The Hong Kong Polytechnic University
香港理工大學

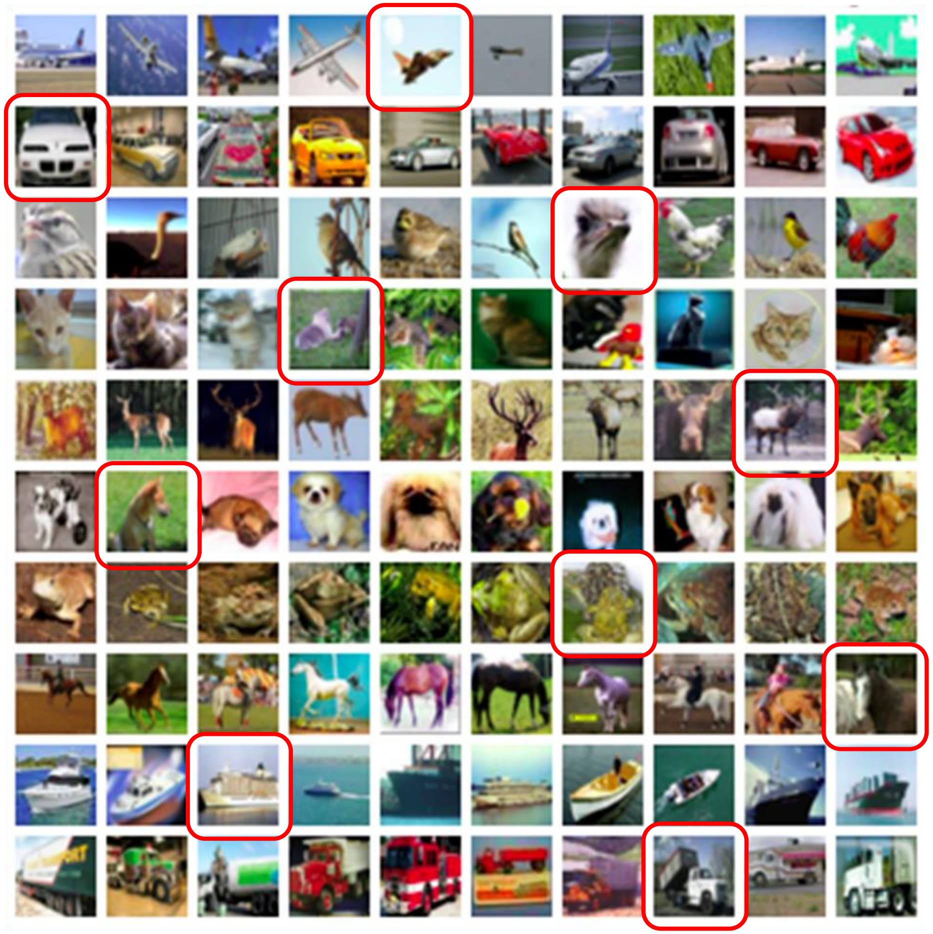
香港城市大學
City University of Hong Kong

Generating Adversarial Examples by Adversarial Networks for Semi-Supervised Learning

Yun Ma*, Xudong Mao*, Yangbin Chen, and Qing Li

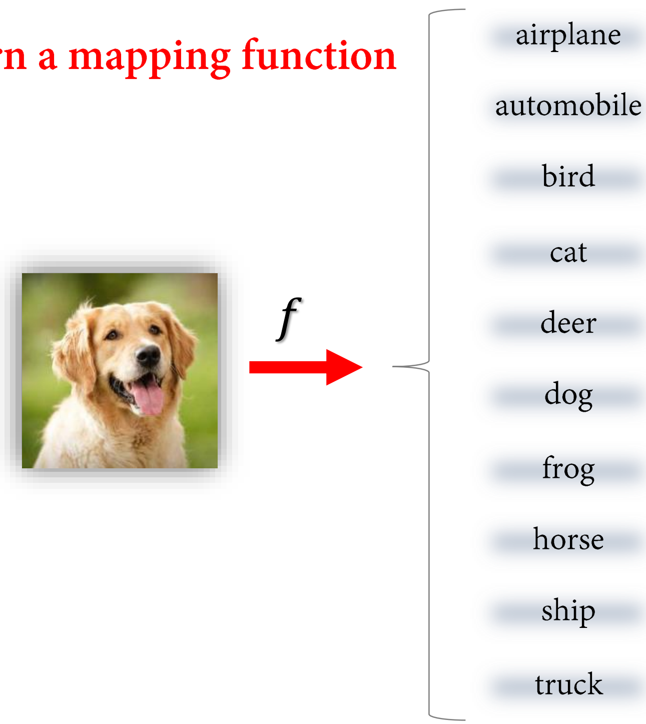
WISE 2019

Semi-Supervised Learning



limited labeled data
+ a large amount of unlabeled data

Learn a mapping function



Semi-Supervised Learning

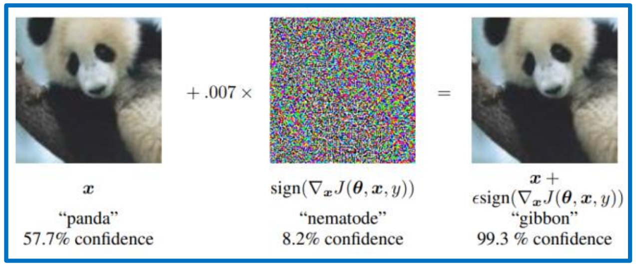
- Deep models for semi-supervised learning
  - ❑Generative models based methods
  - ❑Perturbation based methods
- Virtual Adversarial Training (VAT) [Miyato et al. TPAMI 2018]: a new perturbation based approach
  - ❑Enforce the consistency between the predictions on a sample and its adversarial variant

$$\mathcal{L}_{\text{nll}}(C) + \beta \mathcal{L}_{\text{vat}}(C)$$

$$\mathcal{L}_{\text{nll}}(C) = \mathbb{E}_{(x,y)\sim\mathcal{D}_l}\left[-\log C(y|x)\right],$$

$$\mathcal{L}_{\text{vat}}(C) = \mathbb{E}_{x\sim\mathcal{D}_l\cup\mathcal{D}_u}\left[D[C(\cdot|x), C(\cdot|x + r_{\text{vadv}})]\right],$$

$$\boxed{r_{\text{vadv}} = \underset{\|r\|_2\leq\epsilon}{\arg\max}\, D\left[C(\cdot|x), C(\cdot|x + r)\right]}$$

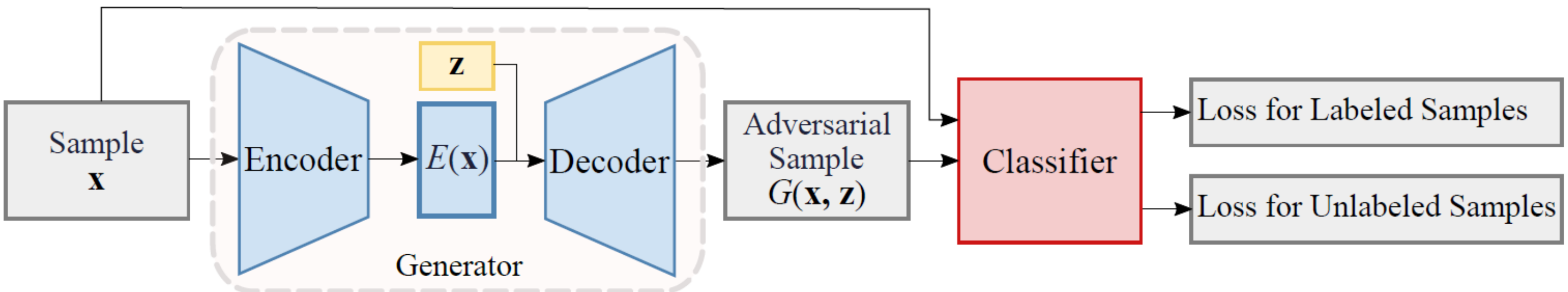


Adversarial Examples
[Goodfellow et al. ICLR 2015]

Our Solution

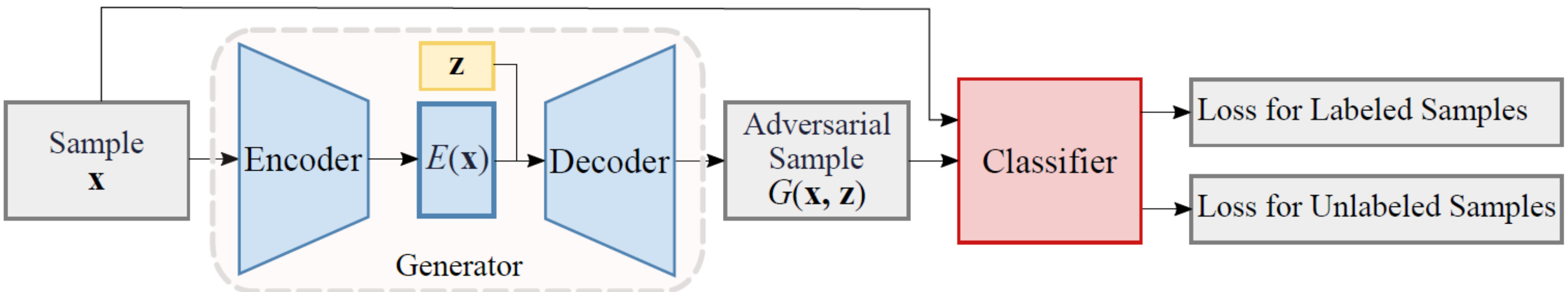
- Motivation
  - ❑VAT only considers adversarial samples with pixel-wise perturbations
  - ❑Other types of adversarial samples can also be useful
- Contribution
  - ❑Propose to adversarially train a classifier and an adversarial sample generator
  - ❑Design an encoder-decoder architecture generator to create adversarial samples from the latent space

Generating Adversarial Examples by Adversarial Networks



- **Generator:** Aims to generate adversarial examples to fool the classifier
- **Classifier:** Aims to classify the original samples and the adversarial examples consistently

Generating Adversarial Examples by Adversarial Networks



Adversarial Loss

$$\min_C \max_G \mathcal{L}_{\text{adv}}(G, C) = \mathbb{E}_{x \sim \mathcal{D}_l \cup \mathcal{D}_u} \left[D[C(\cdot|x), C(\cdot|G(x,z))] \right]$$

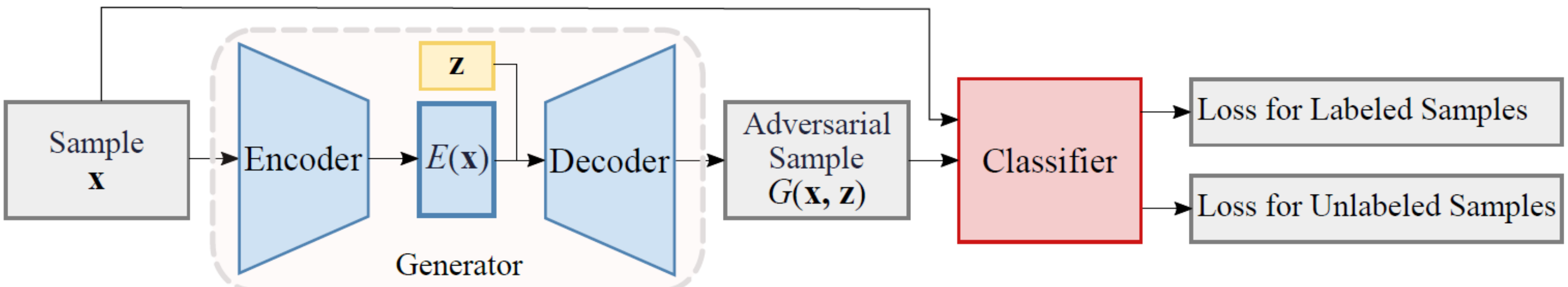
Reconstruction Loss

$$\min_G \mathcal{L}_{\text{reconst}}(G) = \mathbb{E}_{x \sim \mathcal{D}_l \cup \mathcal{D}_u} [\|x - G(x,z)\|_2^2]$$

Full Loss

$$\min_C \max_G \mathcal{L}(G, C) = \mathcal{L}_{\text{nll}}(C) + \alpha \mathcal{L}_{\text{adv}}(G, C) - \lambda \mathcal{L}_{\text{reconst}}(G)$$

Generating Adversarial Examples by Adversarial Networks



Latent Space based Adversarial Example Generation: Go beyond pixel-wise adversarial perturbations

- To generate adversarial samples semantically close to the original samples
- Implement by **differentiating the responsibilities** of the encoder and the decoder

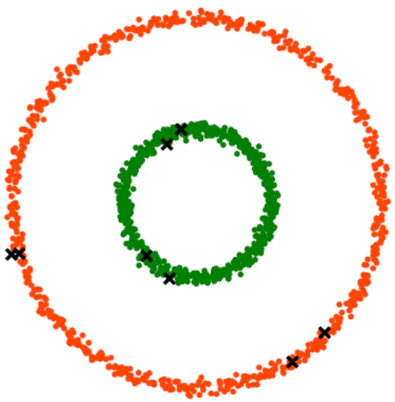
encoder decoder

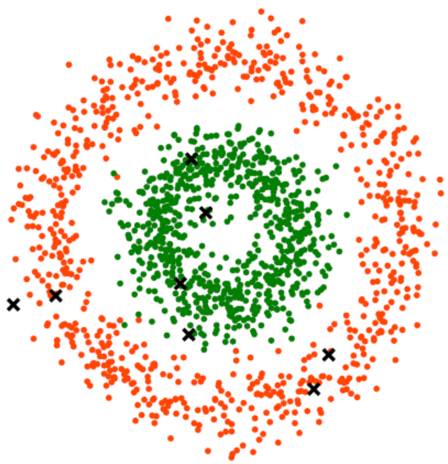
$$\min_{C} \max_{G} \mathcal{L}(G, C) = \mathcal{L}_{\text{nll}}(C) + \boxed{\alpha \mathcal{L}_{\text{adv}}(G, C) \boxed{- \lambda \mathcal{L}_{\text{reconst}}(G)}}$$

Experiments

Case Study on Synthetic Data

Semi-Supervised Learning on Two Circles





Circle (0.01)

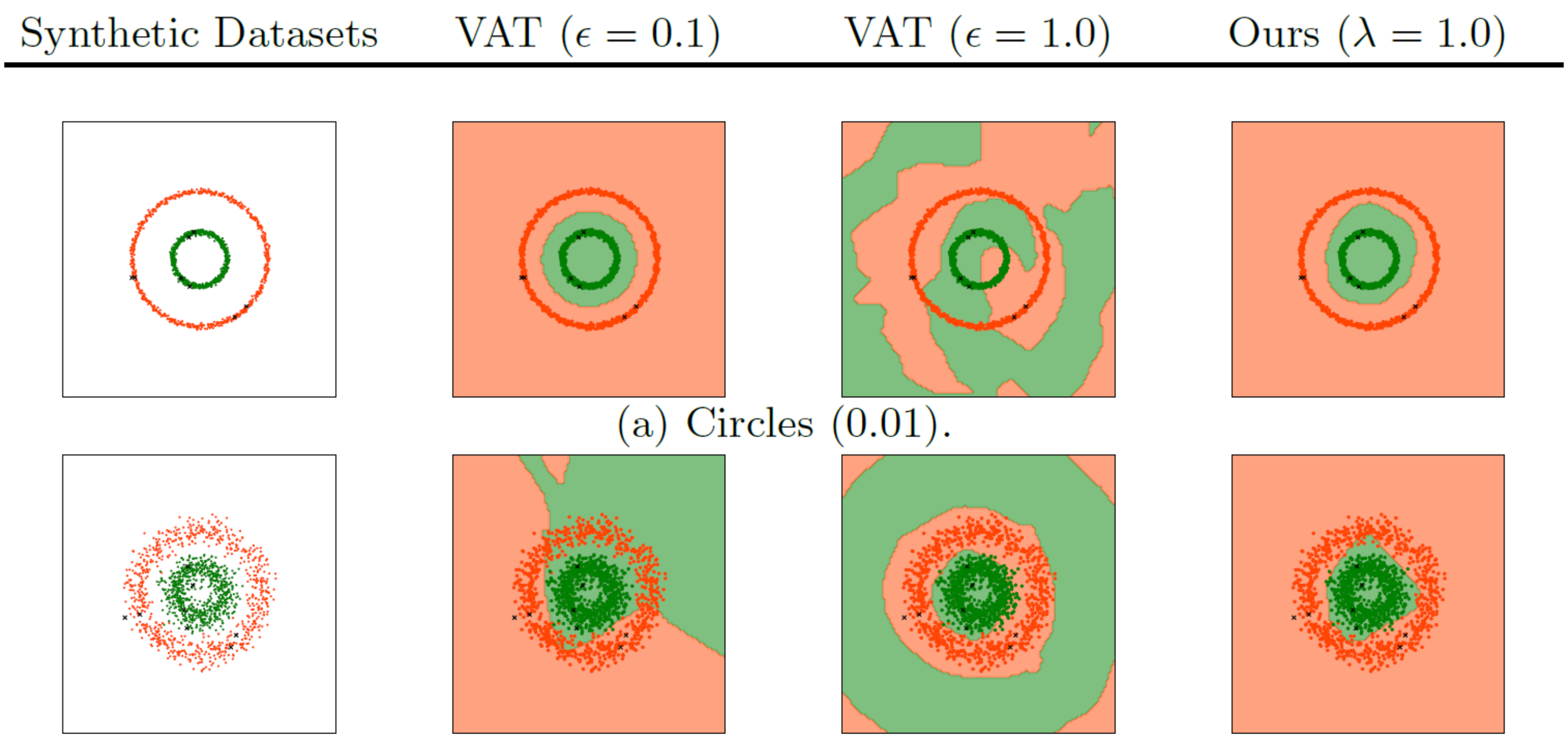
Circle (0.06)

Two classes
8 labeled data points
1500 unlabeled data points

MLP 2-100-50-2 as classifier

Models	Test error rate(%) Circles (0.01)	Circles (0.06)
VAT ($\epsilon = 0.1$)	**0.00** ($\pm$0.00)	24.61 ($\pm$4.58)
VAT ($\epsilon = 1.0$)	4.59 ($\pm$5.64)	5.10 ($\pm$4.28)
Ours ($\lambda = 1.0$)	**0.00** ($\pm$0.00)	**4.75** ($\pm$5.31)

Semi-Supervised Learning on Two Circles



(b) Circles (0.06).

- ➢ Our model shows better robustness in terms of hyper-parameters;
- ➢ The decision boundary from our model shows better reasonableness.

Experiments

Image Classification on MNIST, SVHN, and CIFAR-10

Semi-Supervised Learning on MNIST, SVHN, and CIFAR-10

Models	Test error rate(%) MNIST	SVHN	CIFAR-10
TSVM [6]	5.38	-	-
Pseudo Ensembles Aggreement [3]	2.87	-	-
Deep Generative Model [15]	2.40 ($\pm$0.02)	-	-
Ladder Networks [23]	**0.84** ($\pm$0.08)	-	20.4 ($\pm$0.47)
CatGAN [26]	1.73 ($\pm$0.18)	-	19.58 ($\pm$0.58)
ALI [8]	-	7.42 ($\pm$0.65)	17.99 ($\pm$1.62)
Improved GAN [25]	-	8.11 ($\pm$1.3)	18.63 ($\pm$2.32)
Triple GAN [18]	-	5.77 ($\pm$0.17)	16.99 ($\pm$0.36)
Π model [17]	-	5.43 ($\pm$0.25)	16.55 ($\pm$0.29)
FM-GAN+Jacob.-reg+Tangents [16]	-	4.39 ($\pm$1.2)	16.20 ($\pm$1.6)
GoodSSLwithBadGAN [7]	-	4.25 ($\pm$0.03)	14.41 ($\pm$0.03)
VAT [20]	1.27 ($\pm$0.11)	4.28 ($\pm$0.10)	13.15 ($\pm$0.21)
Our Model	1.17 ($\pm$0.10)	**3.93** ($\pm$0.07)	**12.97** ($\pm$0.10)

MNIST: 60000 samples, 1000 samples are labeled
SVHN: 73257 samples, 1000 samples are labeled
CIFAR-10: 50000 samples, 4000 samples are labeled

Supervised Learning on MNIST, SVHN, and CIFAR-10

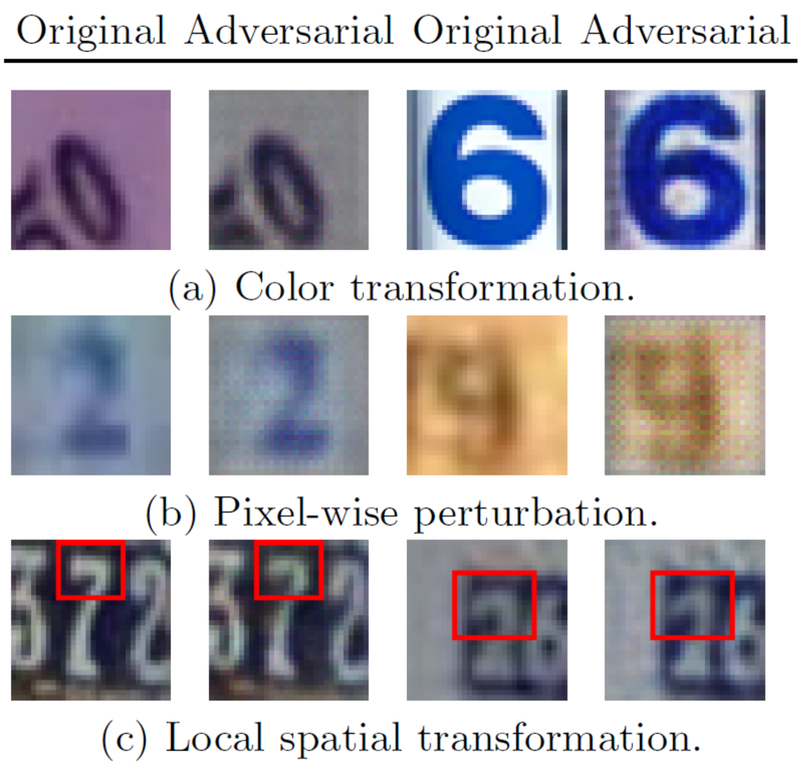
Models	Test error rate(%)		
	MNIST	SVHN	CIFAR-10
Superivised-only	1.09 ($\pm$0.02)	2.79 ($\pm$0.08)	6.58 ($\pm$0.10)
Ladder Networks [23]	**0.57** ($\pm$0.02)	-	-
Π model [17]	-	2.54 ($\pm$0.04)	5.56 ($\pm$0.10)
Temporal Ensembling [17]	-	2.74 ($\pm$0.06)	5.60 ($\pm$0.10)
Adversarial Training [11]	0.78	-	-
RPT [20]	0.84 ($\pm$0.03)	-	6.30 ($\pm$0.04)
VAT [20]	0.64 ($\pm$0.05)	-	5.81 ($\pm$0.02)
Our Model	0.61 ($\pm$0.04)	**2.49** ($\pm$0.06)	**5.51** ($\pm$0.02)

MNIST: 60000 samples
SVHN: 73257 samples
CIFAR-10: 50000 samples

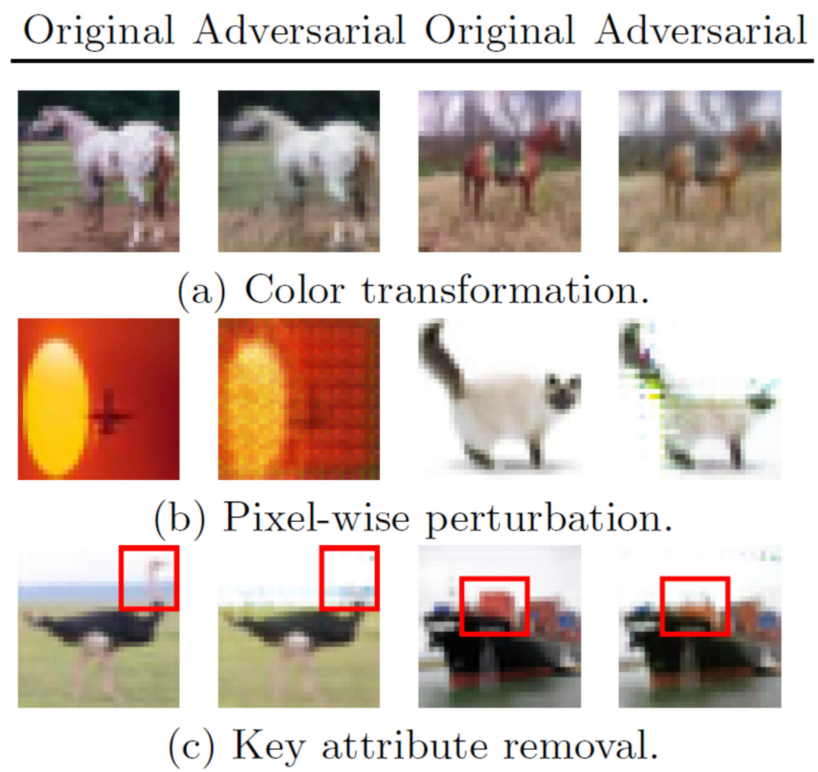
Experiments

Visualization of Adversarial Samples

Visualization of Adversarial Samples

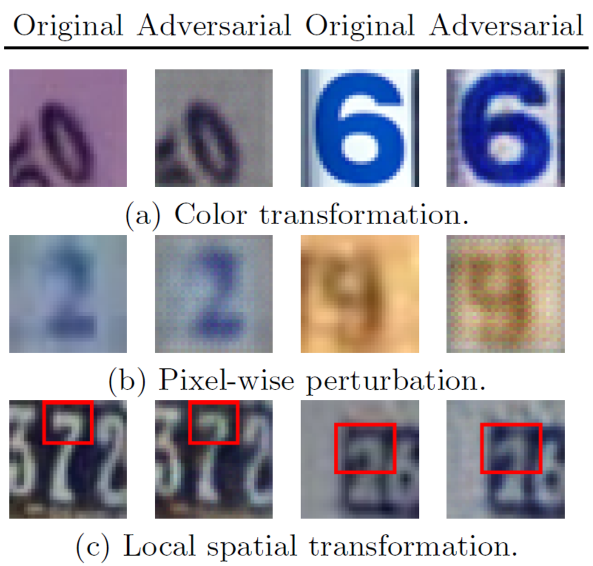


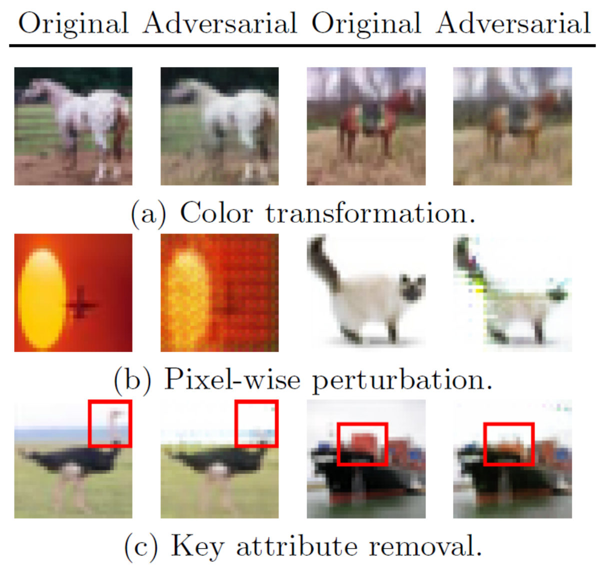
SVHN



CIFAR-10

Visualization of Adversarial Samples





Adversarial Samples

VAT [Miyato et al. TPAMI 2018]

Our adversarial example types are more diverse than those in VAT

implies better regularization and explains our advantage over VAT in semi-supervised and supervised learning.

Conclusion

- An semi-supervised learning framework regularizing the classifier with generated adversarial samples
  - ❑Adversarially training
  - ❑Latent space based adversarial sample generation
  - ❑Better regularization power with more types of adversarial samples
- Future work
  - ❑Integrating our framework with other GAN-based models to further enhance model robustness
  - ❑Generalize the framework to other domains

Thank you !