Measuring Neural Net Robustness with Constraints

Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos,

Dimitrios Vytiniotis, Aditya Nori, Antonio Criminisi

Verification of Learning-Based Systems

robustness/stability

fairness

Neural Net Robustness

f :  $\mapsto$ “school bus”

Adversarial Examples

(Szegedy et al. 2014)

Adversarial Examples

“school bus”

(Szegedy et al. 2014)

Adversarial Examples

“school bus”

“ostrich”

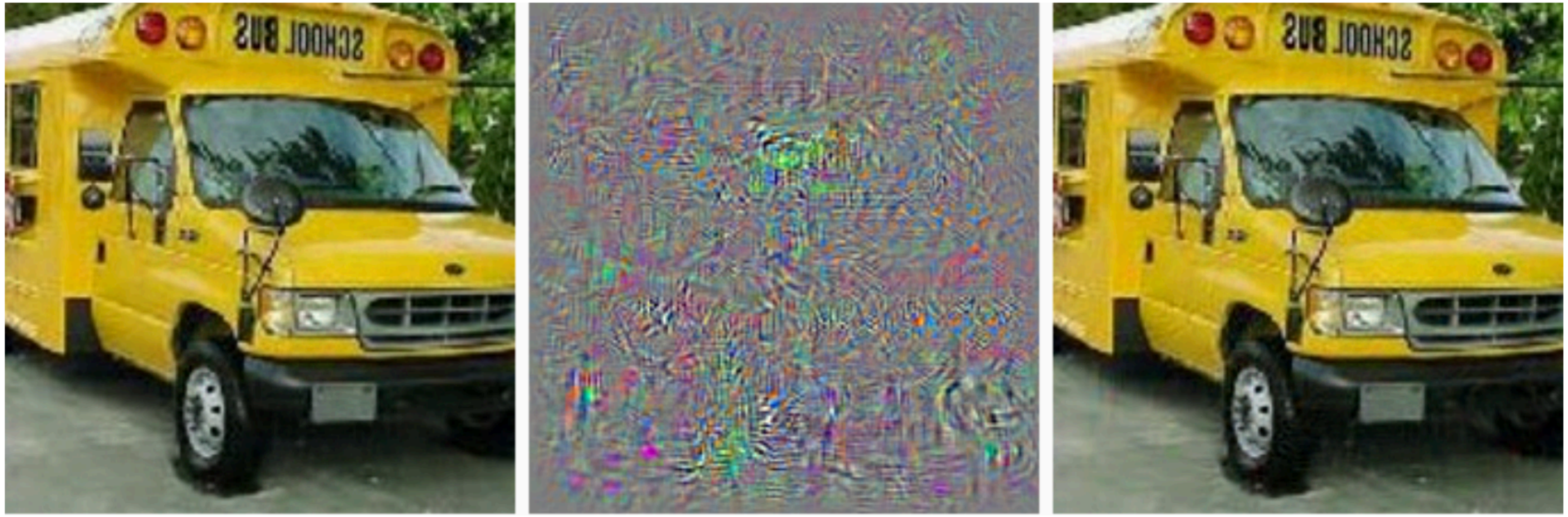
(Szegedy et al. 2014)

Adversarial Examples

“school bus” perturbation (10×) “ostrich”

(Szegedy et al. 2014)

Adversarial Examples

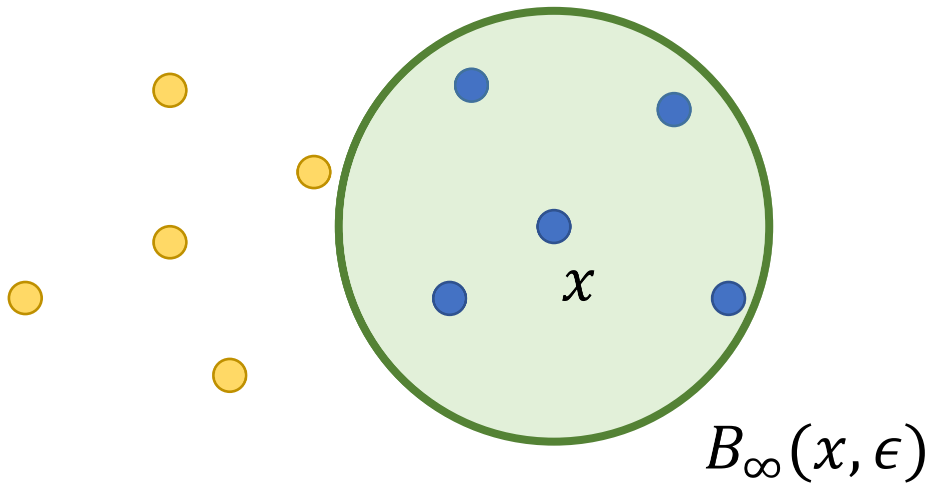


“school bus” perturbation (10×) “ostrich”

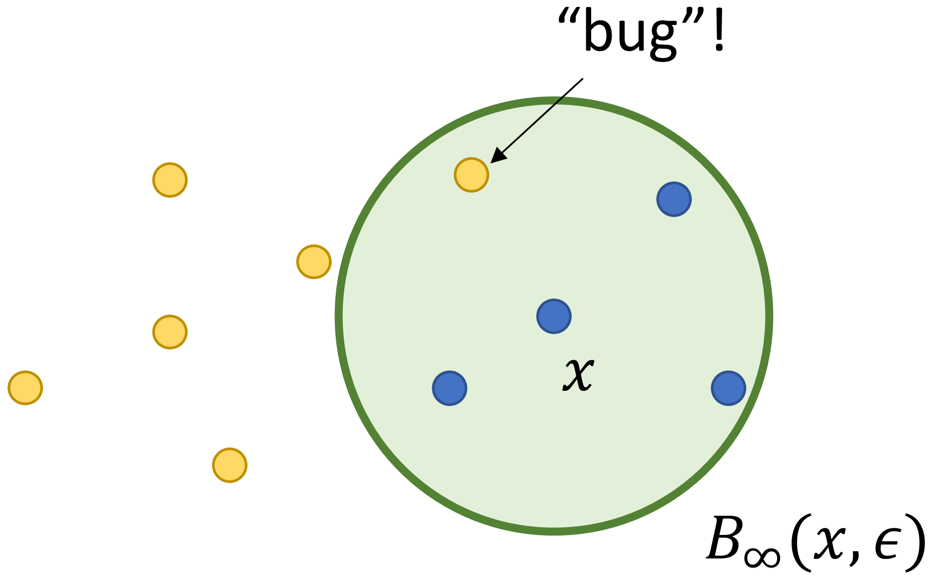
(Szegedy et al. 2014)

robustness: similar images $\Rightarrow$ same label

robustness: $\|x - x'\|_\infty \leq \epsilon \Rightarrow$ same label

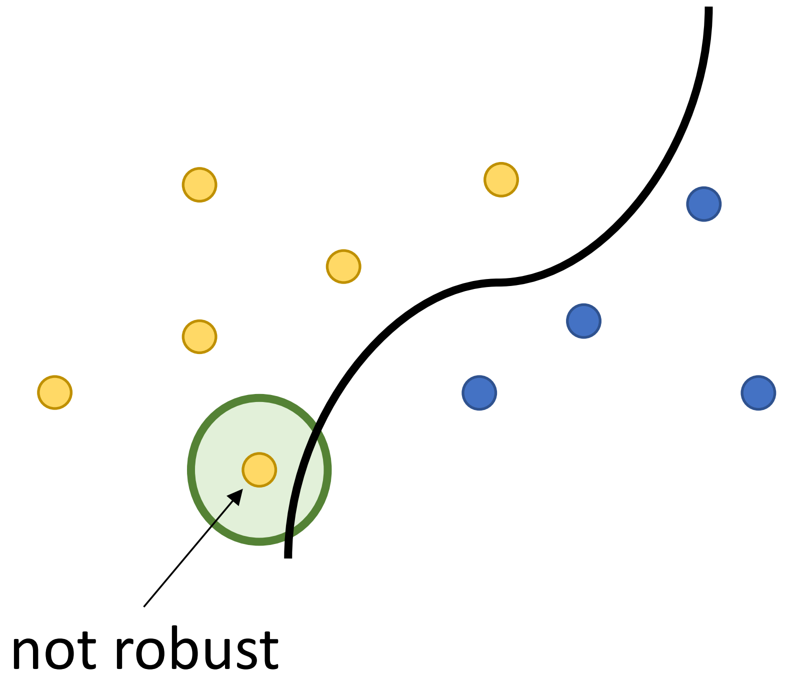


ϵ-robust at x



not ϵ-robust at x

Can we **verify** robustness?


not robust

Can we **quantify** robustness?

$$\psi(f, \epsilon) = \Pr_x[\, f \text{ not } \epsilon\text{-robust at } x]$$

$$\psi(f, \epsilon) = \Pr_x[\, f \text{ not } \epsilon\text{-robust at } x]$$

input distribution (e.g., held-out test set)

$$\psi(f, \epsilon) = \Pr_{x}[\, f \text{ not } \epsilon\text{-robust at } x] \in [0, 1]$$

input distribution (e.g., held-out test set)

Disclaimer

Disclaimer

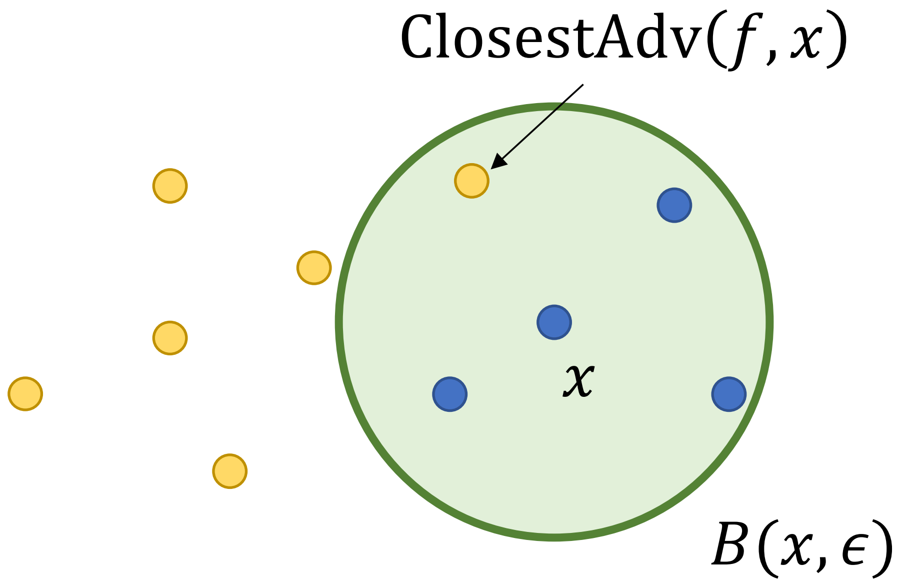
- Exactly quantifying neural net robustness did not scale

Disclaimer

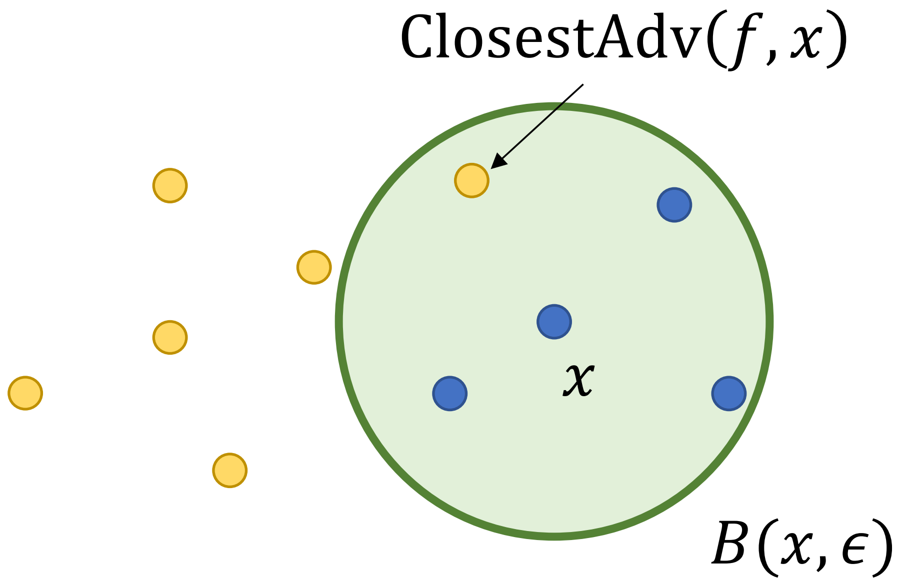
- Exactly quantifying neural net robustness did not scale
- We produce **useful approximations**

Disclaimer

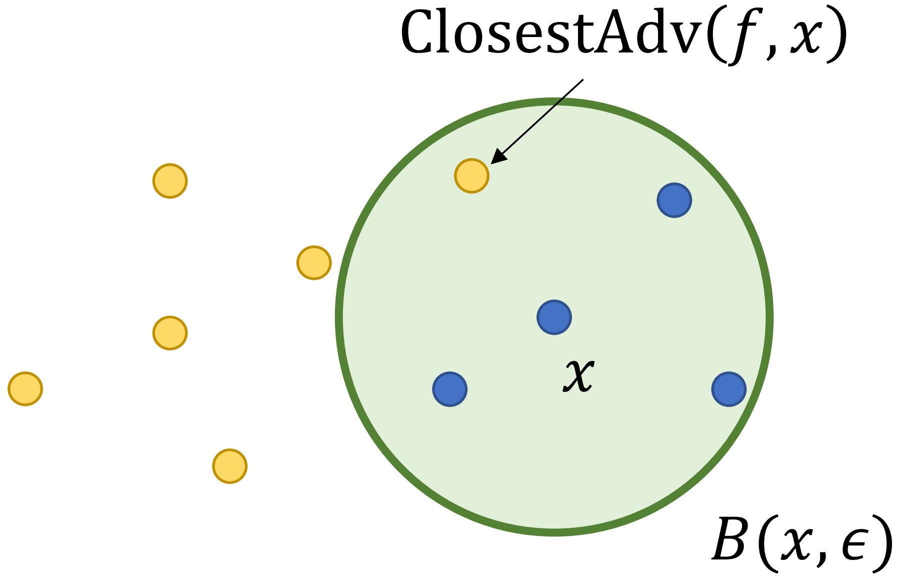
- Exactly quantifying neural net robustness did not scale
- We produce **useful approximations**
- At the end: possible paths forward?



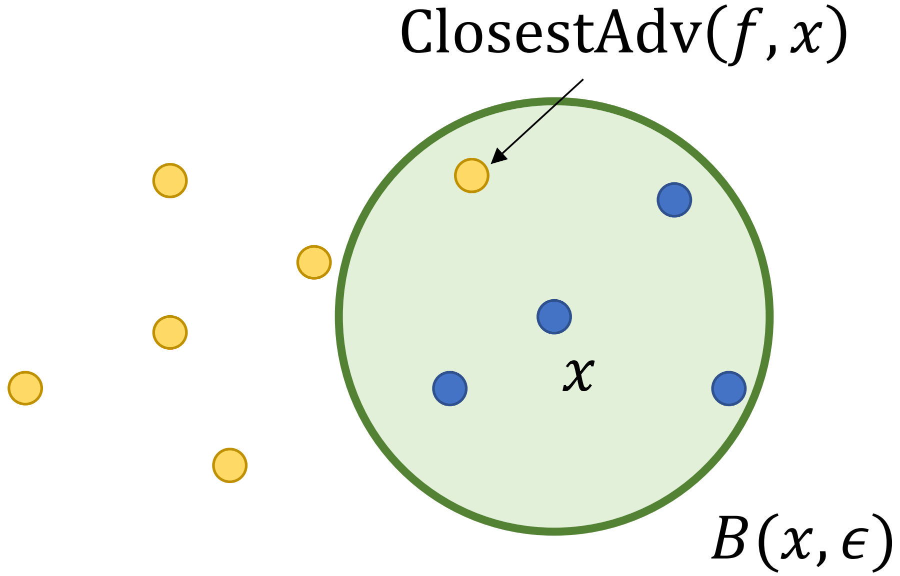
$$\psi(f, \epsilon) = \Pr_x[\, f \text{ not } \epsilon\text{-robust at } x]$$



$$\psi(f,\epsilon) = \Pr_x[\,\|\mathrm{ClosestAdv}(f,x) - x\| \le \epsilon]$$



$$\psi(f,\epsilon) = \frac{1}{|X|}\sum_{x\in X}\mathbb{I}[\|\mathrm{ClosestAdv}(f,x) - x\| \leq \epsilon]$$



$$\psi(f,\epsilon) = \frac{1}{|X|}\sum_{x\in X} \mathbb{I}[\|\mathrm{ClosestAdv}(f,x) - x\| \leq \epsilon]$$

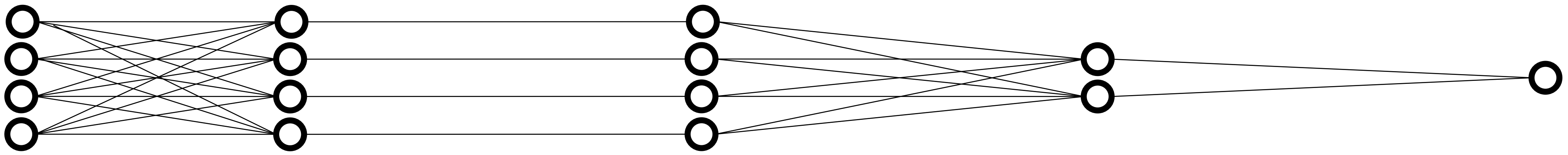
$$
\begin{aligned}
\arg\min_{x'} \quad & \|x' - x\|_\infty \\
\text{subj. to} \quad & f(x') \neq f(x)
\end{aligned}
$$

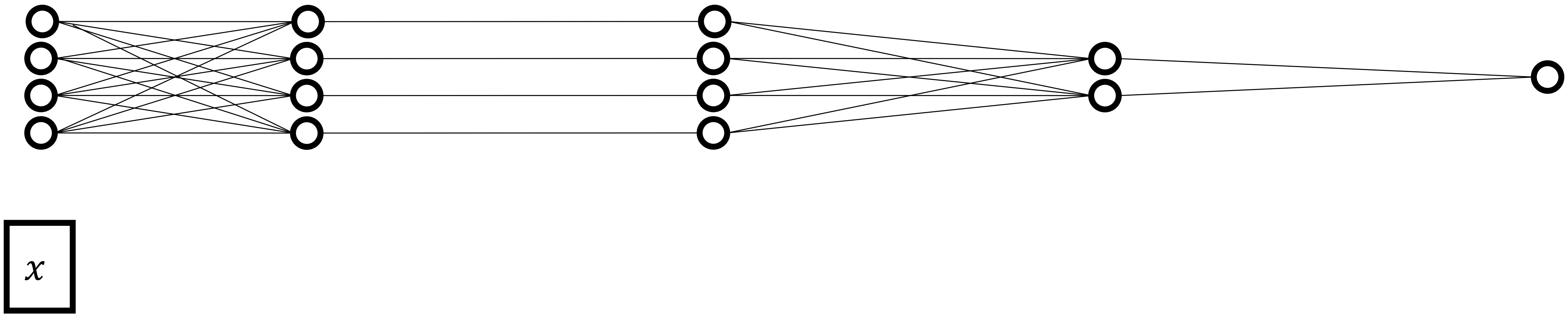
$$
\begin{aligned}
&\arg\min_{x'} \quad && \|x' - x\|_\infty \\
&\text{subj. to} \quad && \bigvee_{\ell \neq f(x)} f(x') = \ell
\end{aligned}
$$

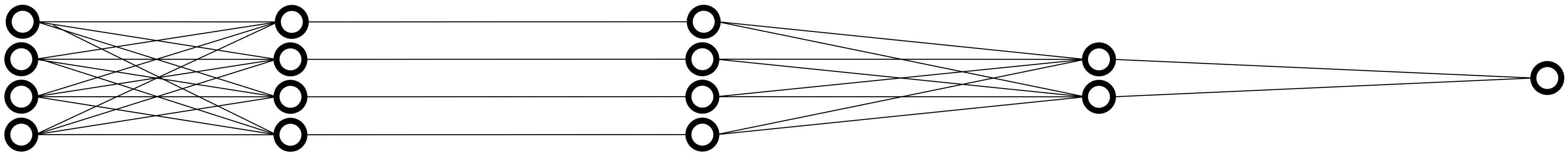
$$\arg\min_{x'} \quad \|x' - x\|_\infty$$

$$\text{subj. to} \quad \bigvee_{\ell \neq f(x)} f(x') = \ell$$

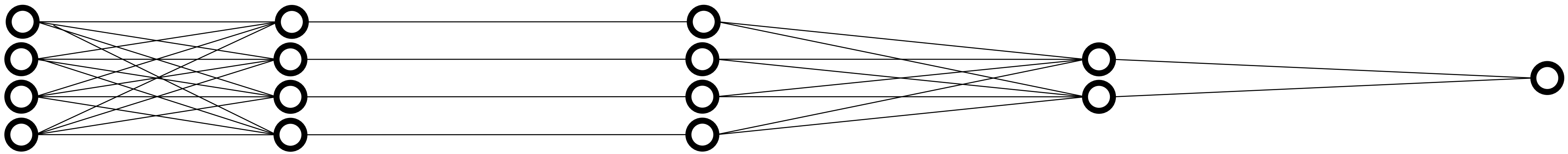
$$\phi_f(x, \ell) = \mathbb{I}[f(x) = \ell]$$




x

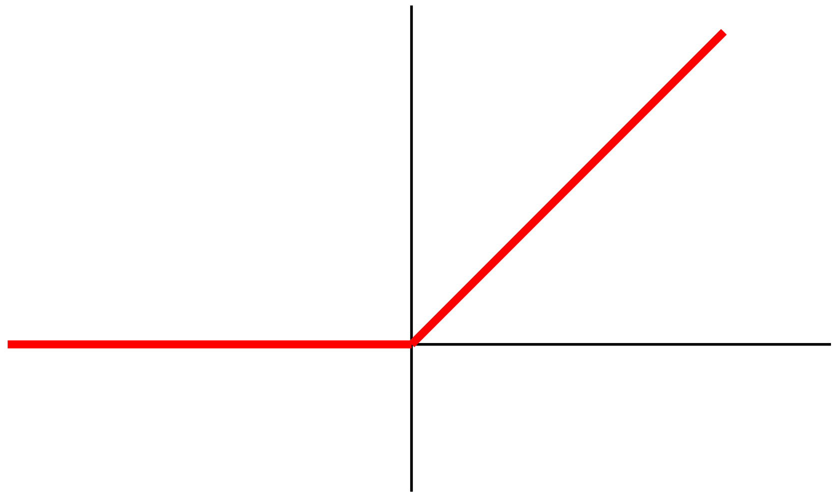


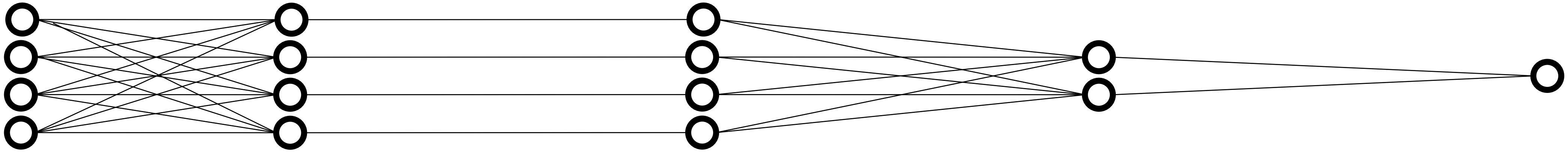
x ➡ $z^{(1)} = W^{(1)}x$



x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$

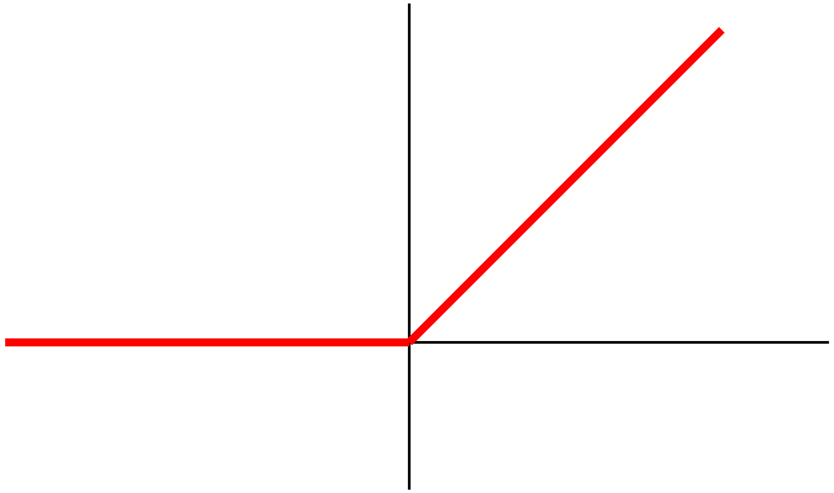
$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

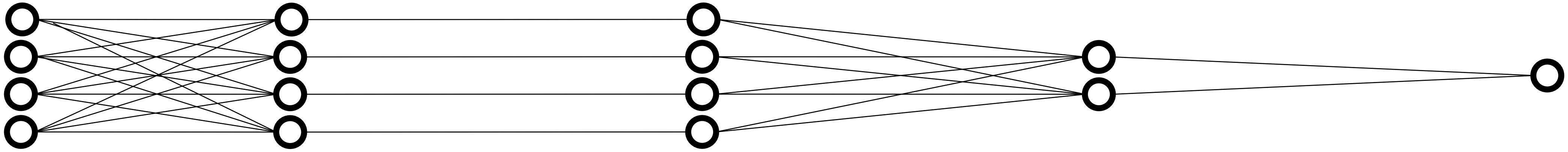




x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$ → $z^{(2)} = W^{(2)}x^{(2)}$

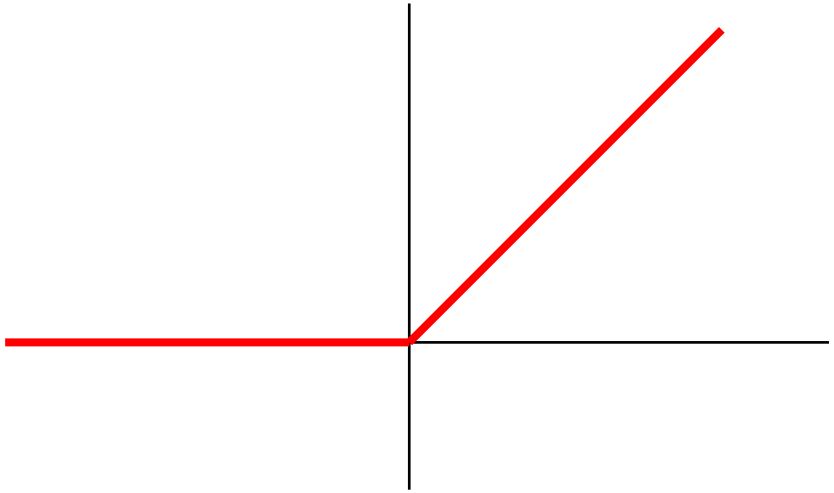
$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

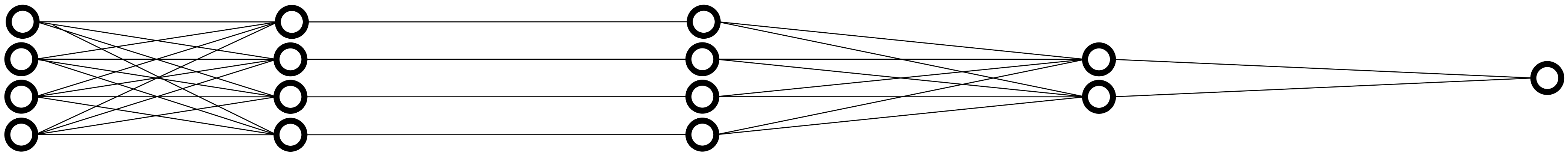




x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$ → $z^{(2)} = W^{(2)}x^{(2)}$ → $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

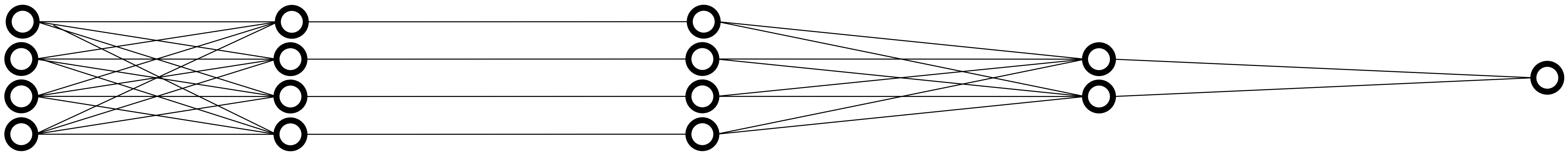




x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$ → $z^{(2)} = W^{(2)}x^{(2)}$ → $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

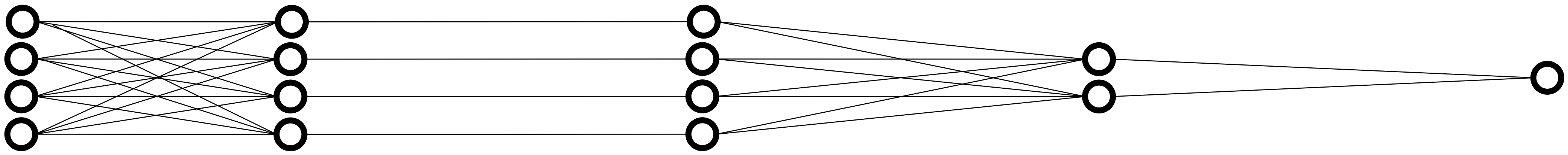
$$\phi_f(x, \ell) =$$



x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$ → $z^{(2)} = W^{(2)}x^{(2)}$ → $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

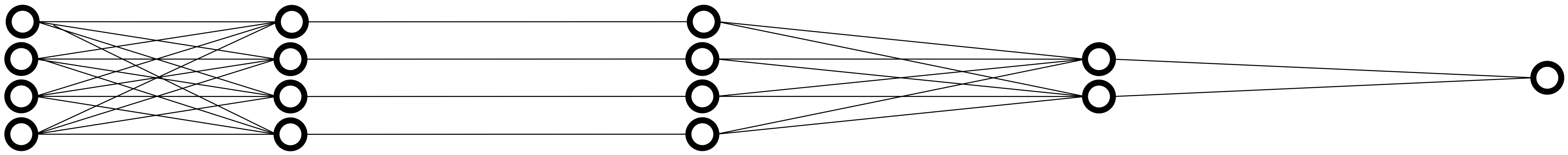
$$\phi_f(x, \ell) = \left(z^{(1)} = W^{(1)}x\right)$$



x ➡ $z^{(1)} = W^{(1)}x$ ➡ $x_i^{(2)} = g\left(z_i^{(1)}\right)$ ➡ $z^{(2)} = W^{(2)}x^{(2)}$ ➡ $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

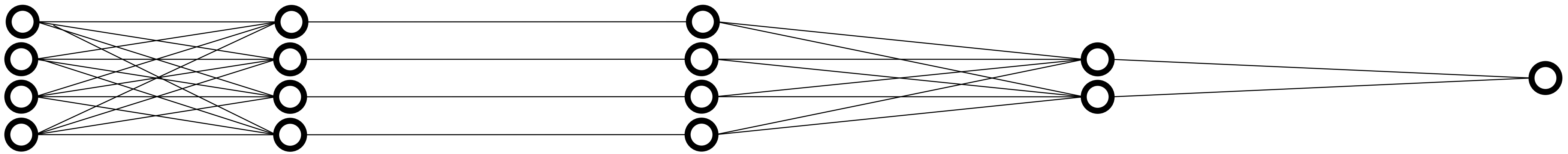
$$\begin{aligned} \phi_f(x, \ell) = {} & \left(z^{(1)} = W^{(1)}x\right) \\ & \land \forall i. \left[\left(z_i^{(1)} \leq 0 \land x_i^{(2)} = 0\right) \lor \left(z_i^{(1)} \geq 0 \land x_i^{(2)} = z_i^{(1)}\right)\right] \end{aligned}$$



x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$ → $z^{(2)} = W^{(2)}x^{(2)}$ → $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

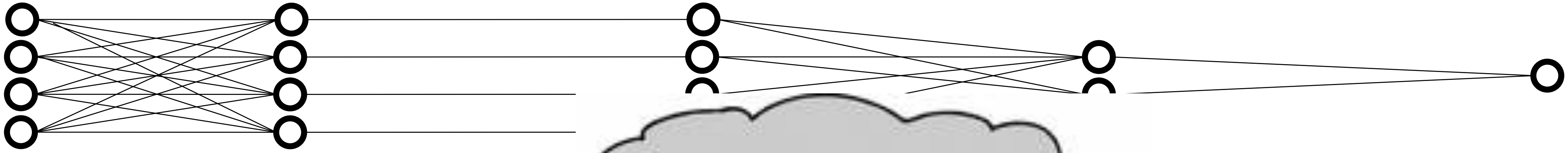
$$\phi_f(x, \ell) = \left(z^{(1)} = W^{(1)}x\right) \land \left(z^{(2)} = W^{(2)}x^{(2)}\right)$$
$$\land\, \forall i. \left[\left(z_i^{(1)} \leq 0 \land x_i^{(2)} = 0\right) \lor \left(z_i^{(1)} \geq 0 \land x_i^{(2)} = z_i^{(1)}\right)\right]$$



x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$ → $z^{(2)} = W^{(2)}x^{(2)}$ → $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

$$\begin{aligned} \phi_f(x,\ell) = & \left(z^{(1)} = W^{(1)}x\right) \wedge \left(z^{(2)} = W^{(2)}x^{(2)}\right) \\ & \wedge \forall i.\left[\left(z_i^{(1)} \leq 0 \wedge x_i^{(2)} = 0\right) \vee \left(z_i^{(1)} \geq 0 \wedge x_i^{(2)} = z_i^{(1)}\right)\right] \\ & \wedge \forall i.\left[z_\ell^{(2)} \geq z_i^{(2)}\right] \end{aligned}$$



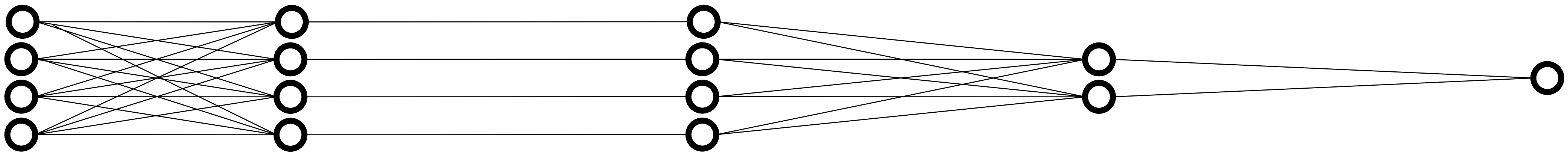
x ➡ $z^{(1)} = W^{(1)}x$ ➡ … $x^{(2)}$ ➡ $y = \arg\max_i \{z_i^{(2)}\}$

$g($ …

$\phi_f(x, \ell) = (z^{(1)} = W$ …

$\wedge \forall i. [(z_i^{($ … $0 \wedge x_i^{(2)} = z_i^{(1)})]$

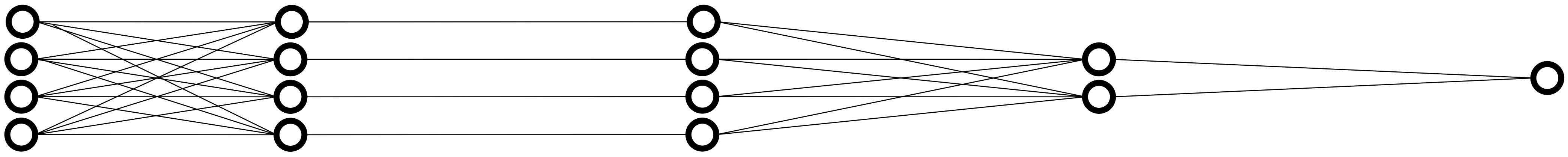
$\wedge \forall i. [z_\ell^{(2}$ …



$$x \Rightarrow z^{(1)} = W^{(1)}x \Rightarrow x_i^{(2)} = g\left(z_i^{(1)}\right) \Rightarrow z^{(2)} = W^{(2)}x^{(2)} \Rightarrow y = \arg\max_i \left\{z_i^{(2)}\right\}$$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

$$\begin{aligned} \phi_f(x, \ell) = &\left(z^{(1)} = W^{(1)}x\right) \wedge \left(z^{(2)} = W^{(2)}x^{(2)}\right) \\ &\wedge \forall i. \left[\left(z_i^{(1)} \leq 0 \wedge x_i^{(2)} = 0\right) \vee \left(z_i^{(1)} \geq 0 \wedge x_i^{(2)} = z_i^{(1)}\right)\right] \\ &\wedge \forall i. \left[z_\ell^{(2)} \geq z_i^{(2)}\right] \end{aligned}$$



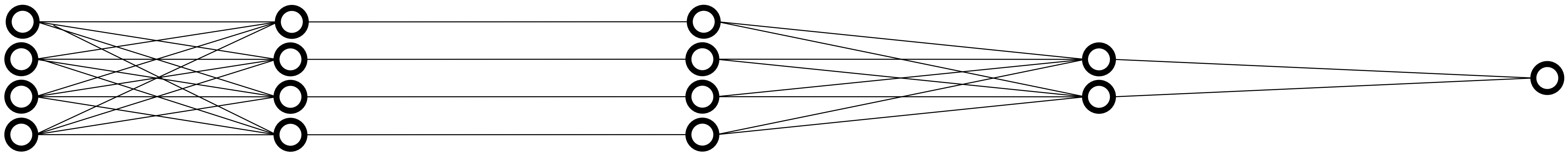
x → $z^{(1)} = W^{(1)}x$ → $x_i^{(2)} = g\left(z_i^{(1)}\right)$ → $z^{(2)} = W^{(2)}x^{(2)}$ → $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

$$\phi_f(x, \ell) = \left(z^{(1)} = W^{(1)}x\right) \land \left(z^{(2)} = W^{(2)}x^{(2)}\right)$$

$\land \forall i.$ [~~$\left(z_i^{(1)} \leq 0 \land x_i^{(2)} = 0\right) \lor$~~ $\left(z_i^{(1)} \geq 0 \land x_i^{(2)} = z_i^{(1)}\right)$]

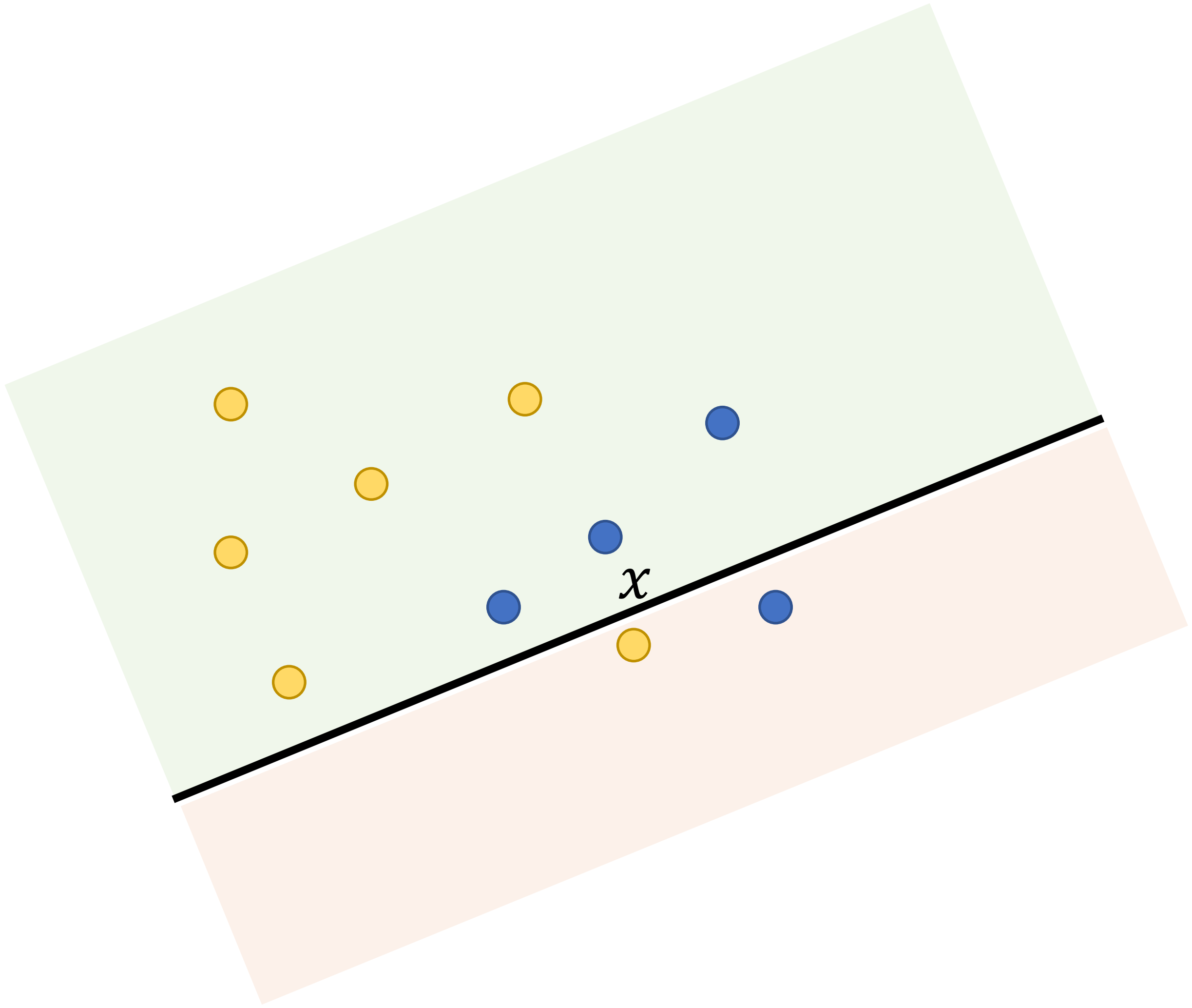
$$\land \forall i. \left[z_\ell^{(2)} \geq z_i^{(2)}\right]$$



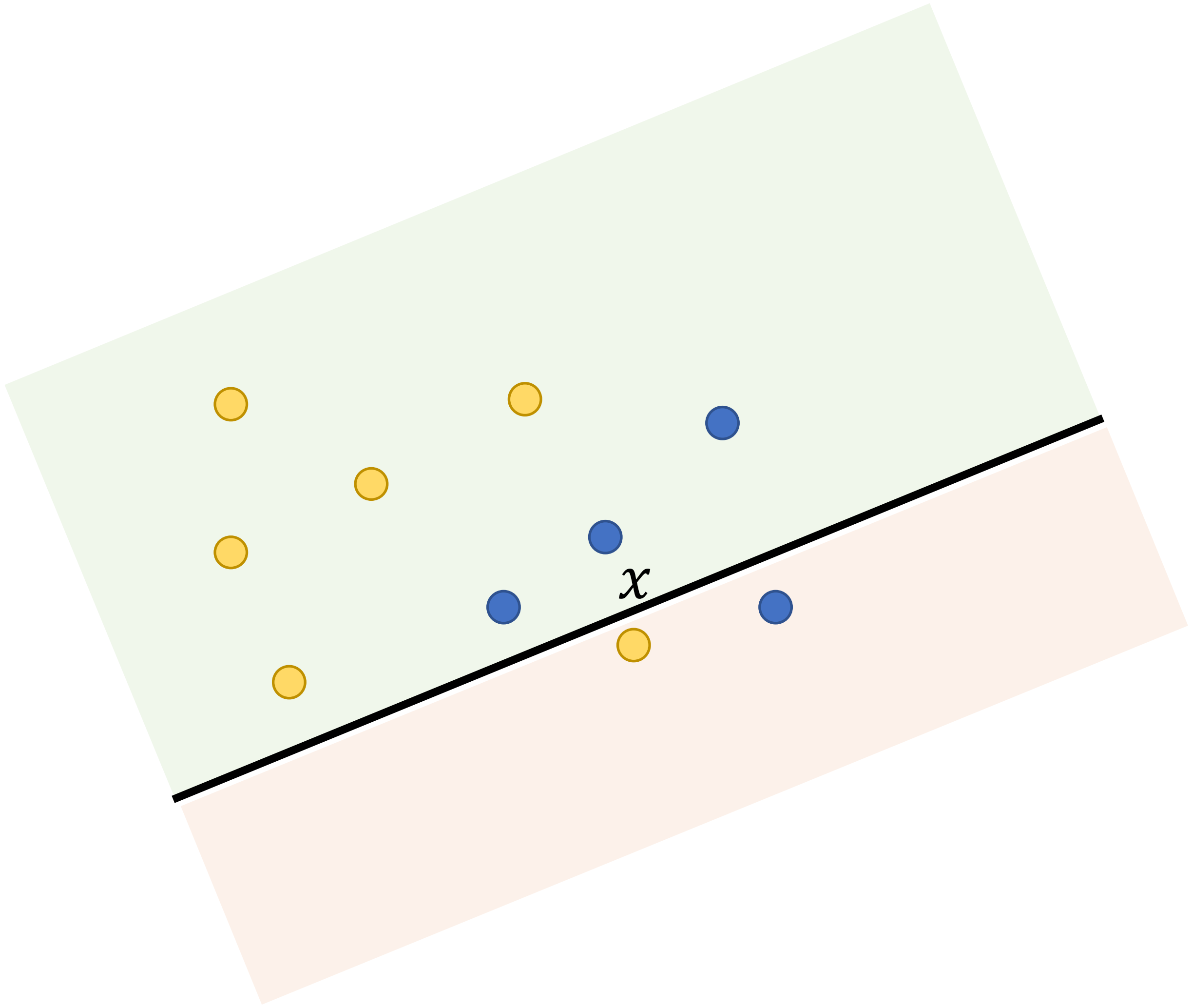
x ➡ $z^{(1)} = W^{(1)}x$ ➡ $x_i^{(2)} = g\left(z_i^{(1)}\right)$ ➡ $z^{(2)} = W^{(2)}x^{(2)}$ ➡ $y = \arg\max_i \left\{z_i^{(2)}\right\}$

$$g(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z \leq 0) \end{cases}$$

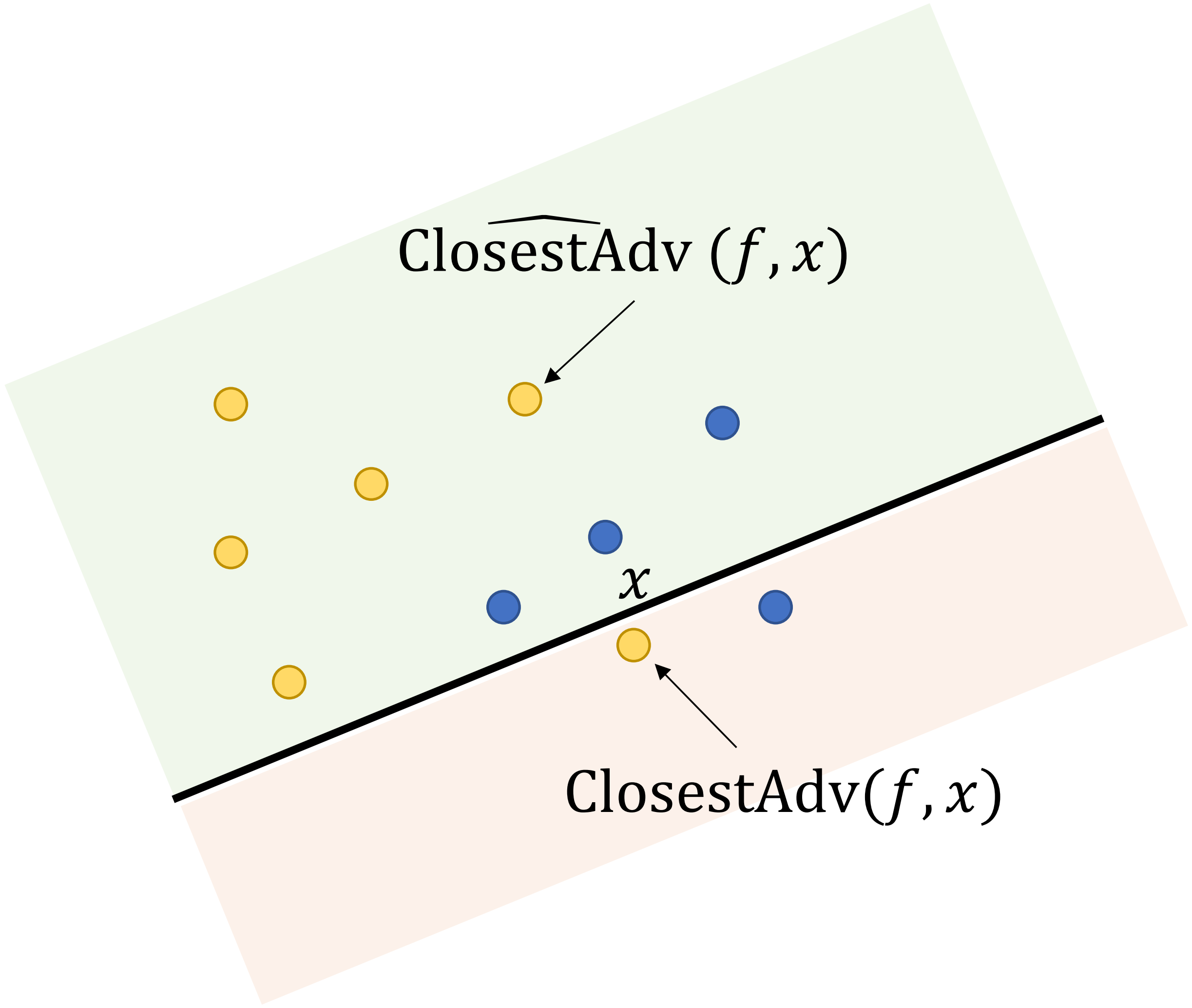
$$\begin{aligned} \phi_f(x, \ell) = &\left(z^{(1)} = W^{(1)}x\right) \wedge \left(z^{(2)} = W^{(2)}x^{(2)}\right) \\ &\wedge \forall i. \left[z_i^{(1)} \geq 0 \wedge x_i^{(2)} = z_i^{(1)}\right] \\ &\wedge \forall i. \left[z_\ell^{(2)} \geq z_i^{(2)}\right] \end{aligned}$$



$$\left(z_i^{(1)} \leq 0 \wedge x_i^{(2)} = 0\right) \vee \left(z_i^{(1)} \geq 0 \wedge x_i^{(2)} = z_i^{(1)}\right)$$

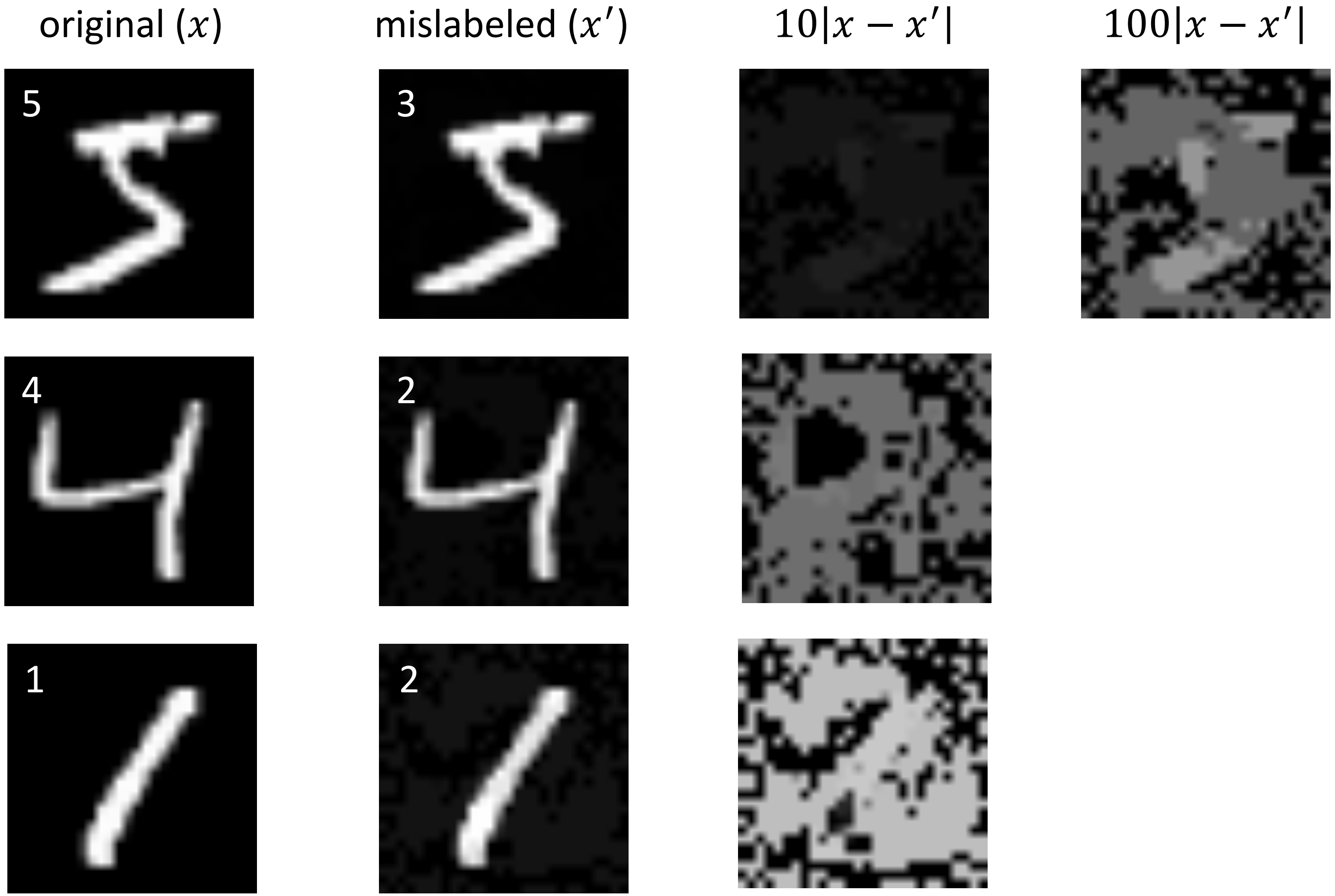


$$\cancel{\left(z_i^{(1)} \leq 0 \land x_i^{(2)} = 0\right)} \lor \left(z_i^{(1)} \geq 0 \land x_i^{(2)} = z_i^{(1)}\right)$$



~~$\left(z_i^{(1)} \leq 0 \wedge x_i^{(2)} = 0\right)$~~ $\vee \left(z_i^{(1)} \geq 0 \wedge x_i^{(2)} = z_i^{(1)}\right)$

MNIST



MNIST

Neural Net	Accuracy (%)	Adversarial Frequency (%)	
		Baseline	Our Algo.
LeNet (Original)	99.08	1.32	7.15
Baseline ($T = 1$)	99.14	1.02	6.89
Baseline ($T = 2$)	99.15	0.99	6.97
Our Algo. ($T = 1$)	99.17	1.18	5.40
Our Algo. ($T = 2$)	99.23	1.12	5.03

$\epsilon = 20$ pixels

Fine Tuning (Goodfellow 2015)

input: X_{train}
while **true:**
    $f =$ train neural network on X_{train}
    $X_{\text{adv}} =$ find adversarial examples
    $X_{\text{train}} = X_{\text{train}} \cup X_{\text{adv}}$

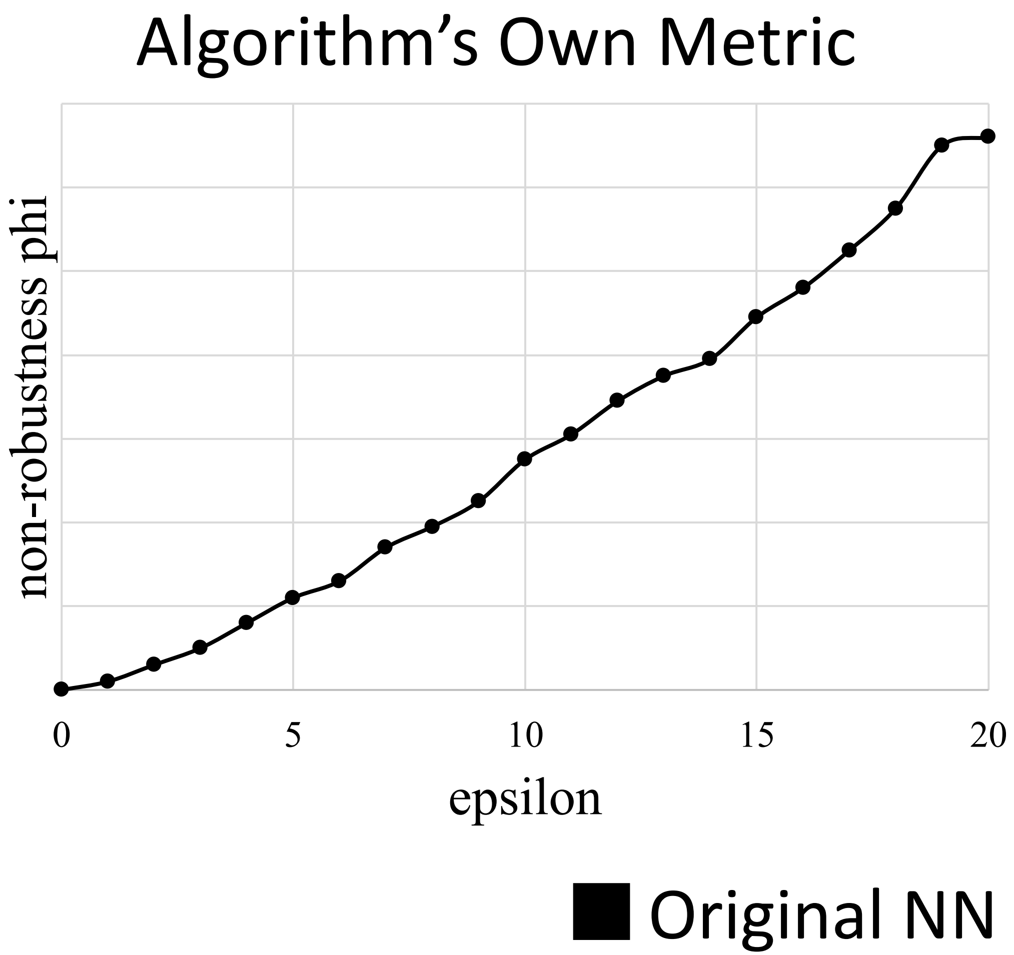
MNIST

Neural Net	**Accuracy (%)**	**Adversarial Frequency (%)**	
		Baseline	Our Algo.
LeNet (Original)	99.08	1.32	7.15
Baseline ($T = 1$)	99.14	1.02	6.89
Baseline ($T = 2$)	99.15	0.99	6.97
Our Algo. ($T = 1$)	99.17	1.18	5.40
Our Algo. ($T = 2$)	99.23	1.12	5.03

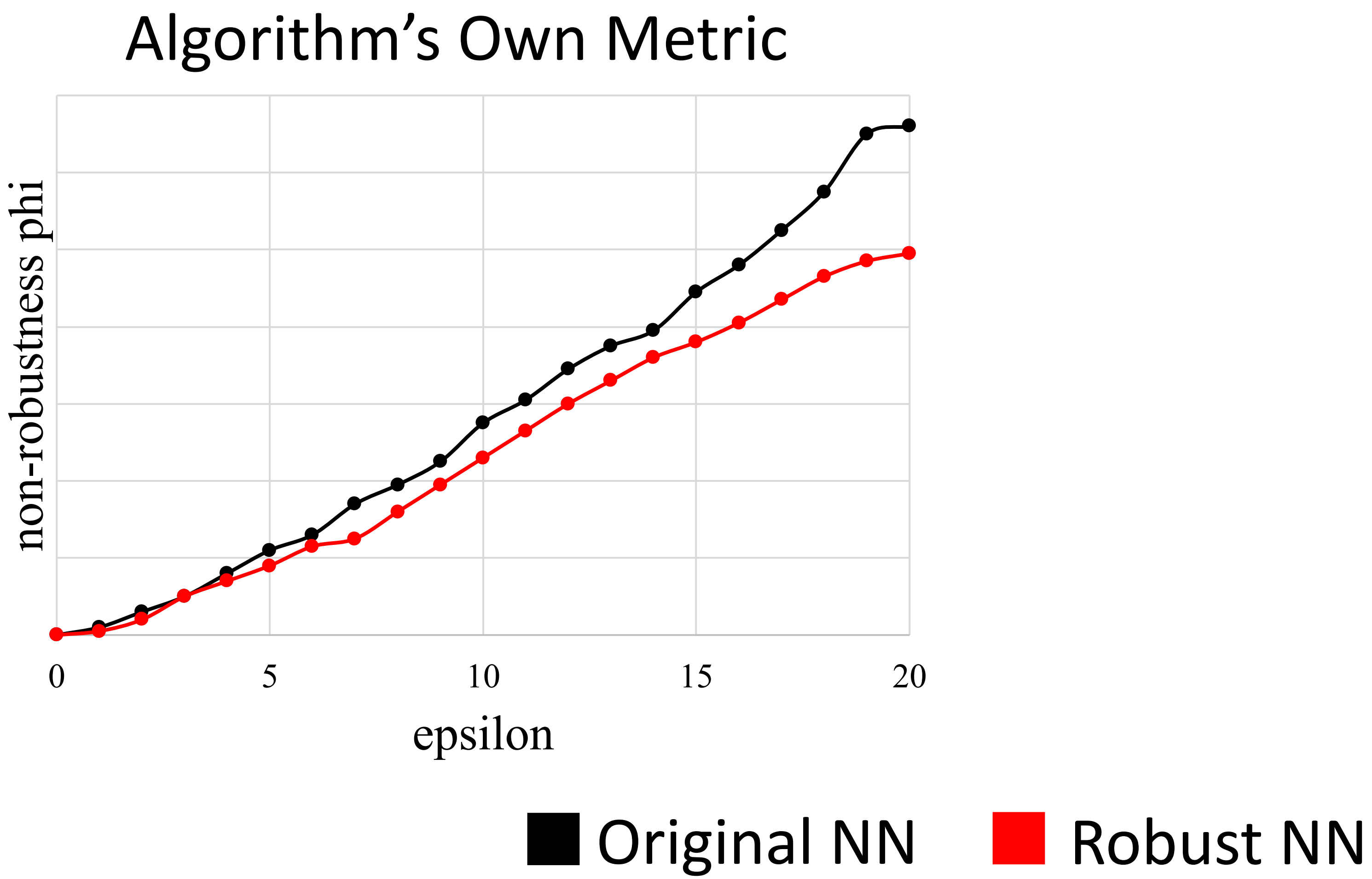
$\epsilon = 20$ pixels

Improving Robustness?

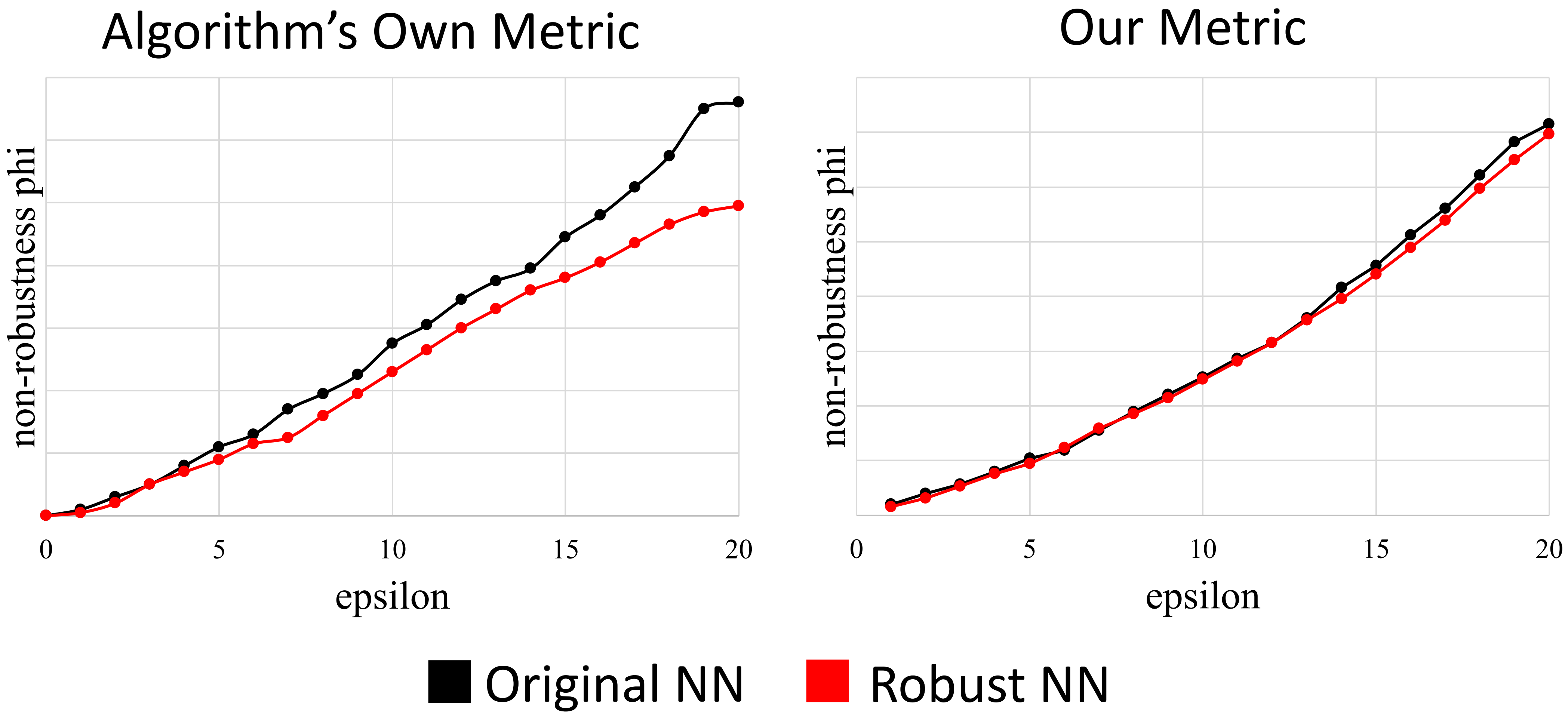
Improving Robustness?



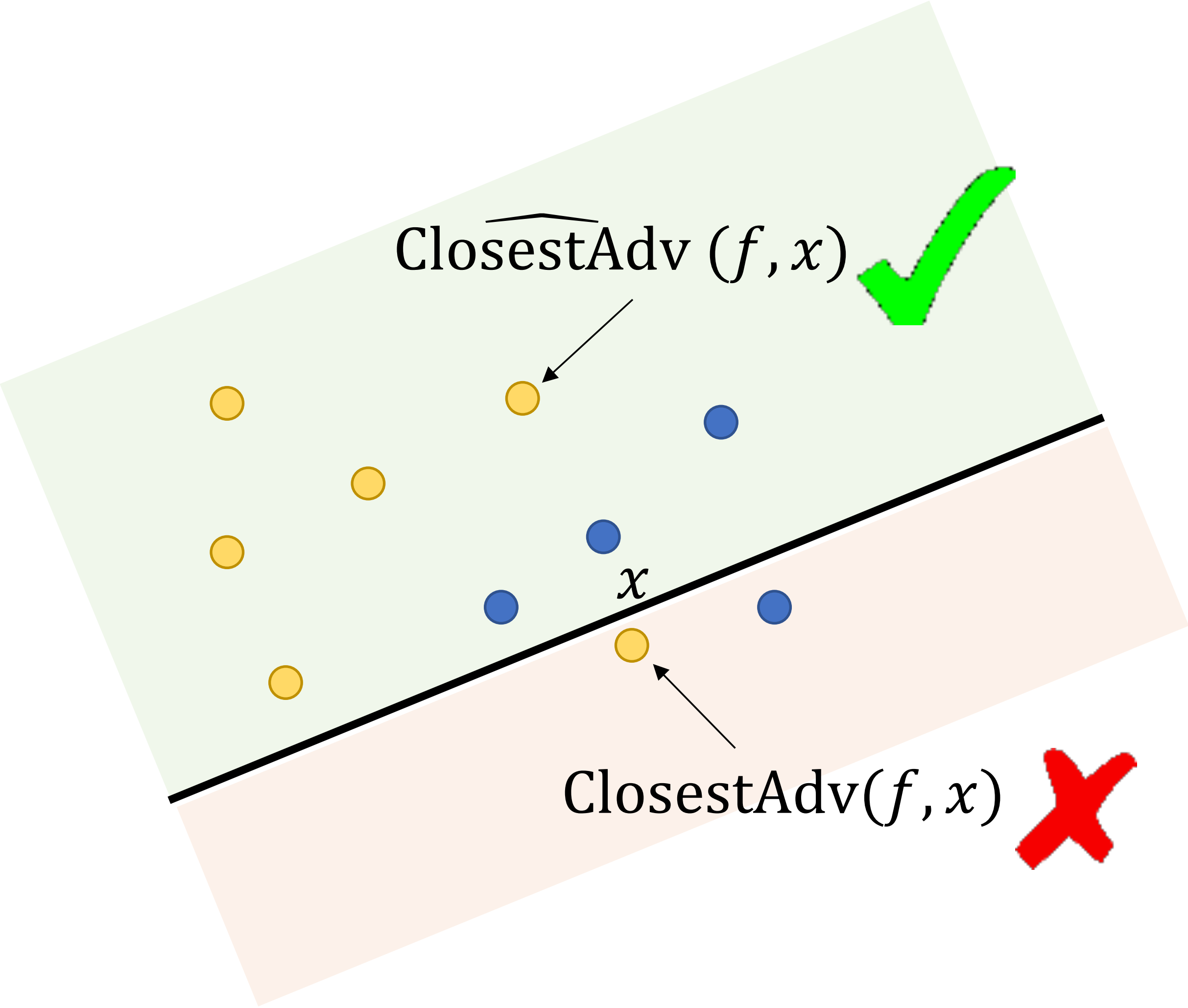
Improving Robustness?



Improving Robustness?



Improving Robustness?



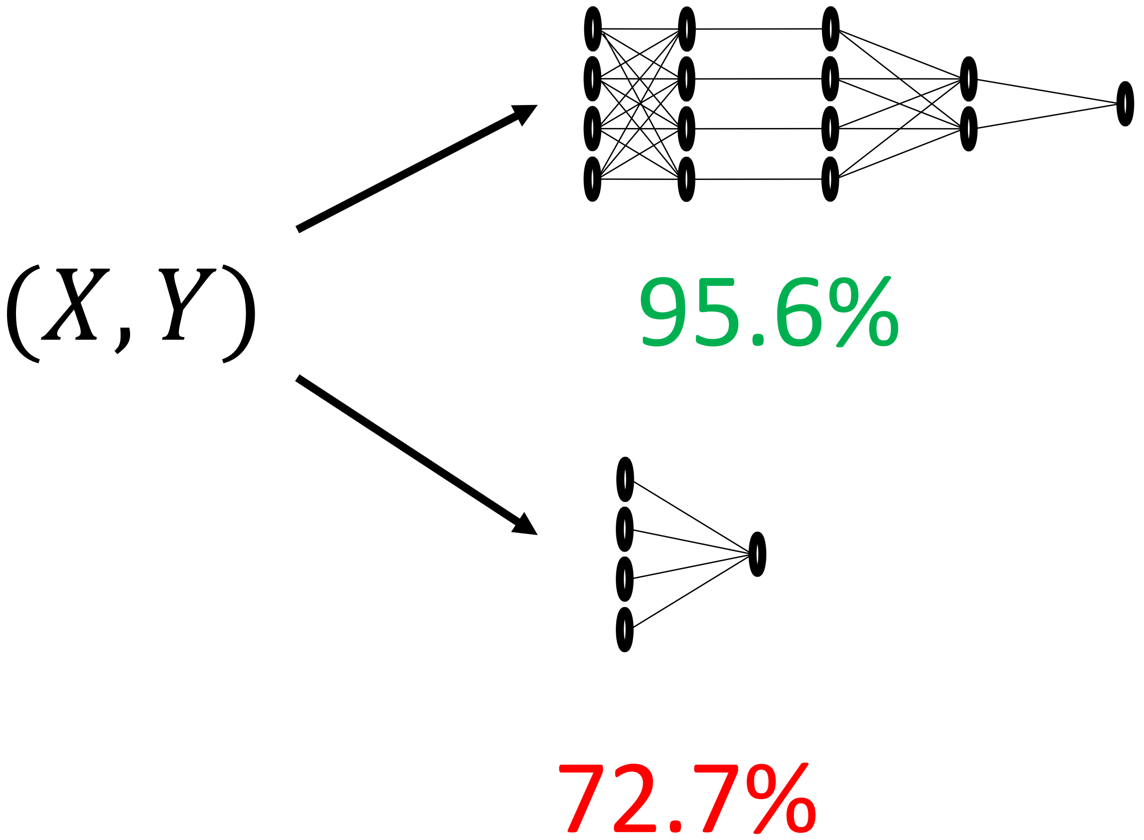
Related Work

- Finding adversarial examples
  - Szegedy et al. 2014, Goodfellow et al. 2015, …

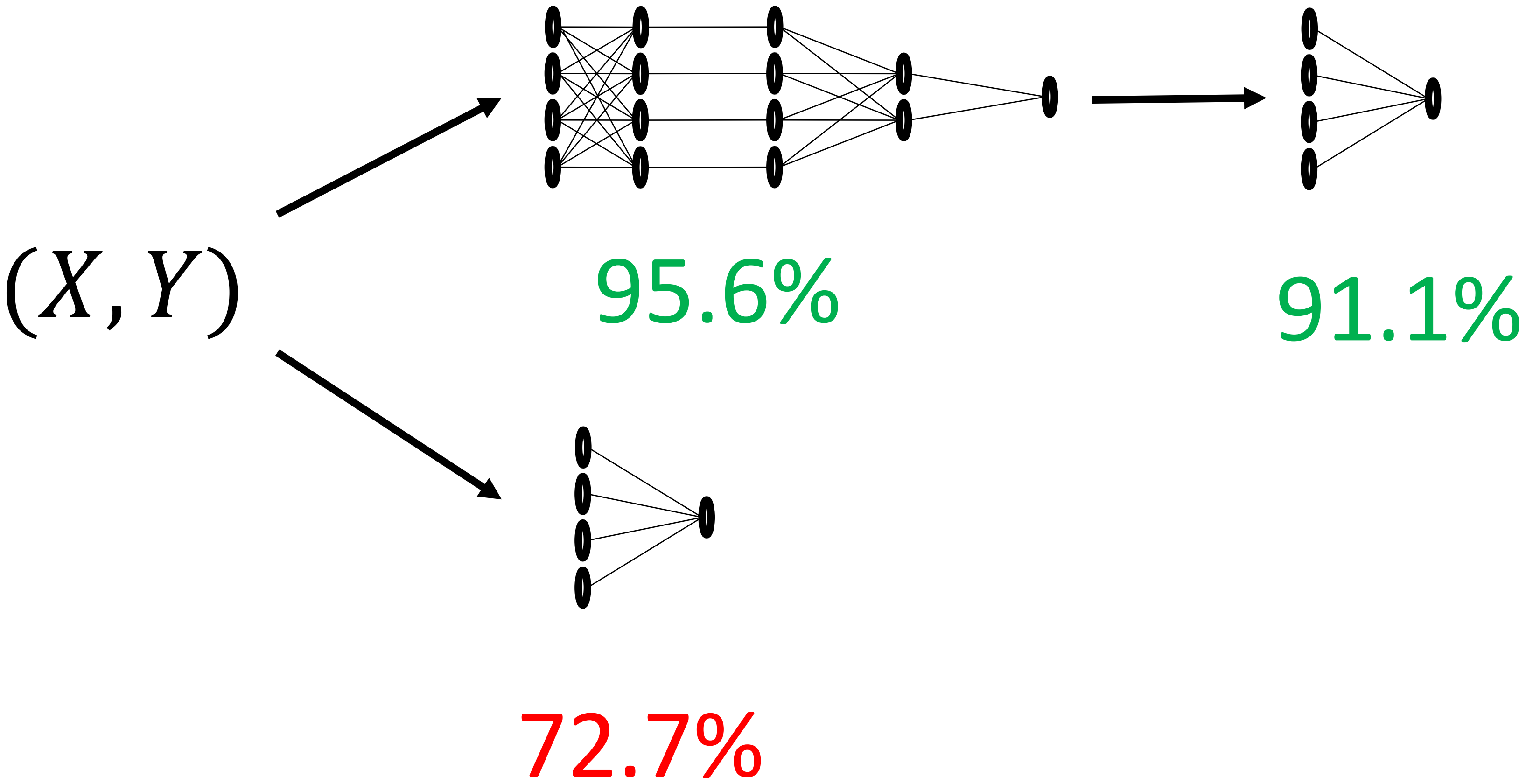
- Verifying robustness
  - Reluplex: SMT solver for theory of ReLU (Katz et al. 2017)
  - DLV: Discretize search space (Huang et al. 2017)
  - Mixed integer programs (Tjeng et al. 2017)
  - AI^2: Abstract interpretation (Gehr et al. 2018)

Verifiable Models via Model Compression (Ongoing)

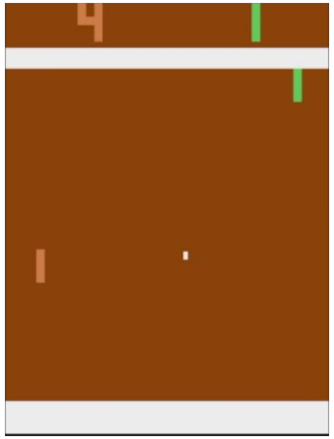
Osbert Bastani, Evan Pu, Armando Solar-Lezama



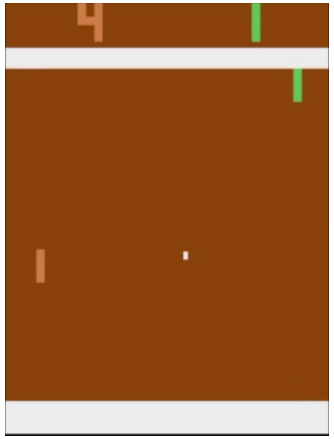
(Ba et al. 2014)

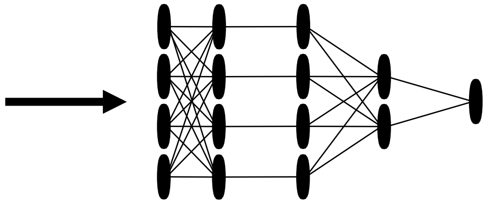


(Ba et al. 2014)



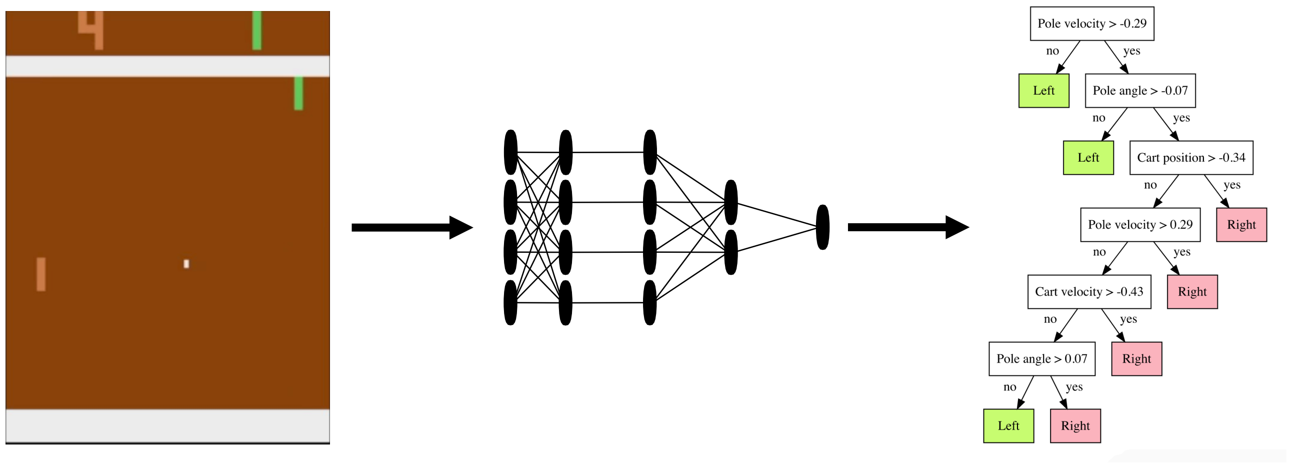
control problem





control problem

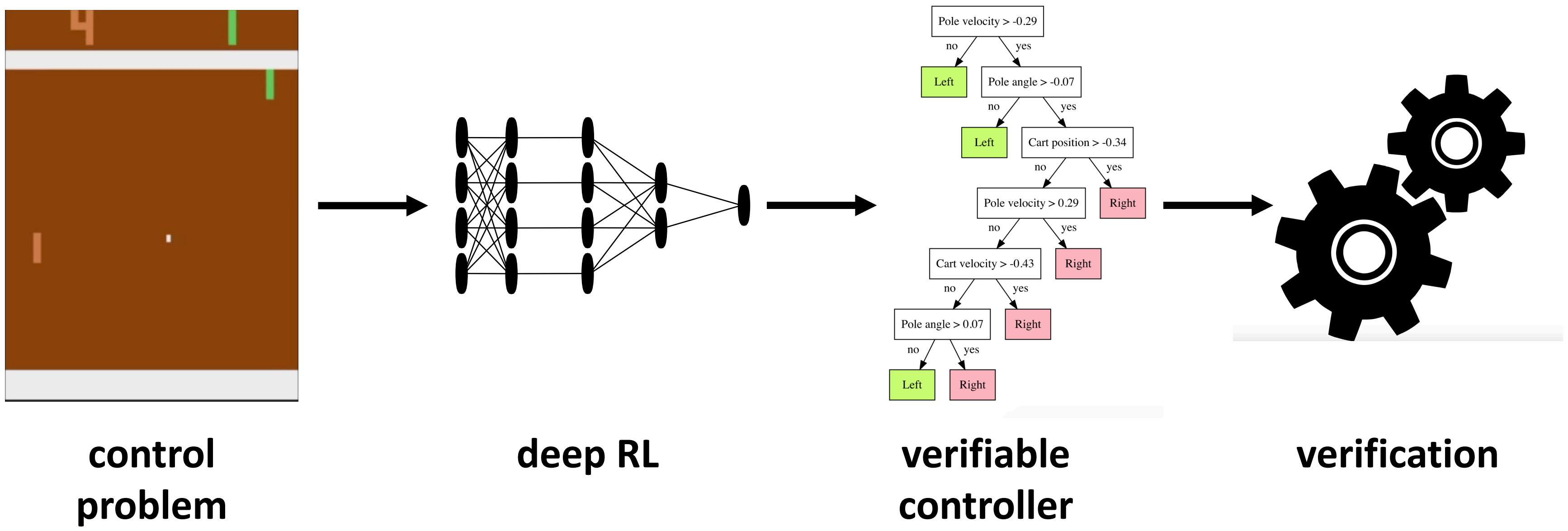
deep RL



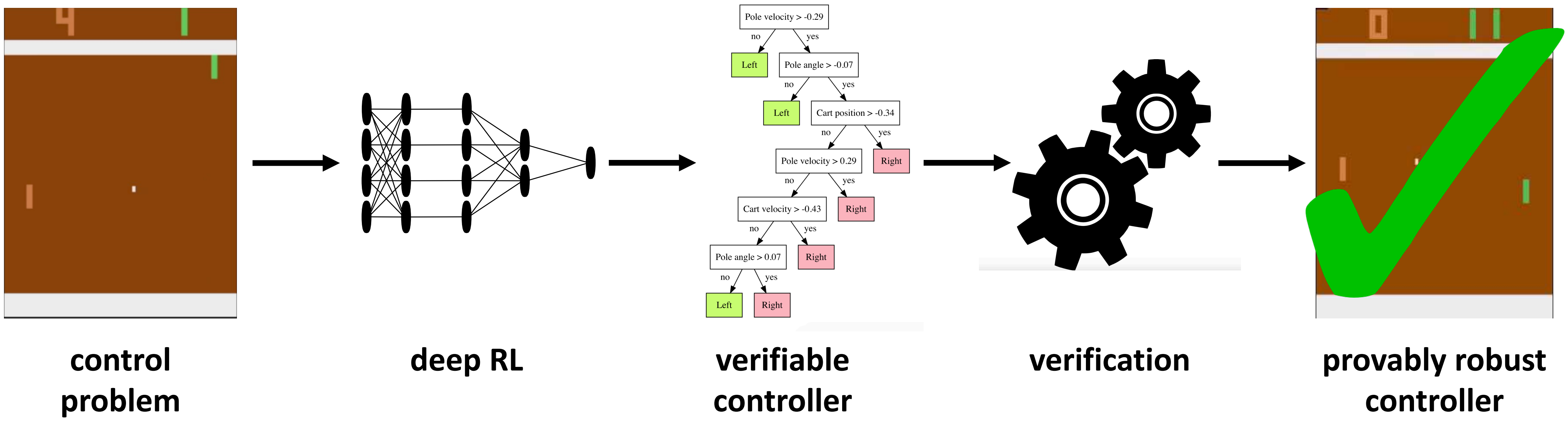
control problem

deep RL

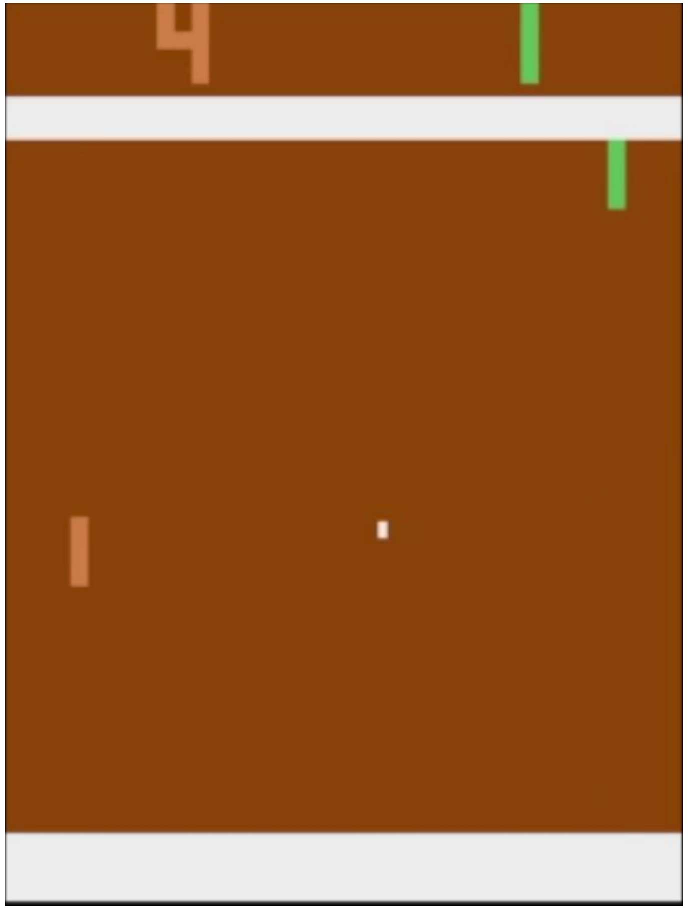
verifiable controller



Pole velocity > -0.29
no
yes
Left
Pole angle > -0.07
no
yes
Left
Cart position > -0.34
no
yes
Pole velocity > 0.29
Right
no
yes
Cart velocity > -0.43
Right
no
yes
Pole angle > 0.07
Right
no
yes
Left
Right
control problem
deep RL
verifiable controller
verification

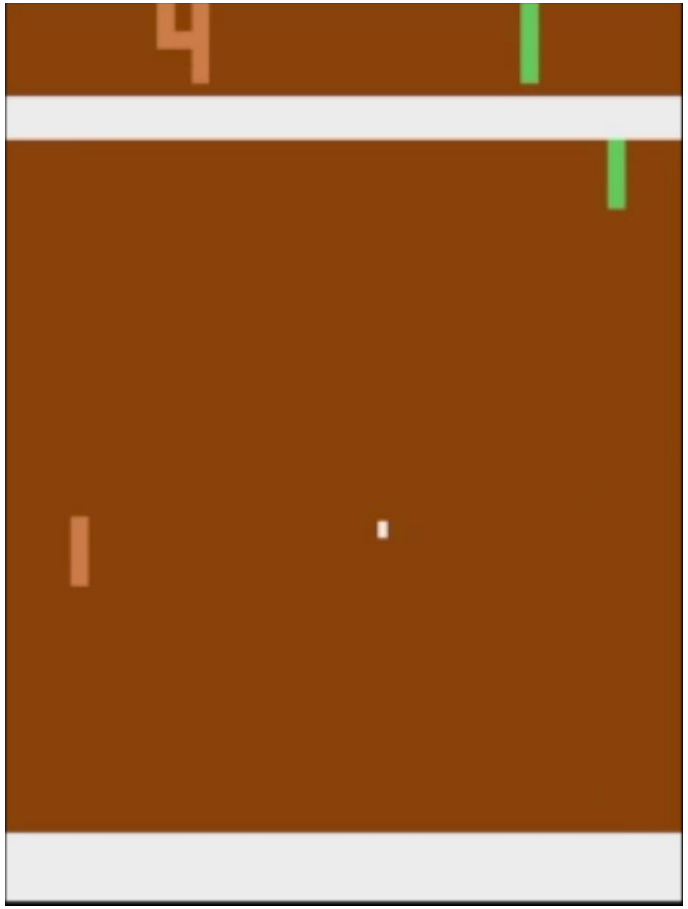

Pole velocity > -0.29
no
yes
Left
Pole angle > -0.07
no
yes
Left
Cart position > -0.34
no
yes
Pole velocity > 0.29
Right
no
yes
Cart velocity > -0.43
Right
no
yes
Pole angle > 0.07
Right
no
yes
Left
Right
control problem
deep RL
verifiable controller
verification
provably robust controller

- **Problem setup**
  - State space: 12 dimensional
  - Action space: Move paddle up/down

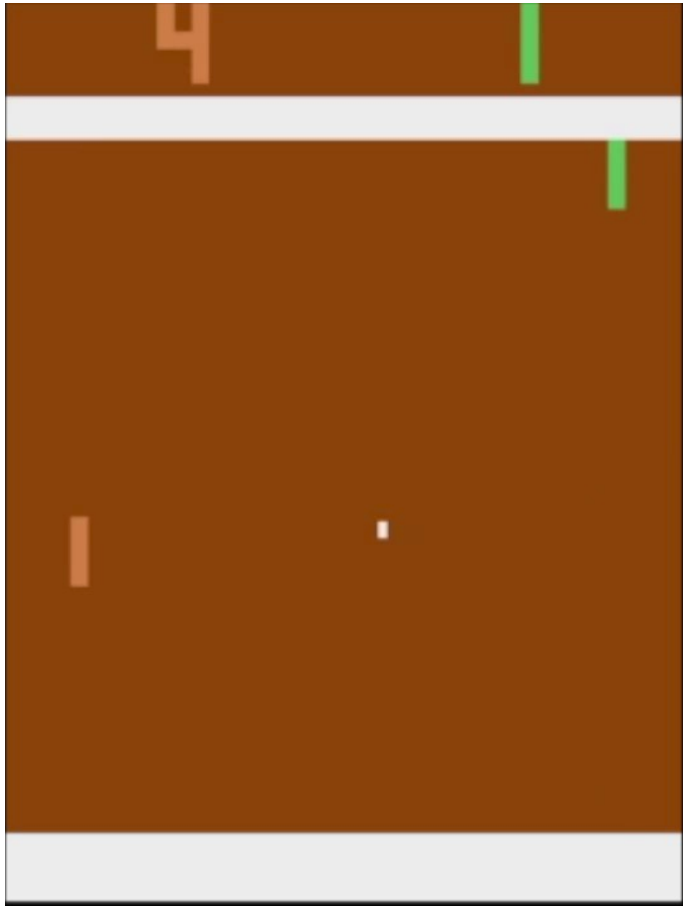


- **Problem setup**
  - State space: 12 dimensional
  - Action space: Move paddle up/down

- **Policy**
  - Decision tree
  - 4354 leaf nodes



- **Problem setup**
  - State space: 12 dimensional
  - Action space: Move paddle up/down
- **Policy**
  - Decision tree
  - 4354 leaf nodes
- **Verification (Robust at one point)**
  - 4354 linear program calls
  - 52.8 seconds



Verification for learning-based systems

- Robustness, stability, etc.