논문 2021-16-01

적대적 데이터 혼합: 분포 외 데이터에 대한 강건성과 추론 결과에 대한 신뢰성 향상 방법

(Adversarial-Mixup: Increasing Robustness to Out-of-Distribution Data and Reliability of Inference)

권 경 필, 유 준 혁*
(Kyungpil Gwon, Joonhyuk Yoo)

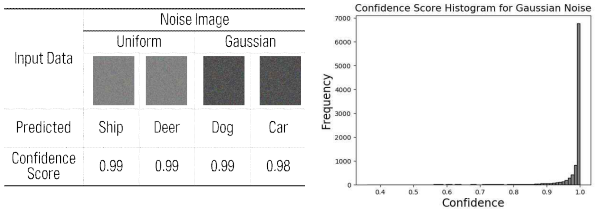
Abstract : Detecting Out-of-Distribution (OOD) data is fundamentally required when Deep Neural Network (DNN) is applied to real-world AI such as autonomous driving. However, modern DNNs are quite vulnerable to the over-confidence problem even if the test data are far away from the trained data distribution. To solve the problem, this paper proposes a novel Adversarial-Mixup training method to let the DNN model be more robust by detecting OOD data effectively. Experimental results show that the proposed Adversarial-Mixup method improves the overall performance of OOD detection by 78% comparing with the State-of-the-Art methods. Furthermore, we show that the proposed method can alleviate the over-confidence problem by reducing the confidence score of OOD data than the previous methods, resulting in more reliable and robust DNNs.

Keywords : Out-of-distribution detection, Adversarial-mixup training, Deep neural networks

Ⅰ. 서 론

심층신경망 (DNNs, Deep Neural Networks)의 기술적인 발전과 함께 이를 산업, 경제 및 사회 전반적인 인프라에 적용하려는 많은 시도가 이루어지고 있다. 현실 세계의 의사결정 시스템에 심층신경망을 적용하기 위해서는 높은 정확성은 물론 판단 결과에 대한 신뢰성 (Reliability) 또한 중요하다. 예를 들어, 의학적 진단 (Medical Diagnosis)에 사용되는 심층신경망은 양성을 음성으로 판단하는 위음성 (False Negative)을 낮추기 위해 음성으로 판단된 데이터 중 신뢰 점수 (Confidence Score)가 낮은 데이터를 사람이 다시 검토하는 등의 과정을 거치게 된다. 이처럼 현실 세계에서는 정답의 정확도뿐만 아니라 추론한 결과가 신뢰할 수 있는지를 판별하는 작업이 우선시 되는 경우가 많다. 따라서 심층신경망의 신뢰성은 정확도 못지않은 중요한 요소이다.

최근의 연구에 따르면 근래의 심층신경망은 적대적 공격 (Adversarial Attack) [1]에 취약하며, 학습 단계에 사용된 데이터와 유사한 특징이 있거나 심지어 학습에 사용되지 않은 데이터와 같은 전혀 엉뚱한 분포 외 (Out-of-Distribution) 데이터가 입력되어도 이를 높은 확신으로 (Over-Confidence) 정답이 아닌 항목을 마치 정답처럼 판단하는 문제점이 발견되었다 [2]. 심층신경망의 학습에 사용되는 데이터 셋은 한정된 데이터와 클래스를 포함하고 있다. 이러한 특성으로 인해 높은 분류 정확도에도 불구하고 약간의 변형된 데이터 혹은 아무 관련 없는 데이터에 대해 이를 잘 설명하지 못할 만큼 학습데이터에 의존적인 (Training-Data-Dependent) 성향을 보인다. 예를 들어, 자연 영상을 포함하고 있는 CIFAR10 데이터 셋으로 학습된 네트워크에 잡음 (Noise) 이미지를 입력하면 그림 1 (a)와 같이 분포 내 데이터 (In-Distribution) 중 하나로 잘못 분류하고, 그림 1 (b)와 같이 1만 개의 잡음 데이터를 입력하였을 때 심층신경망의 신뢰 점수 분포가 95~100 사이에 집중되어 높은 확신으로 분포 외 데이터를 분포 내 데이터로 잘못 분류함을 확인할 수 있다.



(a) (b)
그림 1. 분포 외 데이터에 대한 심층신경망의 (a) 잘못된 분류 결과 (b) 신뢰 점수 분포
Fig. 1. (a) Misclassification Result (b) Confidence Score Distribution for Out-of-Distribution Data

분포 외 데이터 탐지와 관련된 문제를 해결하기 위해 다양한 연구가 진행되고 있으며, 사전학습된 (Pre-Trained) 모

*Corresponding Author (joonhyuk@daegu.ac.kr)
Received: Nov. 26, 2020, Revised: Jan. 5, 2021, Accepted: Jan. 13, 2021.
Kyungpil Gwon: Daegu University (M.S.)
Joonhyuk Yoo: Daegu University (Prof.)
※ 이 연구는 2020년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원과 2020년도 정부 (산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (NRF-2020R1A2C1014768 중견연구자지원사업, P0012724, 산업혁신인재성장지원사업).

델의 출력값을 사후처리 (Post-Processing)하여 분포 외 데이터를 탐지하는 방법과 모델의 학습을 통한 분포 외 데이터 탐지 방법이 있다. 해당 방법은 모델이 학습을 통해 분포 외 데이터에 대해 강건성을 가질 수 있도록 하는 최적의 파라미터를 찾고 이를 바탕으로 추론 과정에 입력되는 분포 외 데이터와 분포 내 데이터 집단과의 거리를 멀어지게 하거나, 분포 외 데이터에 대한 신뢰 점수를 낮추어 심층신경망이 더욱 신뢰성 높은 결과를 도출할 수 있도록 한다. 본 논문에서는 데이터 혼합 (Mixup) 기법 [3]을 기반으로 한 적대적 데이터 혼합 (Adversarial-Mixup) 학습법을 제안한다. 데이터 혼합 기법은 데이터 증강 (Augmentation) 효과와 함께 학습데이터 의존적인 특성을 해결할 수 있고, 제안된 방법은 적대적 데이터 혼합 학습을 통해 모델이 분포 외 데이터에 대한 신뢰 점수 분포를 낮아지게 유도하였다. 또한, 신뢰 점수를 기반으로 한 다양한 평가 지표를 활용하여 기존방법 대비 평균 78% 증가율로 성능이 향상됨을 입증하였다.

본 논문의 구성은 다음과 같다. 2장에서 분포 외 데이터 탐지와 관련된 선행 연구를 소개하고, 3장에서 기존의 데이터 무작위 혼합 방식과 제안하는 적대적 데이터 혼합방법의 개념 및 차이점을 설명한다. 4장에서 다양한 상황 가정과 변화하는 임계값을 이용한 분포 외 데이터 탐지기를 이용하여 5장에서는 제안된 방법이 기존방법 대비 향상된 성능을 입증하고, 분포 외 데이터에 대한 신뢰 점수 분포를 낮은 쪽으로 이동시켜 심층신경망의 추론 결과에 대한 신뢰성이 향상되는 것을 보여준다.

Ⅱ. 관련 연구

분포 외 데이터 탐지에 관한 선행 연구는 크게 네트워크 학습을 통한 방식과 사전 학습된 네트워크를 활용하는 두 가지 방법이 있다. 네트워크 학습 방식에는 학습데이터 라벨의 접근 여부에 따라 지도학습, 비지도/준지도 학습 기반 분포 외 데이터 탐지 방법이 있고, 이는 네트워크 학습을 통해 심층신경망이 분포 외 데이터에 대해 강건성을 가질 수 있도록 하는 최적의 파라미터를 학습을 통해 찾는다. 사전학습된 네트워크를 활용하는 방식은 분포 외 데이터 탐지를 위해 데이터 분포 간 거리 측정법을 적용하거나 경험에 의한 방법으로 사후 출력값을 스케일링하여 분포 외 데이터 탐지를 위한 최적의 하이퍼-파라미터를 찾는 방식이다.

1. 학습기반 분포 외 데이터 탐지 방법

지도학습 기반 방법에는 적대적 생성모델 (GAN, Generative Adversarial Networks)을 사용하여 분포 내 데이터 결정 경계 (Decision Boundary) 주변에 분포 외 데이터를 임의로 생성하여 해당 데이터의 신뢰 점수를 균일 분포 (Uniform Distribution)에 수렴하도록 유도하는 방법이다 [4]. OE (Outlier Exposure) [5]는 네트워크 구조와 손실 함수를 변경하고 학습 단계에 보조 데이터 셋만을 이용하는 비교적 간단한 방법으로 분포 외 데이터 성능을 향상할 수 있음을 입증하였다. 지도학습 기반 방법은 모델의 학습을 통하여 분포 외 데이터 탐지를 위한 최적의 파라미터를 자동으로 찾아준다는 장점이 있지만 현실 세계의 관점에서 분포 외 데이터 수집이 어렵다는 점과 모든 데이터에 대한 라벨링이 되어야 한다는 점에서 시간 비용적 측면의 문제가 있다.

오토인코더를 활용하여 데이터 차원 축소와 복원 과정에서 재구성되는 오류 탐지를 기반으로 분포 외 데이터를 탐지하는 방법 [6], 단일 클래스 분류 (One-Class Classification) 방법론을 사용하여 정상 샘플을 둘러싸는 경계면을 최대한 좁혀 경계면 밖에 있는 샘플을 모두 비정상으로 간주하는 방식인 Deep SVDD [7], Pixel CNN++모델을 활용하여 특정 픽셀 주변의 등장 확률을 계산하여 분포 외 데이터 탐지에 활용하는 방법 [8] 등은 대표적인 비지도/준지도 학습 기반 분포 외 데이터 탐지 방법이다. 해당 방법은 데이터에 대한 라벨링이 필요하지 않아 효율적인 데이터 활용이 가능하지만, 지도학습 방법 대비 성능이 낮다는 단점이 있다.

2. 사전학습된 네트워크 활용 기반 분포 외 데이터 탐지 방법

Maximum-Softmax 방법은 분포 외 데이터 탐지와 관련된 최초의 연구논문이며, 차후 진행되는 연구의 비교지표로 사용되고 있다 [9]. 해당 방법은 소프트맥스 (Softmax)층의 출력값 중 최대값을 임계값과 비교하여 분포 외 데이터를 판별하는 방식이다. ODIN [10]은 추론 과정에 입력되는 데이터에 대해 손실 값을 증가시키는 방향의 그래디언트를 계산하여 소량의 외란 (Disturbance)을 빼주어 분포 외 데이터에 대한 신뢰 점수를 낮추고, 온도 스케일링 (Temperature Scaling) [11] 방법을 적용하여 분포 내 데이터와 분포 외 데이터 사이의 거리를 멀어지게 한다. 거리 측정 기반 방식은 학습된 네트워크의 특징맵 (Feature Map)이 특정 분포를 이루고 있다는 가정하에 확률적 모델링을 통하여 데이터 분포 간 거리를 측정한다. 심층신경망 마지막 컨볼루션 계층의 특징맵을 활용하는 기존의 연구 방법과는 다르게 상대적으로 추상화 정도가 낮은 입력층에 가까운 컨볼루션 계층의 특징맵을 활용, 이를 데이터클러스터 간 마할라노비스 거리 측정법을 적용하여 분포 외 데이터를 탐지하는 방법과 [12], 네트워크 특징맵이 클래스 조건부 가우시안 분포를 따른다는 조건하에 마할라노비스 거리를 구하여 이를 신뢰 점수로 사용하는 방법이 있다 [13].

사전 학습된 네트워크를 활용하는 방식은 네트워크의 학습이 필요하지 않다는 점에서 시간 비용적인 이점이 있지만, 분포 외 데이터 탐지 성능이 사후처리에 사용되는 하이퍼-파라미터에 민감하다는 단점이 있다. 거리 측정 기반 방법은 확률적인 모델링 과정을 통해 하이퍼-파라미터에 의존적인 기존방법 대비 높은 성능을 보이지만, 특징맵을 추출하는 은닉층 (Hidden Layer)의 위치와 분포 외 데이터의 종류에 따라 변화하는 성능으로 인해 일반화 성능이 낮아질 수 있다.

Ⅲ. 적대적 데이터 혼합 방법

본 논문에서 제안하는 적대적 데이터 혼합 (Adversarial-Mixup)법은 데이터 무작위 혼합 (Mixup)방법과 적대적 공격 (Adversarial Attack)의 두 가지 기법이 적용되었다. 첫 번째로 Mixup 방식은 분포 내 데이터에서 임의로 선택된 두 데이터를 선형가중합 하여 새로운 데이터를 생성하는 방법이다. 해당 방법은 간단하지만, 효과적인 데이터 증강 기법으로 알려져 있으며 Mixup 방법으로 생성된 데이터를 학습에 사용하면 심층신경망의 정확도와 신뢰 점수를 일치하게 하는 교정 (Calibration)에 효과가 있음이 입증되었다 [14]. 두 번째로 적대적 공격은 적대적 샘플 (Adversarial Example)을 생성 후 원본 데이터에 더하여 심층신경망이 잘못된 분류를 하도록 유도하는 것을 뜻한다. x를 모델에 입력되는 원본 데이터, y를 x에 대한 라벨, θ를 모델의 학습 파라미터라 하였을 때 적대적 샘플은 아래 수식 (1)과 같이 정의된다.

$$x_{adv} = x + \epsilon sign(\nabla_x J(\theta, x, y)) \ . \tag{1}$$

여기서, $\nabla_x J(\theta, x, y)$는 역전파 과정에서 얻은 손실 값의 그래디언트, $sign(\cdot)$은 부호함수, ϵ은 샘플의 생성 정도 (Magnitude)이다. 적대적 샘플 생성 방법은 수식 (1)과 같이 FGSM (Fast Gradient Sign Method)방법 외 FGSM을 반복적으로 수행하는 Basic Iteration [15] 등 다양한 방법이 존재한다. 적대적 공격은 다소 간단한 방법으로 심층신경망의 정확도를 감소시킬 수 있다. 이에 [16]에서는 적대적 공격이 적용된 데이터를 원본 데이터와 함께 학습하여 심층신경망을 정규화하였다. 이를 적대적 학습 (Adversarial Training)이라 하고, 해당 방법은 적대적 공격이 적용된 데이터에 대해 모델이 학습을 통해 분포 외 데이터 탐지 성능을 증가시킬 수 있는 가장 기본적인 방법으로 해석할 수 있다.

본 논문에서는 데이터 증강과 심층신경망의 신뢰 점수 교정 효과가 있는 Mixup, 적대적 데이터를 학습 단계에 의도적으로 추가하여 심층신경망의 강건성을 증가시키는 특징이 있는 두 방법을 융합하여 적용한다. 제안된 방법은 분포 내 데이터에서 임의로 선택한 데이터와 해당 데이터에 적대적 공격이 적용된 데이터를 혼합한다. 적대적 샘플은 간단하면서 효율적인 FGSM 방법을 선택하였다. 수식 (1)에서 생성된 적대적 샘플을 원본 데이터와 혼합하여 새롭게 생성된 학습데이터 $\tilde{x_i}$는 아래 수식 (2)와 같다.

$$\tilde{x_i} = \lambda x^i_{ori} + (1-\lambda) x^i_{adv} \ . \tag{2}$$

여기서, x^i_{ori}는 무작위로 선택된 i번째 원본 데이터, x^i_{adv}는 i번째 원본 데이터에 적대적 공격이 적용된 데이터이다. λ는 가중치 계수이며 $\lambda \sim Beta(\alpha,\alpha)$의 특성에 따라 대칭 베타분포를 이루는 값들로 구성되어 있다. 연속형 확률분포

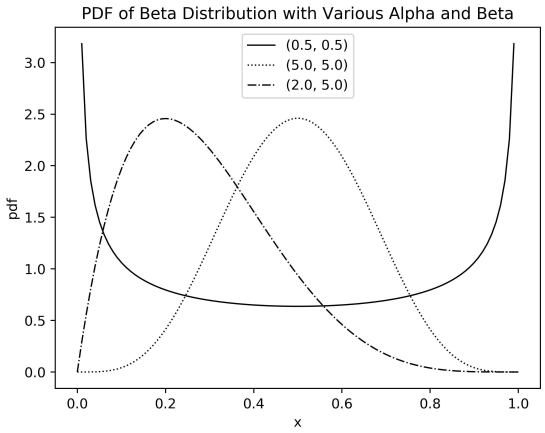


그림 2. 모수에 따른 베타분포의 확률 밀도 함수
Fig. 2. PDF(Probability Density Function) of Beta Distribution with Various Parameters

중 하나인 베타분포는 균등분포 (Uniform Distribution)를 일반화시킨 분포이며, 0에서 1사이의 값을 갖는 특징이 있다. $Beta(\alpha,\alpha)$와 같이 두 모수 α의 값이 같으면 그림 2의 실선처럼 대칭적 구조의 난수가 생성되어 데이터를 일정한 비율로 혼합할 수 있다. Adversarial-Mixup 학습법의 손실 함수는 수식 (3)과 같다.

$$L_{total} = [\lambda * \underbrace{L_{CE}(P_{in}(x_{ori}, y))}_{\text{(a)}}] + [(1-\lambda) * \underbrace{L_{CE}(P_{adv}(\tilde{x}, y))}_{\text{(b)}}] \ . \tag{3}$$

여기서, L_{CE}는 크로스-엔트로피 (Cross-Entropy) 손실 함수, P_{in}, P_{adv}는 각각 분포 내 데이터와 적대적 공격이 적용된 데이터에 대한 모델의 출력값 들이다. (a), (b)는 분포 내 데이터에서 임의로 선택된 원본 데이터 x_{ori}에 대한 손실 값, 적대적 공격이 적용된 데이터 $\tilde{x}$의 손실 값을 뜻한다. 두 손실 값을 혼합하여 사용하였고, $\lambda, (1-\lambda)$로 스케일링하여 적절한 비율로 선형조합을 이루었다. λ값은 데이터 생성에 사용된 스케일링 계수와 동일하다.

Ⅳ. 실험 환경

1. 실험 데이터 구성

현실 세계에서는 분포 외 데이터의 비율이 분포 내 데이터 비율보다 훨씬 높은 특징으로 인하여 모든 분포 외 상황을 포함하지 못한다는 제약이 있다. 따라서, 본 논문에서는 특정 분포 외 데이터가 입력되는 상황의 가정을 위해 학습 단계와 시험 단계에서 서로 다른 클래스를 가지는 데이터 셋을 사용하고, 제안한 방법을 기반으로 학습된 모델의 분포 외 데이터 탐지 성능을 측정하였다.

표 1. 분포 외 데이터 탐지를 위한 상황 가정
Table 1. Scenarios for Detecting Out-of-Distribution Data

In-Distribution (For Training)	Out-of-Distribution (For Testing)	Number of Data
CIFAR 10	SVHN	26K
	iSUN	10K
	ImageNet(Resizing)	
	ImageNet(Crop)	
	LSUN(Resizing)	
	LSUN(Crop)	
	Uniform	
	Gaussian	

2. 실험 데이터 셋

본 논문에서는 표 1과 같이 CIFAR10 데이터 셋을 분포 내 데이터 셋으로 사용하였으며, 분포 외 데이터 셋은 SVHN, iSUN과 Uniform, Gaussian 분포를 가지는 잡음 이미지, ImageNet, LSUN 데이터 셋에서 각각 임의로 잘라내기 (Crop)와 다운샘플링 (Resizing)을 적용한 데이터를 포함하여 총 8개의 데이터 셋을 사용하였다. 각 분포 외 데이터의 수는 세 번째 열에 기재된 수와 같다. 학습에 사용된 모델은 ResNet18, VGGNet16 모델을 사용하였다.

3. 임계값 기반 탐지기

분포 내 데이터, 분포 외 데이터 탐지기$D(x)$는 데이터를 입력하여 소프트맥스 (Softmax) 층에서 얻은 신뢰 점수를 기반으로 특정 임계값과 비교하여 수식 (4)와 같이 In, Out을 구분하는 이진 분류 (Binary Classification) 방식을 사용한다.

$$D(x) = \begin{cases} In & \text{if } C(x) > \delta \\ Out & \text{if } C(x) \leq \delta \end{cases}. \quad (4)$$

여기서, $C(x)$는 데이터를 입력하여 소프트맥스 층에서 얻은 신뢰 점수, δ는 임계값이다. 수식 (4)를 바탕으로 $C(x)$값이 특정 임계값 보다 높으면 분포 내 (In) 데이터, 그 반대의 경우는 분포 외 (Out) 데이터로 판단한다. 임계값의 설정은 수식 (5)와 같이 각 지점을 구한 뒤, 그 사이를 gap만큼 분할하여 그림 3과 같이 설정한다. 예를 들어, start, end, gap이 각각 0.0, 1.0, 100000이라 가정한다면, 0.0과 1.0 사이 10만 개의 임계값이 설정된다.

$$\begin{cases} start = min\,[min(C_{in}(x)), min(C_{out}(x))] \\ end \;= max[max(C_{in}(x)), max(C_{out}(x))] \\ gap \;= 100000 \end{cases}. \quad (5)$$

여기서, $C_{in}(x), C_{out}(x)$는 각각 분포 내, 분포 외 데이터를 학습된 모델에 입력하여 소프트맥스 층에서 얻은 신뢰 점수이다. 다양한 임계값을 이용하여 분포 외 데이터를 탐지하는 과정은 그림 4와 같다.

4. 성능평가 지표

기존의 분류 알고리즘에 대한 성능평가 지표는 전체 데이터에 대해 예측값이 실제 정답과 동일한 데이터의 비율을 나타내는 정확도를 이용하지만, 분포 외 데이터 탐지는 입력 데이터에 대해 분포 내, 분포 외와 같이 두 가지 경우를 판별하는 이진 분류방식을 사용한다. 이진 분류기의 성능은 참양성률 (TPR, True Positive Rate)은 높고 위양성률 (FPR, False Positive Rate)은 낮을수록 좋다. 그림 5의 (a)-(c)는 변화하는 임계값에 따라 TPR, FPR의 관계를 나타낸 것이다. (b)와 같이 임계값을 계속해서 낮추면 실제 분포 외 데이터를 분포 외 데이터로 잘 판별하여 TPR이 높아짐과 동시에 낮은 임계값으로 인하여 실제 분포 내 데이터를 분포 외 데이터로 판별해버려 FPR이 같이 높아지는 특성이 있다. (c)와 같이 임계값을 지속해서 높이면 실제 분포 내 데이터뿐만 아니라 모든 분포 외 데이터를 분포 내 데이터로 판별하여 FPR, TPR이 모두 낮아지게 된다. 이와 같이 다양한 임계값에 따라 변화하는 TPR, FPR을 활용하여 성능지표의 자료로 사용할 수 있다. 본 논문에서는 다양한 임

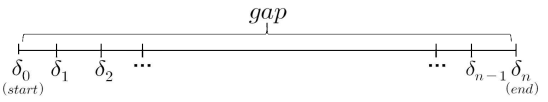


그림 3. 분포 내 데이터, 분포 외 데이터 판별을 위한 임계값 설정
Fig. 3. Determining Threshold Values for In-Distribution and Out-of-Distribution Data Detection

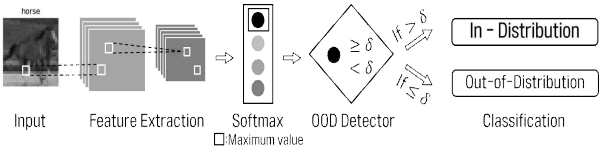


그림 4. 분포 외 데이터 탐지 과정
Fig. 4. Procedure of Detecting Out-of-Distribution Data

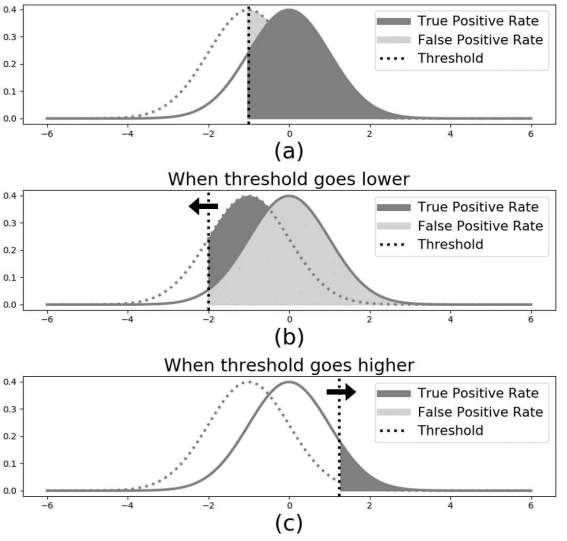


그림 5. 임계값에 따른 TPR, FPR의 변화
Fig. 5. Variation of TPR and FPR According to Threshold

계값과 아래와 같이 TPR, FPR을 활용하는 네 가지의 성능 평가 지표를 활용한다.

FPR@TPR 95는 참양성률 (TPR)이 95%만큼 높을 때 음성 데이터가 양성으로 오판단될 위양성률 (FPR) 값으로 해석할 수 있다. 여기서, TP, FP는 각각 True Positive, False Positive이며, $TPR = TP/(TP+FN)$, $FPR = FP/(FP+TN)$로 계산된다.

AUROC (Area Under Receiver Operating Characteristic) 곡선은 x축과 y축에 각각 TPR, FPR을 두고 다양한 임계값에 대한 이진 분류기의 성능을 한 번에 표시한 것이다. AUROC에서 AU는 ROC곡선의 아래쪽 면적을 의미하며, 100%에 가까울수록 좋은 성능임을 뜻한다.

AUPR (Area Under Precision Recall)은 x축에 Recall (재현율) 값을, y축에 Precision (정밀도) 값을 사용한 그래프이다. 여기서, Recall은 $TP/(TP+FN)$, Precision은 $TP/(TP+FP)$로 구할 수 있다. AUPR에서 AU는 PR곡선의 아래쪽 면적을 의미하며, APUR In/Out은 분포 내/분포 외 데이터를 양으로 판단하였을 때 각각의 PR곡선의 아래쪽 면적을 의미한다.

V. 실험 결과 분석

1. 신뢰 점수 분포의 이동

그림 6은 VGGNet16 모델을 사용하여 CIFAR10이 분포 내 데이터, ImageNet (Resizing)이 분포 외 데이터인 상황에서 서로 다른 분포 외 데이터 탐지 방법을 적용하였을 때의 신뢰 점수 분포이다. Maximum-Softmax 방식은 대부분의 신뢰 점수가 90~100 사이에 집중된 분포를 확인할 수 있고, 이는 분포 외 데이터가 입력되어도 거의 모든 데이터를 과신뢰하여 잘못된 판단을 내림을 뜻한다. ODIN의 경우 소프트맥스 출력값에 온도 스케일링이 적용되어 신뢰 점수가 특정 범위에 편중되어 있으므로 신뢰 점수 분포의 이동을 관측하는 타 방법과의 비교 대상이 될 수 없다. Mixup의 경우 Maximum-Softmax 대비 신뢰 점수의 분포가 편중되지 않고 점수가 낮은 쪽으로 분포가 이동하였지만, 여전히 분포 외 데이터에 대해 높은 신뢰 점수 분포를 가지고 있음을 확인할 수 있다. 이에 반해, 제안한 방법을 적용한 네트워크는 기존방법 대비 신뢰 점수의 분포가 상대적으로 낮은 쪽으로 이동함을 확인할 수 있다. 이는 심층신경망이 현실 세계의 의사결정 시스템에 적용되고 분류 결과에 대한 신뢰 점수가 특정 임계값 이상일 때 작동하도록 설계되었다고 가정하면, 기존방법 대비 제안 방법의 적용을 통한 시스템 설계는 분포 외 데이터 입력과 같은 돌발 상황에 대해 신뢰성 있는 결과를 도출할 수 있음을 의미한다. 이를 통하여 분포 외 데이터에 대한 신뢰 점수를 낮추어 과신뢰 문제가 완화됨을 확인할 수 있다.

2. 분포 외 데이터 탐지 성능 평가

그림 7 (a), (b)는 각각 ResNet18과 VGGNet16 모델을 사용하고, 분포 내 데이터를 CIFAR10, 분포 외 데이터를 ImageNet (Resizing)으로 가정하였을 때의 ROC 곡선이며, 기존의 Maximum-Softmax, Mixup, ODIN, M-TLAT [17] 방법과 비교하였다. M-TLAT 방법은 임의로 선택된 네 개의 데이터에 대해 1차적으로 두 데이터를 선형가중합한 뒤 나머지 두 데이터를 선형가중합하여 해당 데이터에 적대적 공격을 적용한 데이터를 함께 학습하는 방법이다. 분포 외 데이터 탐지에서의 ROC 곡선은 y축을 분포 내 데이터에 대한 TPR, x축을 분포 외 데이터에 대한 FPR로 구성되어 있다. 곡선이 $y=x$를 기준으로 위쪽에 있고 좌측 상단에 가까울수록 좋은 성능임을 미루어볼 때 (a), (b) 모두 기존방법보다 더 좋은 성능을 보임을 확인할 수 있다.

그림 8, 9는 Adversarial-Mixup을 적용하여 학습된 모델의 분포 외 데이터 탐지 성능을 종합한 결과이다. x축은 표 1에서 설정한 분포 외 데이터가 입력되는 상황 가정을 위해 사용된 분포 외 데이터이며, 각 y축의 ★는 낮을수록 ●는 높을수록 좋은 성능을 뜻한다. 그림 안의 모든 값은 % 값이고, 각 막대그래프는 좌측부터 Maximum-Softmax, ODIN,

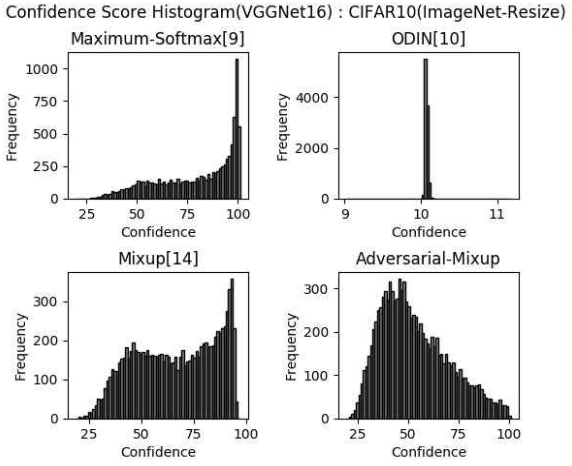


그림 6. 제안 방법에 따른 신뢰 점수 분포 (VGGNet16)
Fig. 6. Confidence Score Distribution According to Proposed Method (VGGNet16)

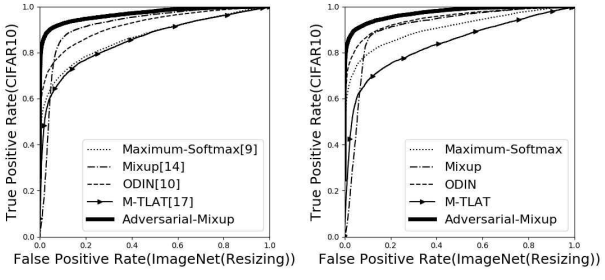


그림 7. 분포 내 데이터 (CIFAR10), 분포 외 데이터 (ImageNet_Resizing)에 대한 (a) ResNet18, (b) VGGNet16의 ROC 곡선
Fig. 7. ROC Curves of (a) ResNet18 (b) VGGNet16 with CIFAR10 and ImageNet (Resizing) as In-Distribution and Out-of-Distribution Dataset Respectively

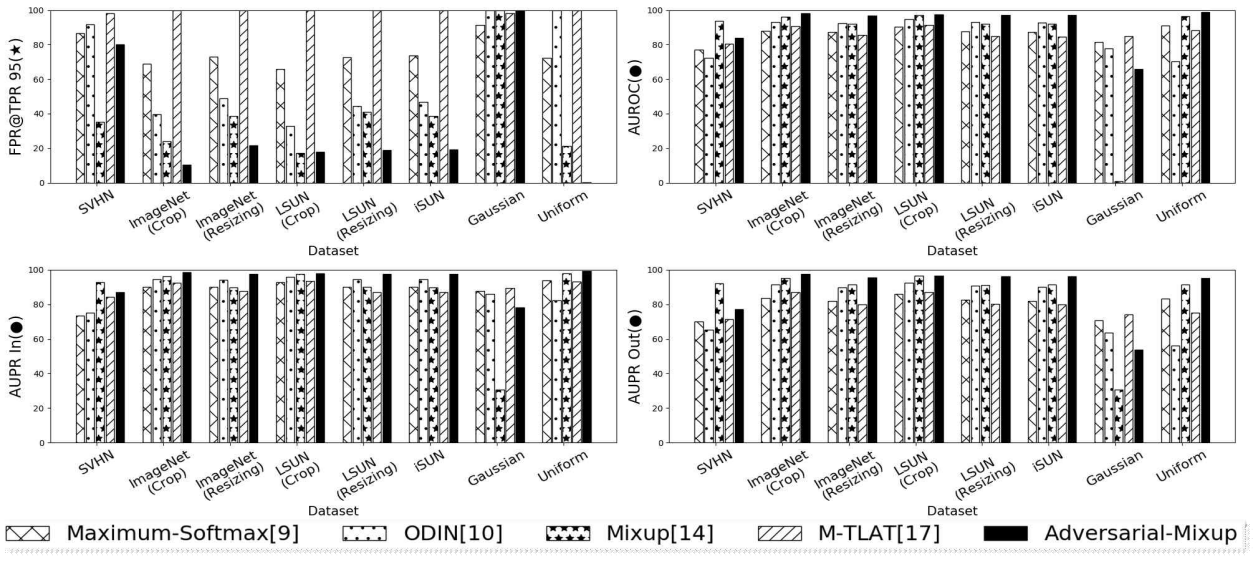


그림 8. ResNet18-CIFAR10에서의 분포 외 데이터 탐지 성능평가
Fig. 8. Out-of-Distribution Data Detection Evaluation Metrics with CIFAR10 as In-Distribution Data in ResNet18 Architecture

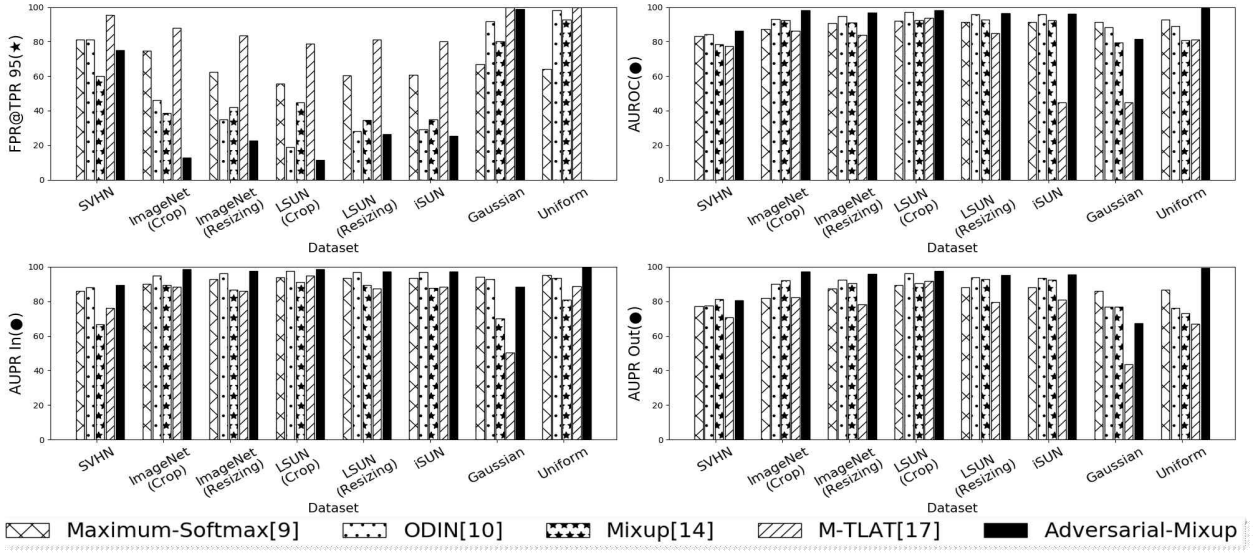


그림 9. VGGNet16-CIFAR10에서의 분포 외 데이터 탐지 성능평가
Fig. 9. Out-of-Distribution Data Detection Evaluation Metrics with CIFAR10 as In-Distribution Data in VGGNet16 Architecture

Mixup, M-TLAT, Adversarial-Mixup 방법을 뜻한다. 그림 8은 ResNet18 모델에 CIFAR10을 분포 내 데이터로 사용한 경우이다. SVHN과 Gaussian이 분포 외 데이터로 사용된 경우를 제외하고 나머지 모든 분포 외 데이터에 대해 제안된 방법이 Mixup 방식 대비 FPR@TPR 95 지표에서 평균 61% 증가율로 향상된 성능을 보였고 특히 LSUN (Crop) 데이터에서 54%, Uniform 데이터에서 최대 100%의 증가율로 향상된 성능을 확인할 수 있다. 그림 9는 VGGNet16 모델에 CIFAR10이 분포 내 데이터로 사용되었을 때의 분포 외 데이터 탐지 성능평가 지표이다. Adversarial-Mixup 방법으로 학습된 모델은 다수의 사례에서 기존의 분포 외 데이터 탐지 성능을 뛰어넘는 것으로 확인되었다. 특히 ImageNet (Crop) 데이터에서 82%, Uniform 데이터에서 최대 98%의 증가율을 보였다.

최근에 제안된 M-TLAT에 따르면 해당 방법은 학습된 모델에 입력되는 분포 내 데이터가 변형 (Corruption) 혹은 왜곡 (Distortion)된 경우에도 이를 정상적인 클래스로 분류하는, 분포 외 데이터에 대한 일반화(Generalization) 성능이 향상됨을 확인할 수 있다. 하지만, 분포 외 데이터 탐지 성능평가에서는 본 논문에서 제안된 Adversarial-Mixup 방법이 더 좋은 성능을 보임을 확인할 수 있다. 또한, 제안된 방법은 입력되는 분포 외 데이터가 학습에 사용된 데이터와

유사한 특성을 포함하고 있는 경우 분포 외 데이터 탐지 성능이 향상되는 것을 확인할 수 있다. 예를 들어, 그림 8, 9 상황에서 학습에 사용된 데이터는 CIFAR10 데이터이며 자연 영상을 포함하고 있다. 학습이 완료된 모델에 ImageNet, LSUN, iSUN 데이터와 같이 자연 영상을 포함하고 있는 데이터가 분포 외 데이터로 입력되었을 때 높은 일반화 성능을 보여 기존방법 대비 최신의 (State-of-the-Art) 성능을 보였다. 반면에 분포 외 데이터로 SVHN이 입력되는 경우 자연 영상에 대한 일반화 성능이 높은 특성과 3채널 숫자데이터를 포함하고 있는 데이터에서 생긴 특성의 차이로 인해 기존방법 대비 낮은 성능이 관측됨을 추정할 수 있다. Mixup 방법은 SVHN에서 가장 좋은 성능을 기록하였고, 이는 3채널 숫자데이터에 대해서는 적대적 공격이 적용된 데이터를 혼합하는 방법보다 분포 내 데이터를 선형가중합 하는 방법이 분포 외 데이터에 대한 강건성을 상대적으로 증가시키는 방법이라 해석할 수 있다.

Ⅵ. 결 론

본 논문에서는 심층신경망 모델이 분포 외 데이터에 대한 강건성을 가질 수 있도록 적대적 데이터 혼합 기법을 활용하여 네트워크를 학습하고, 이를 기반으로 추론 결과에 대한 신뢰성 향상을 목표로 하였다. 실험 결과에서 다양한 유형의 심층신경망에 해당 방법의 적용 가능성을 확인하였고, 기존방법 대비 향상된 성능을 확인할 수 있다. 또한, 제안한 학습 방법이 분포 외 데이터에 대한 신뢰 점수를 낮추어 전체적인 심층신경망 모델의 추론 결과에 대한 신뢰성을 향상하였다. 차후 연구에서는 다양한 모델과 분포 내 데이터 셋을 이용한 확장된 실험을 진행하고, 해당 방법으로 해결하지 못한 경우 (Gaussian 잡음 등)의 해결을 위해 추가적인 연구를 진행할 예정이다.

References

[1] I.J. Goodfellow, J. Shlens, C. Szegedy, “Explaining and Harnessing Adversarial Examples,” International Conference on Learning Representations(ICLR), 2015.

[2] A. Nguyen, J. Yosinski, J. Clune, “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images,” Computer Vision and Pattern Recognition(CVPR), pp. 427-436, 2015.

[3] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, “Mixup: Beyond Empirical Risk Minimization,” International Conference on Learning Representations(ICLR), 2018.

[4] K. Lee, H. Lee, K. Lee, J. Shin, “Training Confidence-Calibrated Classifiers for Detecting Out-of-Distribution Samples,” International Conference on Learning Representations(ICLR), 2018.

[5] D. Hendrycks, M. Mazeika, T. Dietterich, “Deep Anomaly Detection with Outlier Expousre,” International Conference on Learning Representations(ICLR), 2019.

[6] S. Hawkins, H. He, G. Williams, R. Baxter, “Outlier Detection Using Replicator Neural Networks,” International Conference on Data Warehousing and Knowledge Discovery, Springer, pp. 170-180, 2002.

[7] L. Ruff, R.A. Vandermeulen, N. Görnitz, L. Deecke, S.A. Siddiqui, A. Binder, E, Müller, M. Kloft, “Deep One-Class Classification,” Proceedings of the 35th International Conference on Machine Learning(ICML), pp. 4393-4402, 2018.

[8] J. Ren, P.J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, B. Lakshminarayanan, “Likelihood Ratios for Out-of-Distribution Detection,” 33rd Conference on Neural Information Processing Systems(NeurIIPS), pp. 14707-14718, 2019.

[9] D. Hendrycks, K. Gimpel, “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,” International Conference on Learning Representations(ICLR), 2017.

[10] S. Liang, Y. Li, R. Srikant, “Enhancing The Reliability of Out-of-Distribution Image Detection in Neural Networks,” International Conference on Learning Representations(ICLR), 2018.

[11] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, “On Calibration of Modern Neural Networks,” Proceedings of the 34th International Conference on Machine Learning(ICML), 2017.

[12] K. Gwon, J. Yoo, “Out-of-Distribution Data Detection Using Mahalanobis Distance for Reliable Deep Neural Networks,” Proceedings of 2020 IEMEK Symposium on Embedded Technology(ISET 2020), 2020 (in Korean).

[13] K. Lee, K. Lee, H. Lee, J. Shin, “A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks,” 32nd Conference on Neural Information Processing Systems(NeurIIPS), pp. 7167-7177, 2018.

[14] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, S. Michalak, “On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks,” 33rd Conference on Neural Information Processing Systems(NeurIIPS), 2019.

[15] A. Kurakin, I.J. Goodfellow, S. Bengio, “Adversarial Examples in the Physical World,” International Conference on Learning Representations(ICLR), 2017.

[16] C. Szgedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow, R. Fergus, “Intriguing Properties of Neural Networks,” Computer Vision and Pattern Recognition(CVPR), 2014.

[17] A. Laugros, A. Caplier, M. Ospici, “Addressing Neural Network Robustness with Mixup and Targeted

Labeling Adversarial Training," European Conference on Computer Vision(ECCV), 2020.

Kyungpil Gwon(권 경 필)

2013 Embedded System Engineering from Daegu University (B.S.)
2019~Information and Communication Engineering from Daegu University (M.S.)

Career:
2018~ Member, IeMeK
Field of Interests: Deep Learning, Safe/Reliable AI
Email: kyungpil08@daegu.ac.kr

Joonhyuk Yoo (유 준 혁)

1993 Electrical and Electronic Engineering from POSTECH (B.S.)
1995 Electrical and Electronic Engineering from POSTECH (M.S.)
2007 Computer Engineering from University of Maryland at College Park, MD USA (Ph.D.)
2009~Department of Artificial Intelligence at Daegu Univsersity (Professor)

Career:
1995~2000 Embedded System Engineer, Samsung Electronics
2005~2007 Embedded Software Engineer, NASA CREAM
2008~2009 Research Professor, Korea University
Field of Interests: Machine Learning, Computer Vision, On-Device AI, Cyber-Physical Systems
Email: joonhyuk@daegu.ac.kr