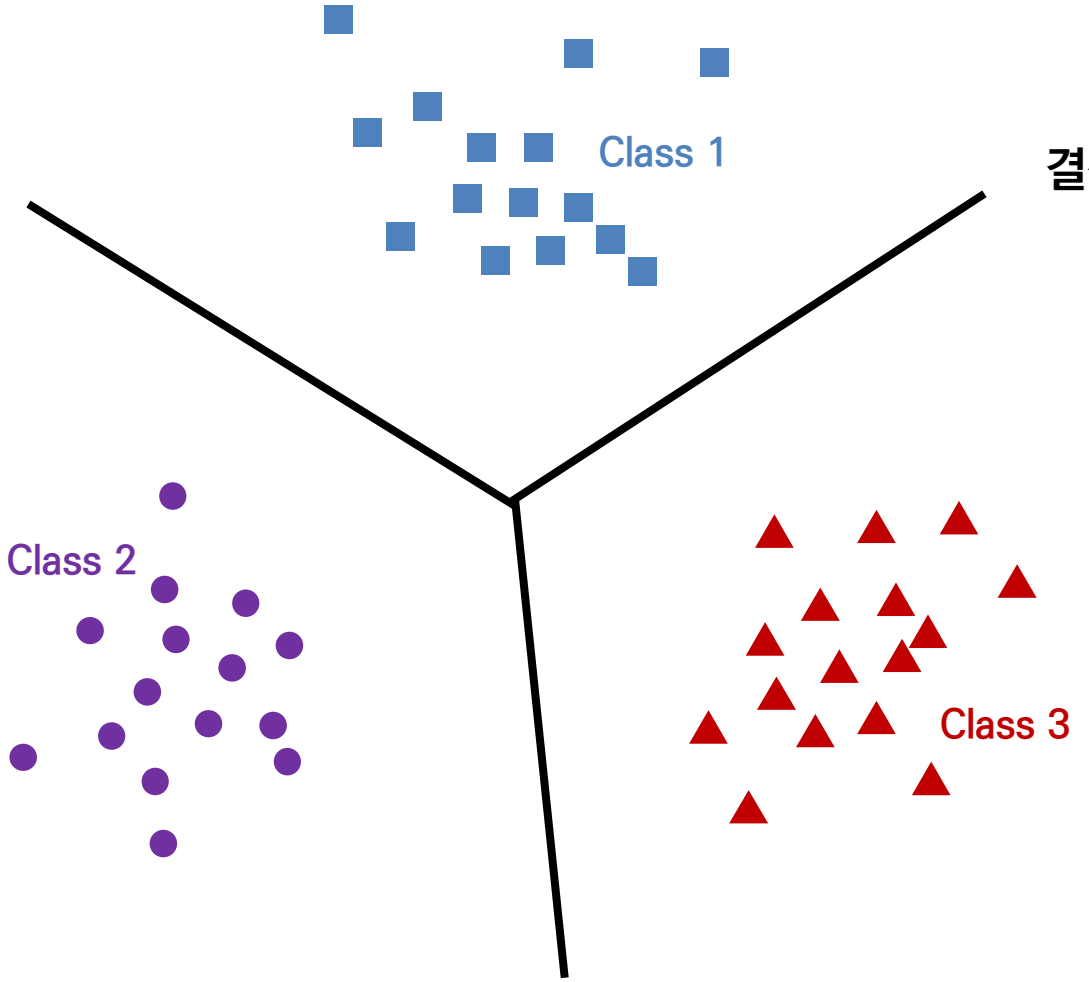
$\frac{\exp(V_3)}{\sum_k \exp(V_k)}$ → $\boldsymbol{P(Y = rabbit)}$ = 0.97

Standard Multi-class Classifier는 Unknown Class의 Data에 대하여 높은 확률로
학습한 Class 중 하나라고 예측한다.*

** Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 427-436).*

Open Set Recognition

(1) Standard Multi-class Classification

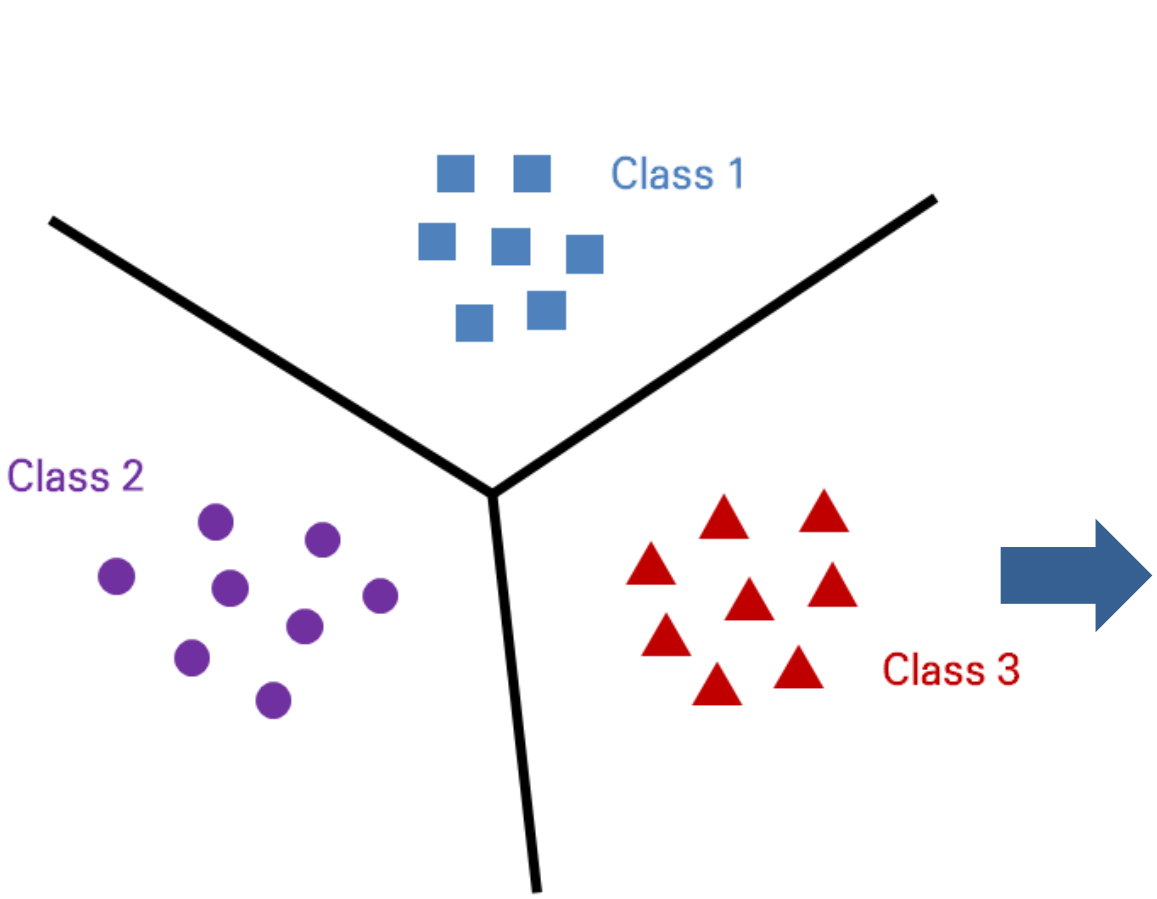


결정 경계 $y^* = \text{argmax}_j \, P(y = j|\mathbf{x})$

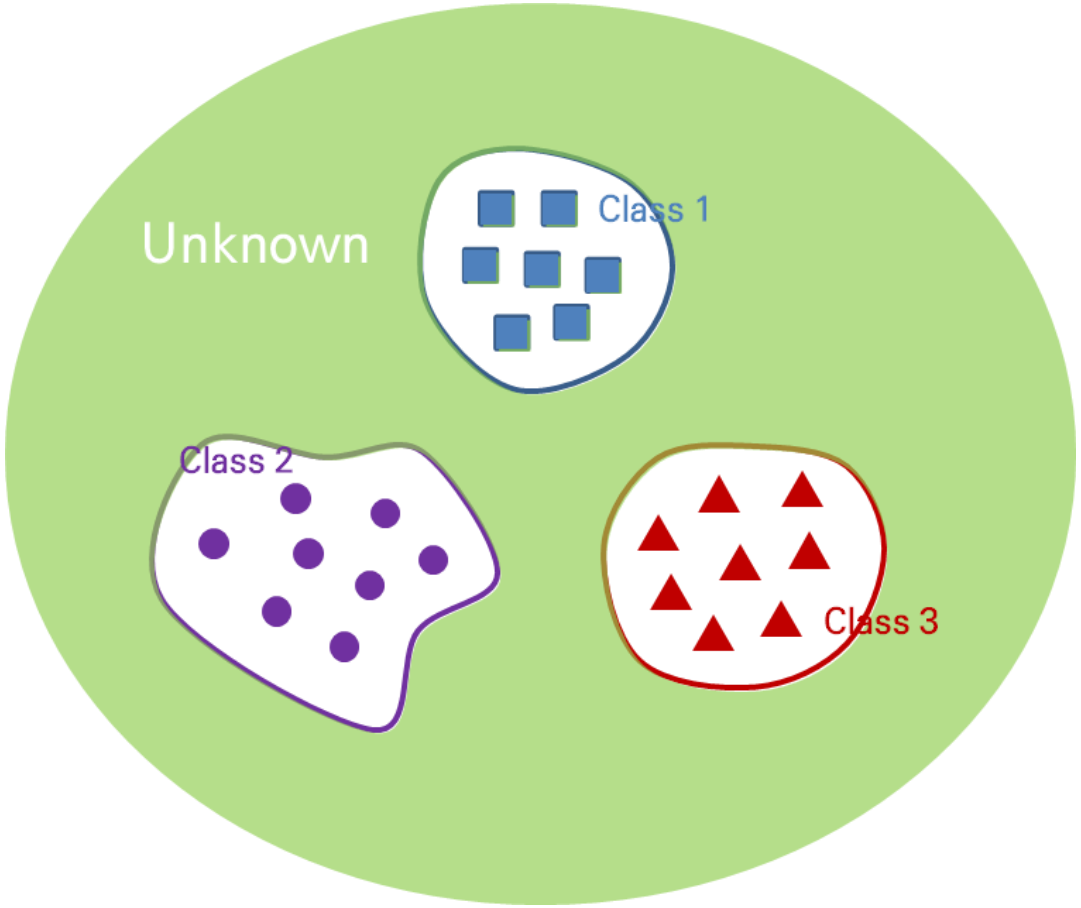
Unknown Data

$$SoftMax\ score = \frac{\exp(V_3)}{\sum_k^C \exp(V_k)}$$

Open Set Recognition



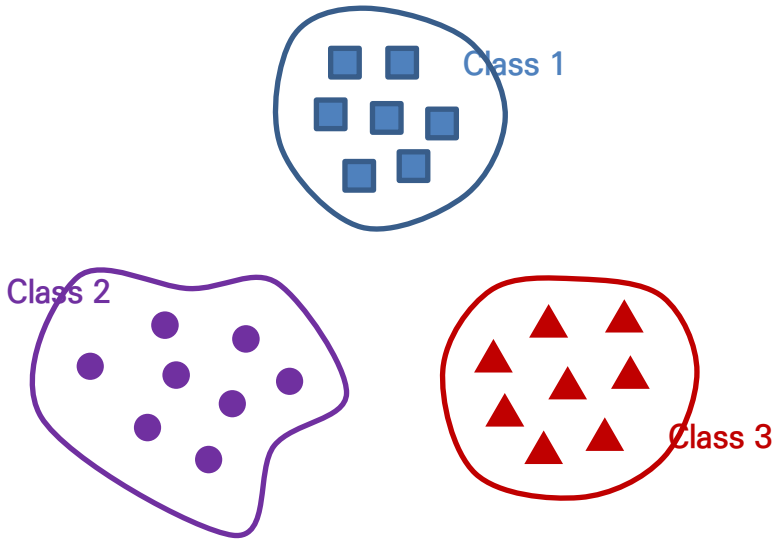
Closed Set Classification



Open Set Recognition

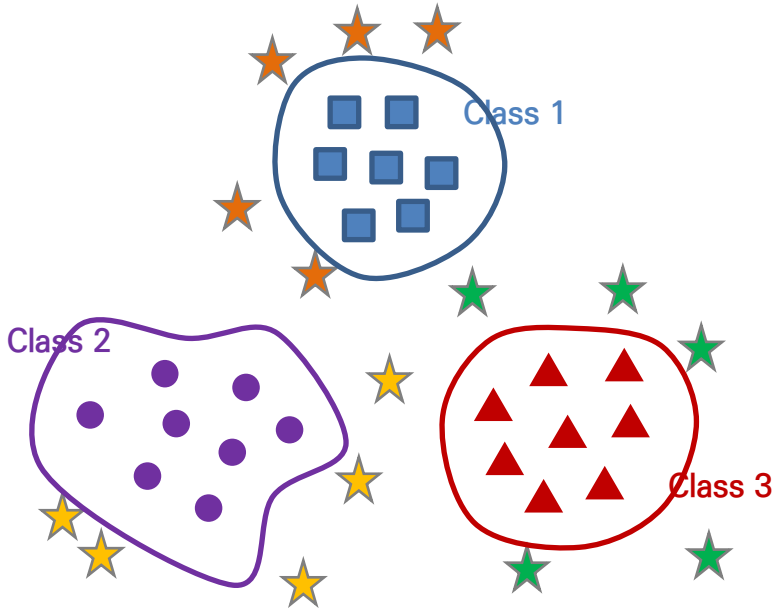
Open Set Recognition

In-distribution 정보 활용



학습 Class(Target Classes)로부터
정보를 추출하여 결정경계를 재구성

Out-distribution(Background Data) 정보 활용

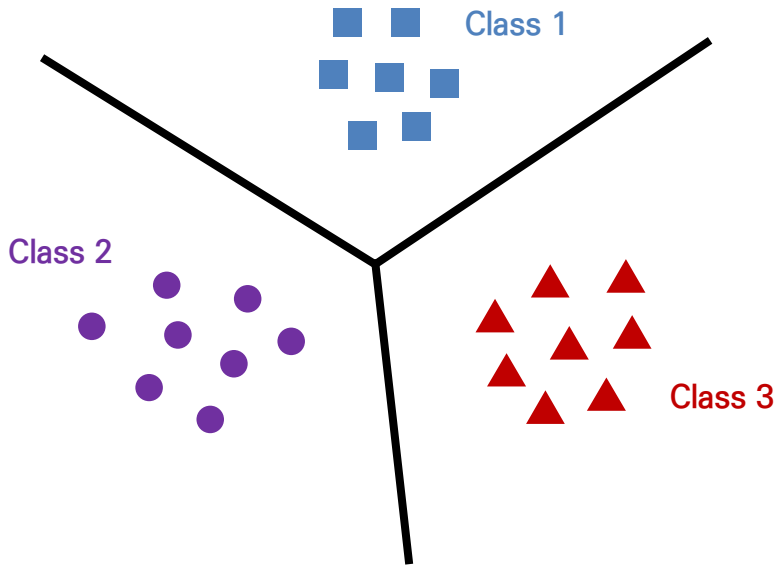


학습 Class가 아닌 Data(Background Data)를
모델의 학습에 반영시켜 결정경계를 재구성

Open Set Recognition

(2) Open Set Recognition with Extreme Value Theorem

In-distribution 정보 활용

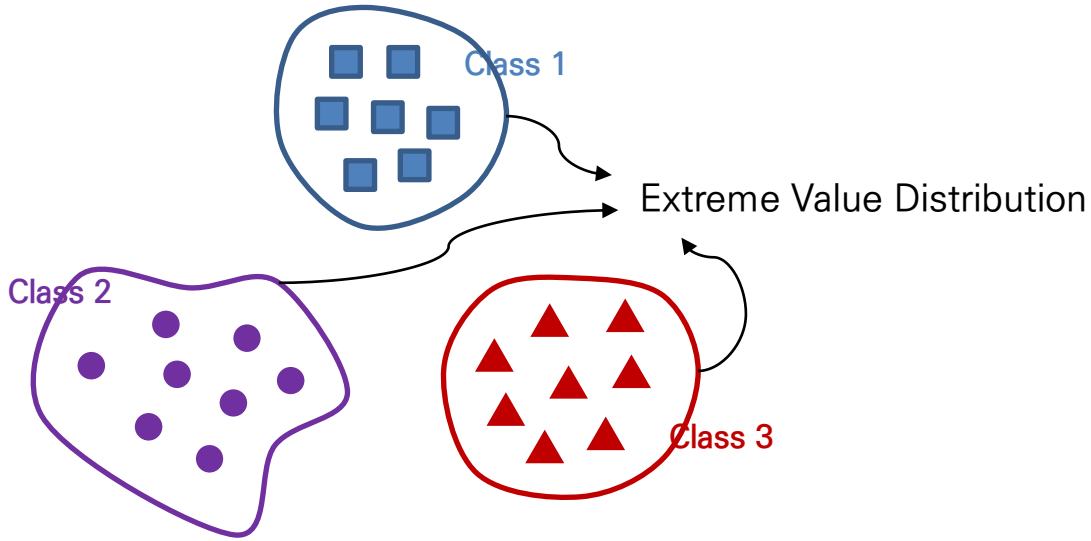


학습된 Classifier의 결정 경계에
각 Class별 학습Data를 기반으로 각 Class별 극단 분포를 도출하고
극단 분포를 통해 결정 경계를 Update

Open Set Recognition

(2) Open Set Recognition with Extreme Value Theorem

In-distribution 정보 활용



학습된 Classifier의 결정 경계에
각 Class별 학습Data를 기반으로 **각 Class별 극단 분포를 도출**하고
극단 분포를 통해 결정 경계를 Update

Open Set Recognition

(2) Open Set Recognition with Extreme Value Theorem

Open Set Recognition in Deep Networks

2020년 2월 15일 오후 6:33 / 조회수: 1146

REFERENCES

Open Set Recognition in Deep Networks_김상훈.pdf

INFORMATION

2020년 2월 21일 오후 1시 ~ 고려대학교 신공학관 221호

발표자:  김상훈

TOPIC

Open Set Recognition in Deep Networks

OVERVIEW

현재까지 Deep Networks에서 Classification/Recognition에 대한 연구는 활발히 진행되어 왔다. 하지만 기존의 Closed Set 기반의 Deep Networks는 학습단계에서 학습되지 않은 테스트 데이터가 들어와도 학습된 클래스 중 하나로만 분류할 수 있다. 금일 세미나에서는 학습되지 않은 데이터를 "unknown"으로 인식할 수 있는 Open Set Recognition의 전반적인 내용과 Open Set Recognition 방법론을 Deep Networks에 적용한 모델에 대한 간략한 소개하고자 한다.

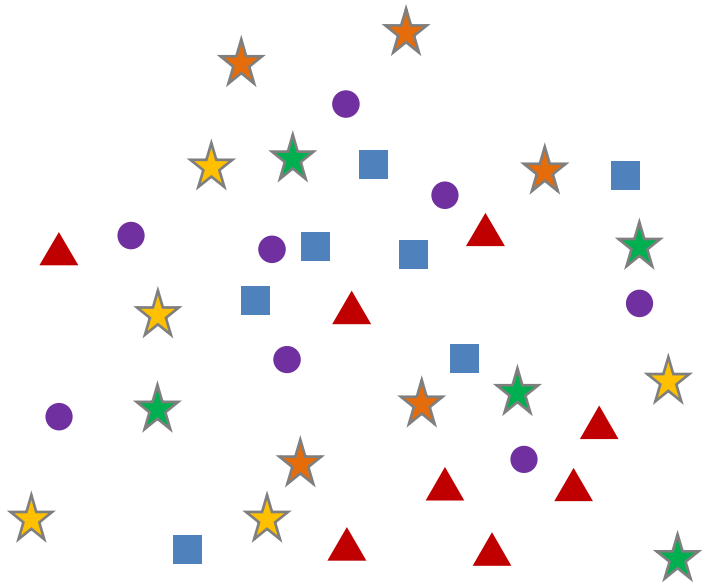
아래는 참고한 자료들을 모아둔 페이지입니다.
https://github.com/iCGY96/awesome_OpenSetRecognition_list

http://dmqa.korea.ac.kr/activity/seminar/281

Open Set Recognition

(3) Open Set Recognition with Background Data

Out-distribution(Background Data) 정보 활용

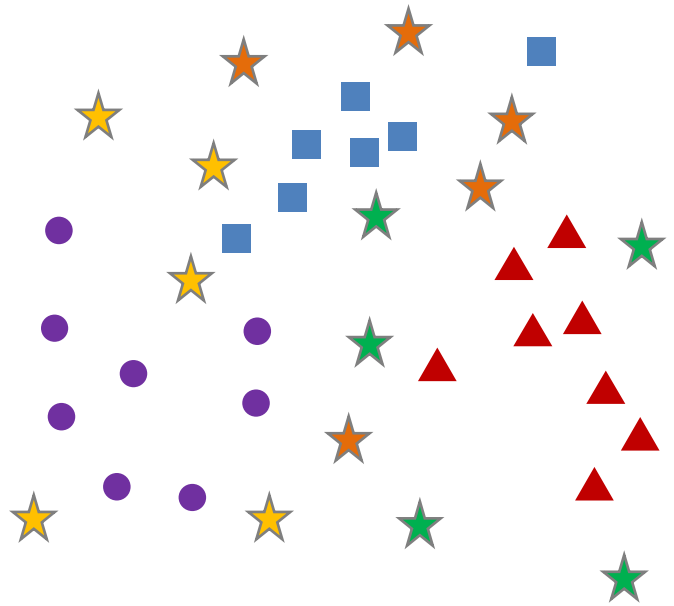


모델의 학습에 Target Classes가 아닌 Class의 Data (Background Data)를 포함하여
Target Classes는 Class별로 뭉치도록
Non-Target Classes (Background Data)는 어떤 Class에도 속하지 않도록 밀어내며 학습

Open Set Recognition

(3) Open Set Recognition with Background Data

Out-distribution(Background Data) 정보 활용

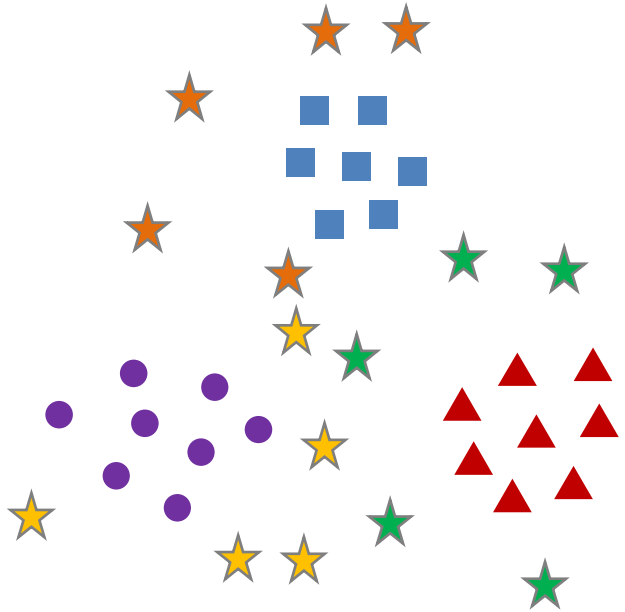


모델의 학습에 Target Classes가 아닌 Class의 Data (Background Data)를 포함하여
Target Classes는 Class별로 뭉치도록
Non-Target Classes (Background Data)는 어떤 Class에도 속하지 않도록 밀어내며 학습

Open Set Recognition

(3) Open Set Recognition with Background Data

Out-distribution(Background Data) 정보 활용

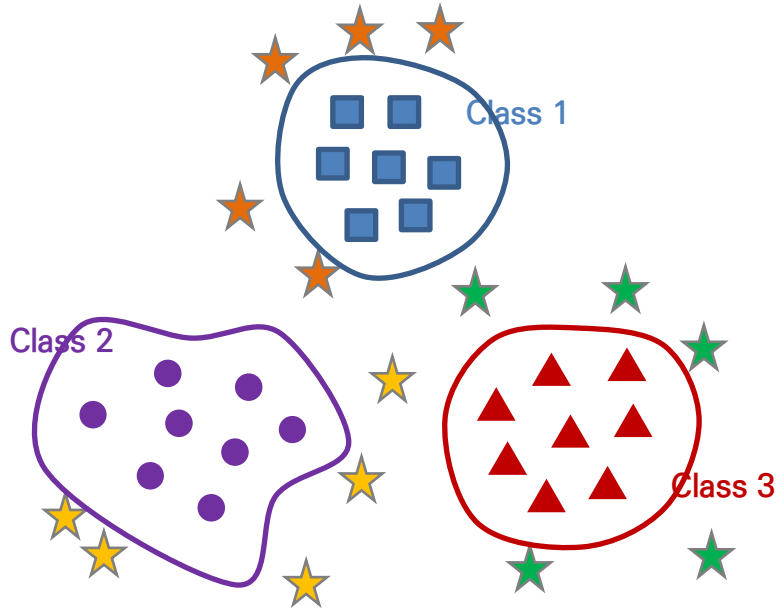


모델의 학습에 Target Classes가 아닌 Class의 Data (Background Data)를 포함하여
Target Classes는 Class별로 뭉치도록
Non-Target Classes (Background Data)는 어떤 Class에도 속하지 않도록 밀어내며 학습

Open Set Recognition

(3) Open Set Recognition with Background Data

Out-distribution(Background Data) 정보 활용



모델의 학습에 Target Classes가 아닌 Class의 Data (Background Data)를 포함하여
Target Classes는 Class별로 뭉치도록
Non-Target Classes (Background Data)는 어떤 Class에도 속하지 않도록 밀어내며 학습

Background Data-based Methods

(1) Reducing Network Agnostophobia

- 2018년 NIPS(Neural Information Processing Systems) 에 발표된 논문
- 2020년 10월 15일 기준 49회 인용

Reducing Network Agnostophobia

Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult
Vision and Security Technology Lab, University of Colorado Colorado Springs
{adhamija | mgunther | tboult} @ vast.uccs.edu

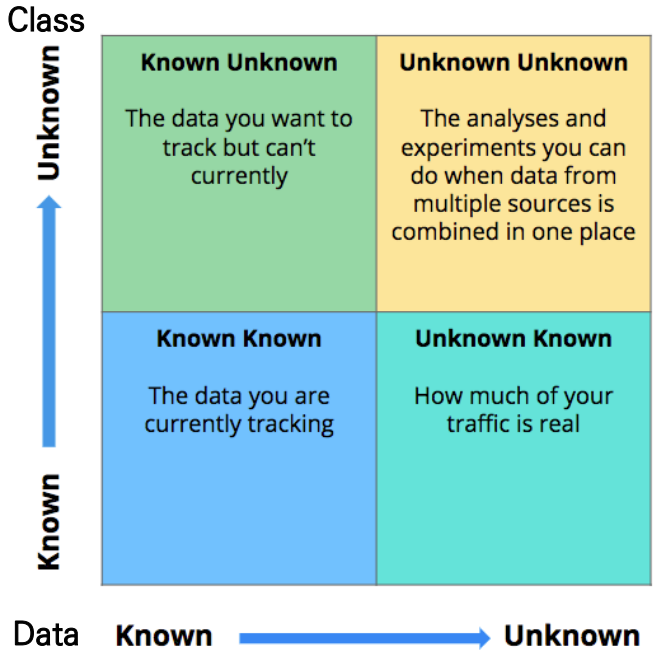
Abstract

Agnostophobia, the fear of the unknown, can be experienced by deep learning engineers while applying their networks to real-world applications. Unfortunately, network behavior is not well defined for inputs far from a networks training set. In an uncontrolled environment, networks face many instances that are not of interest to them and have to be rejected in order to avoid a false positive. This problem has previously been tackled by researchers by either *a*) thresholding softmax, which by construction cannot return *none of the known classes*, or *b*) using an additional background or garbage class. In this paper, we show that both of these approaches help, but are generally insufficient when previously unseen classes are encountered. We also introduce a new evaluation metric that focuses on comparing the performance of multiple approaches in scenarios where such unseen classes or unknowns are encountered. Our major contributions are simple yet effective

Background Data-based Methods

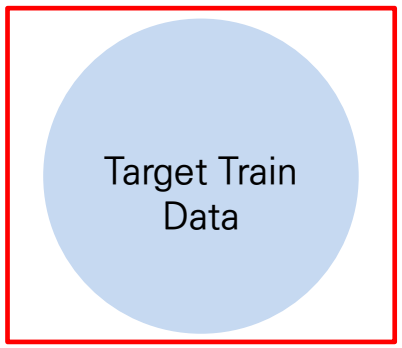
(1) Reducing Network Agnostophobia

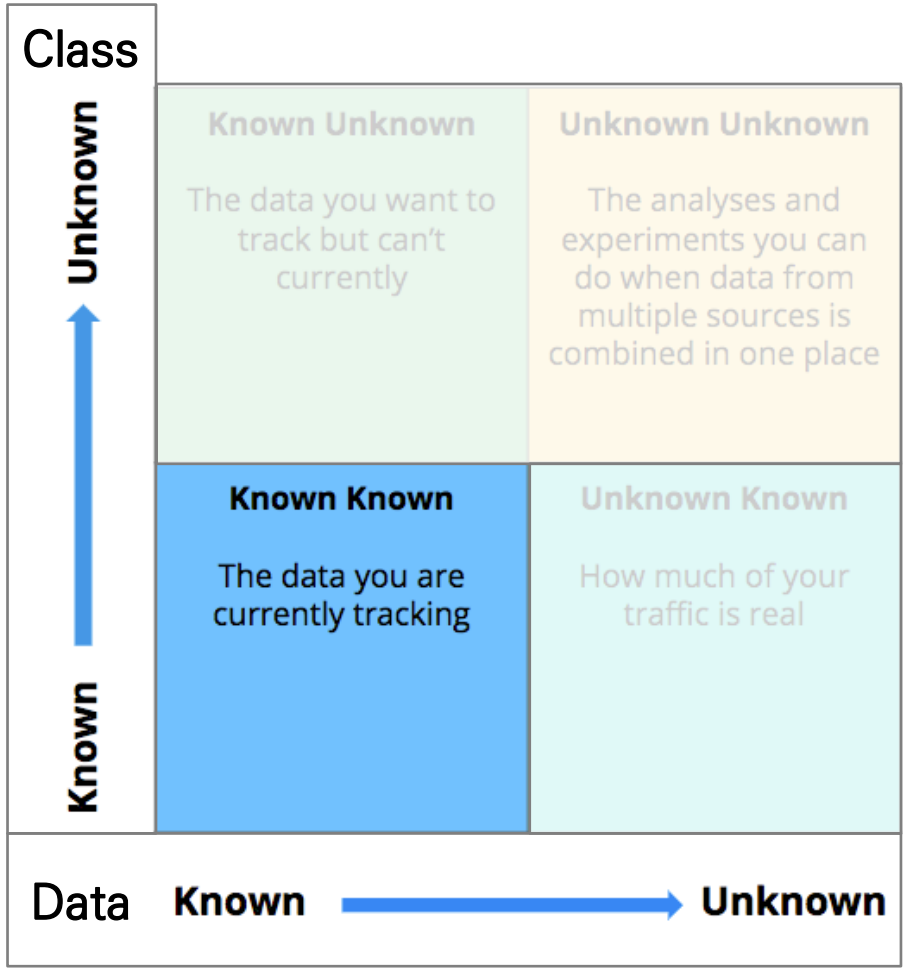
- ❖ Main idea
  - Target Train Data : 모델이 학습해야 할 학습Class의 Data (Known Known Classes)
  - Target Test Data : 학습된 모델이 분류해야 할 학습Class의 Data (Unknown Known Classes)
  - Unknown Classes의 집합 : 모델이 학습한 Class 외의 모든 Class Data의 집합
    - Background Data : Unknown Classes의 집합 중 알고 있는 Data (Known Unknown Classes)
    - Unknown Data : Unknown Classes의 집합 중 B에 속하지 않는 나머지 Data (Unknown Unknown Classes)



Background Data-based Methods

(1) Reducing Network Agnostophobia



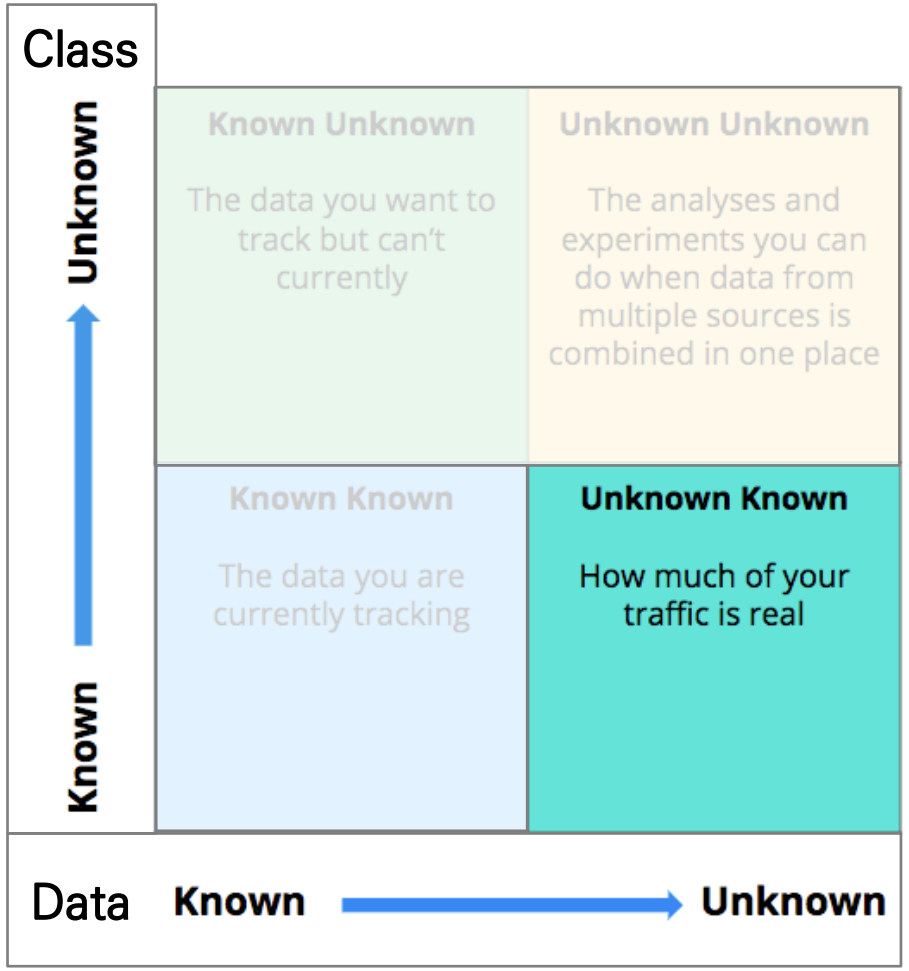


: 수집된 Data로 모델이 학습해야 할 학습 Class의 Data
(Standard Multi-class Classification Task에서 **학습Data**)

Background Data-based Methods

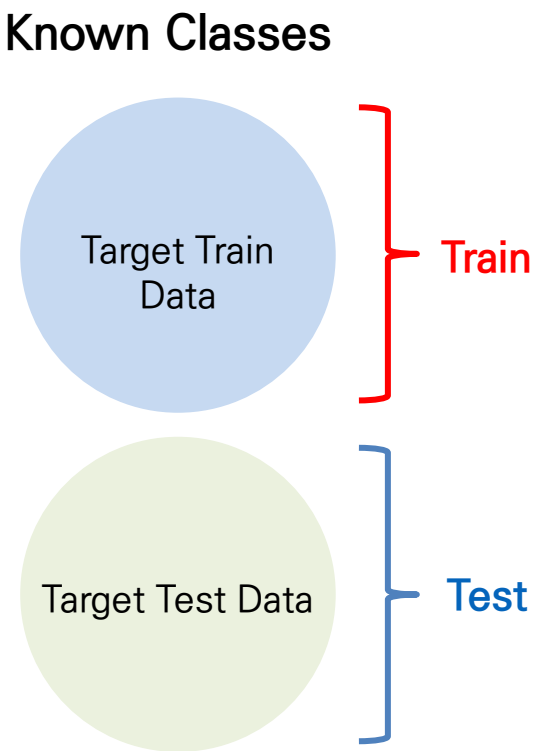
(1) Reducing Network Agnostophobia

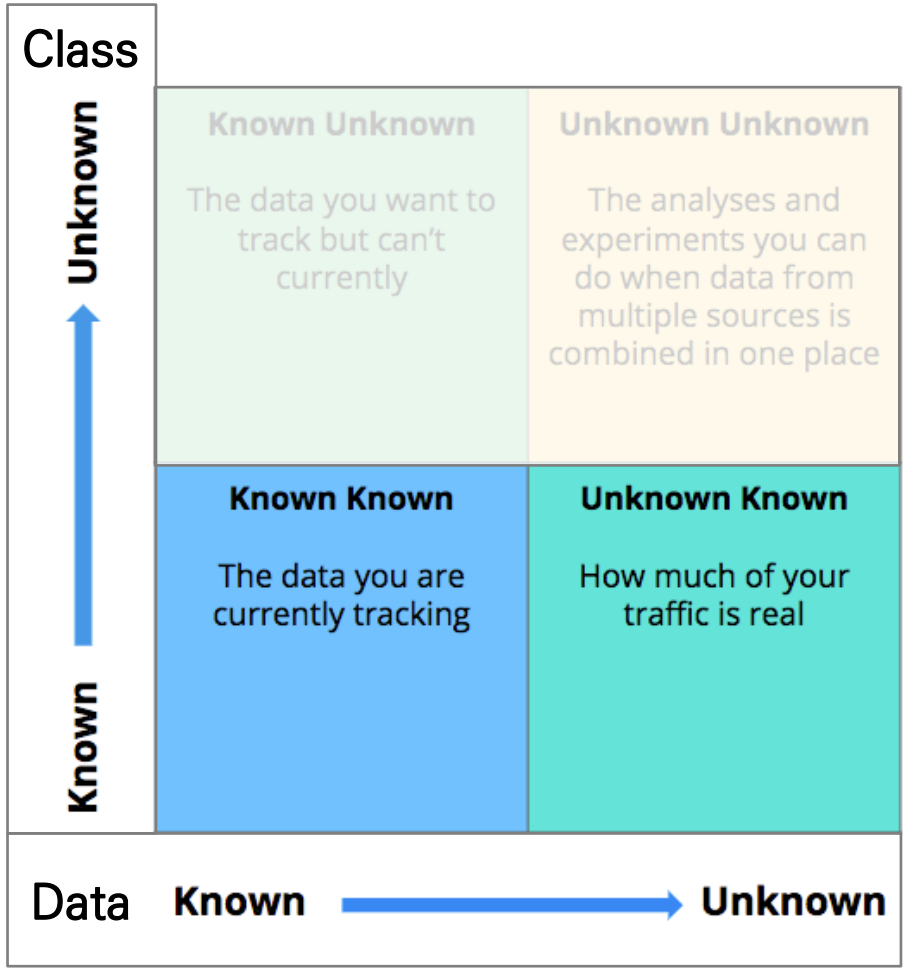
: 수집되지 않은 Data로 학습된 모델이 분류해야 할 학습 Class의 Data
(Standard Multi-class Classification Task에서 **테스트Data**)



Background Data-based Methods

(1) Reducing Network Agnostophobia





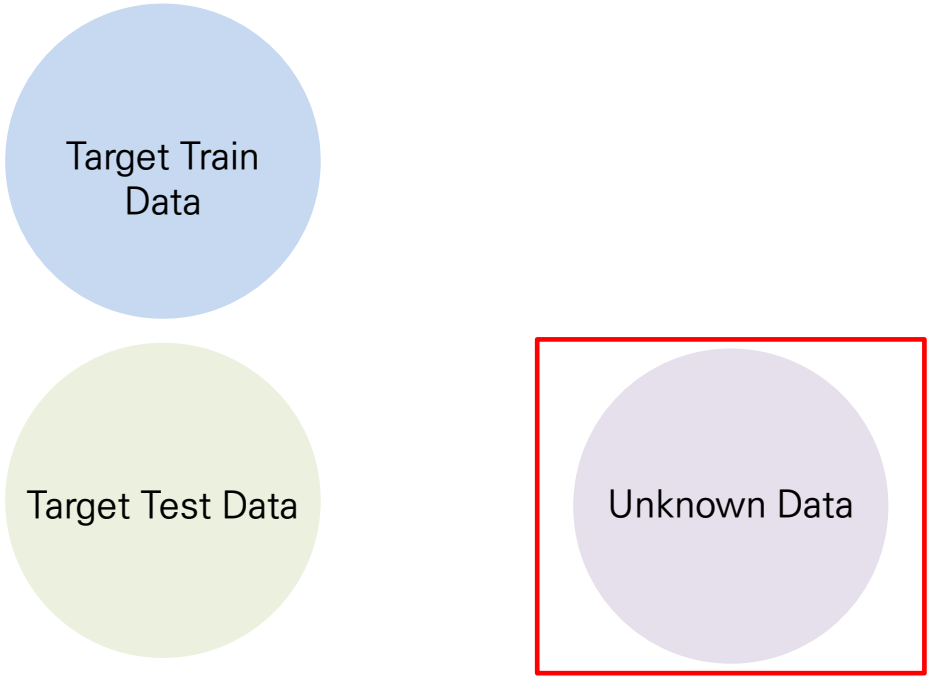
일반적인 Classification Task에서는 수집된 Target Train Data로 모델을 학습하여 앞으로 들어올 Target Test Data를 분류한다.

Background Data-based Methods

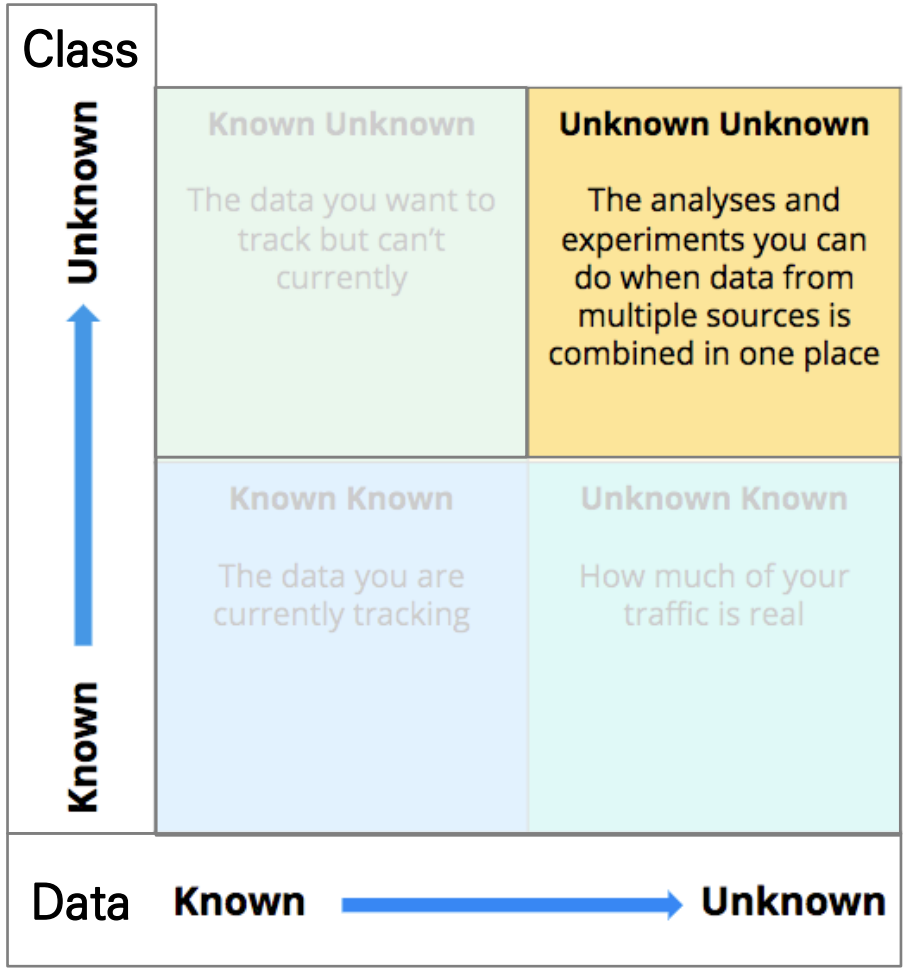
(1) Reducing Network Agnostophobia

❖ Open Set Recognition

Known Classes **Unknown Classes**



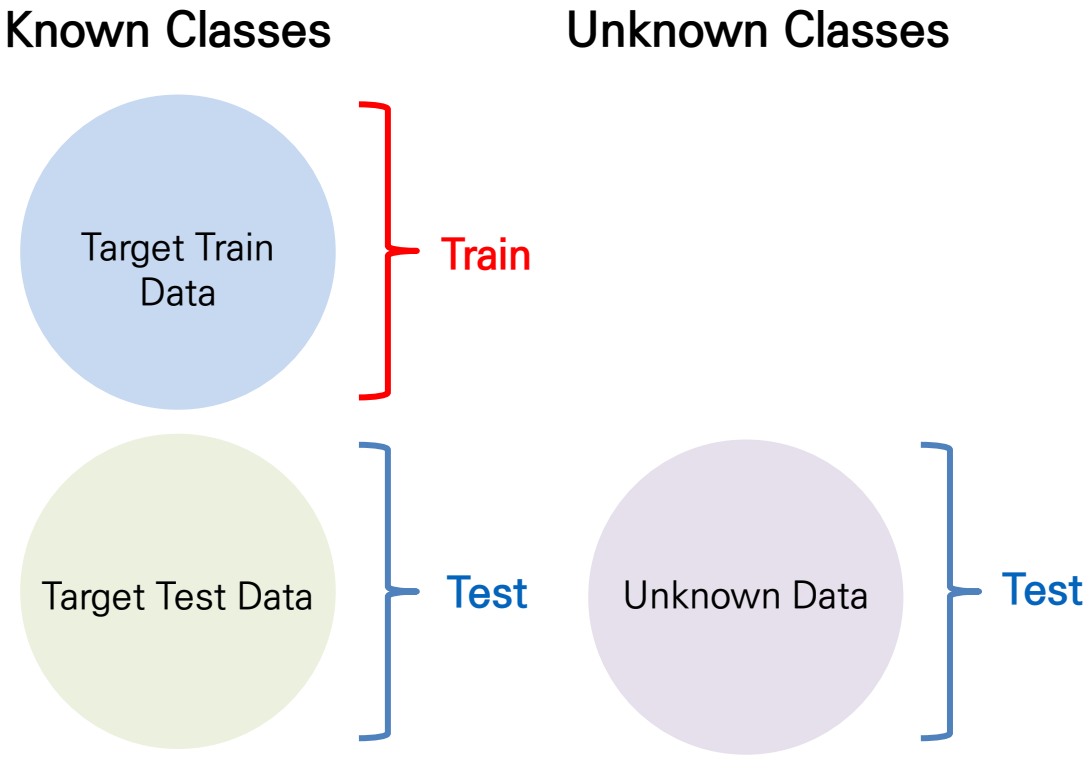
: 수집되지 않았고 모델이 학습한 Classes가 아닌 모르는 Data

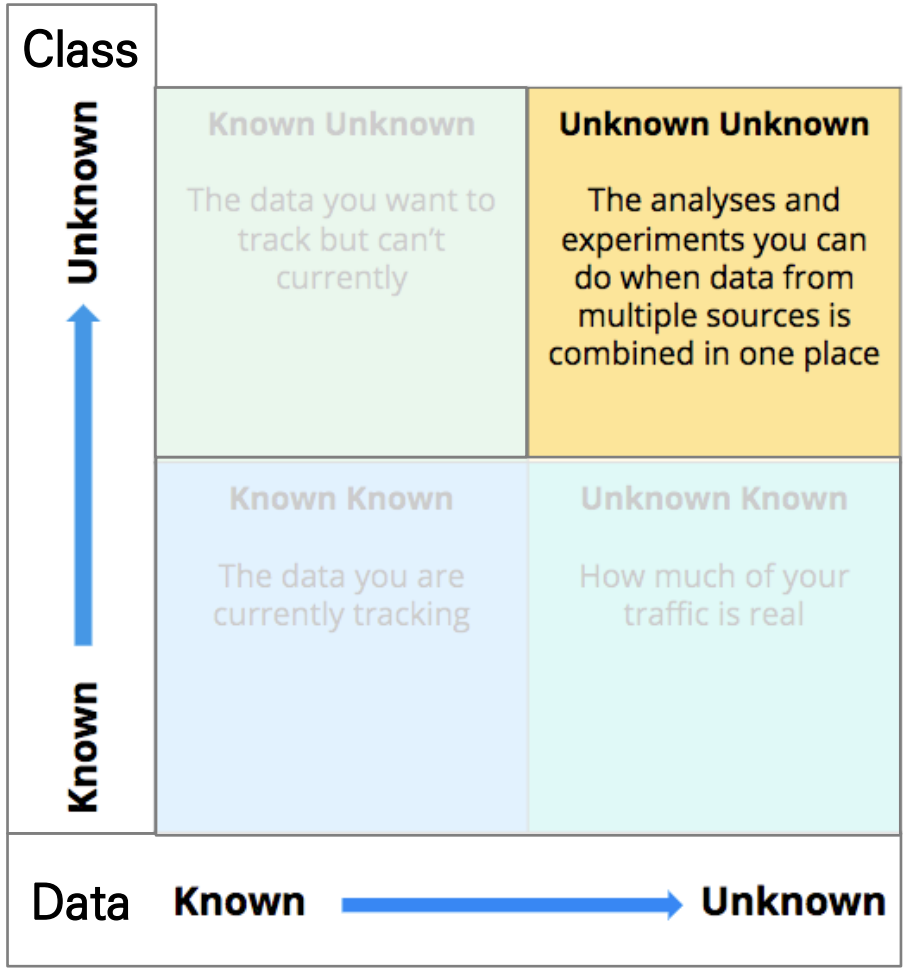


Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ Open Set Recognition



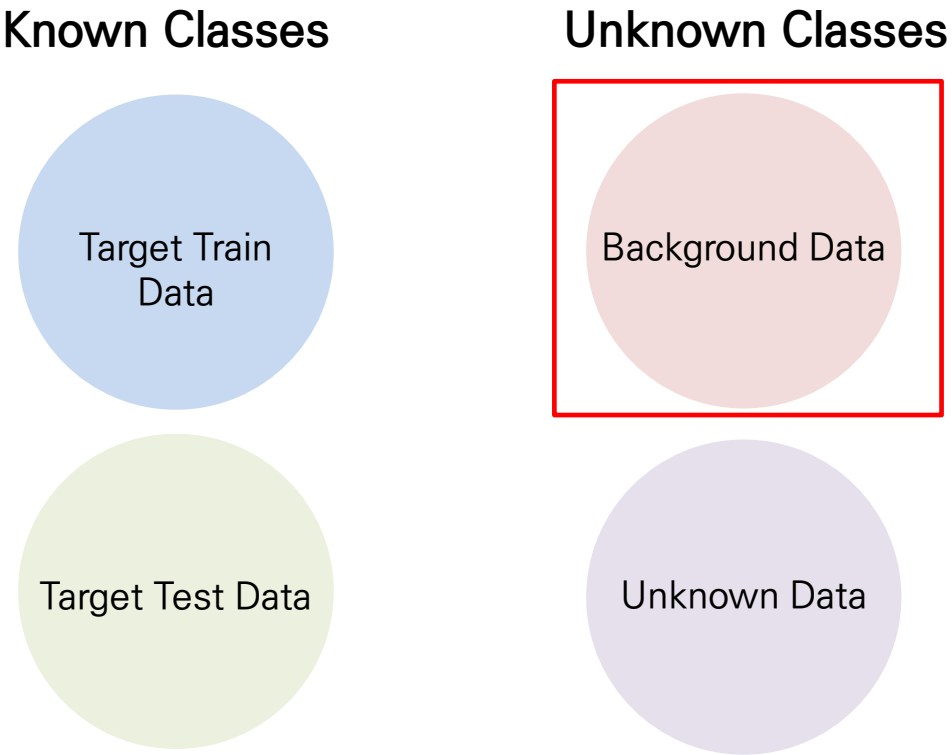


Open Set Recognition의 Task는 Unknown Data를
학습Classes로 **오분류**하지 않고 모르는 Class로 판단할 수 있어야 한다.

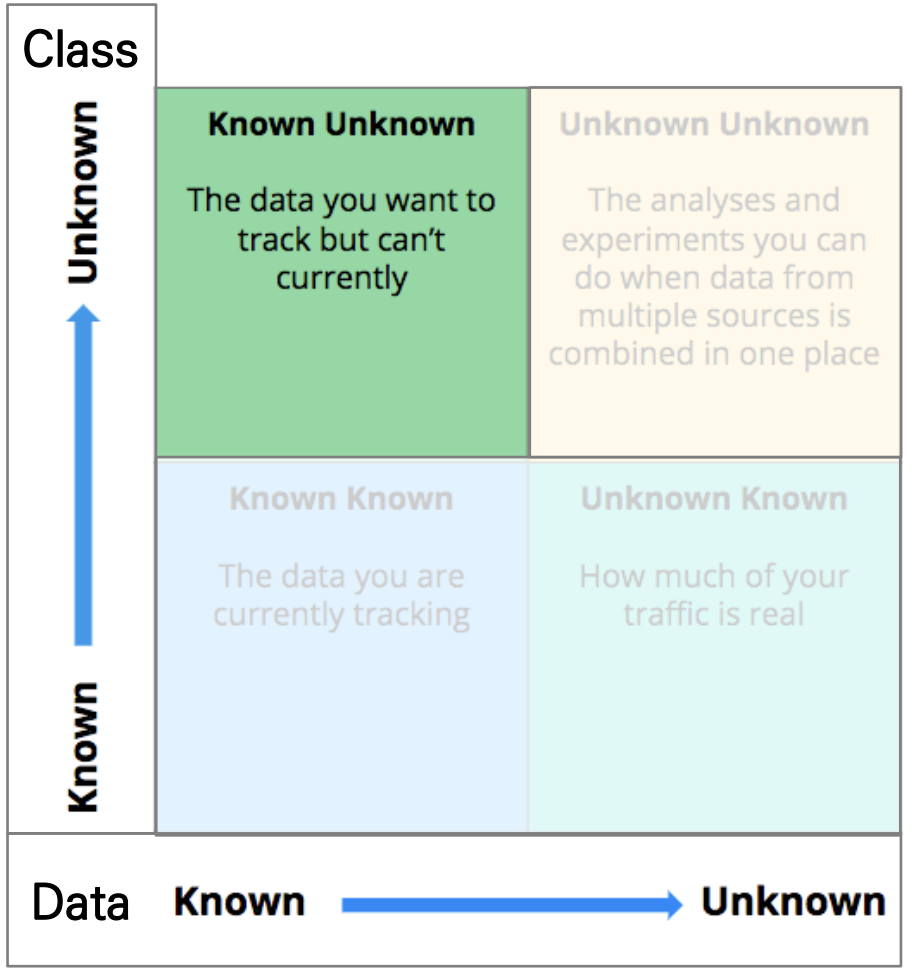
Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ Main idea



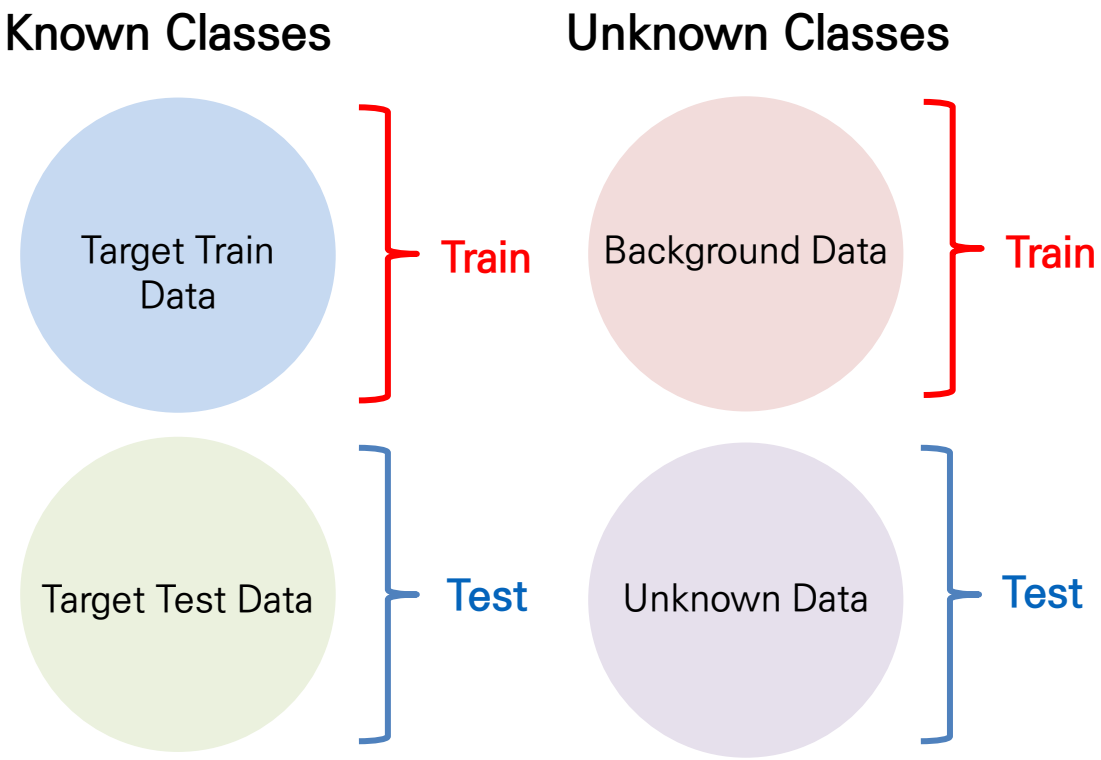
: 수집은 되었지만 모델이 학습해야 할 Class가 아닌 Class의 Data

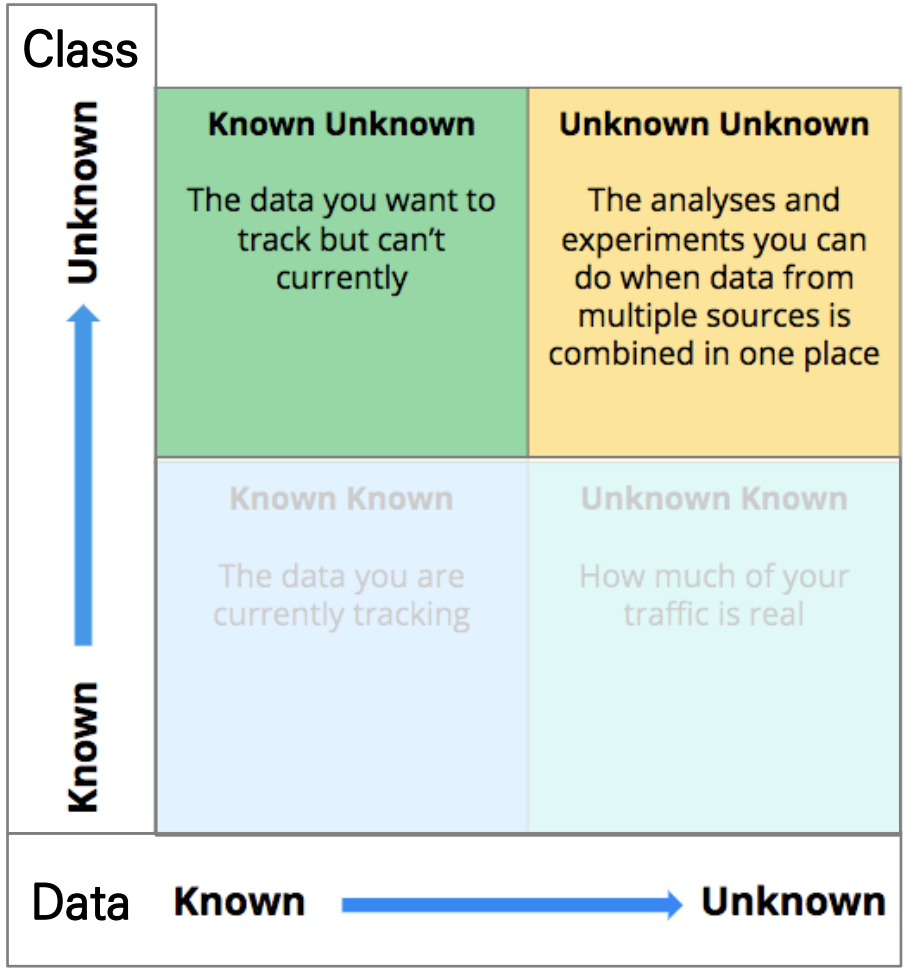


Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ Main idea



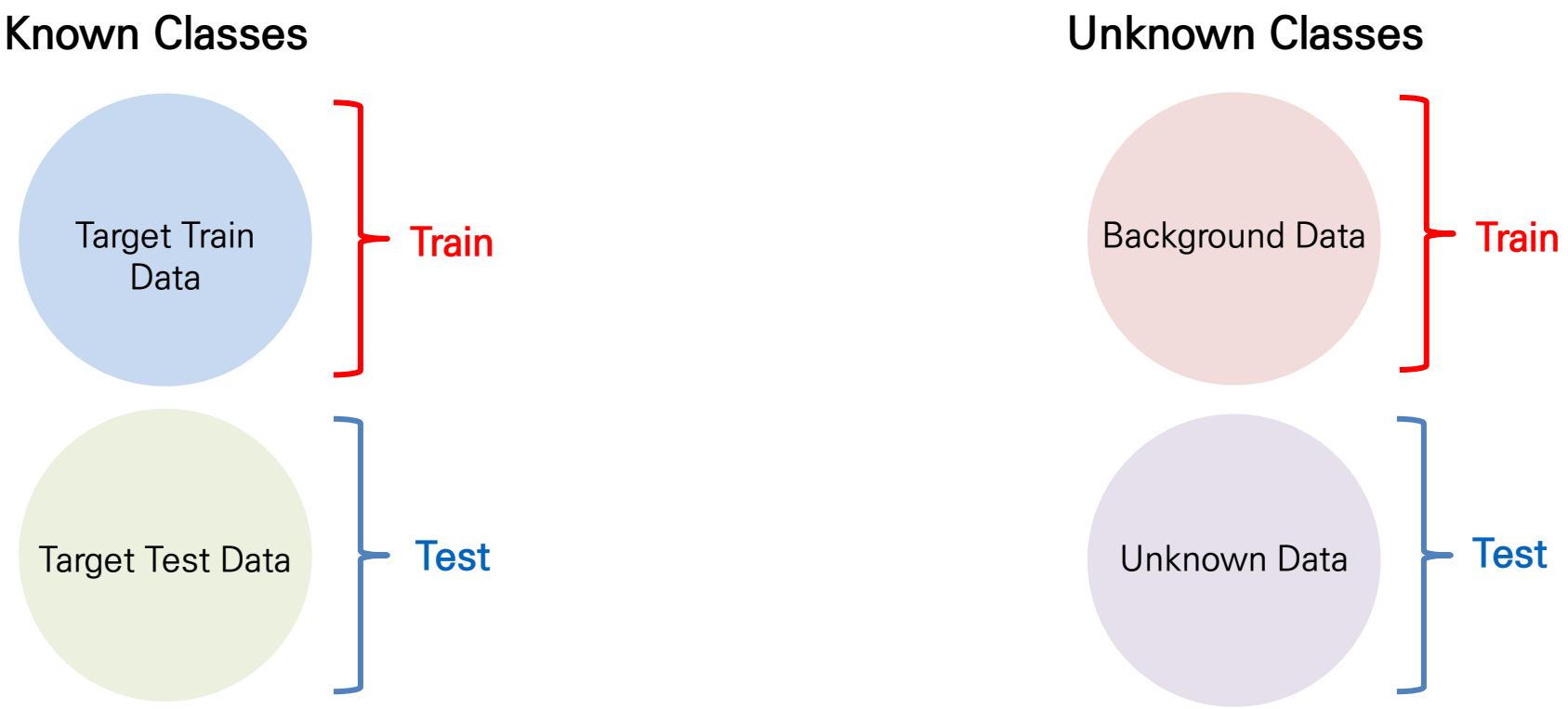


수집되었지만 학습해야 할 Class가 아닌 Class의 Data를 학습에 사용하여 Unknown Data에 대한 Detection 성능 향상을 꾀함.

Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ 논문에서 사용한 Data 소개



Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ 논문에서 사용한 Data 소개

Known Classes

MNIST Data

Unknown Classes

Background Data — Train

Unknown Data — Test

Background Data-based Methods

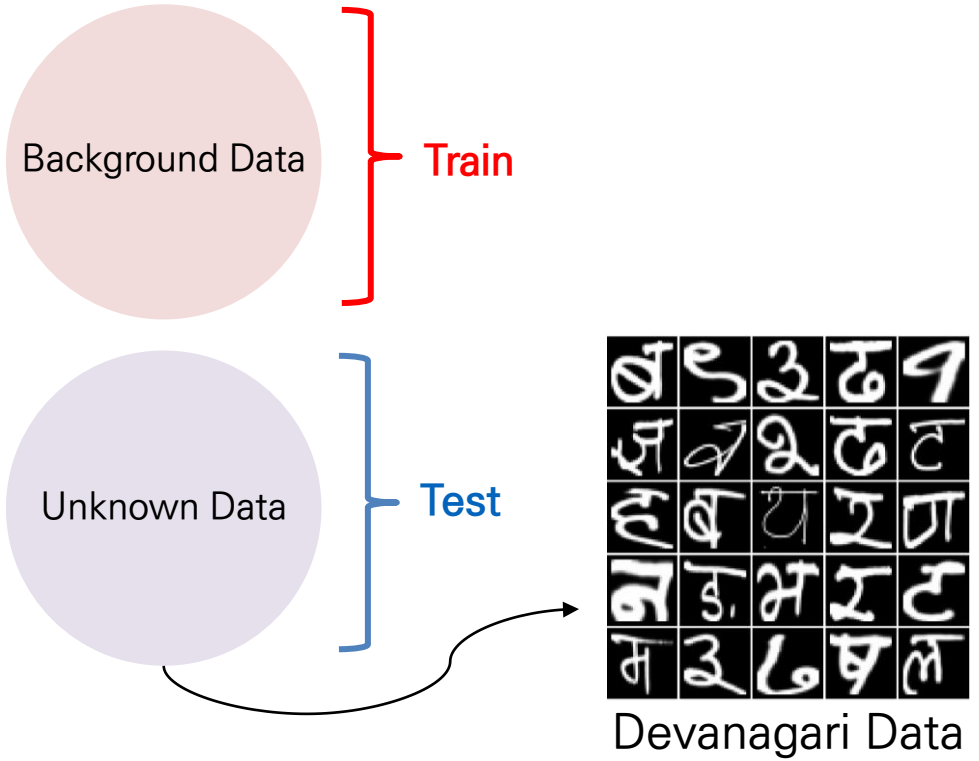
(1) Reducing Network Agnostophobia

❖ 논문에서 사용한 Data 소개

Known Classes

MNIST Data

Unknown Classes



Background Data-based Methods

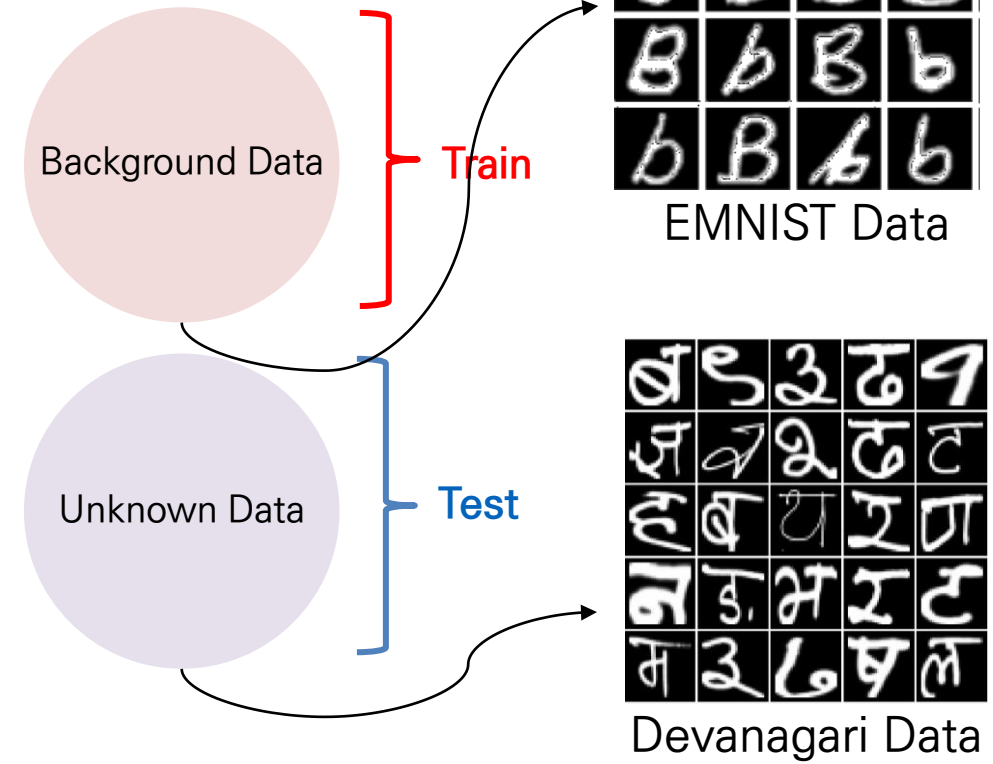
(1) Reducing Network Agnostophobia

❖ 논문에서 사용한 Data 소개

Known Classes

MNIST Data

Unknown Classes



Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ 논문에서 사용한 Data 소개

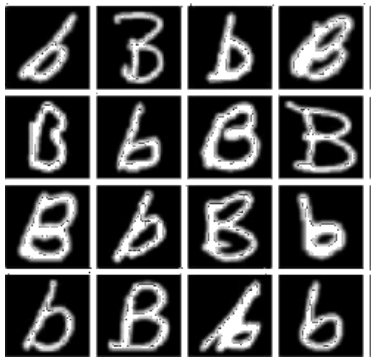
Known Classes

MNIST Data

Unknown Classes

Background Data

Train



EMNIST Data

Background Data와 Unknown Data는
다른 Class의 Data

Unknown Data

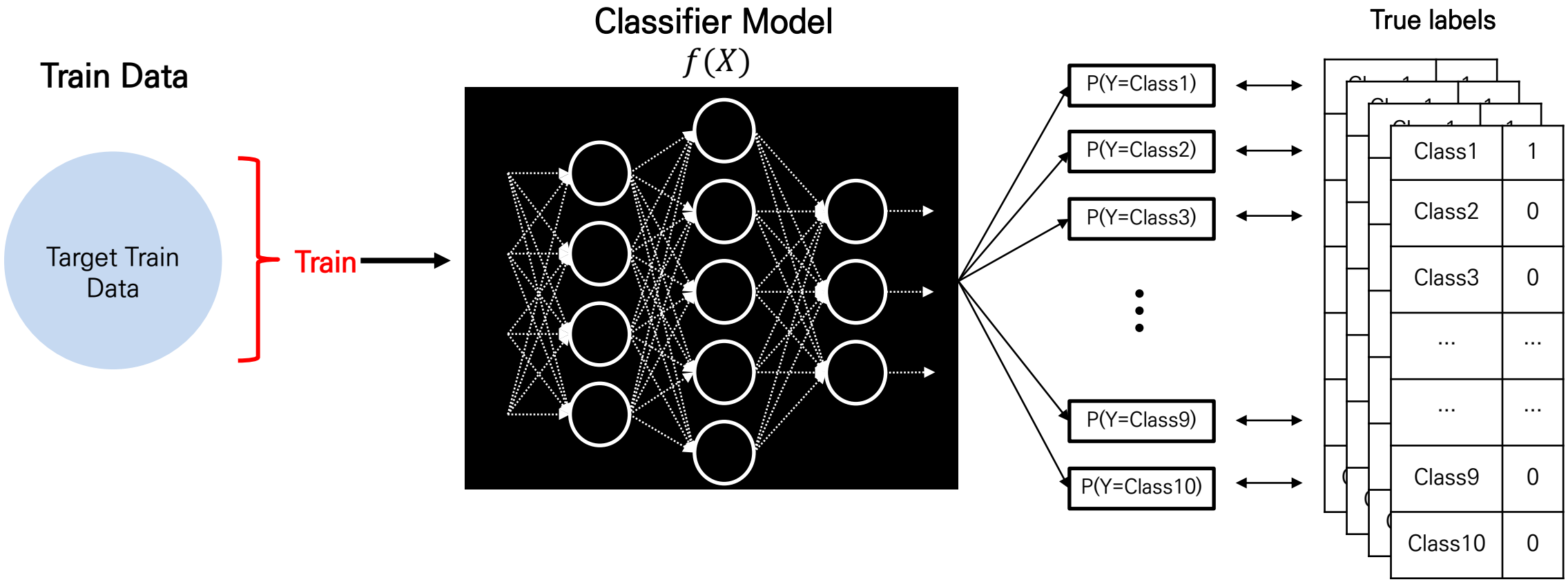
Test

Devanagari Data

Background Data-based Methods

(1) Reducing Network Agnostophobia

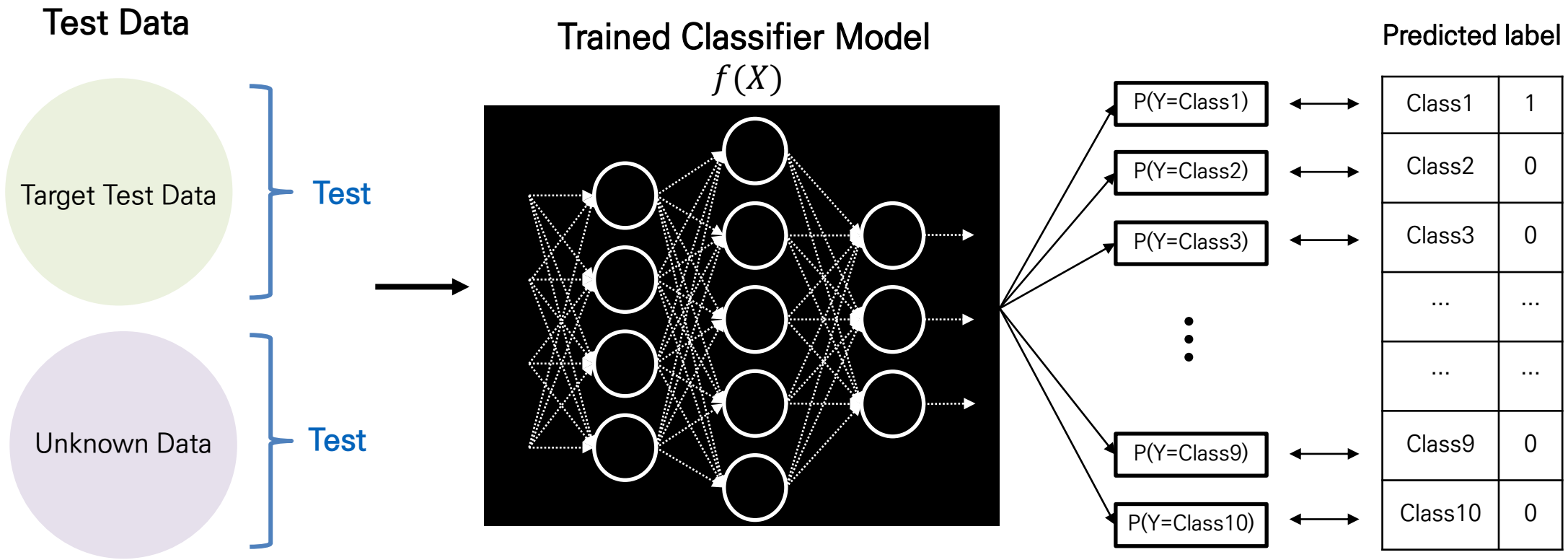
- SoftMax



Background Data-based Methods

(1) Reducing Network Agnostophobia

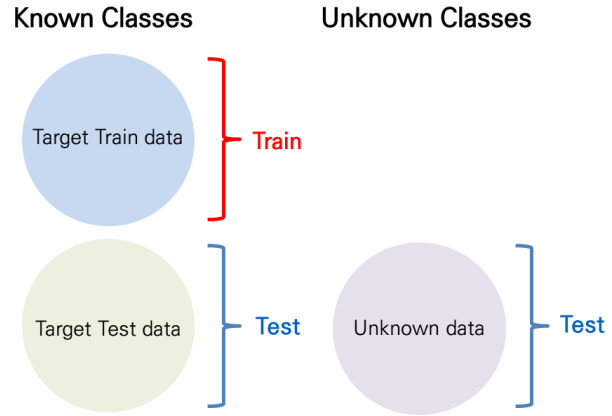
❖ SoftMax



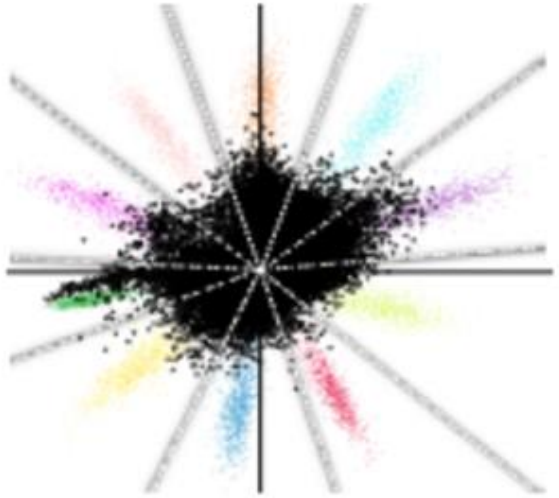
Thresholding 기법 적용

Background Data-based Methods

(1) Reducing Network Agnostophobia



❖ SoftMax



(a) Softmax

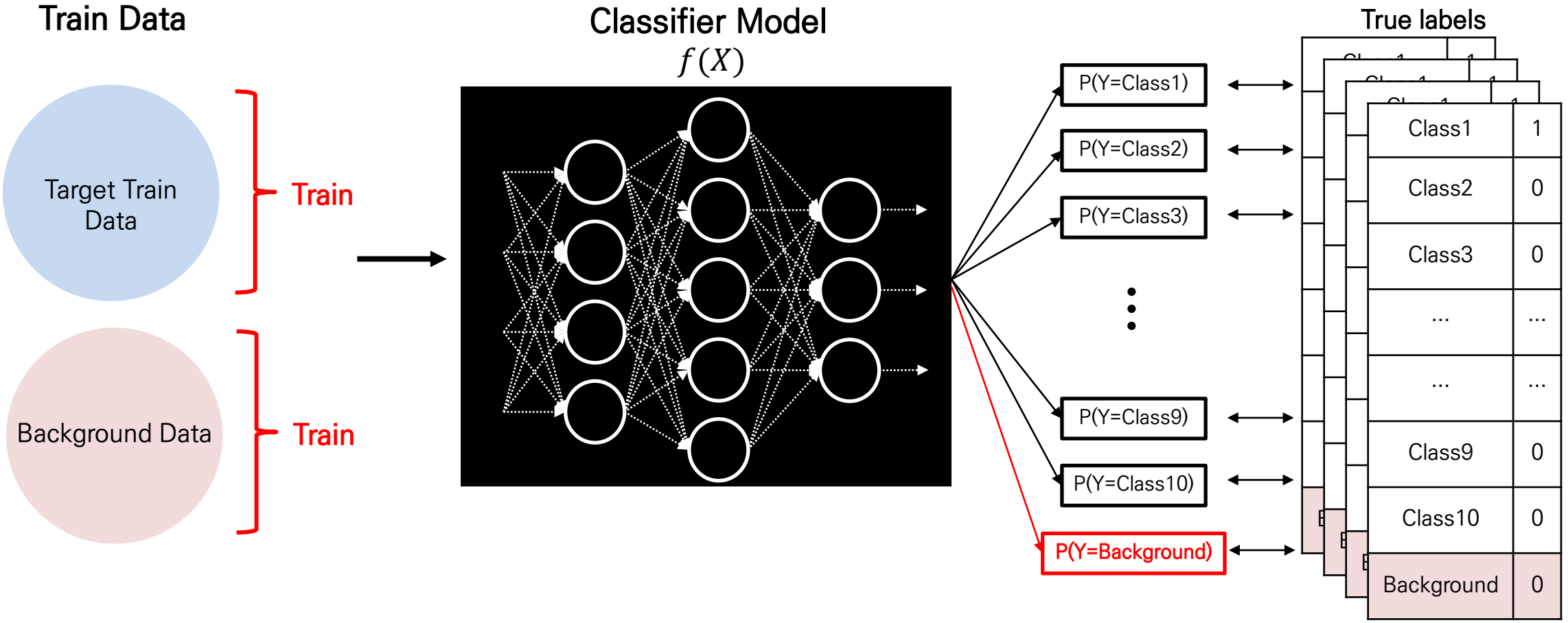
LENET++ network로 테스트 Data를 2차원 특징 공간에 표상하여 시각화

- 검은 색이 아닌 다른 색상으로 표현된 점 : Test Target Data (MNIST Data)
- 검은 색 점 : Unknown Data (Devanagari Data)
- 점선 : SoftMax 점수가 이웃 Class와 동일한 Class 경계

Background Data-based Methods

(1) Reducing Network Agnostophobia

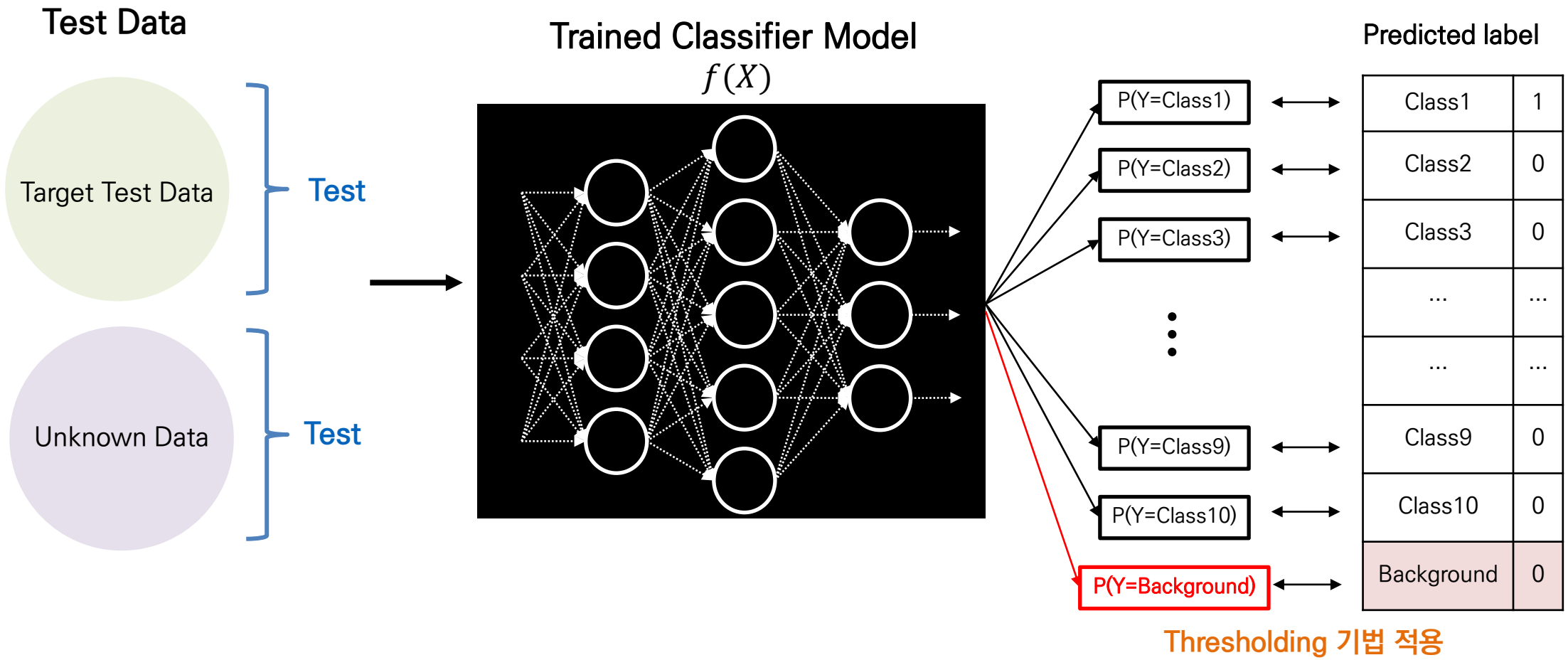
❖ Background



Background Data-based Methods

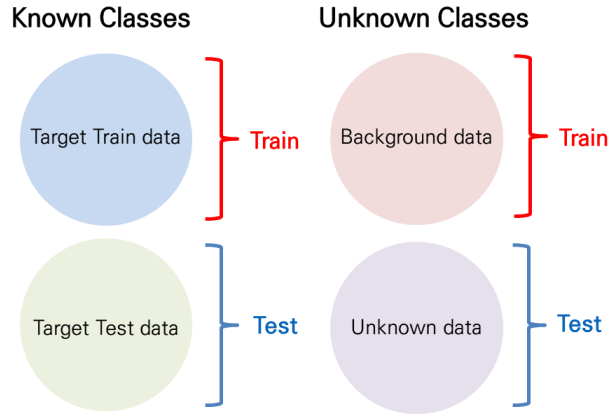
(1) Reducing Network Agnostophobia

❖ Background

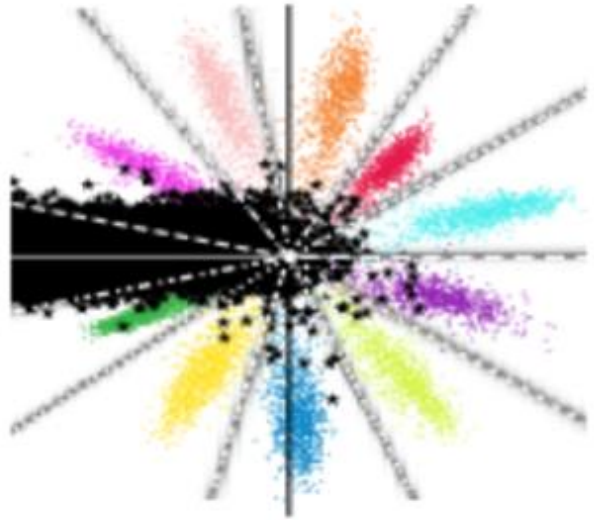


Background Data-based Methods

(1) Reducing Network Agnostophobia



❖ Background



(b) Background

LENET++ network로 테스트 Data를 2차원 특징 공간에 표상하여 시각화

- 검은 색이 아닌 다른 색상으로 표현된 점 : Test Target Data (MNIST Data)
- 검은 색 점 : Unknown Data (Devanagari Data)
- 점선 : SoftMax 점수가 이웃 Class와 동일한 Class 경계

Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ Entropic Open-Set Loss

Cross Entropy Loss

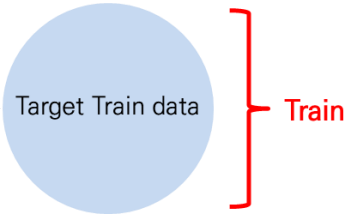
$$Loss = -\sum_{i=1}^{C} t_i \log S_i(x)$$

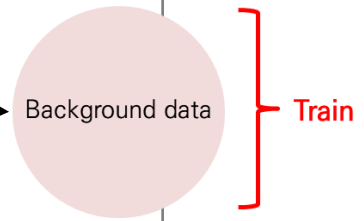
- C : $total\ number\ of\ classes$
- $S_i(x)$: $Softmax\ score\ of\ i-th\ class$
- t_i : $true\ label\ of\ i-th\ class$

Entropic Open-Set Loss

$$Loss = \begin{cases} -\log S_c(x) & if\ x \in D_{targettrain}\ is\ from\ class\ c \\ -\frac{1}{C}\sum_{i=1}^{C} \log S_i(x) & if\ x \in D_{background} \end{cases}$$

- C : $total\ number\ of\ classes$
- $S_i(x)$: $Softmax\ score\ of\ i-th\ class$





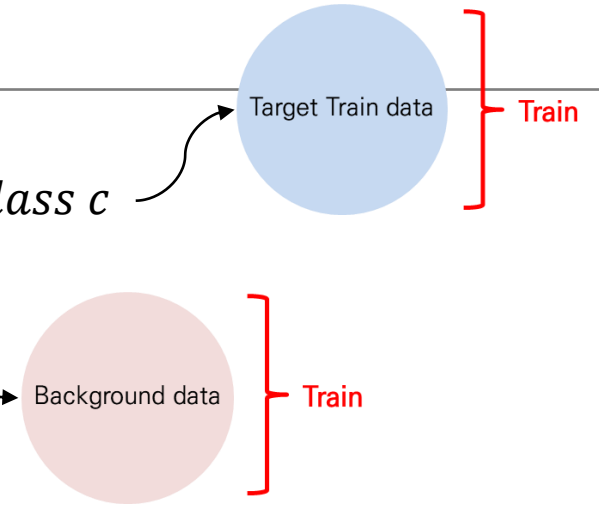
Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ Entropic Open-Set Loss

Entropic Open-Set Loss

$$Loss = \begin{cases} -\log S_c(x) & if\ x \in D_{targettrain}\ is\ from\ class\ c \\ -\frac{1}{C}\sum_{i=1}^{C} \log S_i(x) & if\ x \in D_{background} \end{cases}$$



- C : $total\ number\ of\ classes$
- $S_i(x)$: $Softmax\ score\ of\ i-th\ class$

Lemma 1. *For an input $x \in \mathcal{D}_b'$, the loss $J_E(x)$ is minimized when all softmax responses $S_c(x)$ are equal: $\forall c \in \mathcal{C} : S_c(x) = S = \frac{1}{C}$.*

Lemma 2. *When the logit values are equal, the loss $J_E(x)$ is minimized.*

Background Data-based Methods

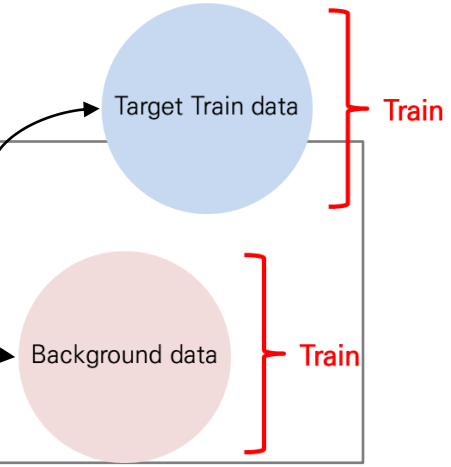
(1) Reducing Network Agnostophobia

❖ Objectosphere Loss

Objectosphere Loss

$$Loss_{Objectosphere} = Loss_{Entropic} + \lambda \begin{cases} \max(\xi - \|F(x)\|, 0)^2 & if\ x \in D_{targettrain} \\ \|F(x)\|^2 & if\ x \in D_{background} \end{cases}$$

- $F(x)$: *deep feature vector that feeds into the logit layer*



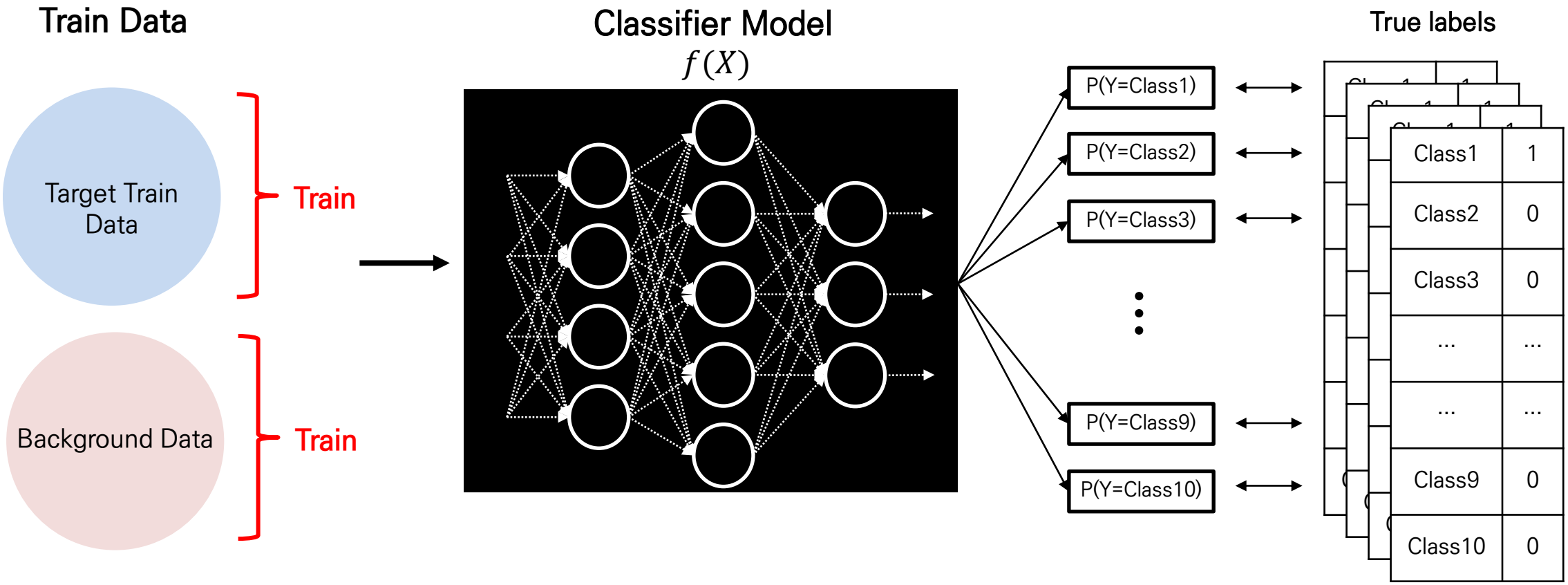
Theorem 1. *For networks whose logit layer does not have bias terms, and for $x \in \mathcal{D}'_b$, the loss $J_E(x)$ is minimized when the deep feature vector $F(x)$ that feeds into the logit layer is the zero vector, at which point the softmax responses $S_c(x)$ are equal: $\forall c \in \mathcal{C} : S_c(x) = S = \frac{1}{C}$ and the entropy of softmax and the deep feature is maximized.*

Theorem 2. *For networks whose logit layer does not have bias terms, given an known unknown input $x \in \mathcal{D}'_b$, the loss $J_R(x)$ is minimized if and only if the deep feature vector $F = \vec{0}$, which in turn ensures the softmax responses $S_c(x)$ are equal: $\forall c \in \mathcal{C} : S_c(x) = S = \frac{1}{C}$ and maximizes entropy.*

Background Data-based Methods

(1) Reducing Network Agnostophobia

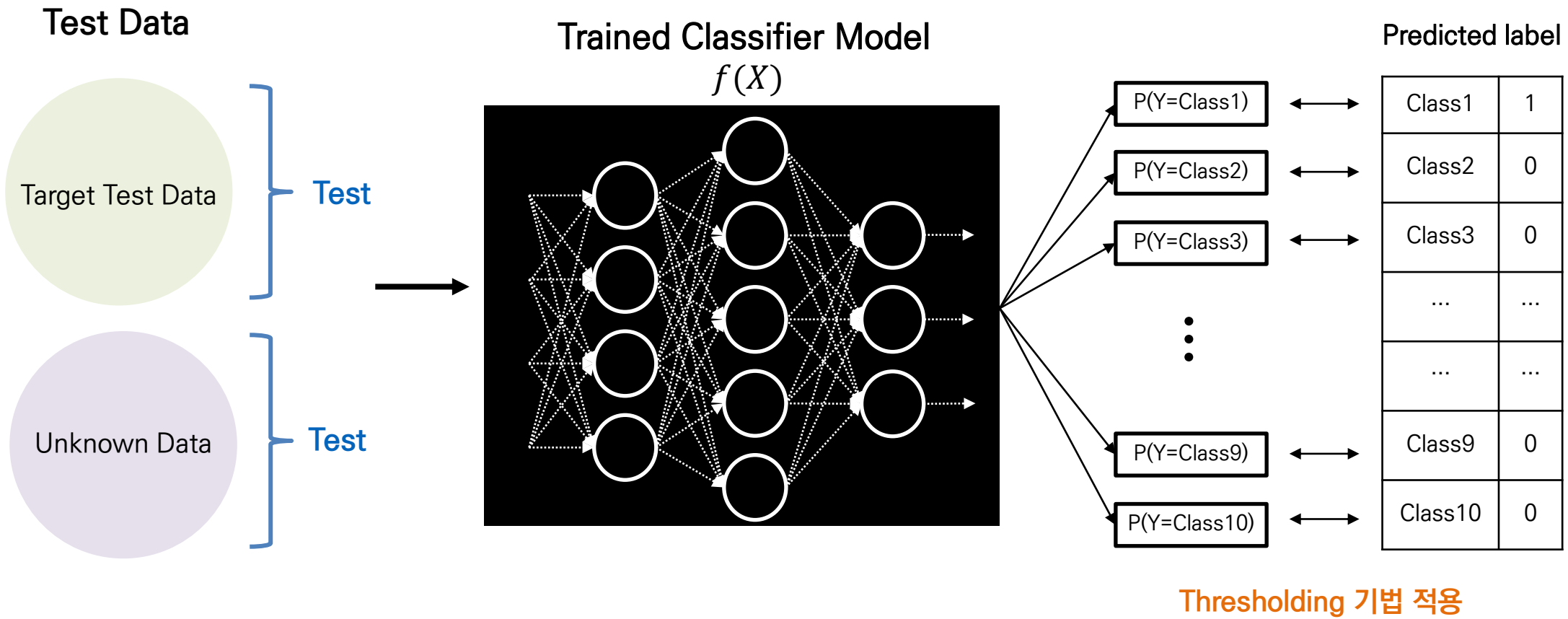
❖ Objectosphere Loss



Background Data-based Methods

(1) Reducing Network Agnostophobia

❖ Objectosphere Loss

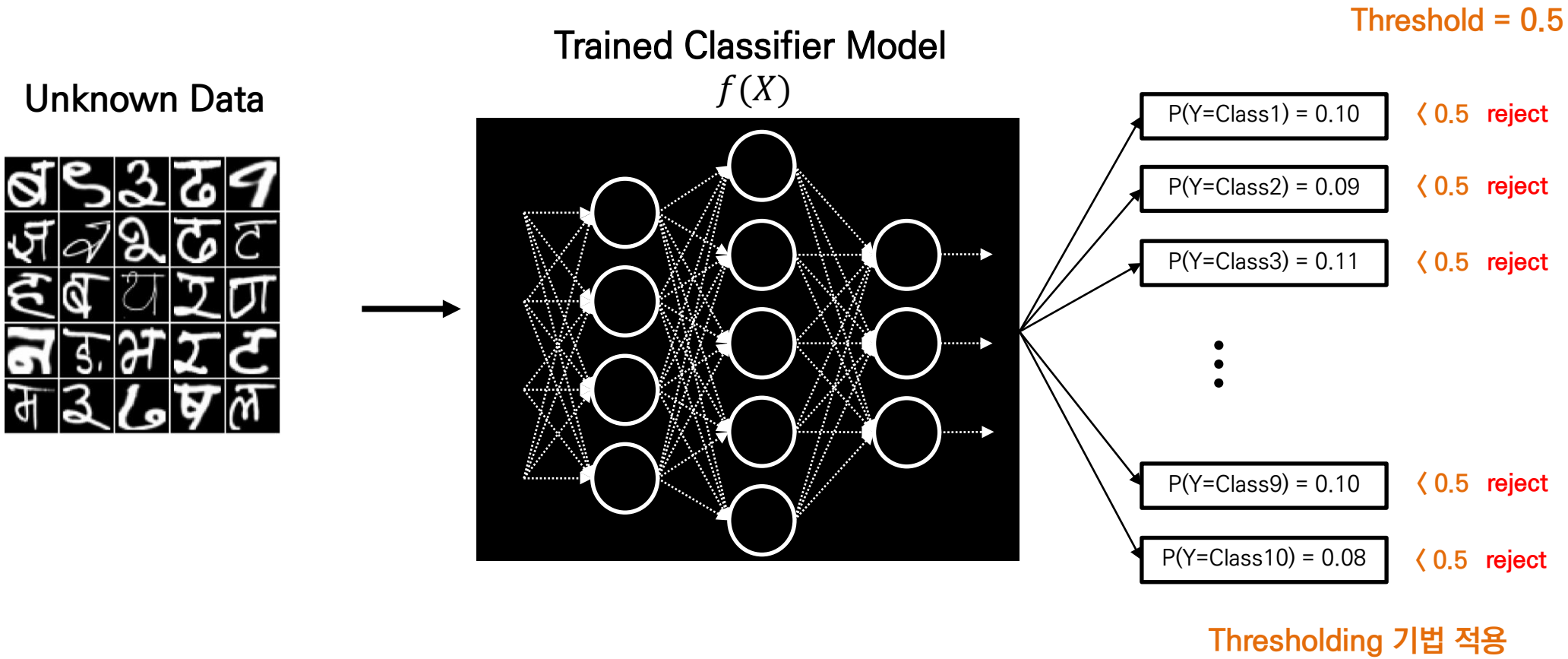


Thresholding 기법 적용

Background Data-based Methods

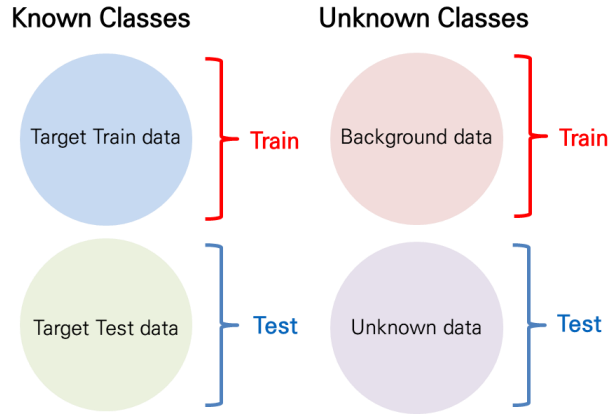
(1) Reducing Network Agnostophobia

❖ Objectosphere Loss

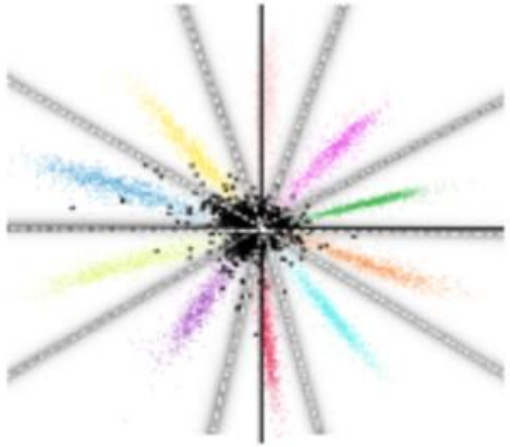


Background Data-based Methods

(1) Reducing Network Agnostophobia



❖ Objectosphere Loss



(c) Objectosphere

LENET++ network로 테스트 Data를 2차원 특징 공간에 표상하여 시각화

- 검은 색이 아닌 다른 색상으로 표현된 점 : Test Target Data (MNIST Data)
- 검은 색 점 : Unknown Data (Devanagari Data)
- 점선 : SoftMax 점수가 이웃 Class와 동일한 Class 경계

Background Data-based Methods

(1) Reducing Network Agnostophobia

Experiment	Unknowns $\lvert\mathcal{D}_a\rvert$	Algorithm	CCR at FPR of 10^{-4}	10^{-3}	10^{-2}	10^{-1}
LeNet++ Architecture Trained with MNIST digits as $\mathcal{D}_c$ and NIST Letters as $\mathcal{D}_b$	Devanagri 10032	Softmax	0.0	0.0	0.0777	0.9007
		Background	0.0	0.4402	0.7527	0.9313
		Entropic Open-Set	0.7142	0.8746	0.9580	0.9788
		Objectosphere	**0.7350**	**0.9108**	**0.9658**	**0.9791**
	NotMNIST 18724	Softmax	0.0	0.3397	0.4954	0.8288
		Background	0.3806	0.7179	0.9068	0.9624
		Entropic Open-Set	0.4201	0.8578	0.9515	**0.9780**
		Objectosphere	**0.512**	**0.8965**	**0.9563**	0.9773
	CIFAR10 10000	Softmax	0.7684	0.8617	0.9288	0.9641
		Background	0.8232	0.9546	0.9726	0.973
		Entropic Open-Set	**0.973**	**0.9787**	**0.9804**	**0.9806**
		Objectosphere	0.9656	0.9735	0.9785	0.9794
ResNet-18 Architecture Trained with CIFAR-10 Classes as $\mathcal{D}_c$ and Subset of CIFAR-100 as $\mathcal{D}_b$	SVHN 26032	Softmax	0.1924	0.2949	0.4599	0.6473
		Background	0.2012	0.3022	0.4803	0.6981
		Entropic Open-Set	0.1071	0.2338	0.4277	0.6214
		Objectosphere	0.1862	0.3387	0.5074	0.6886
		Scaled Objecto	**0.2547**	**0.3896**	**0.5454**	**0.7013**
	CIFAR-100 Subset 4500	Softmax	N/A	0.0706	0.2339	0.5139
		Background	N/A	0.1598	0.3429	0.6049
		Entropic Open-Set	N/A	0.1776	0.3501	0.5855
		Objectosphere	N/A	0.1866	0.3595	0.6345
		Scaled Objecto	N/A	**0.2584**	**0.4334**	**0.6647**

FPR : Unknown Data 중에서 모델이 Target Class라고 예측한 비율

$$Fall\text{-}out(FPR) = \frac{FP}{TN + FP}$$

Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

- 2017년 발표된 arXiv에 있는 논문
- 2020년 10월 15일 기준 239회 인용

TRAINING CONFIDENCE-CALIBRATED CLASSIFIERS FOR DETECTING OUT-OF-DISTRIBUTION SAMPLES

Kimin Lee[*] **Honglak Lee**[§,†] **Kibok Lee**[†] **Jinwoo Shin**[*]
[*]Korea Advanced Institute of Science and Technology, Daejeon, Korea
[†]University of Michigan, Ann Arbor, MI 48109
[§]Google Brain, Mountain View, CA 94043

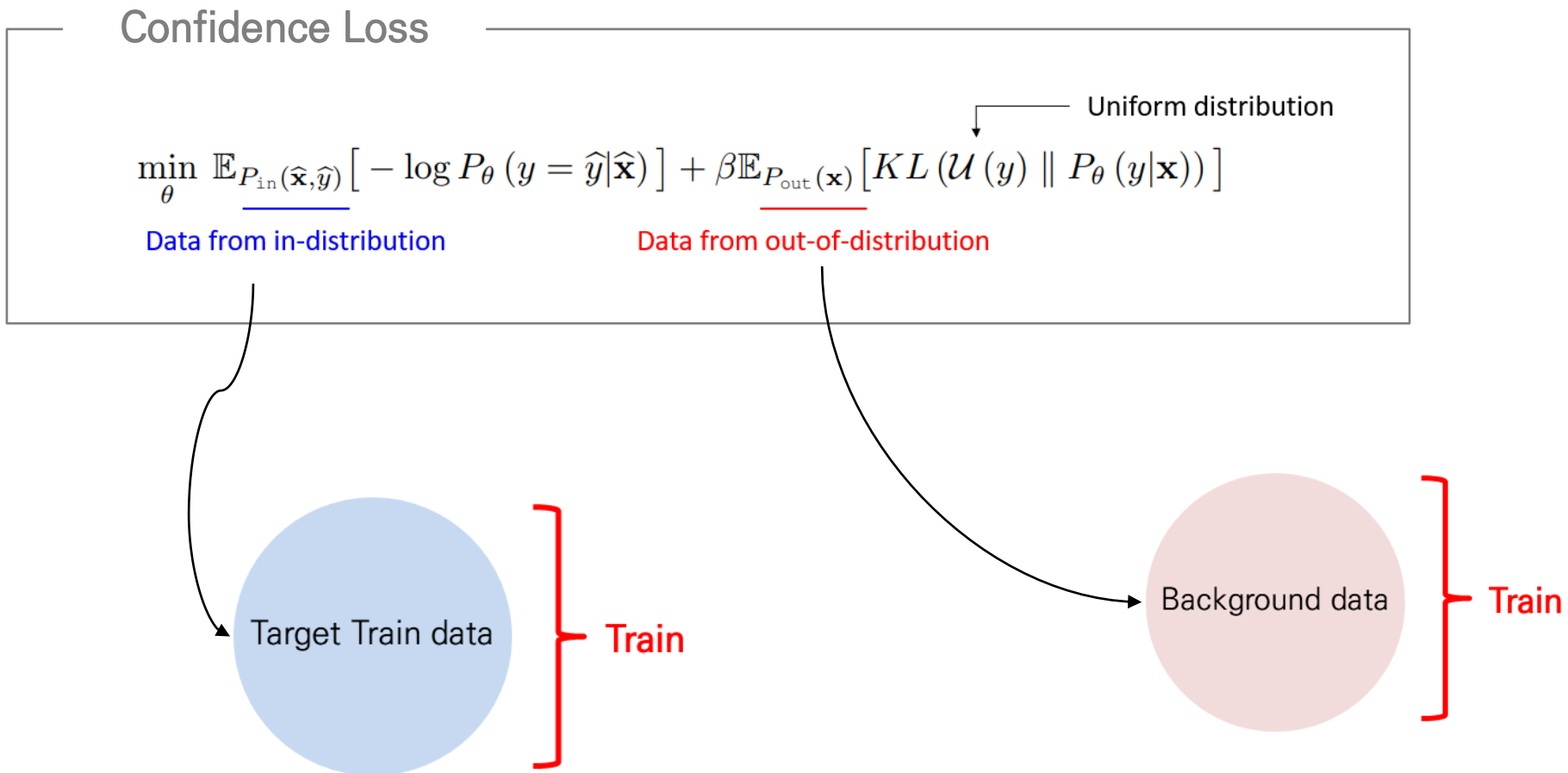
ABSTRACT

The problem of detecting whether a test sample is from in-distribution (i.e., training distribution by a classifier) or out-of-distribution sufficiently different from it arises in many real-world machine learning applications. However, the state-of-art deep neural networks are known to be highly overconfident in their predictions, i.e., do not distinguish in- and out-of-distributions. Recently, to handle this issue, several threshold-based detectors have been proposed given pre-trained neural classifiers. However, the performance of prior works highly depends on how to train the classifiers since they only focus on improving inference procedures. In this paper, we develop a novel training method for classifiers so that such inference algorithms can work better. In particular, we suggest two additional terms added to the original loss (e.g., cross entropy). The first one forces samples from out-of-distribution less confident by the classifier and the second one is for (implicitly) generating most effective training samples for the first one. In essence, our method jointly trains both classification and generative neural networks for out-of-distribution. We demonstrate its effectiveness using deep convolutional neural networks on various popular image datasets.

Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

❖ Confidence Loss



Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

❖ Confidence Loss

Confidence Loss

Uniform distribution

$$\min_{\theta} \mathbb{E}_{P_{\text{in}}(\hat{\mathbf{x}},\hat{y})}\left[-\log P_\theta\left(y=\hat{y}|\hat{\mathbf{x}}\right)\right] + \beta\mathbb{E}_{P_{\text{out}}(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]$$

Data from in-distribution

Data from out-of-distribution

Background Data에 대한 모델의 SoftMax 점수가 Uniform Distribution을 따르도록 학습

Entropic Open-Set Loss

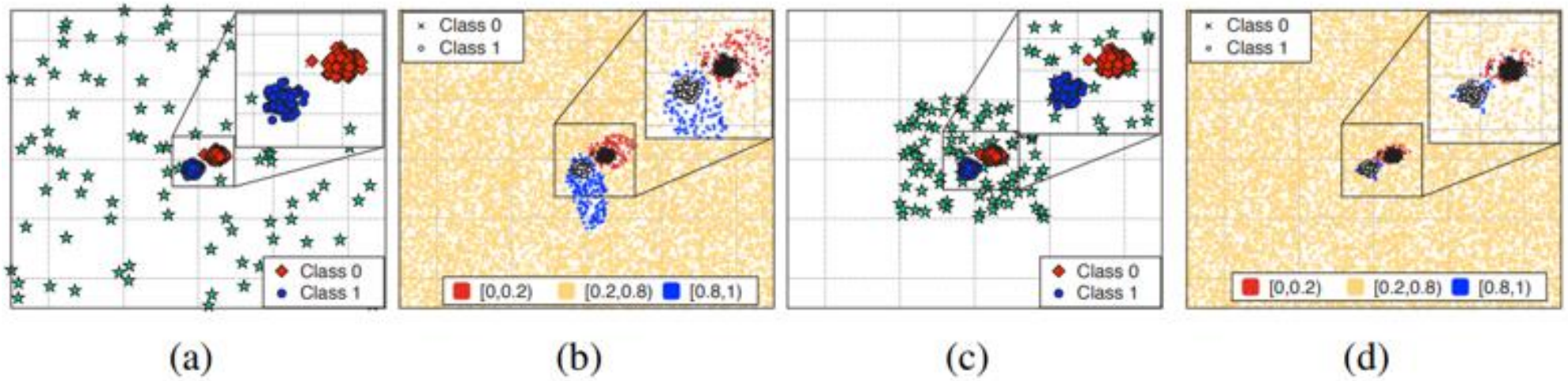
$$Loss = \begin{cases} -\log S_c(x) & if\ x \in D_{targettrain}\ is\ from\ class\ c \\ -\frac{1}{C}\sum_{i=1}^{C}\log S_i(x) & if\ x \in D_{background} \end{cases}$$

$S_c(x)$ *are equal:* $\forall c \in \mathcal{C} : S_c(x) = S = \frac{1}{C}$ *and maximizes entropy.*

Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

❖ Background Data



•(a): Target Train Data (파란색, 빨간색) 와 Background Data (초록색) 을 그림과 같이 설정하면,

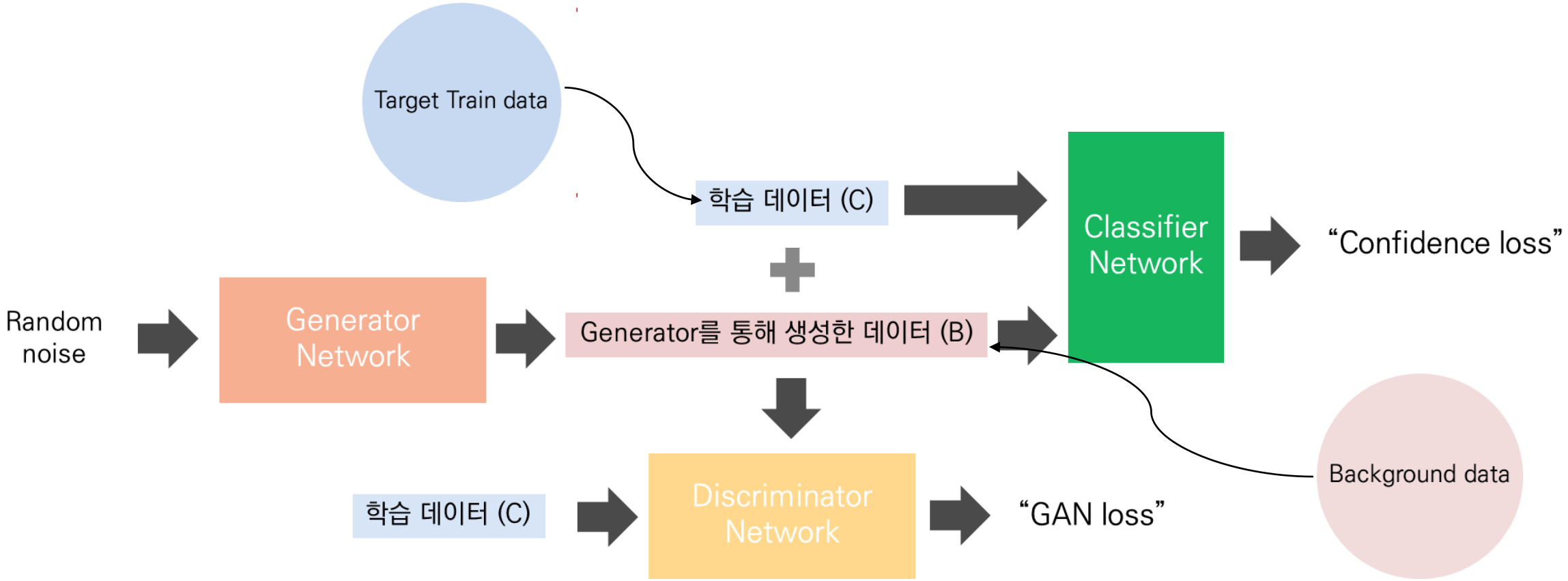
•(b): 모델 학습 이후 결정 경계(검정색)가 그림과 같이, Target Test Data 분포와 동일하게 나타나지 않음.

•(c): 따라서 Background Data를 그림과 같이 Target Train Data**에 최대한 밀접한 Data를 사용**하면,

•(d): 모델의 결정 경계(검정색)를 Target Test Data 분포와 일치하도록 할 수 있음.

Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples



Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

❖ New GAN Loss

New GAN Loss

Need trained $P_\theta(y|x)$

$$\min_G \max_D \; \beta \underbrace{\mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}(y) \parallel P_\theta(y|\mathbf{x})\right)\right]}_{\text{(a)}} + \underbrace{\mathbb{E}_{P_{\text{in}}(\mathbf{x})}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1 - D(\mathbf{x})\right)\right]}_{\text{(b)}}$$

- (a): GAN이 생성하는 Data(Background Data)에 대한 SoftMax 점수가 Uniform distribution을 따르도록 학습.
- (b): GAN이 생성하는 Data가 기존 Data와 비슷해 지도록 생성 모델을 학습 (GAN Loss)

Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

❖ Joint Confidence loss

Joint Confidence loss

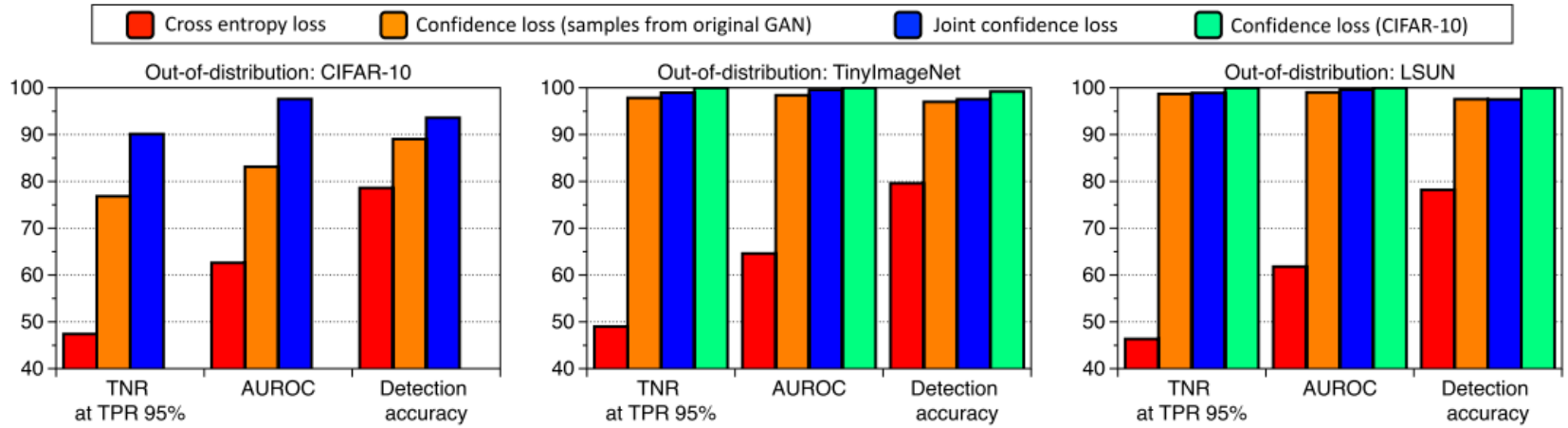
$$\min_G \max_D \min_\theta \underbrace{\mathbb{E}_{P_{\text{in}}(\hat{\mathbf{x}},\hat{y})}\left[-\log P_\theta\left(y=\hat{y}|\hat{\mathbf{x}}\right)\right]}_{\text{(c)}} + \beta \underbrace{\mathbb{E}_{P_G(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]}_{\text{(d)}}$$
$$\underbrace{+\mathbb{E}_{P_{\text{in}}(\hat{\mathbf{x}})}\left[\log D\left(\hat{\mathbf{x}}\right)\right] + \mathbb{E}_{P_G(\mathbf{x})}\left[\log\left(1-D\left(\mathbf{x}\right)\right)\right]}_{\text{(e)}}.$$

Classifier Network과 GAN Network을 번갈아 가며 학습.

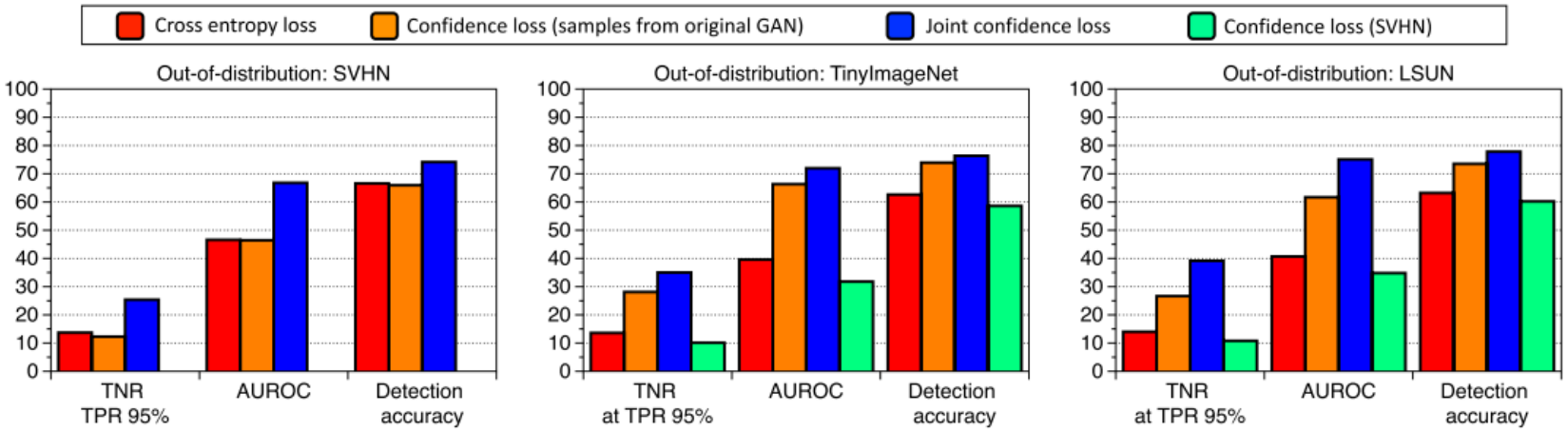
- Classifier Network 학습 시 GAN Network 고정 -> (c) + (d) 사용.
- GAN Network 학습 시 Classifier Network 고정 -> (d) + (e) 사용.

Background Data-based Methods

(2) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples



(a) In-distribution: SVHN

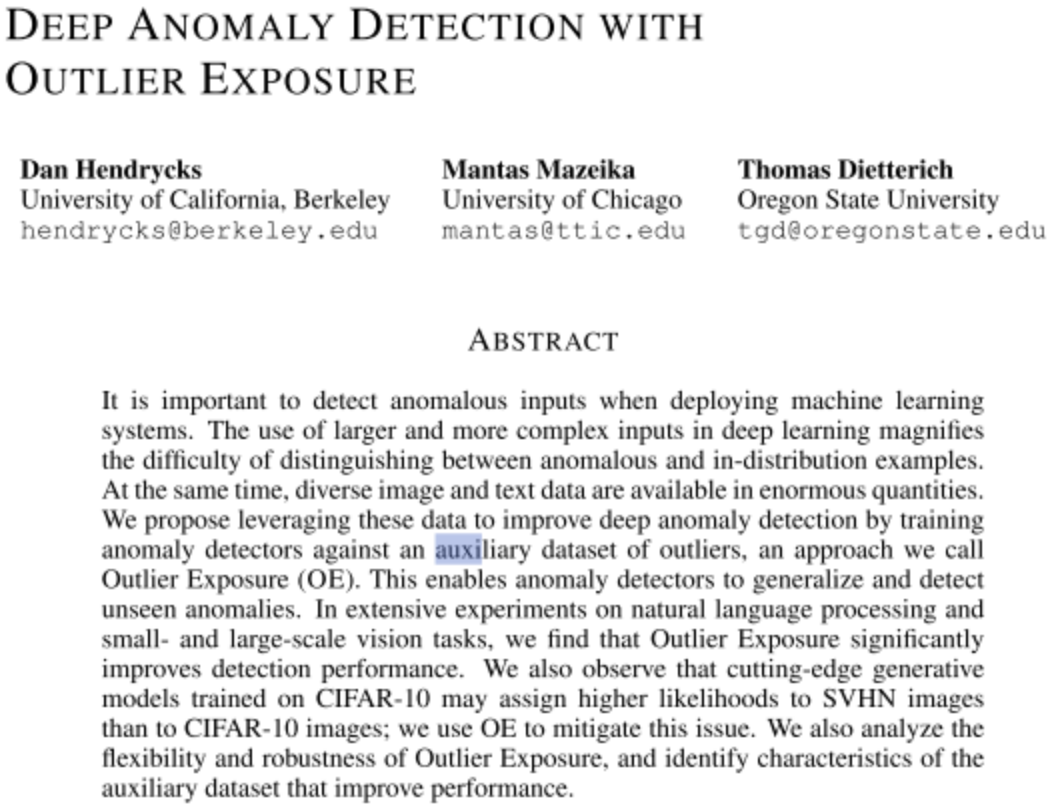


(b) In-distribution: CIFAR-10

Background Data-based Methods

(3) Deep Anomaly Detection with Outlier Exposure

- 2018년 발표된 arXiv에 있는 논문
- 2020년 10월 15일 기준 180회 인용
- 다양한 Classifier 모델과 Target Classes Data, Unknown Data, Background Data의 조합에 따른 다양한 실험을 진행한 논문



DEEP ANOMALY DETECTION WITH
OUTLIER EXPOSURE

Dan Hendrycks
University of California, Berkeley
hendrycks@berkeley.edu

Mantas Mazeika
University of Chicago
mantas@ttic.edu

Thomas Dietterich
Oregon State University
tgd@oregonstate.edu

ABSTRACT

It is important to detect anomalous inputs when deploying machine learning systems. The use of larger and more complex inputs in deep learning magnifies the difficulty of distinguishing between anomalous and in-distribution examples. At the same time, diverse image and text data are available in enormous quantities. We propose leveraging these data to improve deep anomaly detection by training anomaly detectors against an auxiliary dataset of outliers, an approach we call Outlier Exposure (OE). This enables anomaly detectors to generalize and detect unseen anomalies. In extensive experiments on natural language processing and small- and large-scale vision tasks, we find that Outlier Exposure significantly improves detection performance. We also observe that cutting-edge generative models trained on CIFAR-10 may assign higher likelihoods to SVHN images than to CIFAR-10 images; we use OE to mitigate this issue. We also analyze the flexibility and robustness of Outlier Exposure, and identify characteristics of the auxiliary dataset that improve performance.

Background Data-based Methods

(3) Deep Anomaly Detection with Outlier Exposure

Background Data를 사용한 Open Set Recognition 알고리즘의 전반적인 한계 :

어떤 Background Data를 사용하는지에 따라 Open Set Recognition 성능이 크게 달라짐

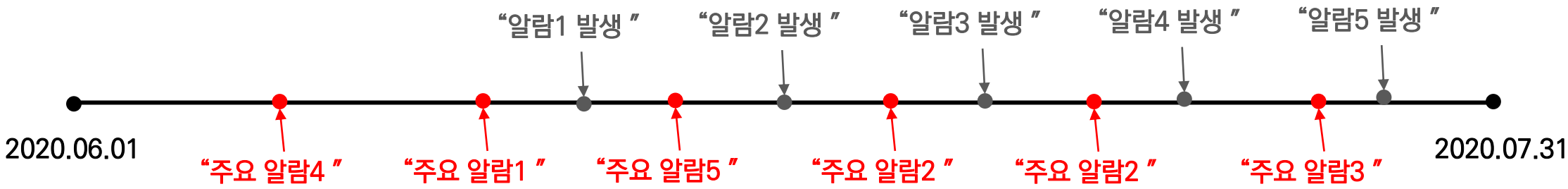
-〉 Background Data 선정이 중요한 이슈

- ❖ Flexibility in Choosing Background Data
  - Gaussian noise나 GAN으로 생성한 Background Data를 활용하는 것은 다양한 Unknown Data에 대해 일반화 성능이 낮음
- ❖ Closeness of Unknown Data, Background Data and Target Test Data
  - Unknown Data와 Background Data간의 유사도가 Open Set Recognition 성능에 큰 영향이 없음
  - Target Test Data와 Background Data간의 유사도가 Open Set Recognition 성능에 중요한 영향을 끼침
- ❖ Diversity of Background Data
  - Diversity가 큰 Background Data를 많이 사용할수록 Open Set Recognition 성능이 향상

Applications

(1) 적용 예시 : 산업 공정의 알람 유형 분류

- ❖ 산업 공정에서는 Data 수집 기간 내에 모든 유형의 알람 Data를 수집하기 힘듦. -〉 학습하지 않은 유형의 알람 Data가 존재
  - Open Set Recognition을 적용하지 않은 알람 유형을 분류기를 구축했을 때, 학습하지 않은 알람 유형에 대해 학습 알람 유형으로오분류
- ❖ 알람 유형에 따라 중요한 알람군과 비중요 알람군으로 나뉨.
  - Open Set Recognition을 적용하지 않은 중요 알람 유형을 분류기를 구축했을 때, 학습하지 않은 비중요 알람 유형에 대해 중요한 알람으로 오분류할 수 있음
- ❖ Background Data를 비중요 알람군의 Data로 활용하면 Target Test Data와 Background Data 모두 시계열 센서
  - Target Test Data와 Background Data Data 간의 유사하기 때문에 높은 Open Set Recognition 성능을 보일 수 있음



Q&A