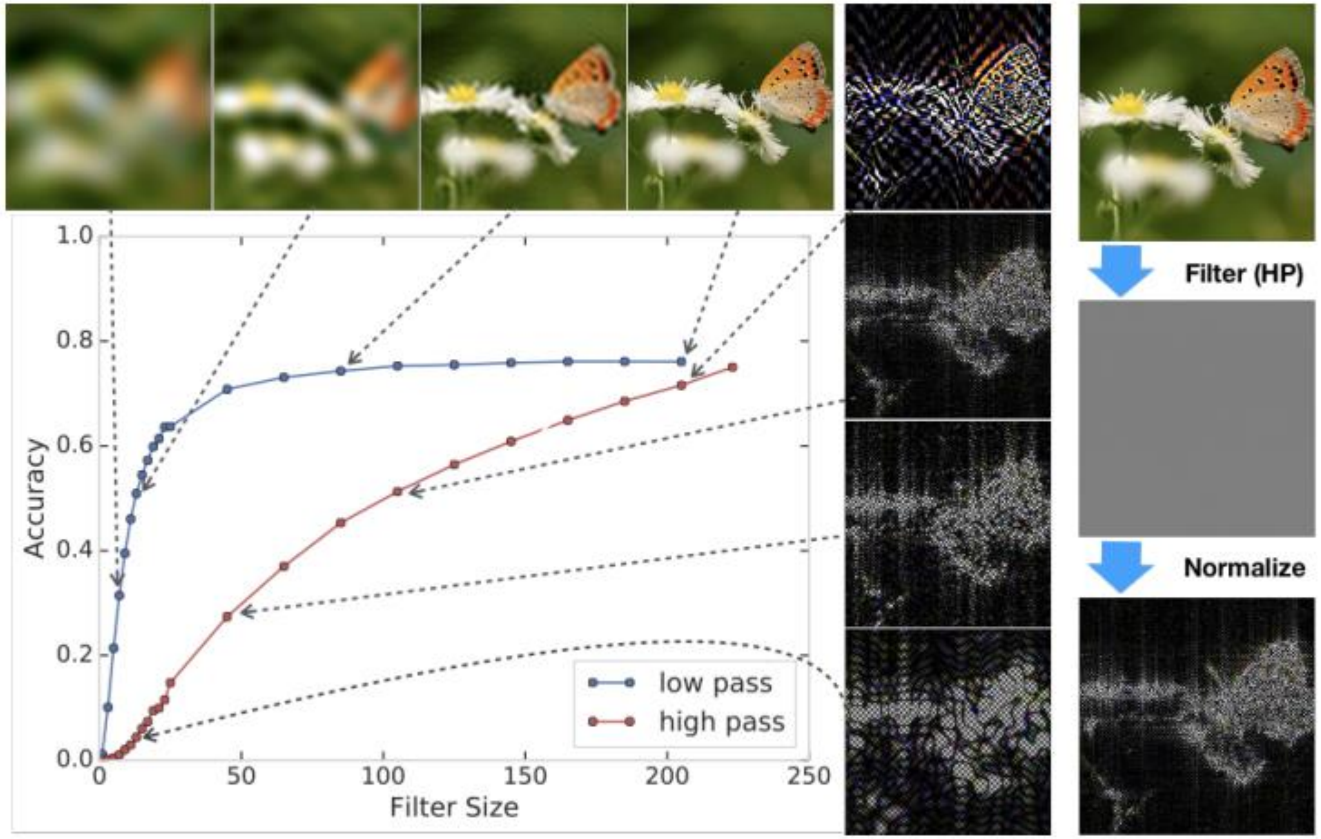
Adversarial Attack & Future Work

Kim Seung Hwan (overnap@khu.ac.kr)

2022. 12. 30.

What I learned from Adversarial Attack

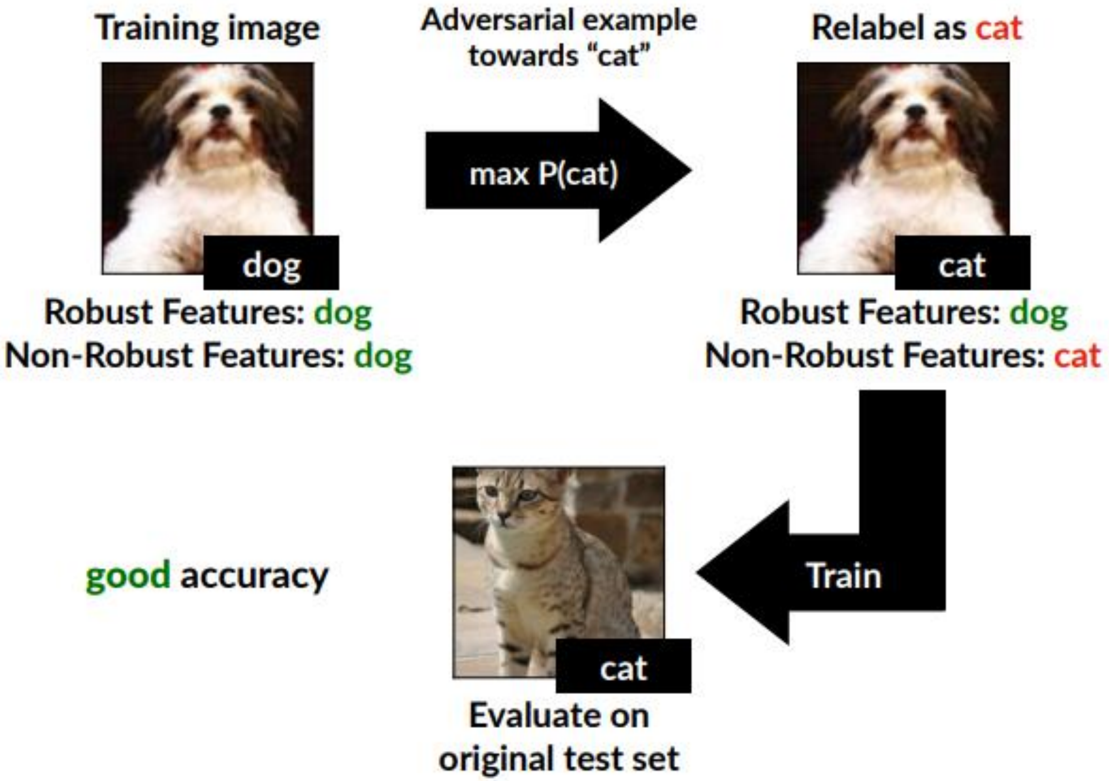
- What is DNN model really learning?



From D. Yin et al. “A Fourier Perspective on Model Robustness in Computer Vision”

What I learned from Adversarial Attack

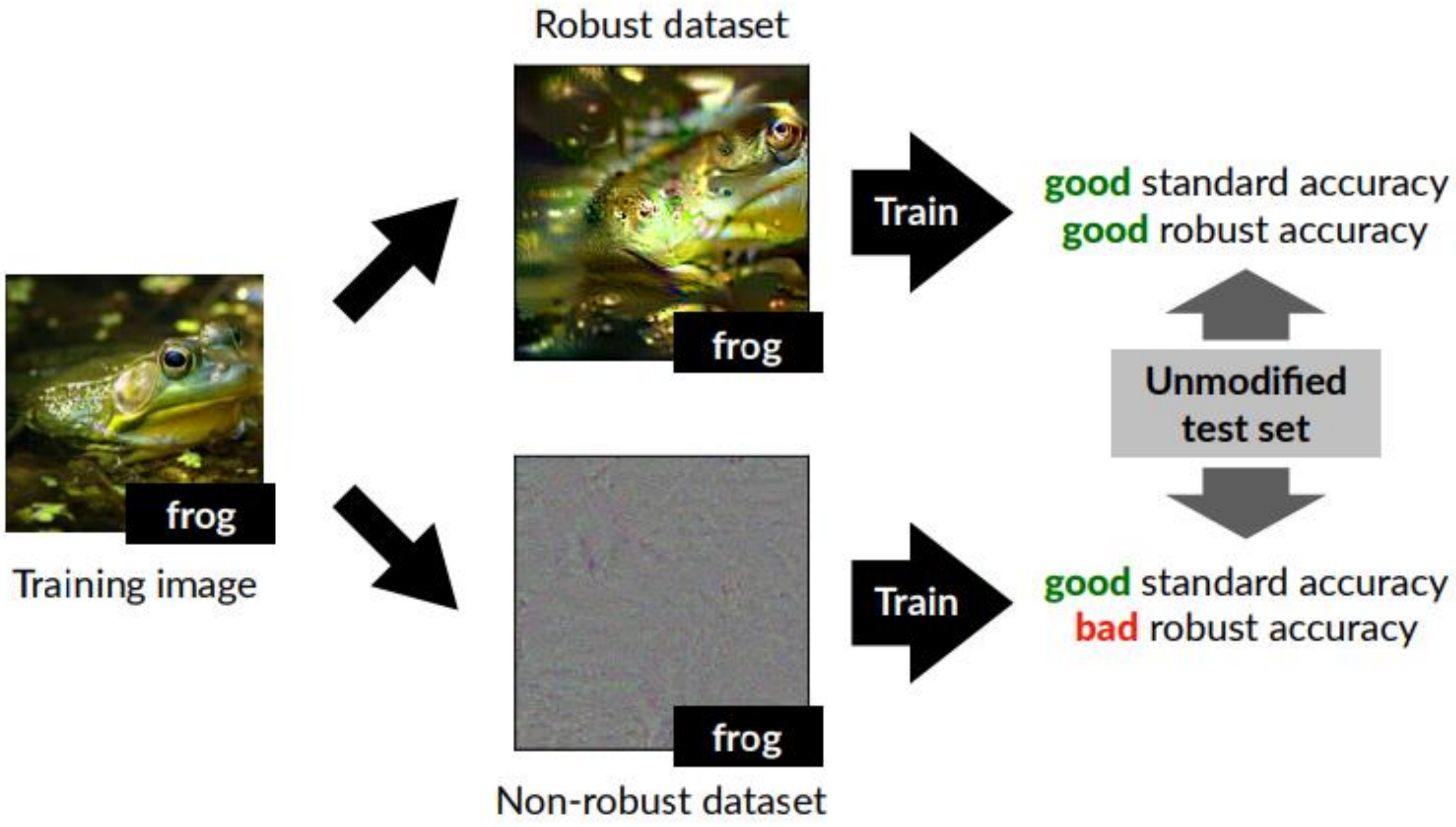
- What is DNN model really learning?



From A. Ilyas et al. "Adversarial Example Are Not Bugs, They Are Features"

What I learned from Adversarial Attack

- What is DNN model really learning?



From A. Ilyas et al. “Adversarial Example Are Not Bugs, They Are Features”

What I learned from Adversarial Attack

- DNN models learn more (or completely different) information than humans
- These are generally different from human perception
- This is because models are forced to decrease loss only

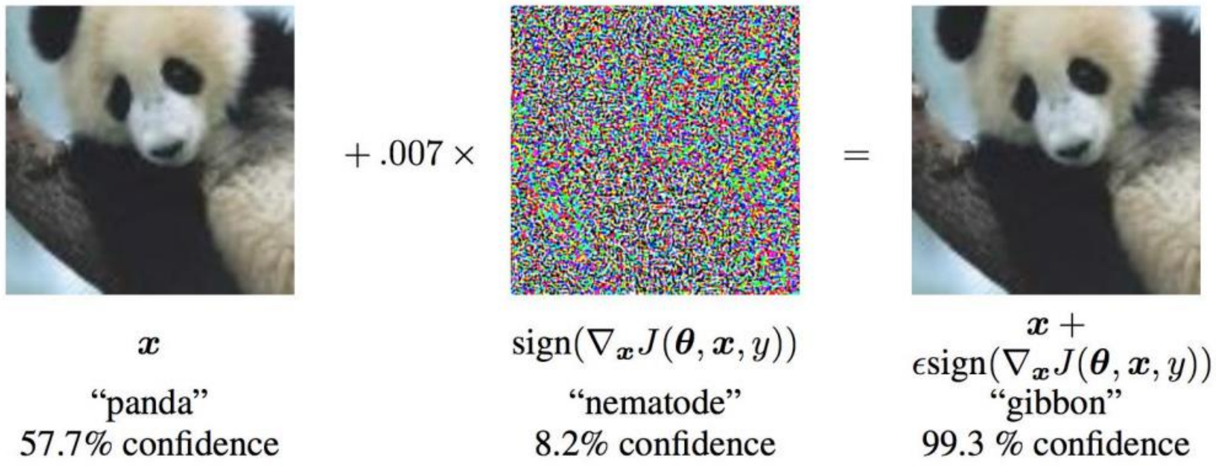
- So, is Adversarial Vulnerability evil?
- Do models really need to mimic human perception?

What I wrote

- “비선형적 이미지 혼합 데이터 증강을 이용한 Deep Neural Network의 적대적 공격 취약성 개선”
- “Adversarial Vulnerability Improvement by Non-linear Image Mixing Data Augmentations”

1. Background

- Adversarial Attack is to fool a DNN model by perturbing inputs that are difficult for humans to perceive


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

- Such vulnerability is an obstacle to the application of DNNs in areas where security or reliability are important (e.g., A.D.)

FGSM method; from I. Goodfellow et al. "Explaining and Harnessing Adversarial Examples"

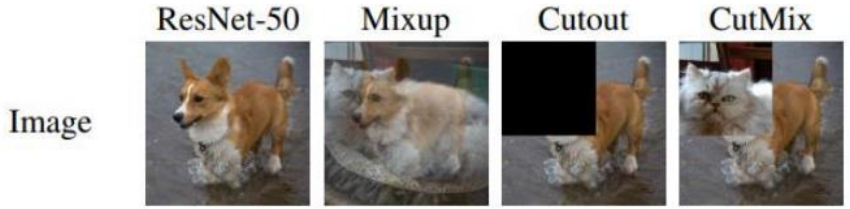
1. Background

- Adversarial Transferability, which means that other models trained on similar datasets can also be fooled, implies a black-box attack
- This suggests that not only the architecture of models, but also the dataset itself has a lot to do with the vulnerability
- Recent studies explain that this is because models mainly extract features that are useful but not robust (i.e., vulnerable to attack)

1. Background

- Data augmentation is essential in modern DNN training
- Data augmentation, however, is also associated with the adversarial vulnerability, as it changes the distribution of datasets

- For this study, let's focus on mixing augmentations (e.g., Mixup, CutMix)
- They perform well and are so universal; they are essential to achieving SOTA



From S. Yun et al. "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features"

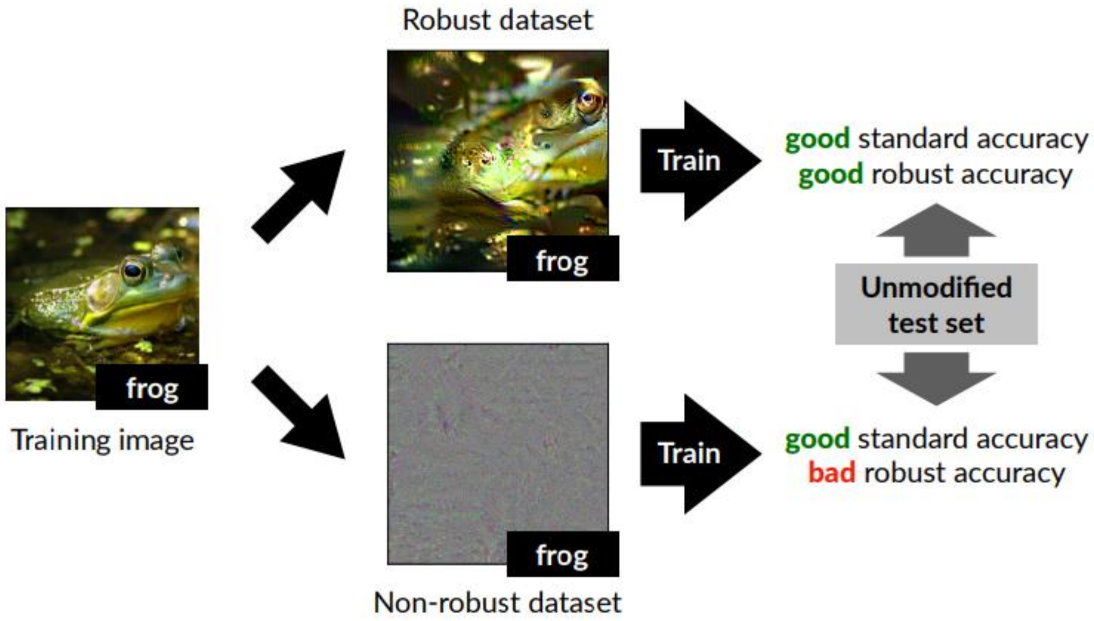
2. Proposal

- Despite the great improvement of mixing augmentations, linear blending of labels is questionable

- Humans barely recognize dogs in 1:9 images of dogs and cats
- But since models need to find 10% of the dogs, it will learn finer features that are difficult for humans to see

2. Proposal

- Useful but less robust features tend not to match human perception, so we need to check them in terms of adversarial vulnerability



From A. Ilyas et al. “Adversarial Example Are Not Bugs, They Are Features”

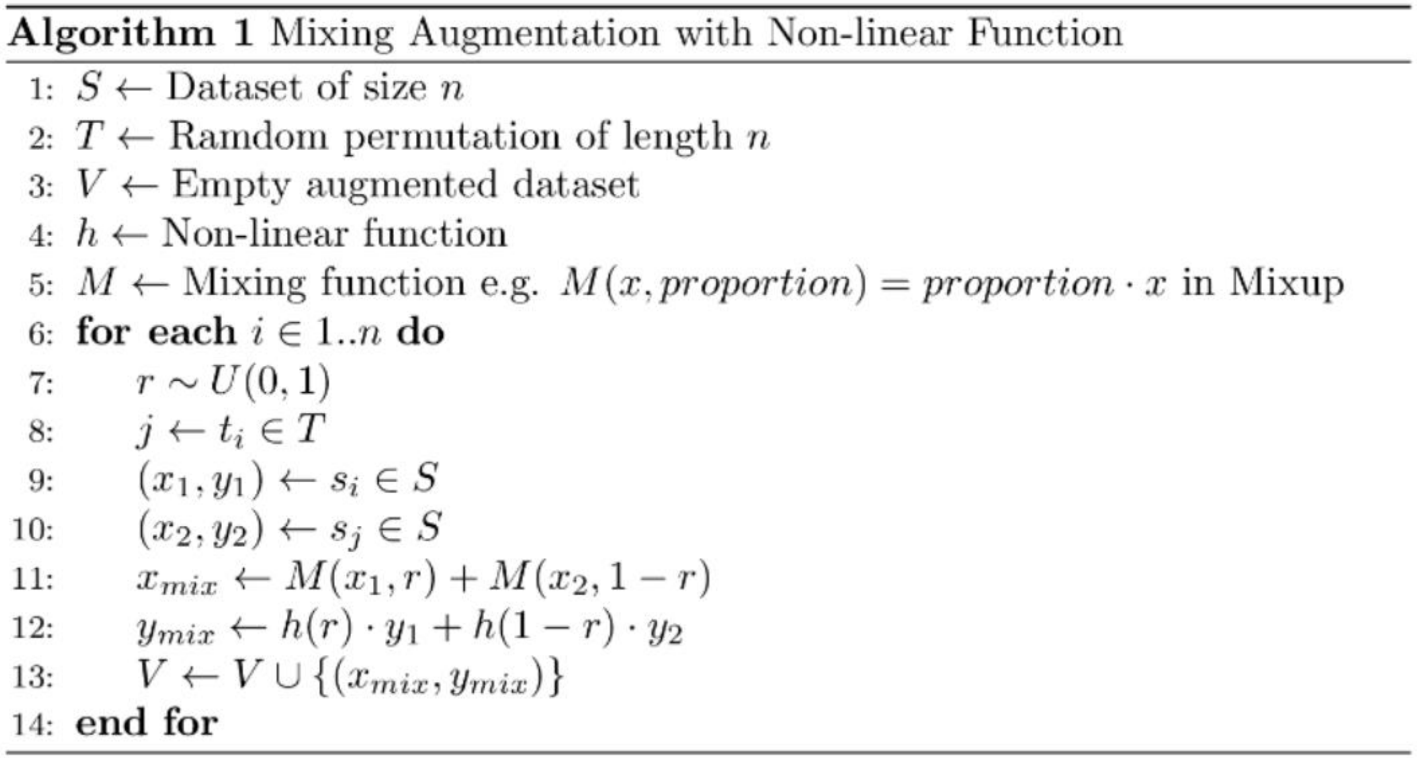
2. Proposal

- Set an appropriate function h to mimic human perception
- h is a function that attenuates small signals and strengthens strong signals
- In this study, h is set as follows (p is a non-linear parameter)

$$h(x) = \begin{cases} 2^{p-1}x^{p} & \text{if } x \leq 0.5 \\ 1 - 2^{p-1}(1-x)^{p} & \text{otherwise} \end{cases}$$

2. Proposal

- Leave the mixing input images as is and pass the label value to merge into h (lines 4 and 12)

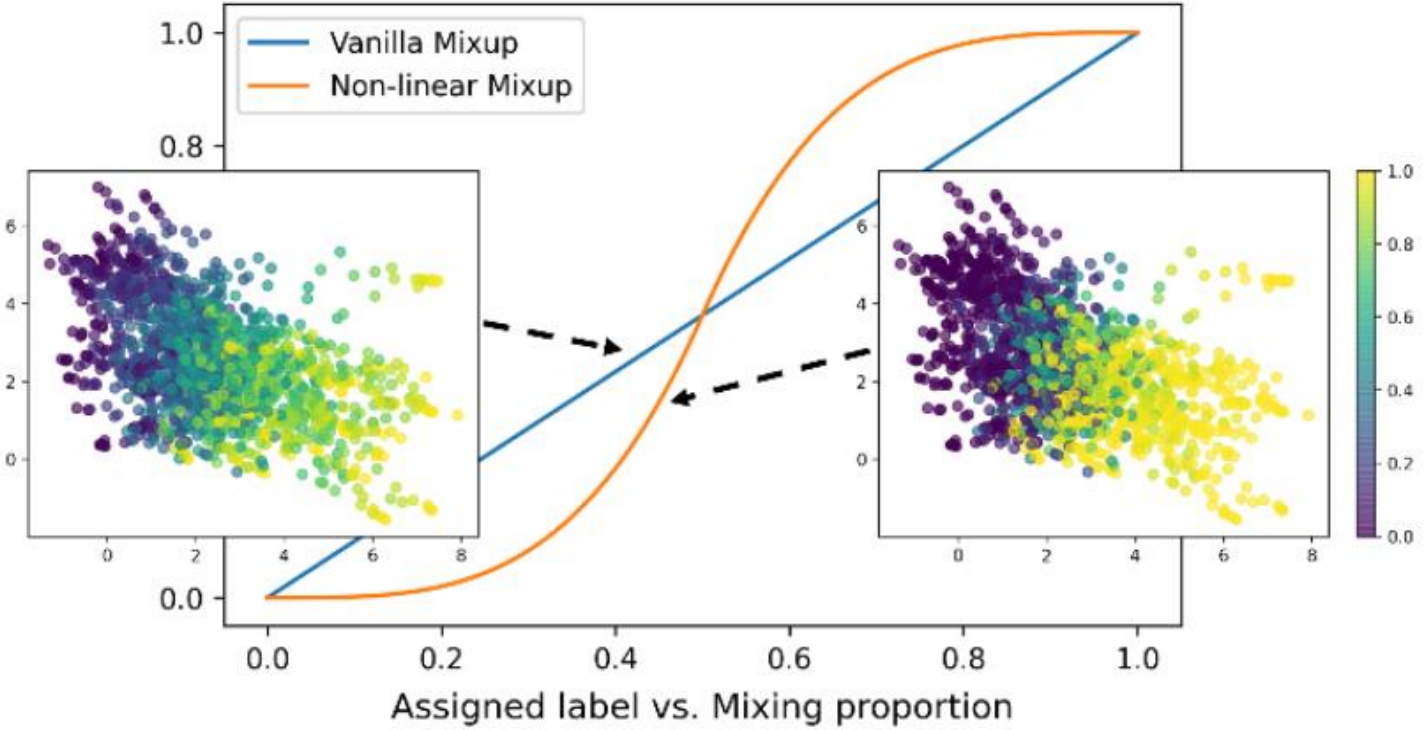


Algorithm 1 Mixing Augmentation with Non-linear Function

```
 1: $S \leftarrow$ Dataset of size $n$
 2: $T \leftarrow$ Ramdom permutation of length $n$
 3: $V \leftarrow$ Empty augmented dataset
 4: $h \leftarrow$ Non-linear function
 5: $M \leftarrow$ Mixing function e.g. $M(x, proportion) = proportion \cdot x$ in Mixup
 6: for each $i \in 1..n$ do
 7:     $r \sim U(0, 1)$
 8:     $j \leftarrow t_i \in T$
 9:     $(x_1, y_1) \leftarrow s_i \in S$
10:     $(x_2, y_2) \leftarrow s_j \in S$
11:     $x_{mix} \leftarrow M(x_1, r) + M(x_2, 1 - r)$
12:     $y_{mix} \leftarrow h(r) \cdot y_1 + h(1 - r) \cdot y_2$
13:     $V \leftarrow V \cup \{(x_{mix}, y_{mix})\}$
14: end for
```

2. Proposal

$$h(x) = \begin{cases} 2^{p-1}x^p & \text{if } x \le 0.5 \\ 1 - 2^{p-1}(1-x)^p & \text{otherwise} \end{cases}$$



3. Experiment

- ResNet-50, CIFAR-10, Mixup, FGSM
- Linear mixing augmentation vs proposed non-linear one (p=6)

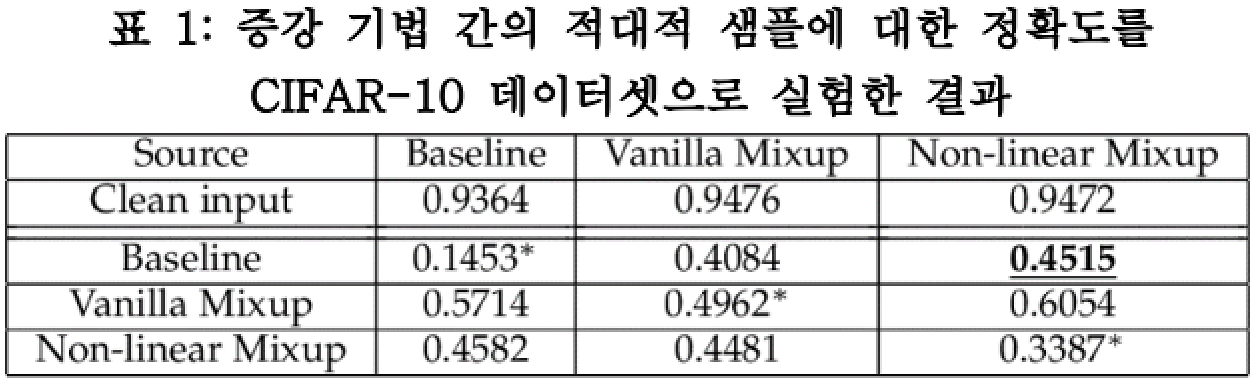


표 1: 증강 기법 간의 적대적 샘플에 대한 정확도를 CIFAR-10 데이터셋으로 실험한 결과

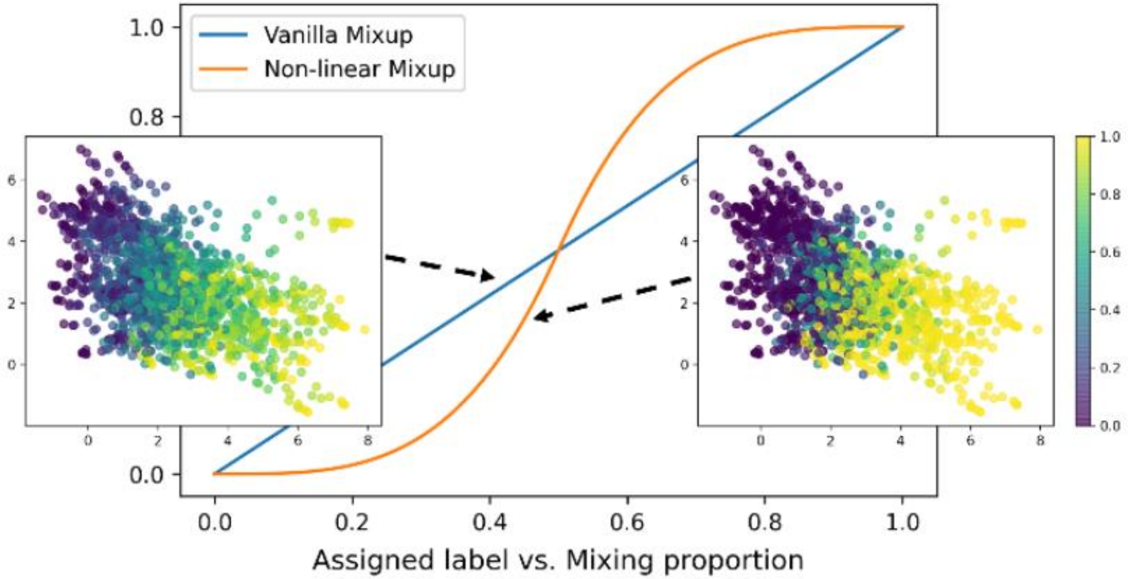
Source	Baseline	Vanilla Mixup	Non-linear Mixup
Clean input	0.9364	0.9476	0.9472
Baseline	0.1453*	0.4084	**0.4515**
Vanilla Mixup	0.5714	0.4962*	0.6054
Non-linear Mixup	0.4582	0.4481	0.3387*

- Maintain the improvement of linear (conventional) mixup
- Superior in the black-box, inferior in the white-box

3. Experiment

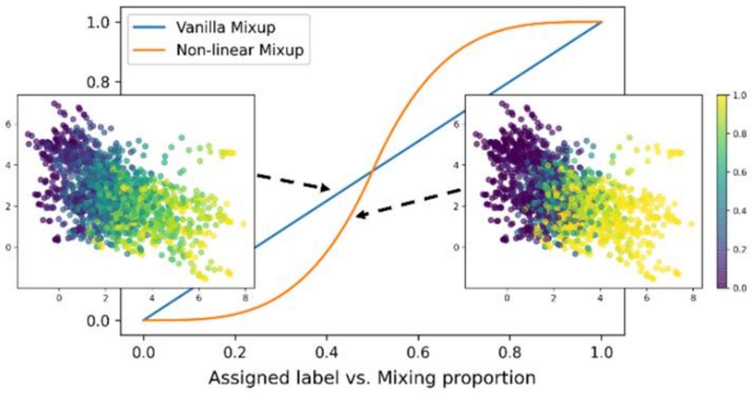
- Empirically verifying the hypothesis

- Let’s look at it from a different perspective
- The proposed non-linearity leads to steeper decision boundaries of the model



3. Experiment

- Adversarial examples are found around or beyond boundaries
- Steep boundaries have little information about the location of decision boundaries and are insensitive to attacks from incorrect directions, which reduces the convergence of attacks
- For an attack in the right direction, a large change is obtained with a small move, which explains the inferiority in the white box



3. Experiment

- Setting h to an arbitrary nonlinear function(e.g. square, sqrt) degrades accuracy or robustness
- The idea of simulating human perception was more important than the nonlinearity itself
- The experiment for parameter p is:

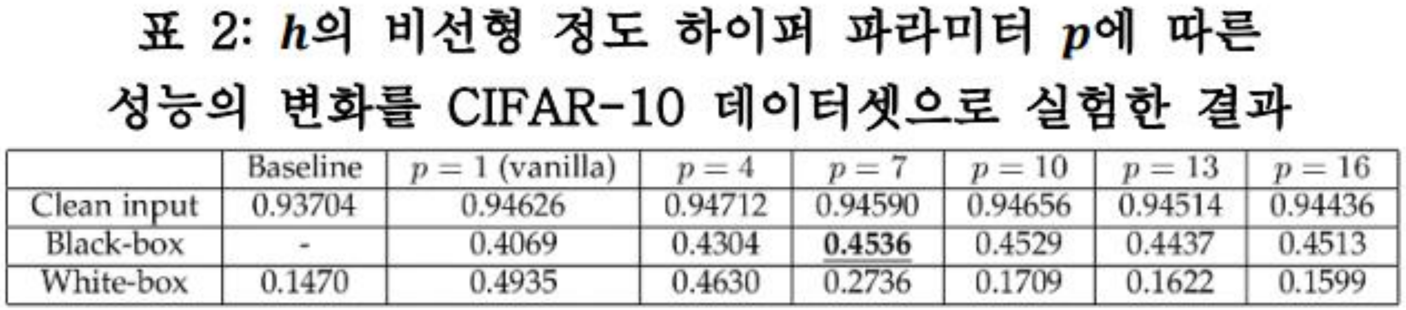


표 2: ***h***의 비선형 정도 하이퍼 파라미터 ***p***에 따른 성능의 변화를 CIFAR-10 데이터셋으로 실험한 결과

	Baseline	$p = 1$ (vanilla)	$p = 4$	$p = 7$	$p = 10$	$p = 13$	$p = 16$
Clean input	0.93704	0.94626	0.94712	0.94590	0.94656	0.94514	0.94436
Black-box	-	0.4069	0.4304	**0.4536**	0.4529	0.4437	0.4513
White-box	0.1470	0.4935	0.4630	0.2736	0.1709	0.1622	0.1599

4. Conclusion

- The proposed algorithm adds only the scalar operation h; it has little cost
- The proposed algorithm can be used generally unlike existing defense methods
- **This study adds non-linearity to a widely used mixing augmentation method to improve vulnerability to black-box attacks while maintaining cost and accuracy**
- The verification of the decision boundary formed when using the proposed algorithm is left for future research

What to improve

- Lack of quantitative comparison with existing studies
- Leap from the question to the solution (the proposed algorithm)
- Good metric?

Topics for winter

- The first is a comparison of samples that are misclassified by the robust models and those that are not
- The two models seem to be learning completely different things
- Then it seems that the samples where the twos are mostly wrong can be qualitatively different

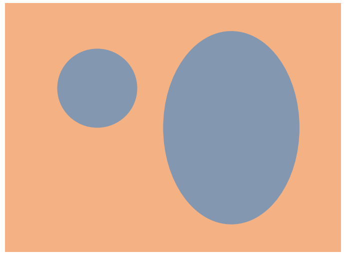
Topics for winter

- The second is a determining if an image has been trained in a generative model (Mr. H.S. talked about)
- I simplify this to determining in a classification task
- It is a difficult problem even under such simple condition

Small experiments

- Adversarial attack perspective, let’s see the distance to the D.B.
- Adversarial attack is about crossing decision boundaries
- Approximate the distance to the decision boundary by running a binary search with the attack

- In fact, the average of the perturbation required by the valid set of images was stronger than that of the train set
- But the variance was large that it was not significant

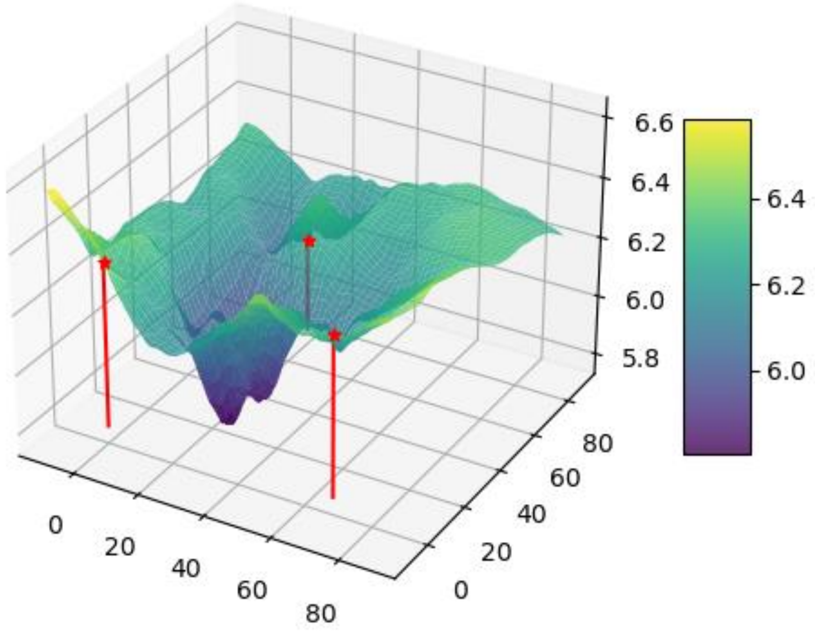


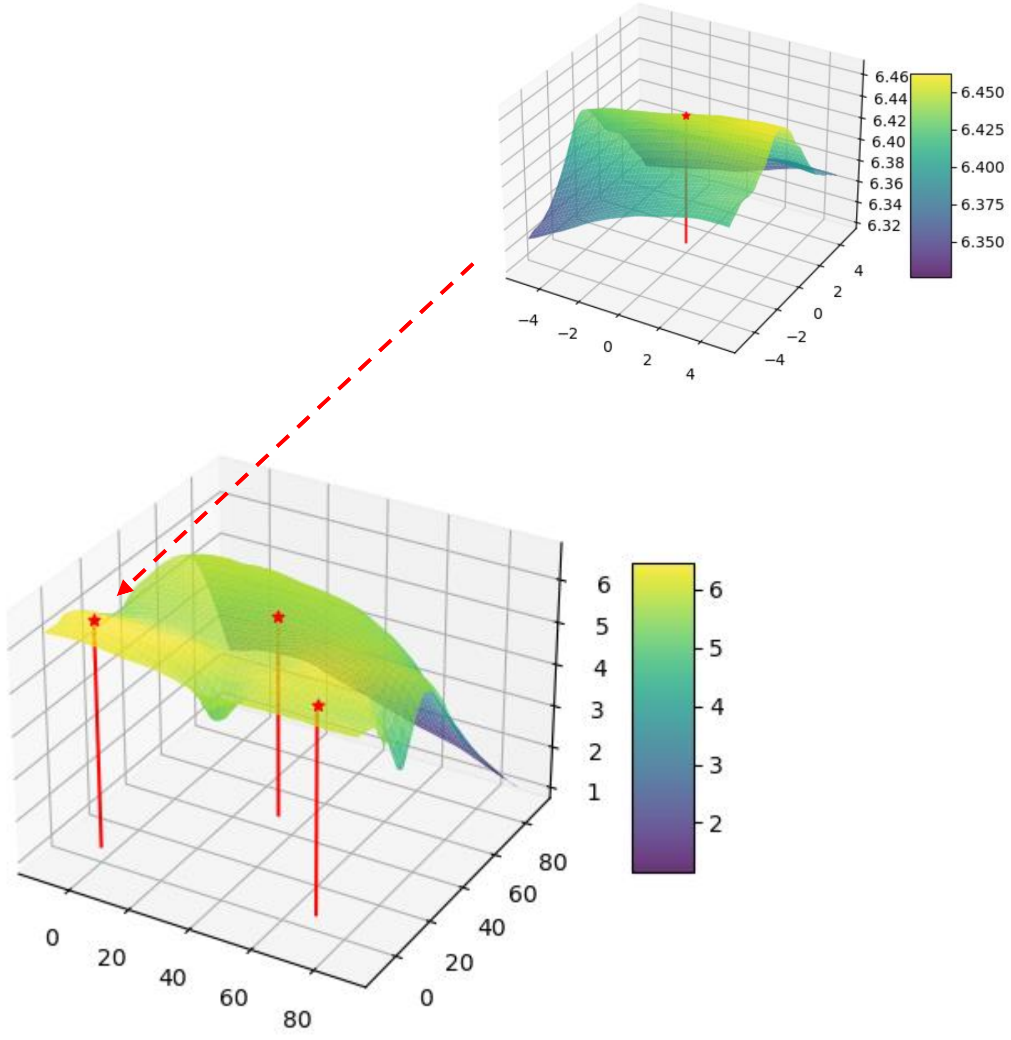
Small experiments

- Decision boundaries fail… What about logit itself?
- I think that data points already learned in the logit plane form a mode (in the context of overfitting)

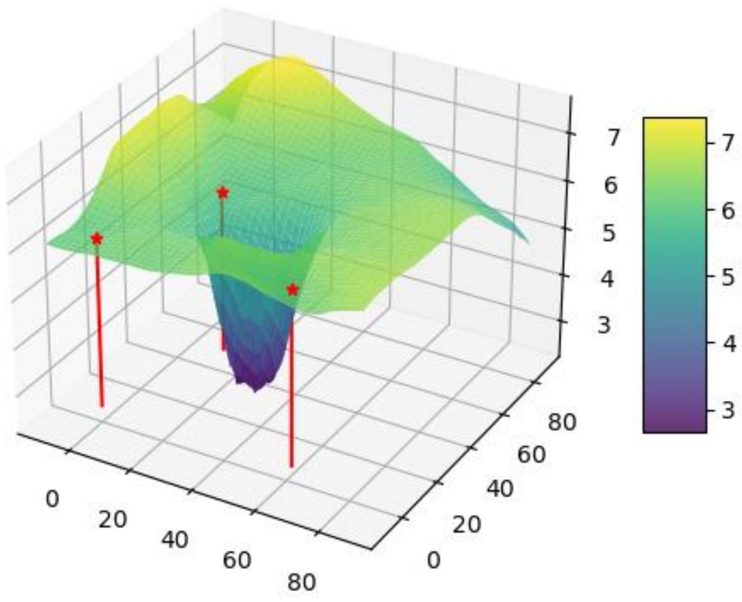
- Think of images as points (i.e., vectors), not tensors
- Then the three images will form a plane in data space
- Observe the logit plane for three images of the same class

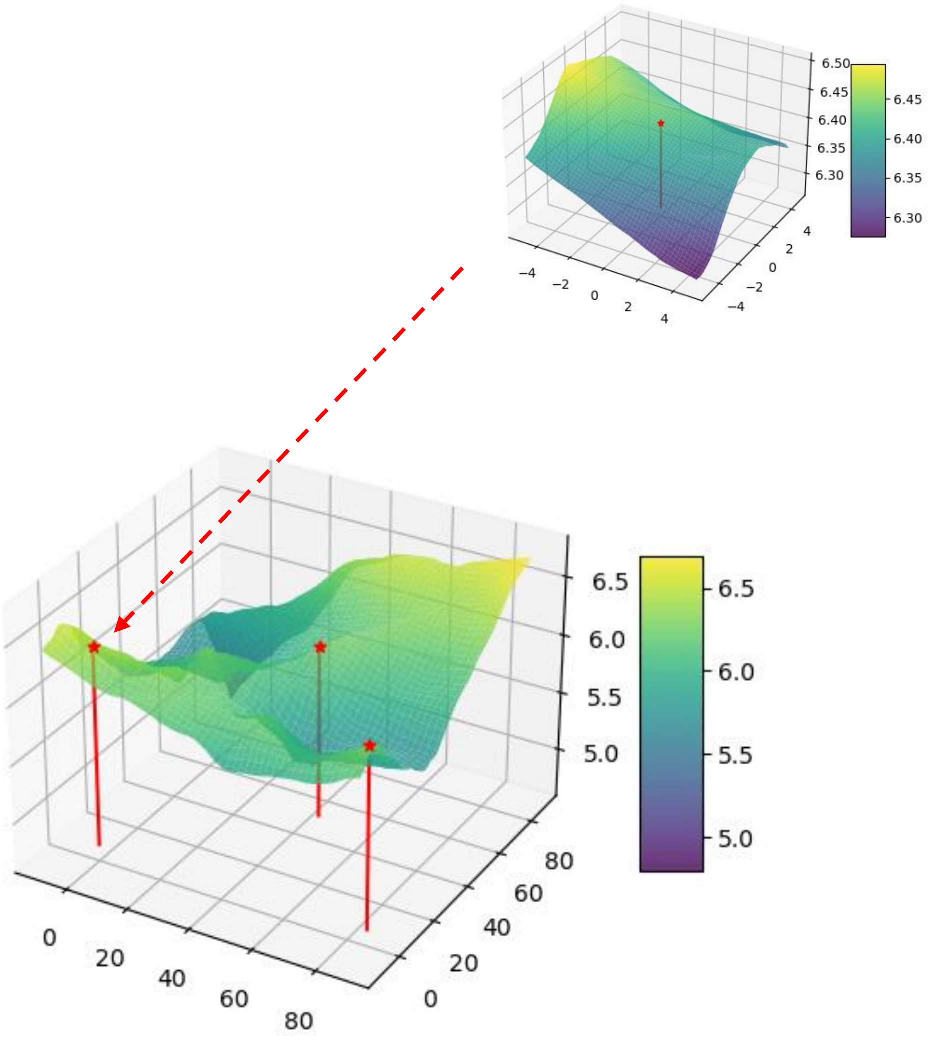
Small experiments





Small experiments





Thank you for listening

Kim Seung Hwan (overnap@khu.ac.kr)

2022. 12. 30.