ConvEBMDefense for Medical Images Analysis

Yuan Du

Committee Members:
Dr. Mitchell Hill (Chair)
Dr. Shunpu Zhang (Vice Chair)
Dr. Hsin-Hsiung Huang
Dr. Yogesh Singh Rawat

Department of Statistics and Data Science
Big Data Analytics

Table of Contents

Vulnerability of Deep Learning Model

1. Imperceptible adversarial attacks can fool Deep Convolutional Neural Networks with high confidence.

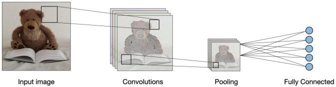


Figure 1: The architecture of CNN, Standford CS 230

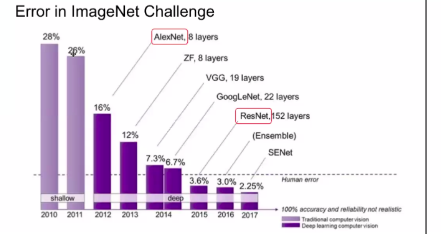


Figure 2: (source: Angshuman Gosh|DLDC 2021)

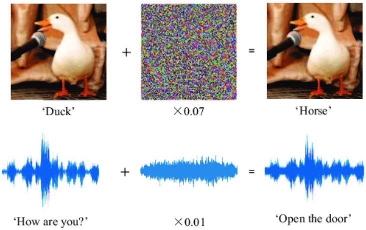


Figure 3: Adversarial examples of image and audio (Gong et al., 2018)

2. Adversarial examples could also attack physical world in 2D and 3D settings.

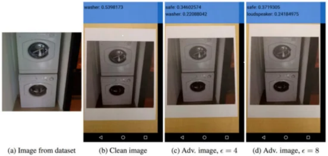


Figure 4: Adversarial examples 2D print (Kurakin et al., 2018)

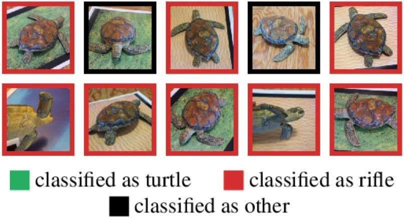


Figure 5: Adversarial examples 3D print (Athalye et al., 2018b)

Adversarial Examples in Healthcare

The United States spent approximately $3.3 trillion (17.8% of GDP) on healthcare in 2016. One study estimated medical fraud to be as high as $272 billion in 2011 (Finlayson et al., 2018).

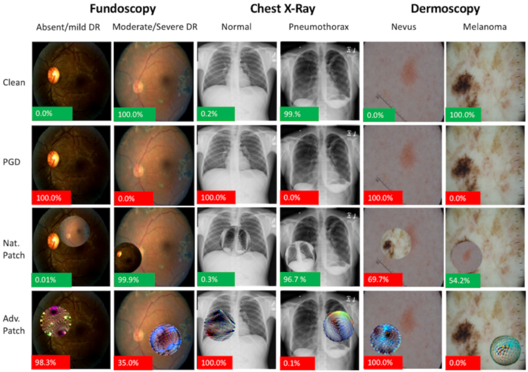


Figure 6: Adversarial examples on medical images (Finlayson et al., 2018)

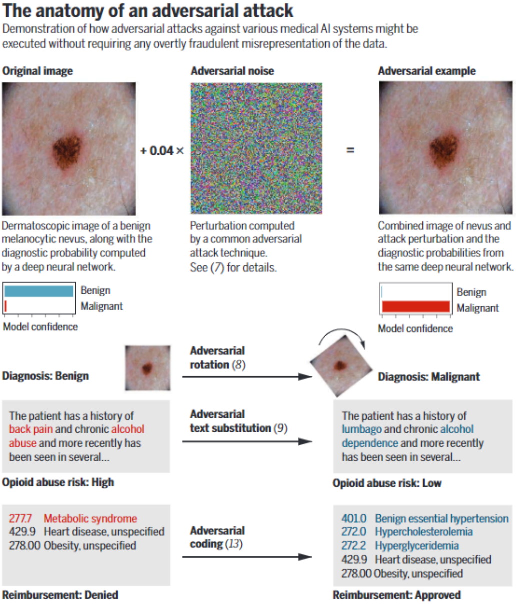


Figure 7: Adversarial examples on medical image, text and coding (Finlayson et al., 2019)

Table of Contents

Definition of Adversarial Attack

1. Given a trained deep learning model f and an original input data sample x, generating an adversarial example x' can generally be described as a box-constrained optimization problem:

$$\begin{aligned} &\min_{x'} \|x - x'\|, \\ \text{s.t. } & f(x') = c', \\ & f(x) = c, \\ & c' \neq c, \\ & x \in [0, 1] \end{aligned} \tag{1}$$

   The distance $d\|.\|$ between $x - x'$ denotes the perturbation added on natural image x.

2. The goal of attack is to fool the model and thus misclassify the labels.

Types of Adversarial Attack

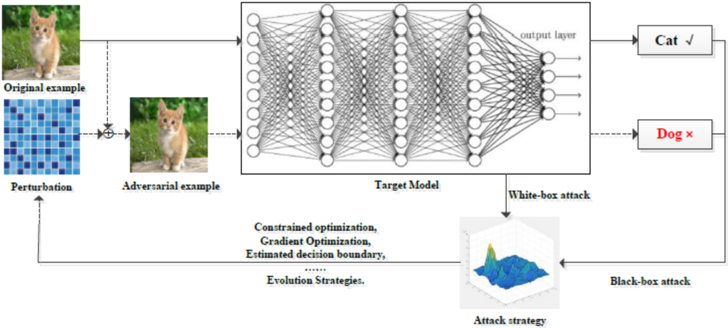


Figure 8: Adversarial example generation and adversarial attack process (Hongshuo Liang et al., 2022)

1. White-box attack: The attacker has complete knowledge of the target model including model training process and weights. It's a stronger attack than black-box attack.
2. Black-box attack: The attacker assumes no knowledge of the target model. One category of black-box attacks allows probing the deployed target models with queries. This setup is more commonly known as query-based attack.

An untargeted white-box intends to increase the loss function with bounded perturbation distance ϵ to generate adversarial examples x':

$$\underset{\delta \in \Delta}{\operatorname{argmax}} \, L(f_\theta(x+\delta), y) \tag{2}$$

where Δ is the ϵ-ball in the l_p-norm.

The common option of perturbation distance are l_∞-norm ϵ ball and l_2-norm ϵ ball around x, where $\epsilon > 0$.

Let θ is the parameters of a model, $L(\theta, x, y)$ be the cost used to train the neural network.

- First generation attack - Fast Gradient Sign Method (FGSM) (Ian J Goodfellow et al., 2014) :

$$x + \epsilon \, \mathrm{sign}(\nabla_x L(\theta, x, y)) \tag{3}$$

- Adaptive attack - Projected Gradient Descent (PGD) (Madry et al., 2017) on the negative loss function:

$$x^{t+1} = \mathrm{Proj}_{x+S}(x^t + \alpha \, \mathrm{sign}(\nabla_{x^t} L(\theta, x^t, y))) \tag{4}$$

More recent stronger adaptive attacks

- Expectation Over Transformation (EOT) aims to constrain the expected effective distance between the adversarial $t(x')$ and original inputs $t(x)$ instead of $x' - x$. PGD is used to iteratively generate the adversarial example by updating the gradient:

$$\nabla_x E_{T(x)}[f(T(x))] = E_{T(x)}[\nabla_x f(T(x))] \tag{5}$$

- Backward Pass Differentiable Approximation (BPDA) can be applied on non-differential network where gradients are not readily available:

$$\nabla_x f(g(x))|_{x=\hat{x}} = \nabla_x f(x)|_{x=g(\hat{x})} \tag{6}$$

where $g(\cdot)$ is neither smooth nor differentiable and can't be backpropagated through to generate adversarial examples.

Table of Contents

Types of Adversarial Defense

- Adversarial Purification (AP): a process that remove/purify adversarial examples before the model training process for adversarial defense.
- Adversarial Training (AT): a process that injects adversarial examples in the training data of a model to make it adversarially robust.

We use Energy based Model(EBM) adversarial purification for defense in this work.

Current Adversarial Defense

Defense evaluation problems: Obfuscated Gradients

- Shattered gradient.
- Stochastic gradient.
- Exploding & Vanishing Gradients.

Defense	Dataset	Distance	Accuracy
Buckman et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%*
Ma et al. (2018)	CIFAR	0.031 (ℓ_∞)	5%
Guo et al. (2018)	ImageNet	0.005 (ℓ_2)	0%*
Dhillon et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%
Xie et al. (2018)	ImageNet	0.031 (ℓ_∞)	0%*
Song et al. (2018)	CIFAR	0.031 (ℓ_∞)	9%*
Samangouei et al. (2018)	MNIST	0.005 (ℓ_2)	55%**
Madry et al. (2018)	CIFAR	0.031 (ℓ_∞)	47%
Na et al. (2018)	CIFAR	0.015 (ℓ_∞)	15%

Table 1. **Summary of Results:** Seven of nine defense techniques accepted at ICLR 2018 cause obfuscated gradients and are vulnerable to our attacks. Defenses denoted with * propose combining adversarial training; we report here the defense alone, see §5 for full numbers. The fundamental principle behind the defense denoted with ** has 0% accuracy; in practice, imperfections cause the theoretically optimal attack to fail, see §5.4.2 for details.

Figure 9: Athalye et al., 2018a

Defense evaluation on medical images is inadequate:

- Papers uses attack methods such as FGSM, BIM, PGD and no paper uses stronger attacks like EOT, BPDA.
- Treat model was not clearly defined.
- Attack statement on l_p-norm, iteration steps, number of evaluated images are not clear or unavailable.

Table of Contents

Modern Deep EBM

EBM (J. Xie et al., 2016) is a Gibbs-Boltzmann density.
A deep EBM has the form:

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\{-U(x;\theta)\} \quad (7)$$

where $x \in R^D$ is an image signal. The energy $U(x;\theta)$ is a ConvNet with weights θ, a scalar output.
Z is intractable normalizing constant:

$$Z(\theta) = \int_{\mathcal{X}} \exp\left[-U(x;\theta)\right] dx \quad (8)$$

In order to find θ such that the parametric model $p_\theta(x)$ is a close approximation of the data distribution $q(x)$. Kullback-Leibler (KL) divergence was used to measure the closeness by solving $\operatorname{argmin}_\theta L(\theta)$:

$$\begin{aligned} &\operatorname*{argmin}_\theta D_{KL}(q(x)||p(x;\theta)) \\ &= \operatorname*{argmin}_\theta E_q[\log \frac{q}{p_\theta}] \end{aligned} \quad (9)$$

EBM (J. Xie et al., 2016) is a Gibbs-Boltzmann density.
A deep EBM has the form:

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\{-U(x;\theta)\} \quad (7)$$

where $x \in R^D$ is an image signal. The energy $U(x;\theta)$ is a ConvNet with weights θ, a scalar output.
Z is intractable normalizing constant:

$$Z(\theta) = \int_{\mathcal{X}} \exp\left[-U(x;\theta)\right] dx \quad (8)$$

In order to find θ such that the parametric model $p_\theta(x)$ is a close approximation of the data distribution $q(x)$. Kullback-Leibler (KL) divergence was used to measure the closeness by solving $\operatorname{argmin}_\theta L(\theta)$:

$$\begin{aligned} &\underset{\theta}{\operatorname{argmin}}\, D_{KL}(q(x)||p(x;\theta)) \\ &= \underset{\theta}{\operatorname{argmin}}\, E_q[\log \frac{q}{p_\theta}] \end{aligned} \quad (9)$$

Main ways to learn probabilistic models:

- $\boxed{\text{MLE learning}}$
- Variational approximation
- Normalizing flow

Maximum Likelihood Estimation

Objective function of MLE learning:

$$\mathcal{L}(\theta) = E_q[-\log p(x;\theta)] \tag{10}$$

The derivative of the loss is:

$$\nabla\mathcal{L}(\theta) = \boxed{\nabla \log z(\theta)} + \nabla E_q[U(X;\theta)] \tag{11}$$

where the $\nabla \log z(\theta)$ can be expressed as:

$$\begin{aligned}
\nabla \log z(\theta) &= \frac{1}{z(\theta)} \nabla U(\theta) \\
&= \frac{1}{z(\theta)} \nabla \int \exp\left[-U(x;\theta)\right] dx \\
&= \frac{1}{z(\theta)} \int \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\
&= \int \frac{1}{z(\theta)} \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\
&= \int p_\theta \nabla[-U(x;\theta)]\, dx \\
&= -E_{p_\theta}[\nabla U(x;\theta)]
\end{aligned} \tag{12}$$

Objective function of MLE learning:

$$\mathcal{L}(\theta) = E_q[-\log p(x;\theta)] \tag{10}$$

The derivative of the loss is:

$$\nabla \mathcal{L}(\theta) = \boxed{\nabla \log z(\theta)} + \nabla E_q[U(X;\theta)] \tag{11}$$

where the $\nabla \log z(\theta)$ can be expressed as:

$$\begin{aligned}
\nabla \log z(\theta) &= \frac{1}{z(\theta)} \nabla U(\theta) \\
&= \frac{1}{z(\theta)} \nabla \int \exp\left[-U(x;\theta)\right] dx \\
&= \frac{1}{z(\theta)} \int \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\
&= \int \frac{1}{z(\theta)} \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\
&= \int p_\theta \nabla[-U(x;\theta)]\, dx \\
&= -E_{p_\theta}[\nabla U(x;\theta)]
\end{aligned} \tag{12}$$

Thus, the gradient used to learn θ becomes:

$$\begin{aligned}
\nabla \mathcal{L}(\theta) &= \nabla E_q[U(X;\theta)] - \nabla E_{p_\theta}[U(X;\theta)] \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta U(X_i^+;\theta) - \underbrace{\frac{1}{m} \sum_{i=1}^{m} \nabla_\theta U(X_i^-;\theta)}_{\text{MCMC sampling}}
\end{aligned} \tag{13}$$

Objective function of MLE learning:

$$\mathcal{L}(\theta) = E_q[-\log p(x;\theta)] \tag{10}$$

The derivative of the loss is:

$$\nabla \mathcal{L}(\theta) = \boxed{\nabla \log z(\theta)} + \nabla E_q[U(X;\theta)] \tag{11}$$

where the $\nabla \log z(\theta)$ can be expressed as:

$$\begin{aligned}\nabla \log z(\theta) &= \frac{1}{z(\theta)} \nabla U(\theta) \\ &= \frac{1}{z(\theta)} \nabla \int \exp\left[-U(x;\theta)\right] dx \\ &= \frac{1}{z(\theta)} \int \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\ &= \int \frac{1}{z(\theta)} \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\ &= \int p_\theta \nabla[-U(x;\theta)]\, dx \\ &= -E_{p_\theta}[\nabla U(x;\theta)]\end{aligned} \tag{12}$$

Thus, the gradient used to learn θ becomes:

$$\begin{aligned}\nabla \mathcal{L}(\theta) &= \nabla E_q[U(X;\theta)] - \nabla E_{p_\theta}[U(X;\theta)] \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta U(X_i^+;\theta) - \underbrace{\frac{1}{m}\sum_{i=1}^{m} \nabla_\theta U(X_i^-;\theta)}_{\text{MCMC sampling}}\end{aligned} \tag{13}$$

Gradient-based MCMC and Langivin Dynamics:

$$X^{(k+1)} = X^{(k)} - \frac{\epsilon^2}{2} \nabla_{X^{(k)}} U(X^{(k)};\theta) + \epsilon Z_k, \tag{14}$$

where ϵ is the step size and $Z_k \sim N(0, I^D)$. The Langevin trajectories are initialized from a set of states $\{X_{i,0}^-\}_{i=1}^n$ obtained from a certain initialization strategy.

Objective function of MLE learning:

$$\mathcal{L}(\theta) = E_q[-\log p(x;\theta)] \tag{10}$$

The derivative of the loss is:

$$\nabla \mathcal{L}(\theta) = \boxed{\nabla \log z(\theta)} + \nabla E_q[U(X;\theta)] \tag{11}$$

where the $\nabla \log z(\theta)$ can be expressed as:

$$\begin{aligned}
\nabla \log z(\theta) &= \frac{1}{z(\theta)} \nabla U(\theta) \\
&= \frac{1}{z(\theta)} \nabla \int \exp\left[-U(x;\theta)\right] dx \\
&= \frac{1}{z(\theta)} \int \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\
&= \int \frac{1}{z(\theta)} \exp\left[-U(x;\theta)\right] \nabla[-U(x;\theta)]\, dx \\
&= \int p_\theta \nabla[-U(x;\theta)]\, dx \\
&= -E_{p_\theta}[\nabla U(x;\theta)]
\end{aligned} \tag{12}$$

Thus, the gradient used to learn θ becomes:

$$\begin{aligned}
\nabla \mathcal{L}(\theta) &= \nabla E_q[U(X;\theta)] - \nabla E_{p_\theta}[U(X;\theta)] \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta U(X_i^+;\theta) - \underbrace{\frac{1}{m} \sum_{i=1}^{m} \nabla_\theta U(X_i^-;\theta)}_{\text{MCMC sampling}}
\end{aligned} \tag{13}$$

Gradient-based MCMC and Langivin Dynamics:

$$X^{(k+1)} = X^{(k)} - \frac{\epsilon^2}{2} \nabla_{X^{(k)}} U(X^{(k)};\theta) + \epsilon Z_k, \tag{14}$$

where ϵ is the step size and $Z_k \sim N(0, I^D)$. The Langevin trajectories are initialized from a set of states $\{X_{i,0}^-\}_{i=1}^n$ obtained from a certain initialization strategy.

Different implementations of the MCMC synthesis step:

1. Contrastive Divergence: runs a finite MCMC from data (Hinton, 2002).
2. Persistent Chain: runs a finite MCMC from the synthesized data from previous epoch.
3. Cooperative Divergence: runs a finite MCMC from a generator in tandem with the energy.
4. Non-persistent short-run MCMC: runs a finite MCMC Gaussian White noise.

Table of Contents

Why do we choose EBM for defense?

- **Simplicity and Stability:** An EBM is the only object that needs to be trained and designed. Separate networks are not tuned to ensure balance (for example, unbalanced training can result in posterior collapse in VAEs or poor performance in GANs).

- **Simplicity and Stability:** An EBM is the only object that needs to be trained and designed. Separate networks are not tuned to ensure balance (for example, unbalanced training can result in posterior collapse in VAEs or poor performance in GANs).
- **Sharing of Statistical Strength:** Since the EBM is the only trained object, it requires fewer model parameters than approaches that use multiple networks. More importantly, the model being concentrated in a single network allows the training process to develop a shared set of features as opposed to developing them redundantly in separate networks.

- **Simplicity and Stability:** An EBM is the only object that needs to be trained and designed. Separate networks are not tuned to ensure balance (for example, unbalanced training can result in posterior collapse in VAEs or poor performance in GANs).
- **Sharing of Statistical Strength:** Since the EBM is the only trained object, it requires fewer model parameters than approaches that use multiple networks. More importantly, the model being concentrated in a single network allows the training process to develop a shared set of features as opposed to developing them redundantly in separate networks.
- **Adaptive Computation Time:** An iterative stochastic optimization process, which allows for a trade-off between generation quality and computation time.

- **Simplicity and Stability:** An EBM is the only object that needs to be trained and designed. Separate networks are not tuned to ensure balance (for example, unbalanced training can result in posterior collapse in VAEs or poor performance in GANs).
- **Sharing of Statistical Strength:** Since the EBM is the only trained object, it requires fewer model parameters than approaches that use multiple networks. More importantly, the model being concentrated in a single network allows the training process to develop a shared set of features as opposed to developing them redundantly in separate networks.
- **Adaptive Computation Time:** An iterative stochastic optimization process, which allows for a trade-off between generation quality and computation time.
- **Flexibility Of Generation**: EBMs directly modeling particular regions as high or lower energy during the generation process to avoid unwanted regions of data, especially for discontinuous data manifolds, unlike VAEs or flow based models.

(Du et al., 2019)

Use EOT as Defense

Understanding EOT attack

1. If $T(x)$ is not differentiable, it will cause exploding or vanishing gradient problem.

$$\nabla_x E_{T(x)}[f(T(x))] = E_{T(x)}[\nabla_x f(T(x))] \tag{15}$$

Understanding EOT attack

1. If $T(x)$ is not differentiable, it will cause exploding or vanishing gradient problem.

$$\nabla_x E_{T(x)}[f(T(x))] = E_{T(x)}[\nabla_x f(T(x))] \tag{15}$$

2. Evaluate stochastic classifiers $f(T(x))$. Let $F(x) = E_{T(x)}[f(T(x))]$. Expectation Over Transformation (EOT) to circumvent stochastic gradient problem that's caused by random classifier (Athalye et al., 2018a).

$$\hat{F}_{H_{\text{adv}}}(x) \approx \frac{1}{H_{\text{adv}}} \sum_{h=1}^{H_{\text{adv}}} f(\hat{x}_h), \quad \hat{x}_h \sim T(x) \text{ i.i.d} \tag{16}$$

where H_{adv} is number of EOT attack samples. Typically around 10 to 30. Small H_{adv} causes random classification.

Convergent EBM Defense

Convergent EBM $T(x)$ and EOT defense. Solve exploding or vanishing gradient problem for T(x) by change of variable $x = h(z)$ where $h(\cdot)$ is differentiable. With large enough MCMC sampling steps K, we can remove the adversarial noise in langevin sampling step.

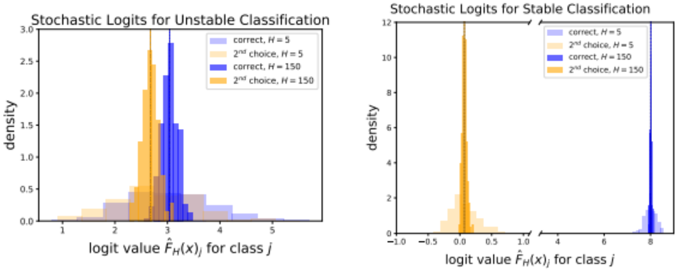


Figure 10: H impact on unstable and stable classification by EOT attack/defense (Hill et al., 2020)

$$\hat{F}_H \approx \frac{1}{H}\sum_{h=1}^{H} f(\hat{x}_h), \quad \hat{x}_h \sim T(x) \text{ i.i.d} \tag{17}$$

Visualization of ConvEBMDefense Model

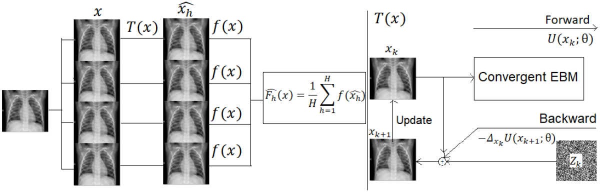


Figure 11: Convergent EBM Defense on Medical images

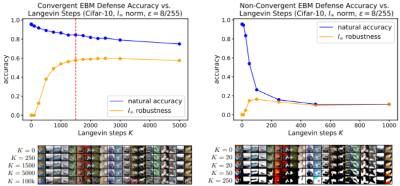


Figure 12: Convergent EBM vs Non-convergent EBM and MCMC steps K (Hill et al., 2020).
we experimented on K=1000 and 2000 on chest-xray

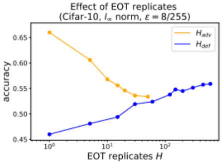


Figure 13: EOT replicates (Hill et al., 2020).
we experimented on H_{def}=64 and 128 on chest-xray with H_{adv}=24

Table of Contents

Dataset and Model Pre-training

We use WideResNet as classifier and have a binary classification accuracy of 92.5%.

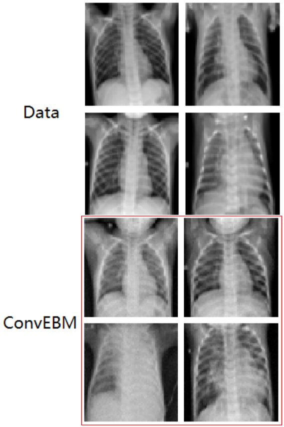


Figure 14: Original images & images Generated by ConvEBM

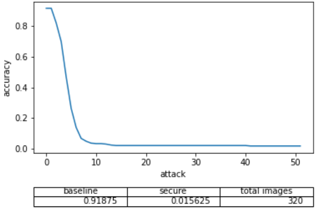


Figure 15: Accuracy over BPDA+EOT24 attack without defnese

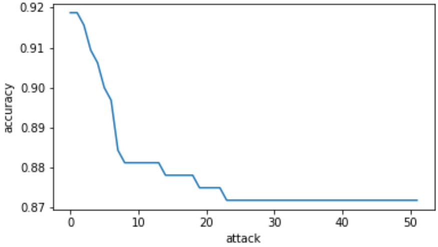
Chest-xray (D. Kermany et al., 2018)	
Train	Test
5,232	624

Table 1: In the training set: 3883 images characterized as depicting pneumonia (2,538 bacterial and 1,345 viral) and 1,349 normal.

Defense Model Training

We use BPDA + EOT attack, which is known as the strongest adaptive attack for the defense evaluation:

$$\Delta_{EOT+BPDA}(x, y) = \frac{1}{H_{\text{adv}}} \sum_{h=1}^{H_{\text{adv}}} \nabla_{\hat{x}_h} L \left(\frac{1}{H_{\text{adv}}} \sum_{h=1}^{H_{\text{adv}}} f(\hat{x}_h), y \right), \quad \hat{x}_h \sim T(x)\text{i.i.d} \tag{18}$$



baseline	secure	total images
0.91875	0.871875	320

Figure 16: Number of adversarial attack steps on Chest x-ray, BPDA + EOT24, K=2000, $H_{\text{def}} = 128$

Experiment Comparison

BPDA+EOT24 attack reduced the robust accuracy to 0.016 without defense.
Our ConvEBMDefense model achieved 86.8% accuracy when bounded by l_∞ distortion with $\epsilon = 0.031$ (8/255) on 320 images, when the attacker has full white-box access.

Dataset	Attack	Defense	Nat	Adv	ϵ	Adv steps	H_{def}	K	samples
Chest-xray	BPDA+EOT24	Ours	0.923($\pm$ 0.003)	0.872($\pm$ 0.007)	8/255	30|50	64|128	2000	320
Chest-xray	BPDA+EOT24	Ours	0.922	0.858	8/255	30	64	1000	320
Chest-xray	BPDA	Ours	0.917($\pm$ 0.011)	0.871($\pm$ 0.002)	8/255	30	64|128	2000	320
Chest-xray	PGD(l_∞)	AT	0.925	0.89	8/255	5|25	NA	NA	NA
Chest-xray14	BIM(l_∞)	Model[1]	0.74	0.650	0.3[?]	5	NA	NA	200
Chest-xray14	PGD	Model[2]	0.862	0.772	-	-	NA	NA	-
Chest-xray14	PGD(l_∞)	AT	0.865	0.839	4/255	4	NA	NA	NA|-

Table 2: Defense for l_∞ against high-power whitebox attacks on Chest Xray. Our robust accuracy with BPDA + EOT attack is averaged at **0.868**. and the natural accuracy preserved the accuracy of pre-trained WideResNet model. None of the evaluated model preserved the accuracy of pre attack or had convincing defense accuracy result. Model[1] (Taghanaki et al., 2019),Model[2] (L. Chen et al., 2021), and AT (Xu et al., 2021) used chest-xray dataset with 14 diseases (Wang et al., 2017). See discussion.

- All evaluations are inadequate with lack of attacking steps.
- The 1st Model used Kernelized manifold mapping to break the local linearity of neural networks. However, their black-box attack is better than white-box attack, which indicates gradient shattering. This should be evaluated by BPDA attack (Athalye et al., 2018a).
- The 2nd Model used pruning and attention layer as a defense method. It's based on random classifier, which should be evaluated by EOT attack (Athalye et al., 2018a).
- The 3rd Model reported PGD 4/255 attack accuracy without AT is 0.455 which indicates the ineffective attack.

Table of Contents

1. Develop universal defense on medical diagnostic system on defense tasks such as Segmentation, Object Detection on any dataset.
2. Improve defense accuracy by improving EBM MCMC sampling.
3. Evaluate and improve the most recent diffusion model defense (Nie et al., 2022) .
4. Use EBM diffusion recovery likelihood model (Gao et al., 2020b) for defense to improve defense accuracy.

Thank You!

References I

Akhtar, Naveed et al. (2021). "Advances in adversarial attacks and defenses in computer vision: A survey". In: *IEEE Access* 9, pp. 155161–155196.

Alayrac, Jean-Baptiste et al. (2019). "Are Labels Required for Improving Adversarial Robustness?" In: *Advances in Neural Information Processing Systems*. Vol. 32.

Apostolidis, Kyriakos D et al. (2021). "A survey on adversarial deep learning robustness in medical image analysis". In: *Electronics* 10.17, p. 2132.

Arjovsky, Martin et al. (2017). "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.

Athalye, Anish et al. (2018a). "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples". In: *International conference on machine learning*. PMLR, pp. 274–283.

References II

Athalye, Anish et al. (2018b). "Synthesizing robust adversarial examples". In: *International conference on machine learning*. PMLR, pp. 284–293.

Biggio, Battista et al. (2013). "Evasion attacks against machine learning at test time". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer, pp. 387–402.

Bortsova, Gerda et al. (2021). "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors". In: *Medical Image Analysis* 73, p. 102141.

Brock, Andrew et al. (2019). "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=B1xsqj09Fm.

Carlini, Nicholas (2022). "A Complete List of All (arXiv) Adversarial Example Papers, 2022 (accessed April, 2022)". In: URL: https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html.

References III

Che, Tong et al. (2020). "Your GAN is secretly an energy-based model and you should use discriminator driven latent sampling". In: *arXiv preprint arXiv:2003.06060*.

Chen, Lun et al. (2021). "Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism". In: *Medical physics* 48.10, pp. 6198–6212.

Chen, Ting et al. (2019). "Self-supervised gans via auxiliary rotation loss". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12154–12163.

Cohen, Jeremy et al. (2019). "Certified Adversarial Robustness via Randomized Smoothing". In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 1310–1320.

Daza, Laura et al. (2021). "Towards robust general medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 3–13.

References IV

- Dhariwal, Prafulla et al. (2021). "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34, pp. 8780–8794.
- Du, Yilun et al. (2019). "Implicit Generation and Modeling with Energy Based Models". In: *Advances in Neural Information Processing Systems*. Vol. 32.
- Du, Yilun et al. (2020). "Improved contrastive divergence training of energy based models". In: *arXiv preprint arXiv:2012.01316*.
- Fang, Wei et al. (2018). "A method for improving CNN-based image recognition using DCGAN". English. In: *Computers, Materials and Continua* 57.1. Funding Information: Acknowledgement: This work was supported in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions. Publisher Copyright: Copyright © 2018 Tech Science Press., pp. 167–178. ISSN: 1546-2218. DOI: 10.32604/cmc.2018.02356.
- Finlayson, Samuel G et al. (2018). "Adversarial attacks against medical deep learning systems". In: *arXiv preprint arXiv:1804.05296*.

References V

Finlayson, Samuel G et al. (2019). "Adversarial attacks on medical machine learning". In: *Science* 363.6433, pp. 1287–1289.

Gao, Ruiqi et al. (2020a). "Flow contrastive estimation of energy-based models". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7518–7528.

Gao, Ruiqi et al. (2020b). "Learning energy-based models by diffusion recovery likelihood". In: *arXiv preprint arXiv:2012.08125*.

Gong, Yuan et al. (July 2018). "Protecting Voice Controlled Systems Using Sound Source Identification Based on Acoustic Cues". In: DOI: 10.1109/ICCCN.2018.8487334.

Goodfellow, Ian et al. (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Goodfellow, Ian J et al. (2014). "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572*.

References VI

- Goodfellow, Ian J. (2017). "NIPS 2016 Tutorial: Generative Adversarial Networks". In: *ArXiv* abs/1701.00160.
- Grathwohl, Will et al. (2020). "Your classifier is secretly an energy based model and you should treat it like one". In: *International Conference on Learning Representations*.
- Grathwohl, Will Sussman et al. (2021). "No {MCMC} for me: Amortized sampling for fast and stable training of energy-based models". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ixpSxO9flk3.
- Han, Tian et al. (2019). "Divergence triangle for joint training of generator model, energy-based model, and inferential model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8670–8679.
- Hill, Mitch et al. (2020). "Stochastic security: Adversarial defense using long-run dynamics of energy-based models". In: *arXiv preprint arXiv:2005.13525*.

References VII

- Hill, Mitch et al. (2022). "EBM Life Cycle: MCMC Strategies for Synthesis, Defense, and Density Modeling". In: *arXiv preprint arXiv:2205.12243*.
- Hinton, Geoffrey E. (2002). "Training Products of Experts by Minimizing Contrastive Divergence". In: *Neural Computation* 14.8, pp. 1771–1800.
- Hirano, Hokuto et al. (2020). "Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks". In: *Plos one* 15.12, e0243963.
- Hirano, Hokuto et al. (2021). "Universal adversarial attacks on deep neural networks for medical image classification". In: *BMC medical imaging* 21.1, pp. 1–13.
- Ho, Jonathan et al. (2020). *Denoising Diffusion Probabilistic Models*. arXiv: 2006.11239 [cs.LG].
- Isensee, Fabian et al. (2018). "nnu-net: Self-adapting framework for u-net-based medical image segmentation". In: *arXiv preprint arXiv:1809.10486*.

References VIII

Karras, Tero et al. (June 2019). "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kermany, D et al. (2018). *Large dataset of labeled optical coherence tomography (OCT) and chest x-ray images, Mendeley data, v3 (2018)*.

Kermany, Daniel S et al. (2018). "Identifying medical diagnoses and treatable diseases by image-based deep learning". In: *Cell* 172.5, pp. 1122–1131.

Khan, Muhammad Zubair et al. (2021). "Deep Neural Architectures for Medical Image Semantic Segmentation: Review". In: *IEEE Access* 9, pp. 83002–83024.

Kurakin, Alexey et al. (2018). "Adversarial examples in the physical world". In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, pp. 99–112.

References IX

- Lee, Kwot Sin et al. (2021). "Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3942–3952.
- Lee, Kyungmin et al. (2021). *Efficient randomized smoothing by denoising with learned score function*. URL: https://openreview.net/forum?id=sI4SVtktqJ2.
- Liang, Hongshuo et al. (2022). "Adversarial Attack and Defense: A Survey". In: *Electronics* 11.8, p. 1283.
- Ma, Xingjun et al. (2021). "Understanding adversarial attacks on deep learning based medical image analysis systems". In: *Pattern Recognition* 110, p. 107332.
- Madry, Aleksander et al. (2017). "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083*.
- Miyato, Takeru et al. (2018). "Spectral normalization for generative adversarial networks". In: *arXiv preprint arXiv:1802.05957*.

References X

Nie, Weili et al. (2022). "Diffusion Models for Adversarial Purification". In: *arXiv preprint arXiv:2205.07460*.

Nijkamp, Erik et al. (2020). "On the Anatomy of MCMC-based Maximum Likelihood Learning of Energy-Based Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34.

Pan, Sinno Jialin et al. (2010). "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

Papernot, Nicolas et al. (2017). "Practical black-box attacks against machine learning". In: *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519.

Paul, Rahul et al. (2020). "Mitigating Adversarial Attacks on Medical Image Understanding Systems". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1517–1521. DOI: 10.1109/ISBI45749.2020.9098740.

References XI

Ronneberger, Olaf et al. (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

Song, Yang et al. (2019). "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems* 32.

— (2020). *Improved Techniques for Training Score-Based Generative Models*. arXiv: 2006.09011 [cs.LG].

Song-, Zhu et al. (1998). "Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling". In: *International Journal of Computer Vision* 27.2, pp. 107–126.

Szegedy, Christian et al. (2013). "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199*.

References XII

Taghanaki, Saeid Asgari et al. (2019). "A kernelized manifold mapping to diminish the effect of adversarial perturbations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11340–11349.

Tieleman, Tijmen (2008). "Training restricted Boltzmann machines using approximations to the likelihood gradient". In: *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071.

Torrey, Lisa et al. (2009). "Transfer Learning". In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*.

Tramer, Florian et al. (2020). "On adaptive attacks to adversarial example defenses". In: *Advances in Neural Information Processing Systems* 33, pp. 1633–1645.

References XIII

Wang, Xiaosong et al. (2017). "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.

Xie, Cihang et al. (2017). "Adversarial examples for semantic segmentation and object detection". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1369–1378.

Xie, Jianwen et al. (2016). "A theory of generative convnet". In: *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2635–2644.

Xie, Jianwen et al. (2018). "Cooperative learning of energy-based model and latent variable model via mcmc teaching". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32.

Xu, Mengting et al. (2021). "Towards evaluating the robustness of deep diagnostic models by adversarial attack". In: *Medical Image Analysis* 69, p. 101977.

References XIV

Yoon, Jongmin et al. (2021). "Adversarial purification with Score-based generative models". In: *arXiv preprint arXiv:2106.06041*.

Yu, Lantao et al. (2020). *Training Deep Energy-Based Models with f-Divergence Minimization*. arXiv: 2003.03463 [cs.LG].

Yuan, Xiaoyong et al. (2019). "Adversarial examples: Attacks and defenses for deep learning". In: *IEEE transactions on neural networks and learning systems* 30.9, pp. 2805–2824.

Zagoruyko, Sergey et al. (2016). "Wide Residual Networks". In: *arXiv preprint arXiv:1605.07146*.

Zhu, Jun-Yan et al. (Oct. 2017). "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.