Out mixup: Out of distribution detection using mixup with out-of-distribution data

Hangeol Chang(Yonsei university) Songkuk Kim*(Yonsei university)

Abstract

우리는 classifier training에서 in-distribution image(cifar 10) 과 out-of-distribution image(isun)의 mixup을 통하여 out-of-distribution detection 성능 향상을 이루어 냈다. 두 dataset을 mixup하고, out-of-distribution image의 mixup 정도에 따라 Target variable을 uniform distribution에 가깝게 함으로서, out-of-distribution data input에 대한 유의미한 confidence 감소를 발생시켰고, out-of-distribution detection 성능을 향상시켰다. 또한 이러한 방법으로 훈련한 network는 낮은 정답률을 가진 경우 낮은 confidence를 출력하여 overconfidence 문제의 해결에도 효과적임을 알 수 있었다.

1 Introduction

Machine learning classifier는 많은 문제를 효과적으로 해결하며, 현대는 많은 dataset에 대해서 매우 높은 정확도로 정답을 예측한다. 하지만 machine learning classifieer는 때때로 틀린 정답을 높은 confidence로 예측한다. (Goodfellow et al.,2015; Hendrycks et al. 2017). 또한 학습시키지 않은 data인 out-of-distribution data에 대해서도 높은 confidence로 하나의 정답을 선택한다. 왜냐하면, Machine learning classifier에서는 accuracy 상승을 최우선으로 network가 학습되었기 때문이다. 이러한 잘못된 높은 confidence 예측은 classifier를 활용할 때 위험한 상황을 발생시키거나, 큰 손해를 불러올 수 있다. 그래서 현대에는 Machine learning classifier에서 학습에 사용하지 않은 label에 속하는 dataset인 out-of-distribution data가 input으로 들어올 때, 이를 선별해 내는 out-of-distribution detection 성능을 발전시키기 위한 다양한 연구가 이루어 지고 있다.

Out-of-distribution detection문제는 2016년 Dan Hendrycks의 논문 “A BASELINE FOR DETECTING MISCLASSIFIED AND OUT-OF-DISTRIBUTION EXAMPLES IN NEURAL NETWORKS(ICLR2017)”에서 처음 제안되었다. 해당 논문에서, Hendrycks는 out-of-distribution detection문제를 정의하고 몇가지 성능평가 방법을 제시하였고, Baseline성능을 제시하였다. 이후 2017년에, Shiyu Liang의 논문” Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks(ICLR2018)”에서는, inference단계에서, temperature scaling 및 input preprocessing을 사용하여, out-of-distribution detection성능을 향상시켰다. 2017년 Dan Hendrycks는 논문” A BASELINE FOR DETECTING MISCLASSIFIED AND OUT-OF-DISTRIBUTION EXAMPLESIN NEURAL NETWORKS”에서, training단계에서 GAN loss를 이용하여 out-of-distribution dataset을 생성하고, out-of-distribution data의 output이 uniform distribution이 되도록 학습시켜 out-of-distribution성능을 향상시켰다. 해당 논문에서는 out-of-

distribution data의 output이 uniform distribution을 가지도록 유도하는 훈련방법이 효과가 있음을 보여준다.

2 purpose of study

본 논문은 Training 단계에서 보조out-of-distribution dataset을 사용하여 out-of-distribution detection 성능을 높이는 방법을 제안한다. 본 논문에서는 Training 단계에서, Hongyi(2018)가 accuracy 향상을 위해 제안한 mixup기법을 활용한다. classifier의 Training에 사용된 in-distribution data와 out-of-distribution data를 mixup하여 input으로 넣어주고, out-of-distribution data가 mixup된 비율대로, uniform distribution과 가까운 output이 나오도록 loss를 설정한다.

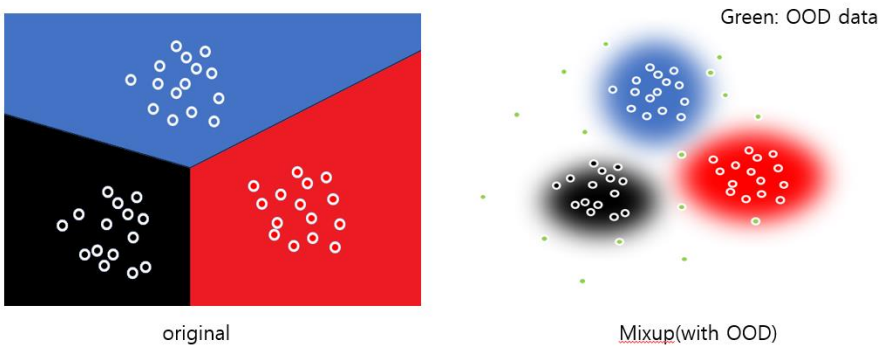

figure 1:mixup with out of distribution data의 기대 효과

해당 논문에서도, 2017년 Dan Hendrycks의 논문과 마찬가지로, out-of-distribution data가 들어온 경우 uniform noise를 가지는 것을 이상적이라고 가정하고, 이를 유도한다. 하지만, 단순히 학습에 사용되는 out-of-distribution data가 uniform noise를 가지도록 학습할 경우, 학습에 사용되는 종류와 비슷한, out-of-distribution data에 대해서만, detection이 용이한 문제점이 발생할 수 있다. 이에 대한 해결책으로, 우리는 mixup을 사용하여, out-of-distribution data로 in-distribution data의 성질을 약하게 하는 방법을 고안하였다.

out-of-distribution data가 어디에 있던지, In-distribution data에 out-of-distribution data를 섞는 비율을 증가시킬수록, data는 in-distribution에서 멀어질 것이다. 따라서, In-distribution data에 out-of-distribution data를 섞는 비율을 증가시킬수록 confidence가 감소하도록 training 된 경우, 각각의 label 영역에 smoothing된 연속적인 경계가 만들어져, label영역과 거리가 멀어질수록, 낮은 confidence를 출력하여, out-of-distribution detection 성능의 향상을 기대할 수 있다.

3 Experiment method

3-1. Image set and network

실험은 CIFAR-10 data의 classification task를 ResNet-18을 사용하여 진행하였다. 훈련은 momentum(momentum=0.9,lr=0.1)방법을 기본으로 하여,100epoch및 150epoch에서 learning rate를 각각 1/10씩 감소시키는 방법을 사용하였다. (Hongyi et al.2018)

실험의 training의 input으로는 in-distribution data(CIFAR-10)와 out-of-distribution data를 mixup하여 들어간다. Cifar 10 data와 mixup 한 out-of-distribution data는 ISUN data를 사용하였다. 해당 dataset에는 특정 동물이나 사물이 아니라, 배경을 label로 한 사진이 많아 cifar-10 data와 label이 겹치는 사진이 적어 선택하였다.

3-2.mixup

training에서는 다음의 mixup방법을 사용하였다.

-Input data mixup

in-distribution-image data에서 random하게 뽑은 image를 x_1, out-of-distribution image data에서 random하게 뽑은 data를 x_2라고 할 때, training의 input data, x는,

$$x = \lambda \times x_1 + (1-\lambda) \times x_2 \qquad \lambda \sim \text{Beta}(1,1)\ \text{(Hongyi et al.2018)}$$

이다.

-Target variable mixup

Input data에 out-of-distribution data가 섞인 만큼, target variable은 uniform distribution에 가깝게 한다.

x_1의 target variable을 y_1 이라 하고 모든 element의 값이 1인 행렬을 $\vec{1}$이라 할 때, input data, x의 target variable y는,

$$y = \lambda \times y_1 + (1-\lambda)/10 \times \vec{1}$$

이다. Trainig의 loss는 Target variable과 output의 cross entropy loss를 사용하였다.

Training에서는 Mixup한 data의 사용 비율을 0%부터 10%간격으로 100%까지 변화시켜 가며, 11개의 훈련된 network를 얻었다. 그래서 mixup한 data가 몇%를 차지할 때 가장 성능 향상폭이 큰지 검증해 보았다.

3-3 evaluation method

Out-of-distribution data를 detection하는 방법으로는 특정 threshold 값보다 confidence가 낮으면 out-of-distribution data라고 판단하는 방법을 사용한다 (Hendrycks et al.2017). 해당 성능의 평가를 위해서, threshold를 변화시켜가며 성능을 측정하는 방법들을 사용하였다. 성능 측정 지표로는, Liang(2018)이 ODIN 논문에서 제안한 FPR at 95%TPR 및 Hendrycks(2017)이 제안한 AUPR 및 AUROC를 사용하였다. 또한 out-of-distribution detection의 성능 평가를 위한 out-of-distribution data는 Liang(2018)이 Odin 논문에서 사용한 Gaussian Noise, Uniform Noise, Tiny-imagenet(crop,resize), LSUN(crop,resize), iSUN data를 사용하였다.

4. Experiment results

4-1. accuracy and test loss

Mixup rate(%)	test loss	Accuracy(%)
0	1.609968424	86
10	1.608134508	87
20	1.606212616	87
30	1.610428333	87
40	1.616743803	87
50	1.627949238	87
60	1.615610957	87
70	1.611382484	87
80	1.627797723	87
90	1.634374142	87
100	1.623036504	87

figure 2. Acuuracy 및 test loss

Mixup data 비율이 늘어남에 따라 in-distribution data를 충분히 학습하지 못할 것이 우려되었으나, 이번 실험에서는 모든 input data가 mixup된 data일 경우에도 accuracy감소는 발생하지 않았다. 다른 network와 dataset에 대해서는 추가 검증이 필요할 것으로 보이나, 실험한 상황의 경우에는 mixup 으로 인한 성능 저하는 없는 것으로 판단할 수 있다.

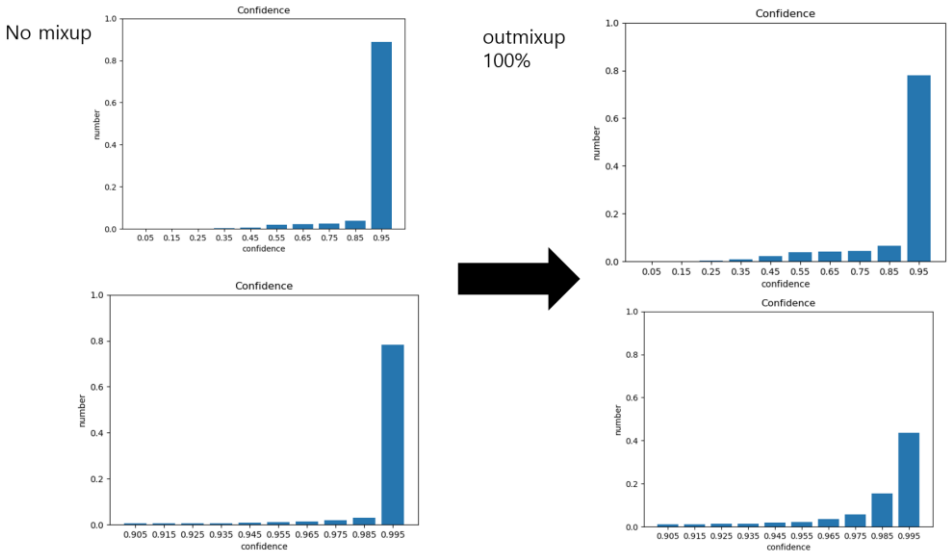

figure 3.resnet 18에서 mixup에 따른 in-distribution-data의 confidence 비율 변화.

실험에서 in-distribution data의 confidence는 살짝 감소하였다. 우리는 training에서, in-distribution training data의 영역과 거리가 멀어질수록, 점차 confidence가 낮아지는 것을 의도하였고, test data와 training data에는 차이가 있으므로, 약간의 confidence감소는 합리적인 결과로 볼 수 있다.

4-2 confidence

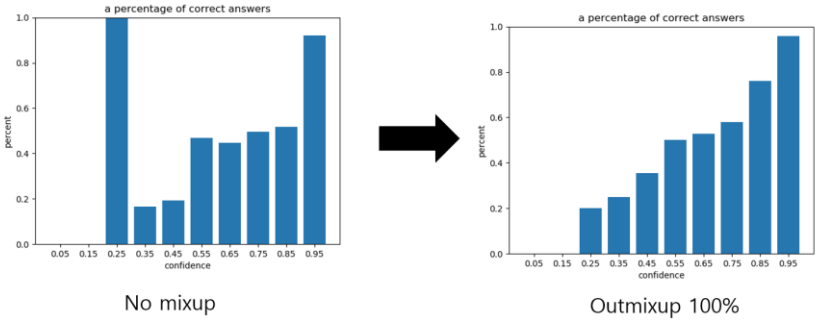

figure 4.resnet 18에서 각 confidence 범위에서 정답률

위 그래프에서는 confidence에 따른 classification task의 정답률을 보여준다. Out-of-distribution data와 mixup을 진행한 경우 confidence가 정답률과 유사해진 것을 볼 수 있다. 정답률이 높은 경우 더 높은 confidence를 출력하며, 정답률이 낮은 경우 더 낮은 confidence를 출력한다. Out-of-distribution data와 mixup을 진행하면, network가 맞추기 어려운 input에 대해 낮은 confiden를 출력하므로, confidence가 낮을 경우 해당 출력값이 틀릴 가능성이 높다고 예측이 가능해진다.

이를 ECE로 점수화 시켜 파악해 볼 수 있다. ECE의 계산 방법은 다음과 같다.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

구간을 100개로 나누어 ECE를 계산해 본 결과 ECE는 mixup을 하지 않았을 때 0.0902에서 모든 data를 mixup한 경우 0.0435까지 감소하였다.

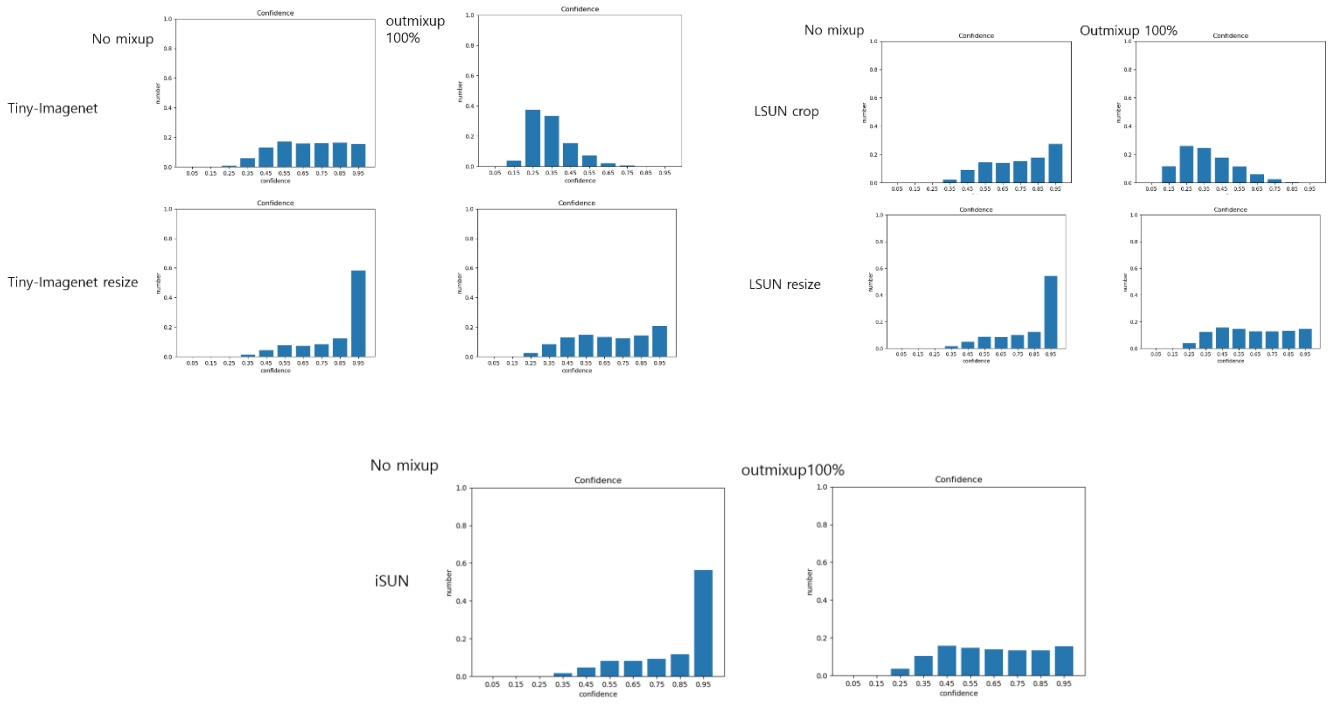

figure 5.mixup에 따른 out-of-distribution data의 confidence 변화

위 그림과 같이 다양한 out-of-distribution data를 network의 input으로 넣었을 때, out-of-distribution data와 mixup을 한 경우 mixup을 하지 않은 경우보다 훨씬 낮은 confidence를 보이는 것을 확인할 수 있었다. Data를 가리지 않고, 실험한 모든 data에 대해 일관적인 confidence 감소를 보였다. 이 감소 폭은 in-distribution-data의 감소 폭보다 크므로, out-of-distribution detection 성능의 향상을 기대할 수 있다.

3.AUROC, AUPR, FPR at TPR 95%

no mixup /mixup with out-of-distribution data	FPR at TPR 95%	AUROC	AUPR In	AUPR Out
Tiny-ImageNet (crop)	0.078/0	0.983/1	0.987/1	0.977/0.999
Tiny-ImageNet (resize)	0.521/0.316	0.922/0.959	0.94/0.968	0.899/0.949
LSUN (crop)	0.107/0	0.979/0.999	0.985/0.999	0.97/0.999
LSUN (resize)	0.475/0.227	0.933/0.97	0.949/0.976	0.914/0.964
iSUN	0.488/0.241	0.93/0.968	0.947/0.975	0.91/0.961
Uniform Noise	0.947/0.931	0.661/0.807	0.698/0.85	0.591/0.702
Gaussian noise	0.891/0.518	0.77/0.928	0.813/0.945	0.693/0.901

figure 6.resnet 18에서 mixup 시 out-of-distribution detection 성능

put-of-distribution을 판단하는 threshold를 변화시켜가며AUROC, AUPR 및 FPR at TPR 95%를 측정한 결과이다.

input이 일반 cifar10 data일때와 비교하여, input이 cifar10과 out-of-distribution data(iSUN)data를 mixup한 경우에 모든 상황에서 더 좋은 수치를 얻을 수 있었다. FPR at TPR 95%는 감소하였고, AUROC 및 AUPR은 모든 상황에서 증가하였다. Out-of-distribution data와의 mixup을 통해, out-of-distribution detection성능의 향상을 얻을 수 있음을 알 수 있다.

4.정답과 오답 구분

	nomixup	**outmixup**
FPR at TPR 95%	61.60%	56.80%
AUROC	87.40%	89.00%
AUPR In	87.60%	89.70%
AUPR Out	85.20%	86.70%

figure 7. mixup 시 정답과 오답을 구분하는 성능

정답과 오답을 out-of-distribution data detection과 같은 방법으로, network가 일정 threshold이하의 confidence를 출력할 때, 오답으로 판정하는 방법을 사용하였을 때, threshold를 변화시켜가며 AUROC, AUPR 및 FPR at TPR 95%를 측정한 결과이다. 약간의 성능 향상을 보였다.

5 Discussion & future work

실험 결과에서 볼 수 있듯 우리는 Out-of-distribution data와의 mixup을 통해 in-distribution data label 영역과 먼 이미지가 들어올수록 점점 낮은 confidence를 가지도록 훈련시켜 out-of-distribution data가 input으로 들어오거나, 정답을 정확하게 판정하기 어려운 data가 들어올 경우 낮은 confidence를 출력하도록 유도하여 out-of-distribution detection 성능의 향상을 이뤄냈다. 해당 실험의 방법을 적절하게 사용하여, 다양한 영역에서 out-of-distribution detection 성능을 향상시킬 수 있을 것으로 기대된다.

이 실험이 image 개수가 적은 dataset과 parameter가 적은 network에서 이뤄진 만큼, imagenet과 같이 더 규모가 큰 dataset에 대해서도 효과적으로 동작하는지 추가 실험이 필요하다. 또한, 해당 실험의 효과가, 기존에 out-of-distribution 성능 향상 기법인 Dean temperature scaling(Hinton et al. 2014; Liang et al.2018), input preprocessing(Goodfellow et al.2014; Liang et al.2018)등과 동시에 동작 가능한지에 대한 실험도 추가적으로 필요하다. 해당 방법이, 다른 out-of-distribution 방법과 동시에 동작 가능하다면, out-of-distribution detection 성능 향상에 큰 성과가 될 것으로 보인다.

Reference

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR), 2015.

Dan Hendrycks and Kevin Gimpel. A BASELINE FOR DETECTING MISCLASSIFIED AND OUT-OF-DISTRIBUTION EXAMPLES IN NEURAL NETWORKS. In International Conference on Learning Representations (ICLR), 2017.

Shiyu Liang, Yixuan Li and R. Srikant. ENHANCING THE RELIABILITY OF OUT-OF-DISTRIBUTION IMAGE DETECTION IN NEURAL NETWORKS In International Conference on Learning Representations (ICLR), 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Network 2016

Geoffrey Hinton, Oriol Vinyals and Jeff. Distilling the Knowledge in a Neural Network (NIPS),2014.

Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy .EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES