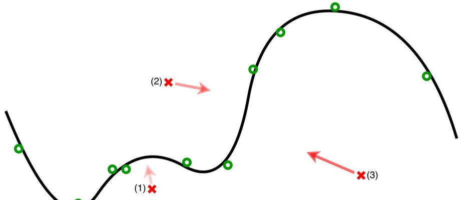
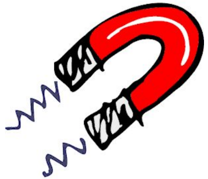
MagNet

A Two-Pronged Defense against Adversarial Examples





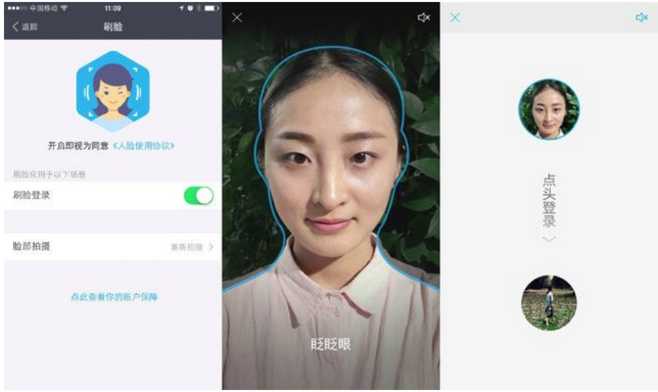
Dongyu Meng

ShanghaiTech University, China

Hao Chen

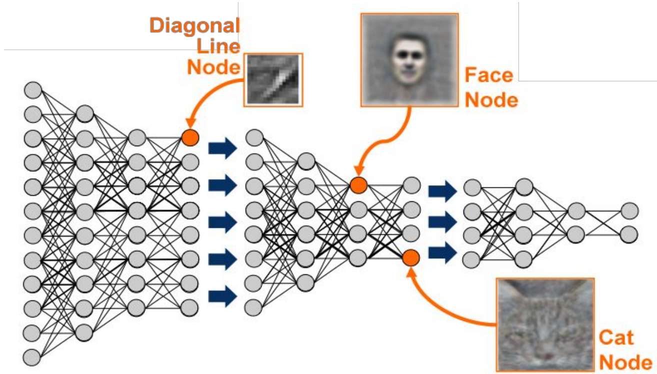
University of California, Davis, USA

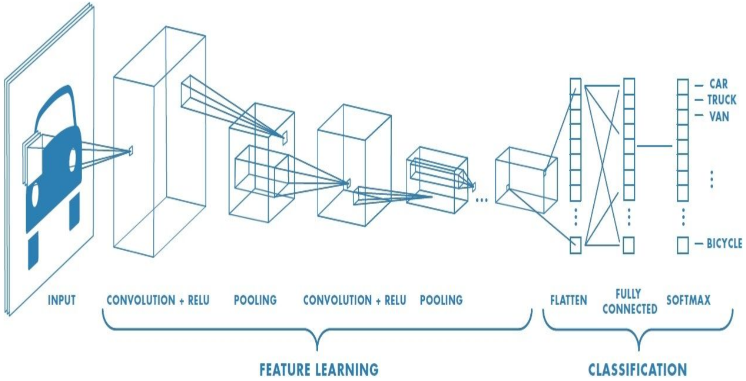
Neural networks in real-life applications



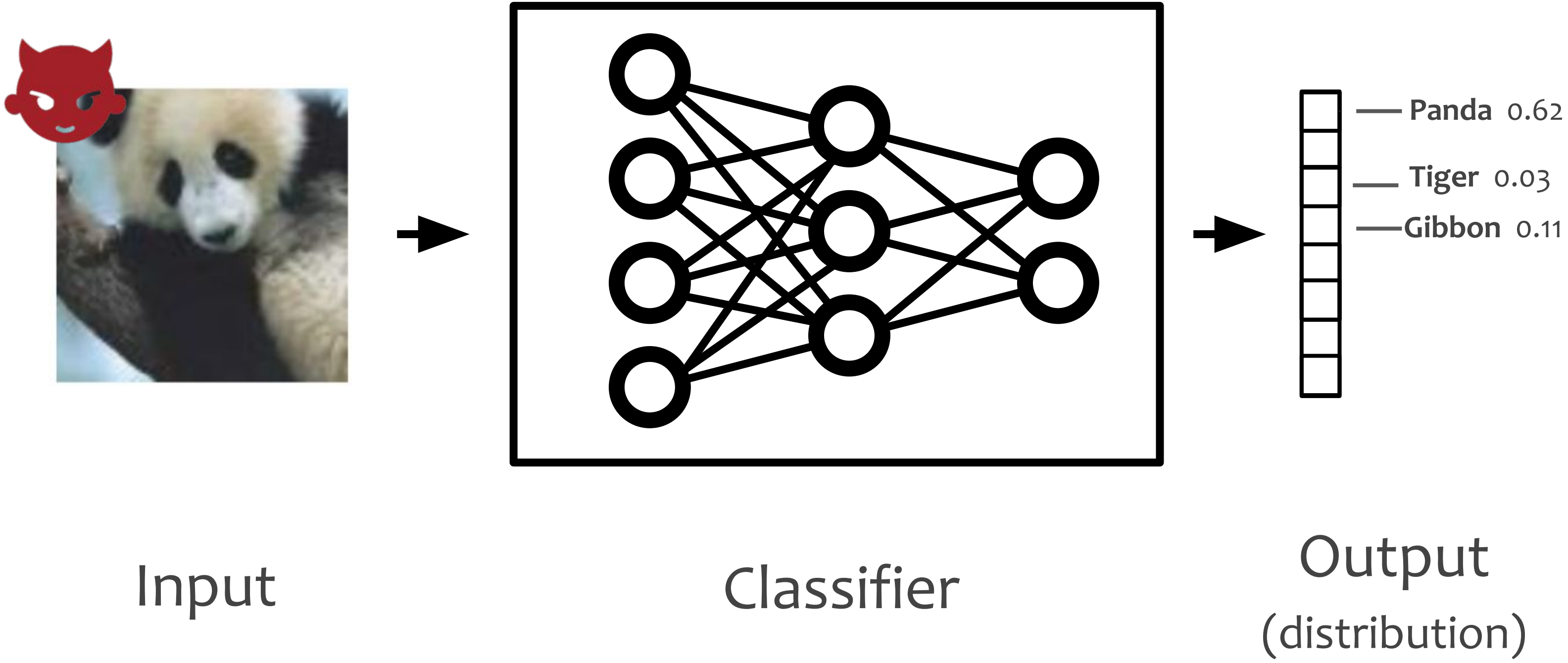
user authentication

autonomous vehicle



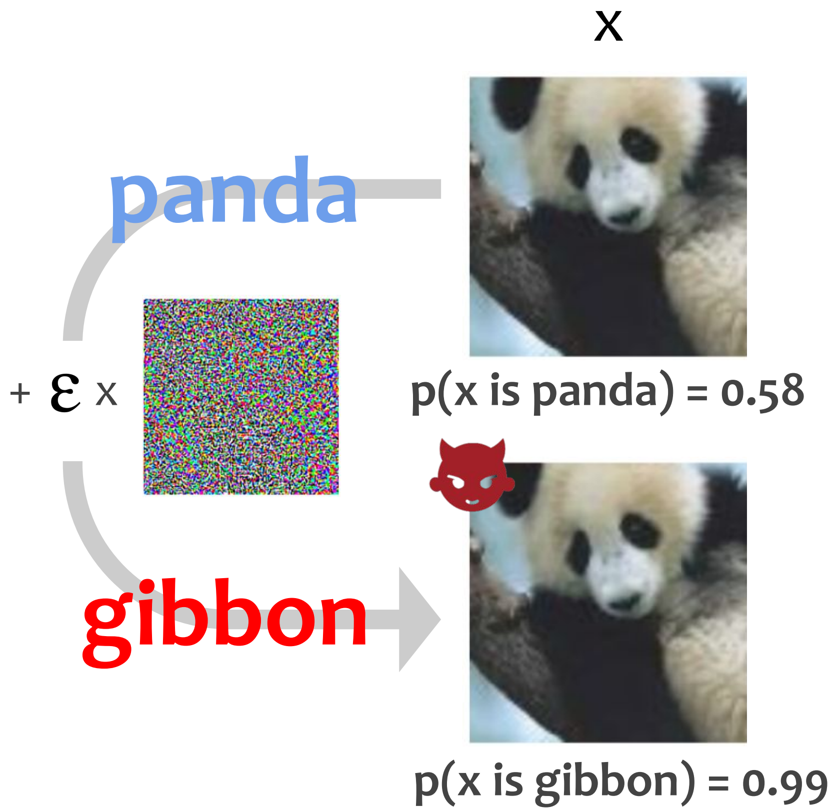


Neural networks as classifier



Adversarial examples

Examples carefully crafted to
- look like normal examples
- cause misclassification



[ICLR 15] Goodfellow, Shlens, and Szegedy. EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Attacks

$$x' = x + \epsilon \cdot sign(\nabla_x Loss(x, l_x))$$

Fast gradient sign method(FGSM)
[Goodfellow, 2015]

Carlini's attack
[Carlini, 2017]

$$\underset{\delta}{\text{minimize}} \quad \|\delta\|_2 + c \cdot \underline{f(x+\delta)}$$
$$\text{such that} \quad x + \delta \in [0,1]^n$$

confidence

$$\underline{f(x') = \max(Z(x')_{l_x} - \max\{Z(x')_i : i \neq l_x\}, \boxed{-\kappa})}$$

Iterative gradient
[Kurakin, 2016]

Deepfool
[Moosavi-Dezfooli, 2015]

……

Defenses

	target specific attack	modify classifier
Adversarial training [Goodfellow, 2015]	Yes	Yes
Defensive distillation [Papernot, 2016]		Yes
Detecting specific attacks [Metzen, 2017]	Yes	
……		

Desirable properties

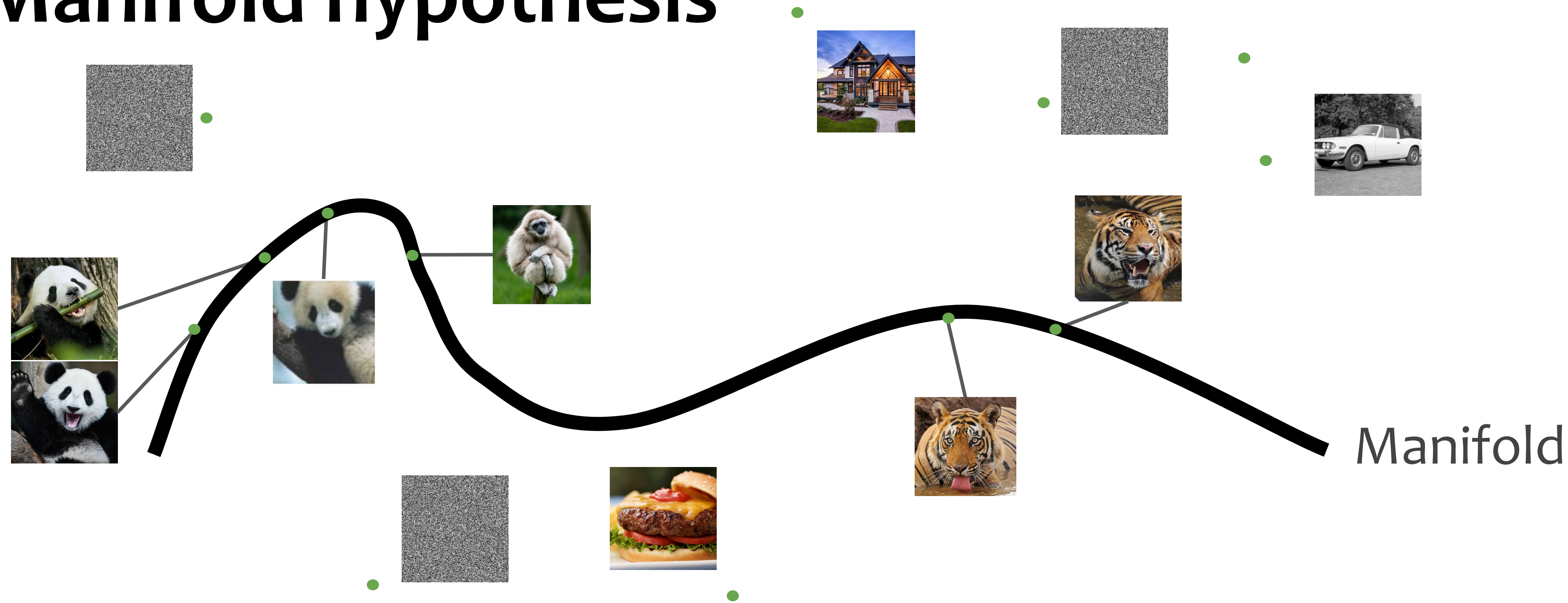
Does not modify target classifier.

- Can be deployed more easily as an add-on.

Does not rely on attack-specific properties.

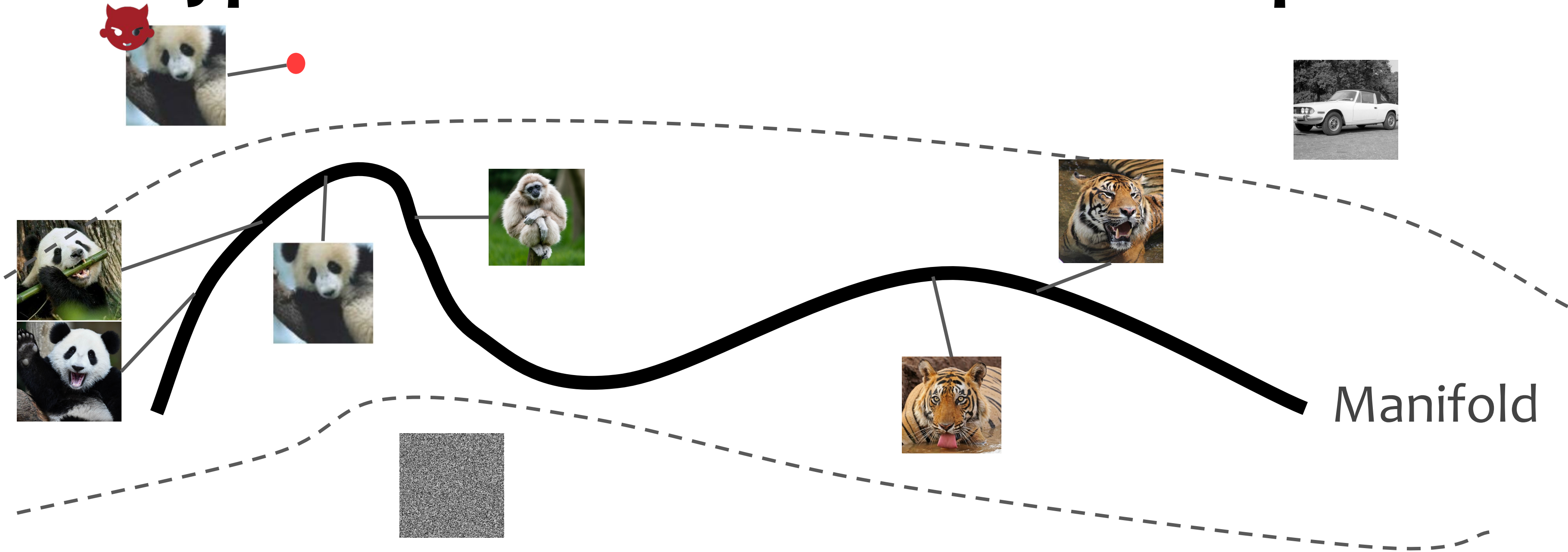
- Generalizes to unknown attacks.

Manifold hypothesis



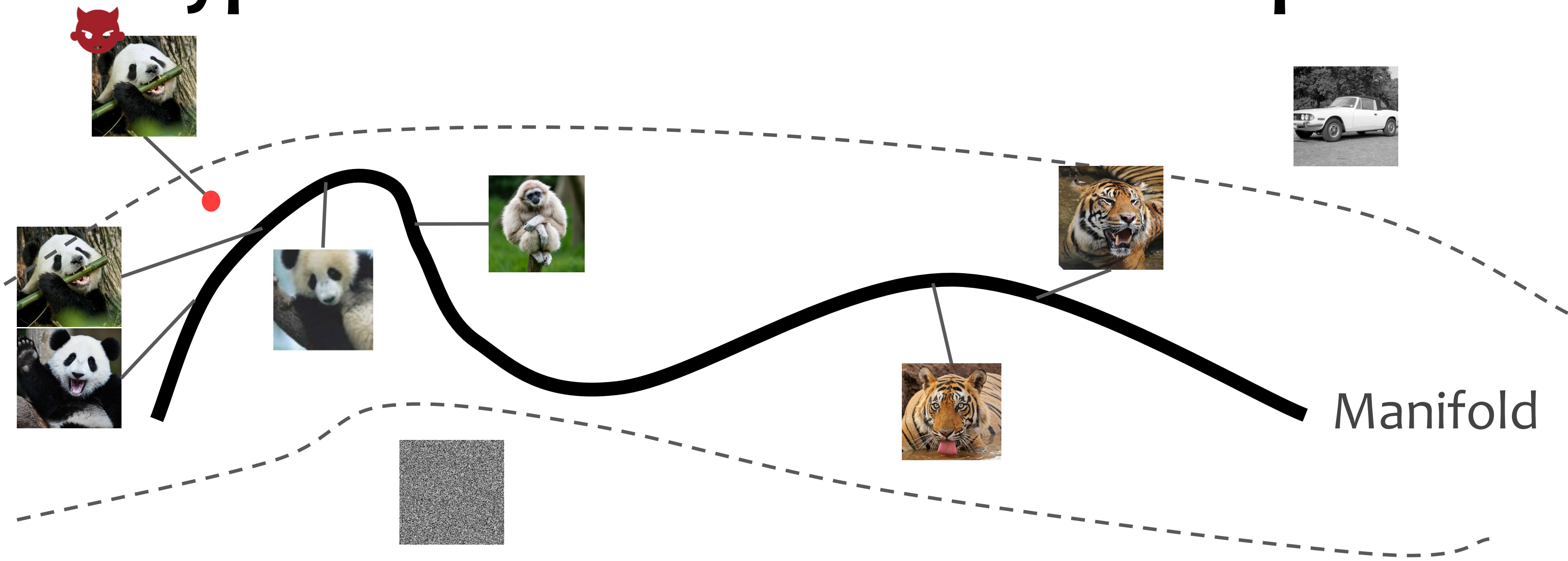
Possible inputs take up dense sample space.
But inputs we care about lie on a low dimensional **manifold**.

Our hypothesis for adversarial examples



Some adversarial examples are **far away** from the manifold.
Classifiers are not trained to work on these inputs.

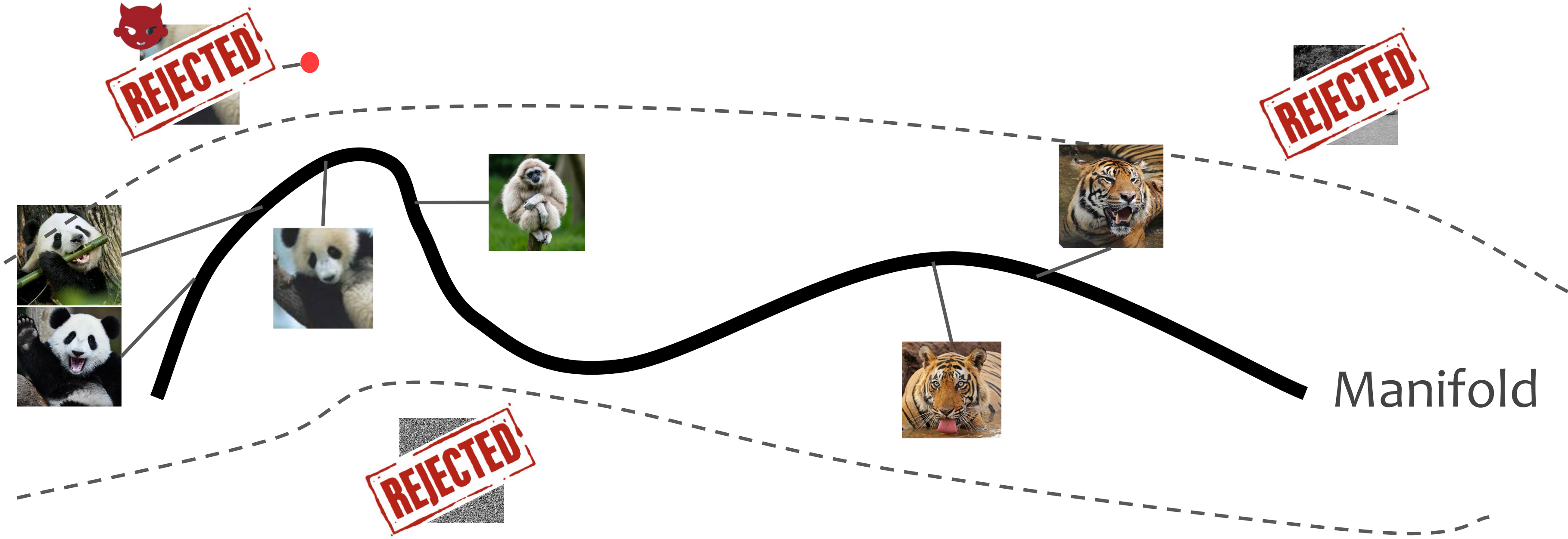
Our hypothesis for adversarial examples



Other adversarial example are **close** to the manifold boundary where the classifier **generalizes poorly.**

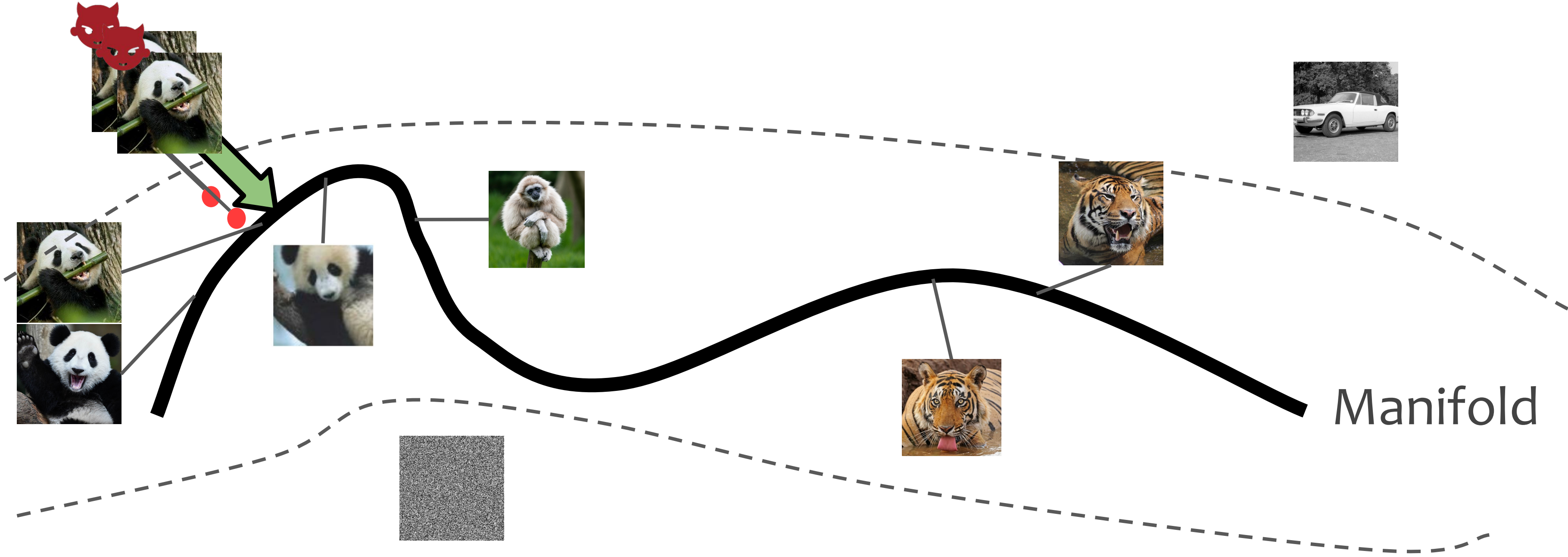
Sanitize your inputs.

Our solution



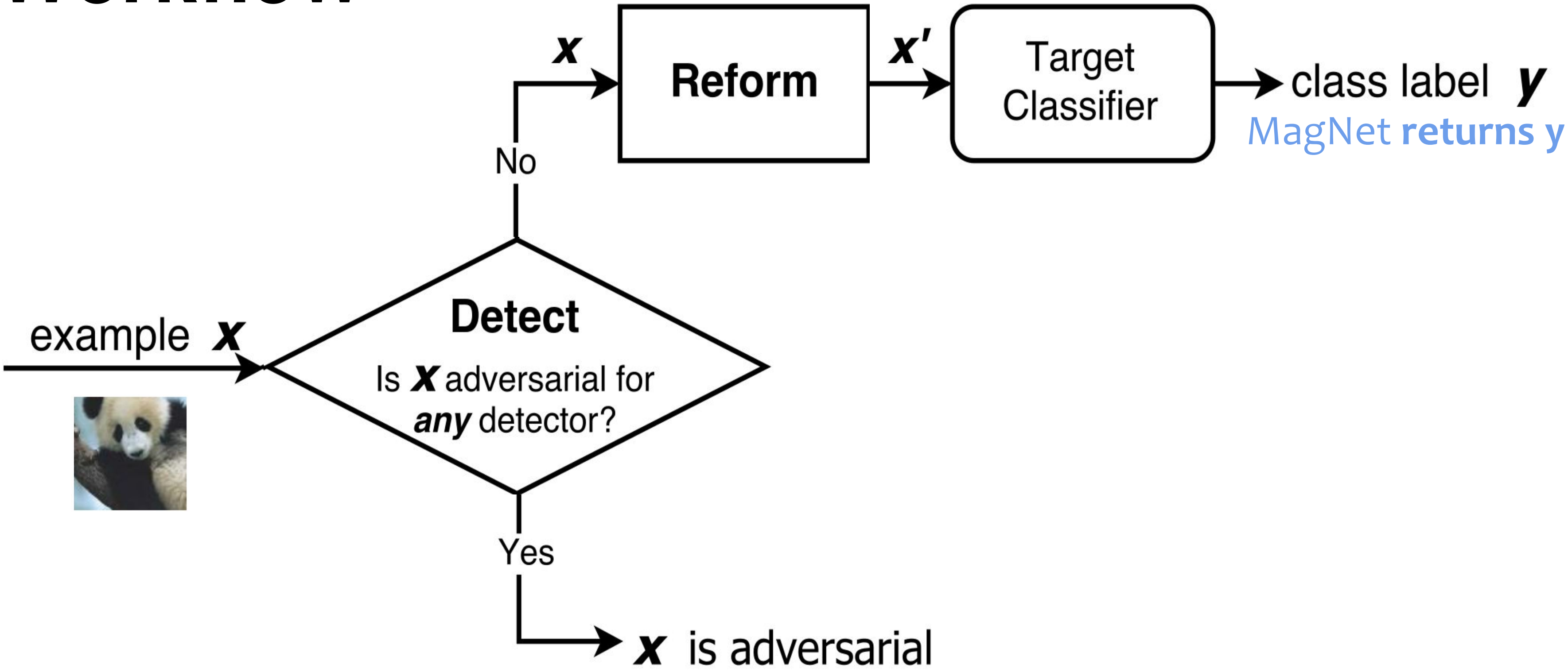
Detector: Decides if the example is far from the manifold.

Our solution



Reformer: Draws the example towards the manifold.

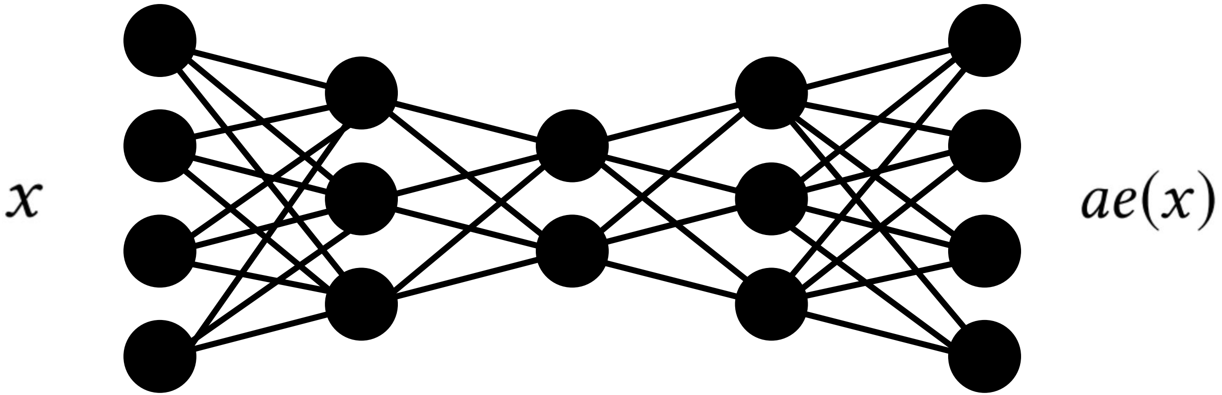
Workflow



MagNet **returns y**

MagNet **rejects** the input

Autoencoder

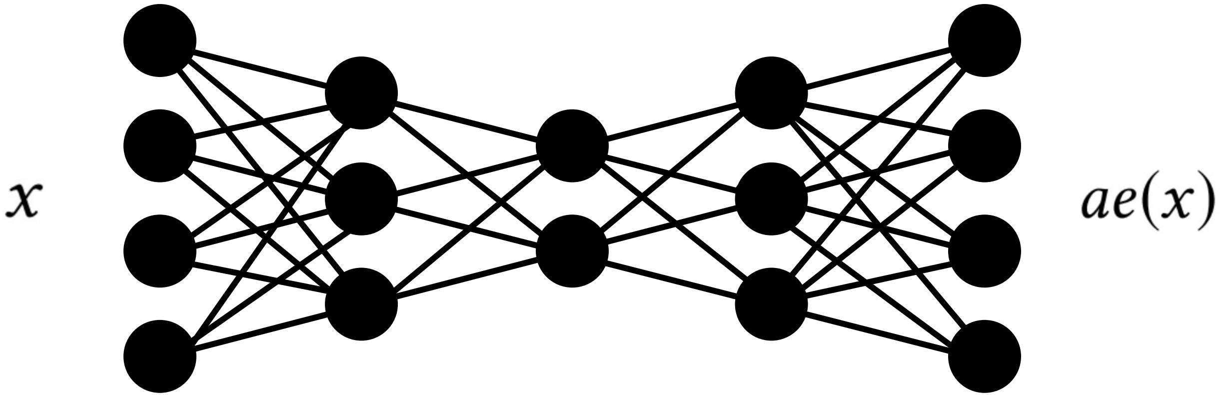


- Neural nets.
- Learn to copy input to output.
- Trained with constraints.

Reconstruction error:

$$||x - ae(x)||_2$$

Autoencoder

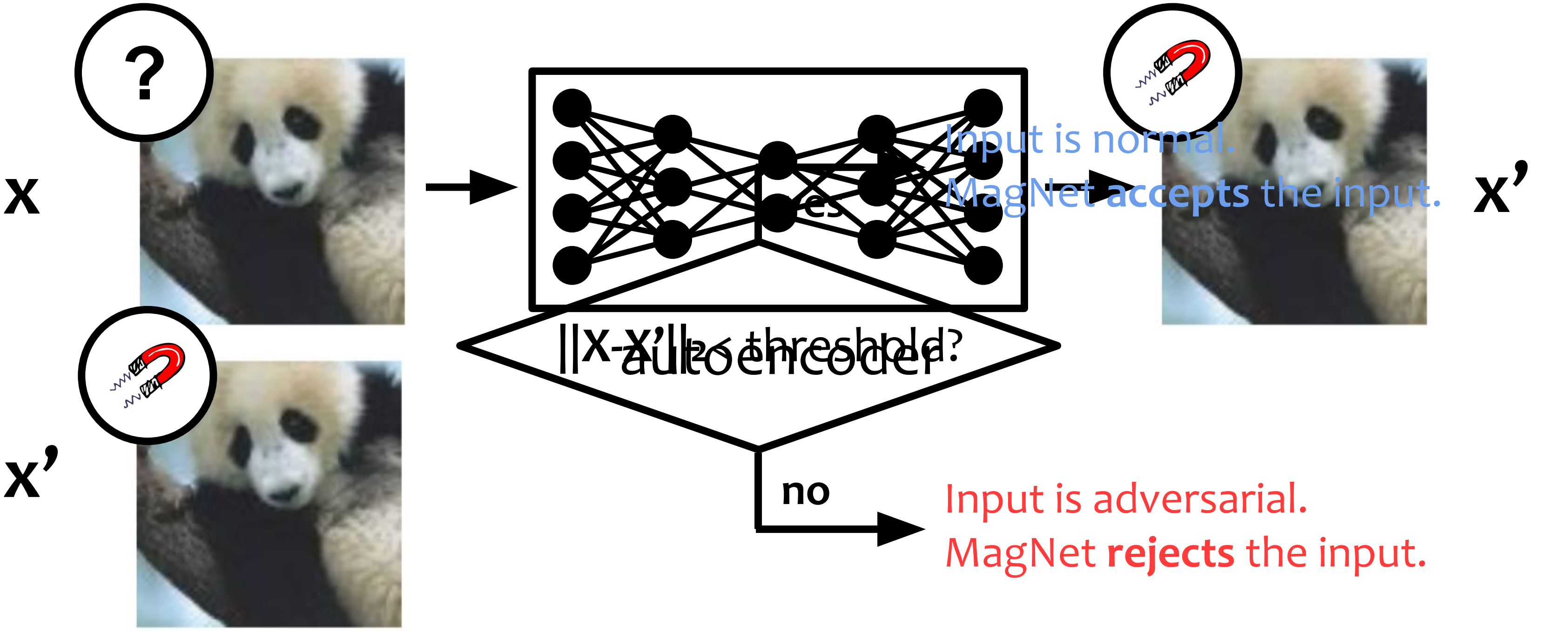


Autoencoders

- learn to map inputs towards manifold.
- approximate input-manifold distance with reconstruction error.

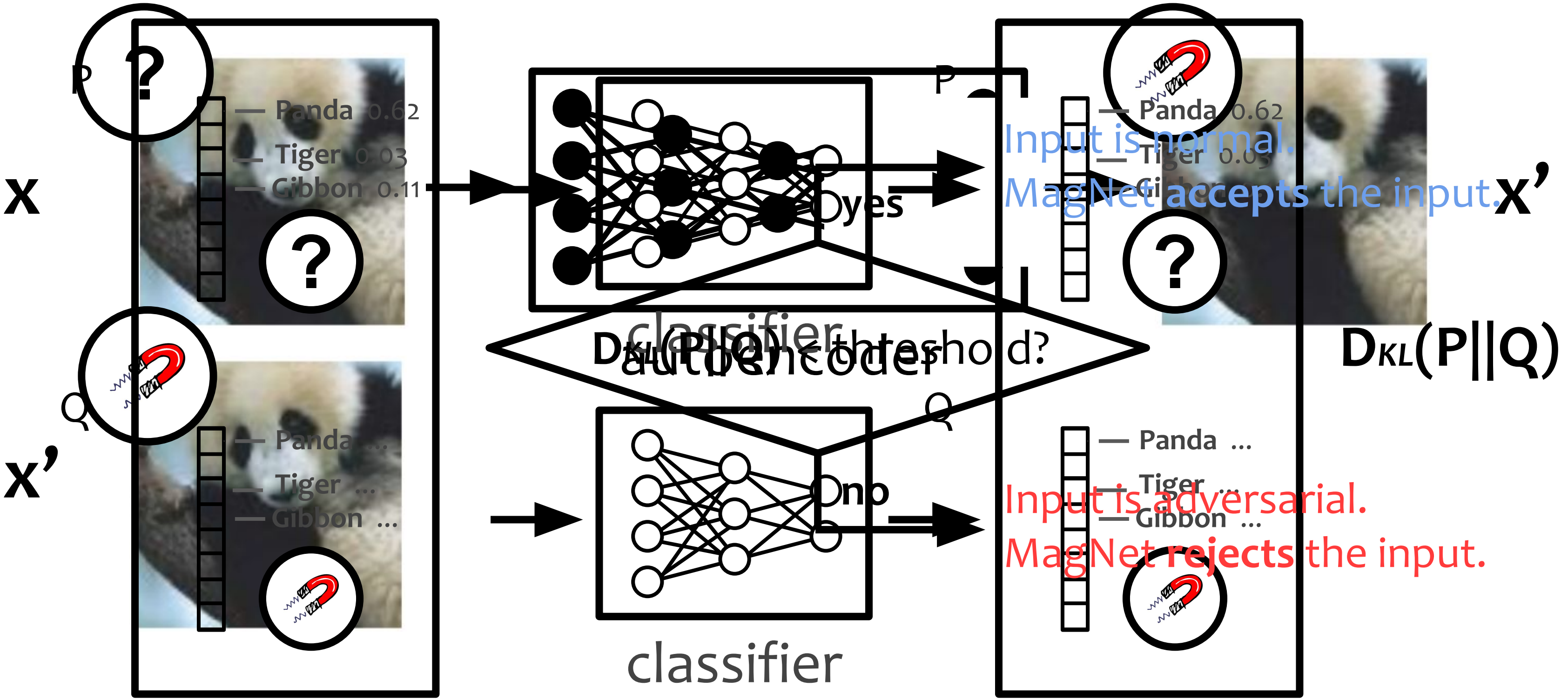
Train autoencoders on **normal examples only** as building blocks.

Detector
-- based on reconstruction error

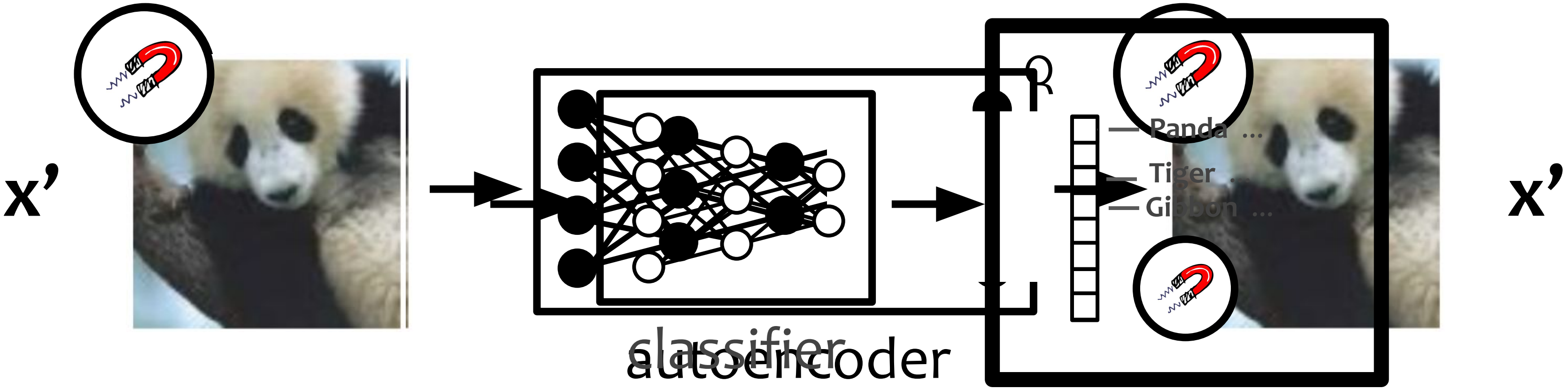


Detector

-- based on probability divergence



Reformer



MagNet **returns** Q as final classification result.

Threat model

 knows the parameters of …

	target classifier	defense
blackbox defense	✓	✗
whitebox defense	✓	✓

Blackbox defense on MNIST dataset

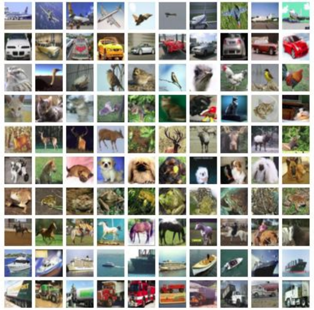
accuracy on adversarial examples

Attack	Norm	Parameter	No Defense	With Defense
FGSM	L^∞	$\epsilon = 0.005$	96.8%	100.0%
FGSM	L^∞	$\epsilon = 0.010$	91.1%	100.0%
Iterative	L^∞	$\epsilon = 0.005$	95.2%	100.0%
Iterative	L^∞	$\epsilon = 0.010$	72.0%	100.0%
Iterative	L^2	$\epsilon = 0.5$	86.7%	99.2%
Iterative	L^2	$\epsilon = 1.0$	76.6%	100.0%
Deepfool	L^∞		19.1%	99.4%
Carlini	L^2		0.0%	99.5%
Carlini	L^∞		0.0%	99.8%
Carlini	L^0		0.0%	92.0%

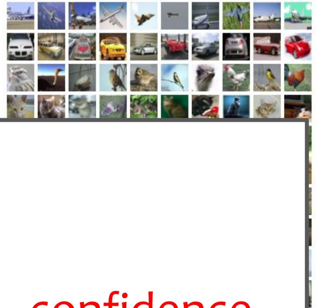
Blackbox defense on CIFAR-10 dataset

accuracy on adversarial examples

Attack	Norm	Parameter	No Defense	With Defense
FGSM	L^∞	$\epsilon = 0.025$	46.0%	99.9%
FGSM	L^∞	$\epsilon = 0.050$	40.5%	100.0%
Iterative	L^∞	$\epsilon = 0.010$	28.6%	96.0%
Iterative	L^∞	$\epsilon = 0.025$	11.1%	99.9%
Iterative	L^2	$\epsilon = 0.25$	18.4%	76.3%
Iterative	L^2	$\epsilon = 0.50$	6.6%	83.3%
Deepfool	L^∞		4.5%	93.4%
Carlini	L^2		0.0%	93.7%
Carlini	L^∞		0.0%	83.0%
Carlini	L^0		0.0%	77.5%

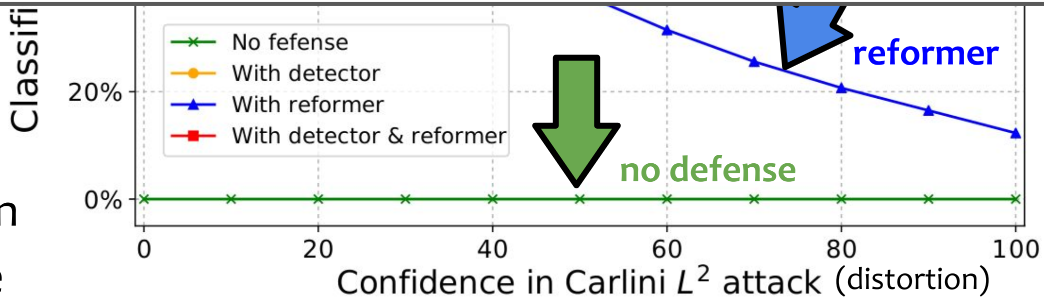


Detector vs. reformer



$$
\begin{aligned}
&\underset{\delta}{\text{minimize}} && \|\delta\|_2 + c \cdot f(x+\delta) \\
&\text{such that} && x + \delta \in [0,1]^n
\end{aligned}
$$

$$f(x') = \max(Z(x')_{l_x} - \max\{Z(x')_i : i \neq l_x\}, \overset{\text{confidence}}{\boxed{-\kappa}})$$



small distortion
less noticeable

large distortion
more transferable

Detector and reformer **complement each other**.

Whitebox defense is not practical

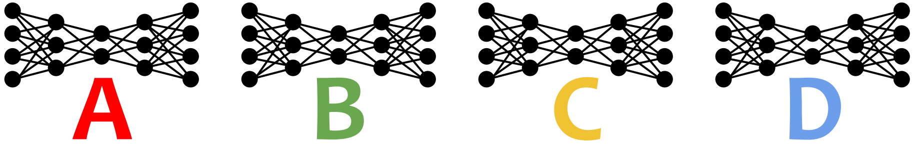
To defeat whitebox attacker, defender has to either

- make it **impossible** for attacker to find adversarial examples,
- or create a **perfect** classification network.

Graybox model

knows the parameters of...

	classifier	defense
blackbox defense	✓	✗
graybox defense	✓	?
whitebox defense	✓	✓



- Attacker knows possible defenses.
- Exact defense is only known at run time.

Defense strategy

- Train diverse defenses.
- Randomly pick one for each session.

Train diverse defenses

With MagNet, this means training diverse autoencoders.

Our Method:

Train n autoencoders at the same time.

$$\text{Minimize } L(x) = \sum_{i=1}^{n} \underbrace{\boxed{\mathrm{MSE}(x, ae_i(x))}}_{\text{reconstruction error}} - \alpha \sum_{i=1}^{n} \underbrace{\boxed{\mathrm{MSE}\Big(ae_i(x), \underbrace{\boxed{\frac{1}{n}\sum_{j=1}^{n} ae_j(x)}}_{\text{average reconstructed image}}\Big)}}_{\text{autoencoder diversity}}$$

Graybox classification accuracy

generate attack on →

defend with ↓

	A	B	C	D	E	F	G	H
A	0.0	92.8	92.5	93.1	91.8	91.8	92.5	93.6
B	92.1	0.0	92.0	92.5	91.4	92.5	91.3	92.5
C	93.2	93.8	0.0	92.8	93.3	94.1	92.7	93.6
D	92.8	92.2	91.3	0.0	91.7	92.8	91.2	93.9
E	93.3	94.0	93.4	93.2	0.0	93.4	91.0	92.8
F	92.8	93.1	93.2	93.6	92.2	0.0	92.8	93.8
G	92.5	93.1	92.0	92.2	90.5	93.5	0.1	93.4
H	92.3	92.0	91.8	92.6	91.4	92.3	92.4	0.0
Random	81.1	81.4	80.8	81.3	80.3	81.3	80.5	81.7

Limitations

The effectiveness of MagNet depends on assumptions that

- detector and reformer functions exist.
- we can approximate them with autoencoders.

We show empirically that these assumptions are likely correct.

Conclusion

We propose MagNet framework:

- **Detector** detects examples far from the manifold
- **Reformer** moves examples closer to the manifold

We demonstrated effective defense against adversarial examples in blackbox scenario with MagNet.

Instead of whitebox model, we advocate **graybox** model, where security rests on model diversity.

Thanks & Questions?

Find more about MagNet:

● https://arxiv.org/abs/1705.09064	Paper
● https://github.com/Trevillie/MagNet	Demo code
● mengdy.me	Author homepage