Towards Robust Neural Networks via Close-loop Control

Zhuotong Chen[1,ℇ], Qianxiao Li[2, 3, ℇ], Zheng Zhang[1]

[1] University of California, Santa Barbara,

[2]National University of Singapore

[3]Insitute of High Performance Computing, SG

([ℇ] Equal contribution)

Robustness Issues of Deep Neural Networks

- FGSM (Goodfellow 2014)
- PGD (Madry 2017)
- CW (Carlini & Wargner, 2017)
- Manifold attack (Jalal 2017)
- And many

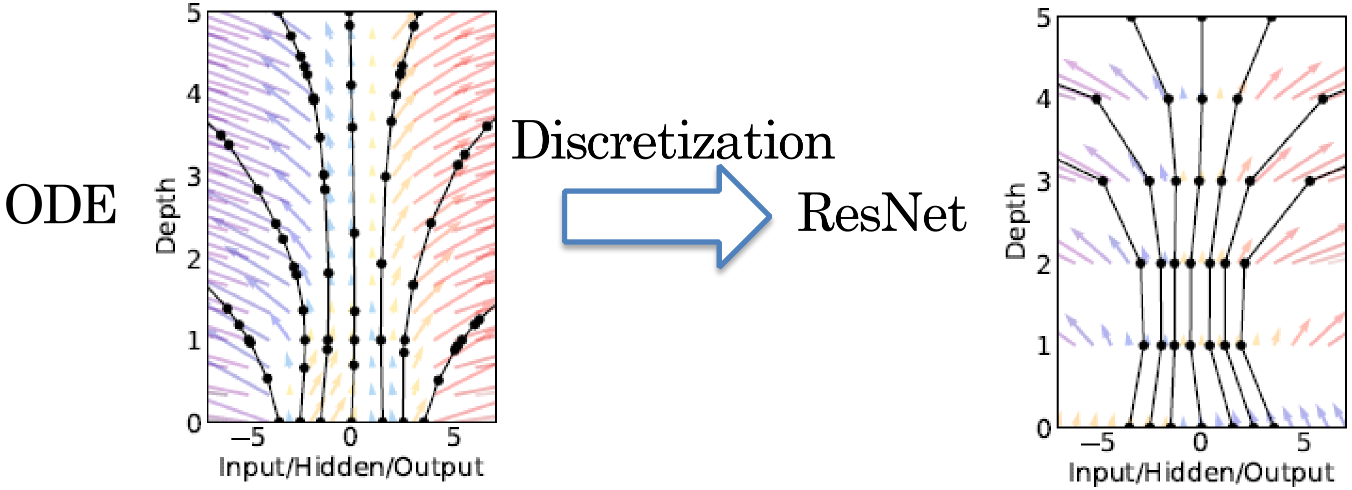
- Adversarial training (so many)
- Grading masking (Liu 2018)
- Data augmentation (Shorten 2019)
- Reactive defense (Metzen 2017, Song 2017)
- And many

- ❖ We propose a close-loop control method to improve robustness of neural networks
  - ❖ Define an objective function to connect close-loop method and neural network robustness
  - ❖ Numerical solver to obtain the solution

Dynamic System Perspective of DNN

- DNN as discretization of ODE (E 2017, Haber 2017, Chen 2018)

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}_t, t, \boldsymbol{\theta}_t) \qquad \mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{f}(\mathbf{x}_t, t, \boldsymbol{\theta}_t)$$

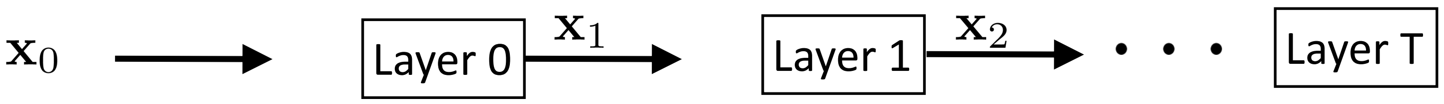


- DNN training as open-loop control (Li et al. 2017)
- Adversarial training formulated as robust open-loop control (Zhang et al. 2019)
- DNN training as trajectory optimization (Liu et al. 2020)

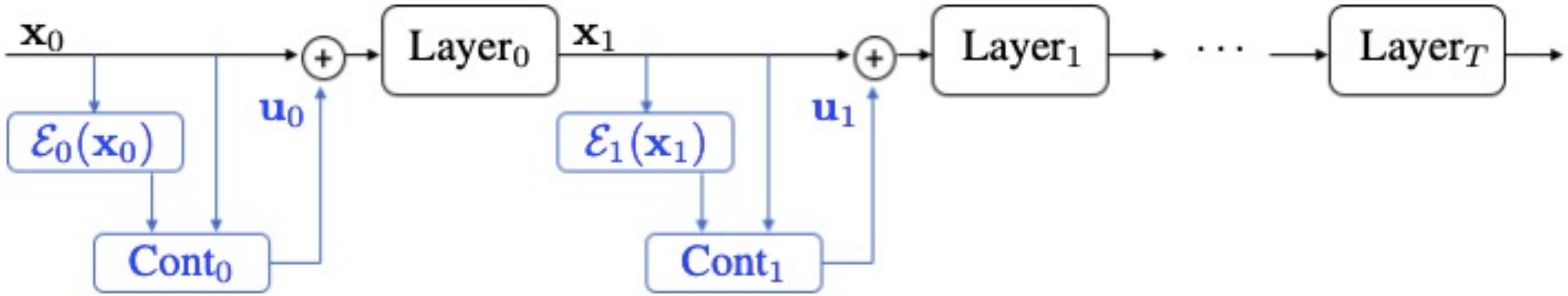
Close-loop Control for Robust Neural Networks

- Consider a feedforward network as discrete dynamic system

$$\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t, \boldsymbol{\theta}_t), \mathbf{x}_0 = \text{input data}, \text{ Label } \mathbf{y} = \Phi(\mathbf{x}_T)$$



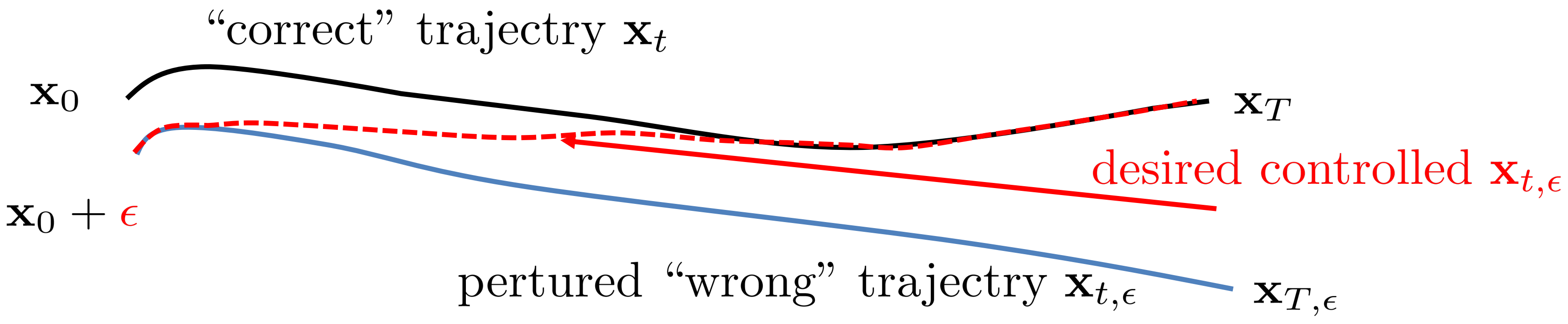
- DNN with close-loop controllers



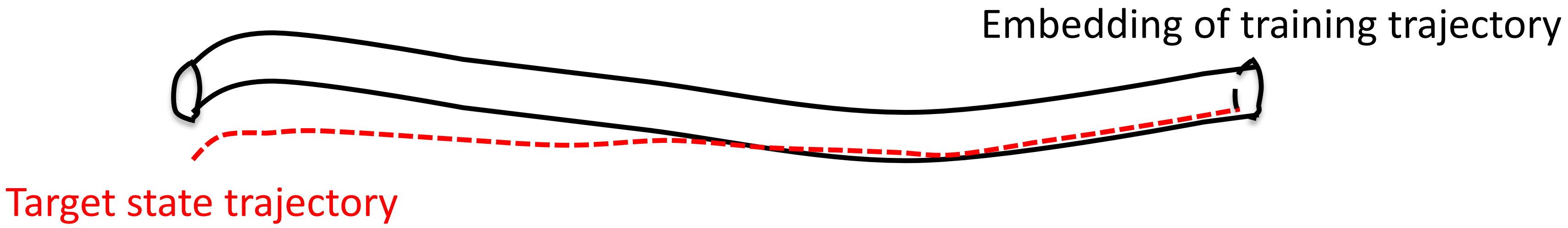
- Difference between open and close-loop control methods
  - Controls are adaptive for each input
  - The network parameters are not modified

Controller Design Based on Embedding

- Idea in classical trajectory optimization

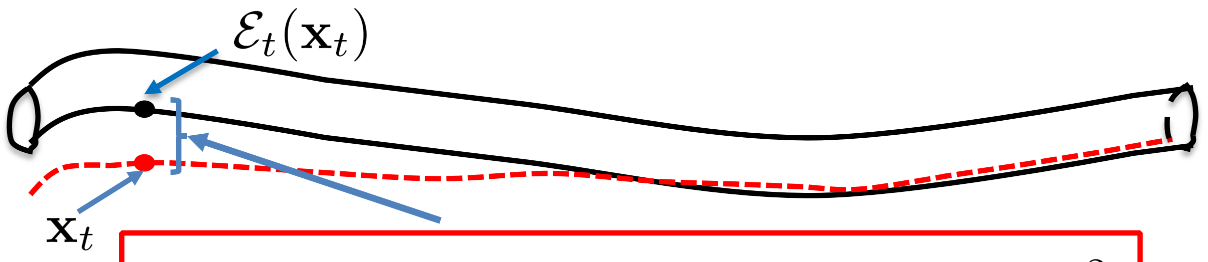


- We do NOT know the true trajectory (require true label)
- We control the manifold of clean trajectories of training dataset



Controller Design Based on Embedding

- Distance with the 'desired' manifold measured by



Reconstruction error $e_t = \|\mathcal{E}_t(\mathbf{x}_t) - \mathbf{x}_t\|^2$

- Running loss of every layer with regularization

$$\mathcal{L}(\mathbf{x}_t, \pi_t(\mathbf{x}_t), \mathcal{E}_t(\cdot)) = \|\mathcal{E}_t(\mathbf{x}_t) - \mathbf{x}_t\|^2 + \pi_t(\mathbf{x}_t)^T \mathbf{R} \pi_t(\mathbf{x}_t)$$

- Overall close-loop control loss:
  - Expected total loss of all layers except the last layer

$$\min_{\{\pi_t(\cdot)\}_{t=0}^{T-1}} \mathbb{E}_{(\mathbf{x}_0, \mathbf{y}) \sim \mathcal{D}} \sum_{t=0}^{T-1} \mathcal{L}(\mathbf{x}_t, \pi_t(\mathbf{x}_t), \mathcal{E}_t(\mathbf{x}_t))$$

Numerical Solver

- Solve the close-loop control objective function is hard
  - Requires solving a super high-dim PDE
- Instead, compute a specific control signal for a given possible perturbed data sample
- Pontryagin's Maximum Principle
- Iterate the following steps:
  - Forward propagation of x_t
  - Backward propagation of p_t
  - Optimize over u_t to maximize the Hamiltonian

Numerical Results

- Control result of a standard trained ResNet-20
  - CLC-NN + Linear: Proposed method with linear embedding
  - CLC-NN + nonlinear: Proposed method with auto-encoder embedding

Dataset	ϵ	Accuracy: original model without CLC / CLC-NN + Linear / CLC-NN + Nonlinear				
		Type of input perturbations				
		None	Manifold	FGSM	PGD	CW
CIFAR-10	2	92 / 88 / 89	24 / 79 / 82	21 / 56 / 56	0 / 50 / 50	8 / 75 / 79
	4		5 / 78 / 81	11 / 40 / 30	0 / 31 / 19	0 / 75 / 79
	8		1 / 78 / 81	8 / 20 / 12	0 / 11 / 2	0 / 76 / 79
CIFAR-100	2	69 / 60 / 58	9 / 51 / 52	9 / 25 / 23	0 / 17 / 22	4 / 47 / 49
	4		3 / 50 / 52	5 / 15 / 9	0 / 6 / 4	1 / 47 / 49
	8		2 / 50 / 52	4 / 9 / 5	0 / 1 / 0	0 / 47 / 49

- Comparison with Reactive Defense (linear embedding)
  - + means outperform, - means underperform

Method	Type of input perturbations				
	None	Manifold	FGSM	PGD	CW
CIFAR-10	-3	+47 / +63 / +66	+27 / +20 / +13	+43 / +35 / +25	+66 / +76 / +77
CIFAR-100	+1	+34 / +37 / +38	+22 / 0 / +9	+44 / +30 / +11	+37 / +30 / +16