Deep neural networks, universal approximation, and nonlinear geometric control.

Paulo Tabuada [1] and Bahman Gharesifard [2]

[1]Vijay K. Dhir Professor of Engineering
Cyber-Physical Systems Laboratory
Department of Electrical and Computer Engineering
University of California at Los Angeles

[2]Department of Mathematics and Statistics
Queen's University

Control system models of ResNets

- We model deep Residual Networks (ResNets) as control systems:

$$\dot{x} = s\Sigma(Wx + b), \tag{1}$$

  with state $x \in \mathbb{R}^n$ and where $(s, W, b) \in \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ are regarded as control inputs.

- Any property of (1) will hold for a (finite-depth) ResNet by time-discretizing the solutions of (1):

  - **A proposal on machine learning via dynamical systems**
    E. Weinan, Communications in Mathematics and Statistics, 5, 2017.
    Stable architectures for deep neural networks
    E. Haber and L. Ruthotto, Inverse Problems, 34(1), 2017.
    Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations
    Y. Lu, A. Zhong, Q. Li, and B. Dong, International Conference on Machine Learning, 2018.

Memorization Capabilities of Residual Neural Networks

CyPhyLab UCLA

Memorization capabilities of ResNets

Problem definition

Problem (Memorization)

Given:

- *a function* $f : E \to \mathbb{R}^n$ *defined on a compact set* $E \subset \mathbb{R}^n$*,*
- *a finite set* $E_{samples} \subset E$*,*
- *the evaluation of* f *on* $E_{samples}$*, i.e.,* $f(x)$ *for each* $x \in E_{samples}$*,*

does there exist a ResNet outputing $f(x)$ *for each input* $x \in E_{samples}$*?*

Memorization capabilities of ResNets

Problem definition

Problem (Memorization)

Given:

- *a function* $f : E \to \mathbb{R}^n$ *defined on a compact set* $E \subset \mathbb{R}^n$*,*
- *a finite set* $E_{samples} \subset E$*,*
- *the evaluation of* f *on* $E_{samples}$*, i.e.,* $f(x)$ *for each* $x \in E_{samples}$*,*

does there exist a ResNet outputing $f(x)$ *for each input* $x \in E_{samples}$*?*

Problem (Memorization)

Does there exist an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n \times n} \times \mathbb{R}^n$ *so that the solution* ξ *of:*

$$\dot{x} = s\Sigma(Wx + b),$$

satisfies $\xi(0) = x$ *and* $\xi(\tau) = f(x)$ *for every* $x \in E_{samples}$*.*

- Ingredients:
  - Lie algebraic controllability techniques;
  - New arguments to handle the ensemble aspect fo the problem.

- Ingredients:
  - Lie algebraic controllability techniques;
  - New arguments to handle the ensemble aspect fo the problem.

Theorem (Controllability)

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then the ensemble control system is controllable everywhere except on a sub-manifold of positive co-dimension.

Memorization capabilities of ResNets

Main memorization result

- Ingredients:
  - Lie algebraic controllability techniques;
  - New arguments to handle the ensemble aspect fo the problem.

Theorem (Controllability)

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then the ensemble control system is controllable everywhere except on a sub-manifold of positive co-dimension.

- Almost any finite data set $\{(x, f(x))\}_{x\in E_{\text{samples}}}$ can be memorized by a deep ResNet with n neurons per layer.

- Ingredients:
  - Lie algebraic controllability techniques;
  - New arguments to handle the ensemble aspect fo the problem.

Theorem (Controllability)

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then the ensemble control system is controllable everywhere except on a sub-manifold of positive co-dimension.

- Almost any finite data set $\{(x, f(x))\}_{x\in E_{\text{samples}}}$ can be memorized by a deep ResNet with n neurons per layer.
- If a finite set data set $\{(x, f(x))\}_{x\in E_{\text{samples}}}$ cannot be memorized, then any arbitrarily small perturbation of it can be memorized.

Memorization capabilities of ResNets

Activation functions

- Are there injective activation functions σ satisfying $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$?

Function name	Definition	Satisfied differential equation
Logistic function	$\sigma(x) = \frac{1}{1+e^{-x}}$	$D\sigma - \sigma + \sigma^2 = 0$
Hyperbolic tangent	$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$D\sigma - 1 + \sigma^2 = 0$
Soft plus	$\sigma(x) = \frac{1}{r}\log(1 + e^{rx})$	$D^2\sigma - rD\sigma + r(D\sigma)^2 = 0$

- Moreover, $\lim_{r\to\infty} \frac{1}{r}\log(1 + e^{rx}) = \text{ReLU}(x) = \max\{0, x\}$.
- Leaky ReLUs can be similarly handled.

Approximation Capabilities of Residual Neural Networks

CyPhyLab UCLA

Problem (Approximation)

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ *be a continuous function and* $E \subset \mathbb{R}^n$ *a compact set. Given a desired accuracy* $\varepsilon \in \mathbb{R}^+$*, do there exist inputs* $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n \times n} \times \mathbb{R}^n$ *resulting in a solution* ξ_x *satisfying* $\xi_x(0) = x$ *and:*

$$\sup_{x \in E} \|f(x) - \xi_x(\tau)\|_\infty \leq \varepsilon.$$

CyPhyLab UCLA

- Ingredients:
  - Previous controllability result;
  - Monotonicity.

- Special case: f is a monotone analytic function.

- Ingredients:
  - Previous controllability result;
  - Monotonicity.

- Special case: f is a monotone analytic function.

Theorem

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then, for every monotone analytic function $f : \mathbb{R}^n \to \mathbb{R}^n$, for every compact set $E \subset \mathbb{R}^n$, and for every $\varepsilon \in \mathbb{R}^+$ there exists an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{n\times n} \times \mathbb{R}^n$ so that the resulting solution ξ_x satisfies $\xi_x(0) = x$ and:

$$\sup_{x\in E} \|f(x) - \xi_x(\tau)\|_\infty \leq \varepsilon.$$

Approximation capabilities of ResNets

Main approximation result

- Ingredients:
  - Previous controllability result;
  - Monotonicity.

- General case: embedded $f : \mathbb{R}^n \to \mathbb{R}^n$ into a monotone function $\tilde{f} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$.

Corollary

Let $n > 1$ and assume the activation function σ is injective, non-negative, and satisfies $D\sigma = a_0 + a_1\sigma + a_2\sigma^2$ for some $a_2 \neq 0$. Then, for every continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$, for every compact set $E \subset \mathbb{R}^n$, and for every $\varepsilon \in \mathbb{R}^+$ there exists an input $(s, W, b) : [0, \tau] \to \mathbb{R} \times \mathbb{R}^{(n+1)\times(n+1)} \times \mathbb{R}^{n+1}$ so that the resulting solution ξ_x satisfies $\xi_x(0) = x$ and:

$$\sup_{x \in E} \|f(x) - \beta \circ \xi_{\alpha(x)}(\tau)\|_\infty \leq \varepsilon.$$