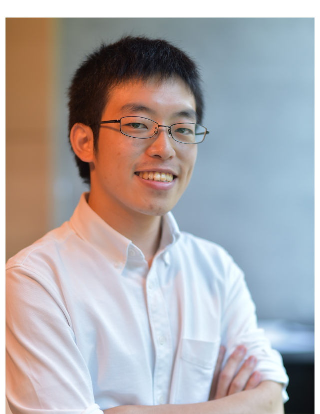
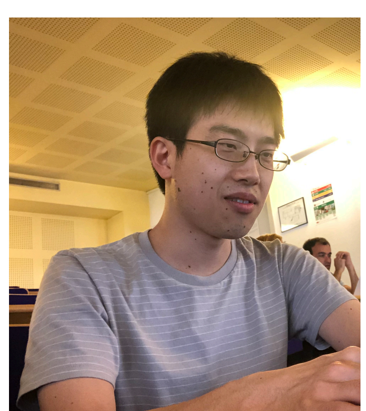
NEURAL INFORMATION
PROCESSING SYSTEMS

Universal Approximation Property of Neural Ordinary Differential Equations

Takeshi Teshima[1,2], Koichi Tojo[2],
Masahiro Ikeda[2], Isao Ishikawa[3,2], Kenta Oono[1]

[1]The University of Tokyo, Japan [2]RIKEN, Japan [3]Ehime University, Japan

NeurIPS2020 Workshop
Differential Geometry meets Deep Learning (DiffGeo4DL)

Take-home message

What did we do?

Theoretically investigated:
How expressive are NODEs?

NODE = Neural Ordinary Differential Equations
[CRBD18]

What is the result?

Universality of NODE + Affine transform for a large class of diffeomorphisms w.r.t. $\sup$-norm.

Why important?

- **Strong ($\sup$-norm) guarantee for a large class of invertible maps.**
- **cf. Previous result: Universality for $C^0(\mathbb{R}^n, \mathbb{R}^m)$ w.r.t. L^p-norm.** [LLS20]

Message

NODE-based invertible neural networks have guaranteed representation power for approximating diffeomorphisms.

Model: Neural Ordinary Differential Equations

NODE layer

$$\mathrm{Lip}(\mathbb{R}^d) := \{f\colon \mathbb{R}^d \to \mathbb{R}^d \mid f \text{ is Lipschitz}\}$$

For each $f \in \mathrm{Lip}(\mathbb{R}^d)$, we define an invertible map $\mathbf{x} \mapsto \mathbf{z}(1)$ via an initial value problem [DJ76]

$\mathbf{x}$ → Solve for $\mathbf{z}(t)$:
$$\begin{cases} \mathbf{z}(0) = \mathbf{x}, \\ \dot{\mathbf{z}}(t) = f(\mathbf{z}(t)) \ (t \in \mathbb{R}). \end{cases}$$
→ $\mathbf{z}(1)$

NODE layers [CRBD18]

Then, for $\mathscr{H} \subset \mathrm{Lip}(\mathbb{R}^d)$, consider the set of NODEs:

$$\mathrm{NODEs}(\mathscr{H}) := \{\mathbf{x} \mapsto \mathbf{z}(1) \mid f \in \mathscr{H}\}$$

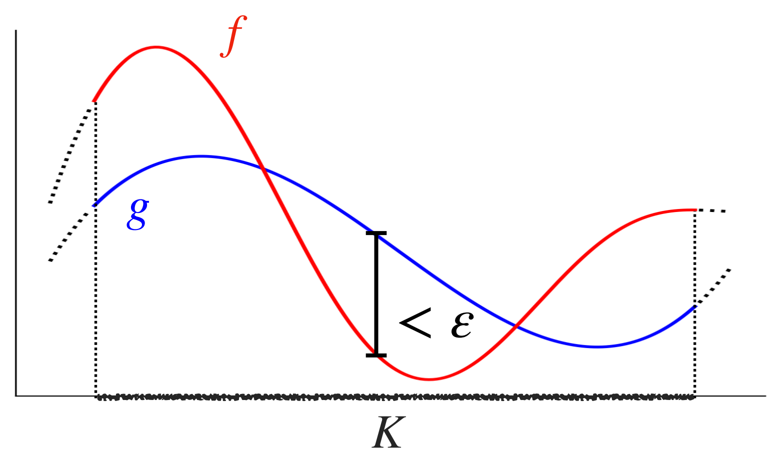
Model (composition of NODEs and affine transform)

$$\mathrm{INN}_{\mathscr{H}\text{-NODE}} := \{W \circ \psi_k \circ \cdots \circ \psi_1 \mid \psi_1, \ldots, \psi_k \in \mathrm{NODEs}(\mathscr{H}), W \in \mathrm{Aff}, k \in \mathbb{N}\}$$

Main result

Definition (Universality) [C89,HSW89]

sup-**universal approximator**: the model can approximate any target function w.r.t. sup-norm on a compact set.



Definition (Approximation target $\mathscr{D}^2$)

Fairly **large set** of smooth invertible maps.

$$\mathscr{D}^2 := \{C^2\text{-diffeo of the form } f : U_f \to f(U_f)\}$$

$(U_f \subset \mathbb{R}^d$: open C^2-diffeo to $\mathbb{R}^d)$

$\mathscr{D}^2$

General

$\{C^2\text{-diffeo on } \mathbb{R}^d\}$

...and more

Theorem

$d \geq 2$

If $\mathscr{H}$ is a sup-universal approximator for $\mathrm{Lip}(\mathbb{R}^d)$, then $\mathrm{INN}_{\mathscr{H}\text{-NODE}}$ is a sup-universal approximator for $\mathscr{D}^2$.

Ex. for $\mathscr{H}$: multi-layer perceptron [LBH15], Lipschitz Networks [ALG19].

Take-home message

What did we do?

Theoretically investigated:
How expressive are NODEs?

NODE = Neural Ordinary Differential Equations [CRBD18]

What is the result?

Universality of NODE + Affine transform for a large class of diffeomorphisms w.r.t. $\sup$**-norm.**

Why important?

- **Stronger (**$\sup$**-norm) guarantee for a smaller but large class of maps,**
- **cf. previous result: Universality for** $C^0(\mathbb{R}^n, \mathbb{R}^m)$ **w.r.t.** L^p**-norm.** [LLS20]

Message

NODE-based invertible neural networks have guaranteed representation power for approximating diffeomorphisms.

References

[C89] Cybenko, G. (1989).
Approximation by superpositions of a sigmoidal function.
Mathematics of Control, Signals, and Systems, 2, 303–314.

[HSW89] Hornik, K., Stinchcombe, M., & White, H. (1989).
Multilayer feedforward networks are universal approximators.
Neural Networks, 2(5), 359–366.

[CRBD18] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. (2018).
Neural ordinary differentialequations.
Advances in Neural Information Processing Systems 31, 6571–6583.

[LLS20] Q. Li, T. Lin, and Z. Shen. (2020).
Deep learning via dynamical systems: an approximation perspective.
arXiv:1912.10382 [cs, math, stat].

[DJ76] W. Derrick and L. Janos. (1976).
A global existence and uniqueness theorem for ordinary differentialequations.
Canadian Mathematical Bulletin, 19(1), 105–107.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. (2015).
Deep learning.
Nature, 521(7553), 436–444.

[ALG19] C. Anil, J. Lucas, and R. Grosse. (2019).
Sorting out Lipschitz function approximation.
Proceedings of the 36th International Conference on Machine Learning, PMLR 97, 291–301.