ICML | 2022
Thirty-ninth International Conference on
Machine Learning

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Towards Theoretical Analysis of Transformation Complexity of ReLU DNNs





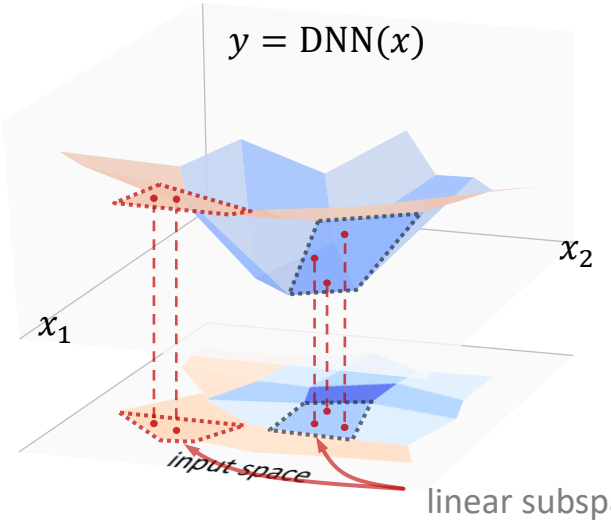
Jie Ren[1*], Mingjie Li[1*], Meng Zhou[2], Shih-Han Chan[3], Quanshi Zhang[1]

[1]Shanghai Jiao Tong University
[2]Carnegie Mellon University
[3]University of California San Diego

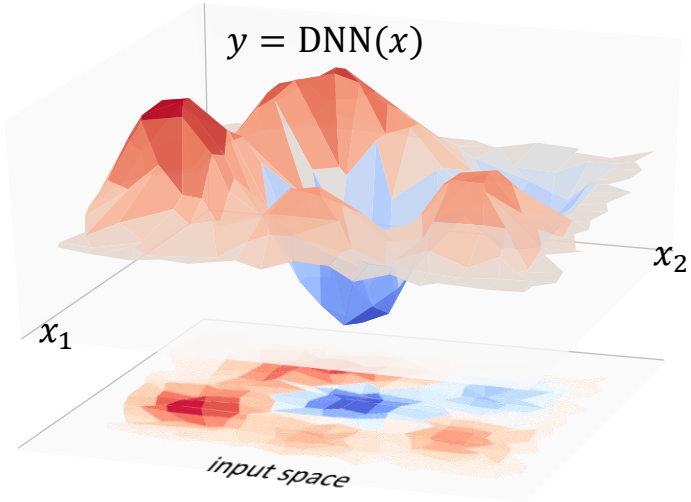
Overview

This study analyzes ***the diversity of transformations*** in piecewise linear DNNs.

A **simple** DNN
*(**low** complexity)*



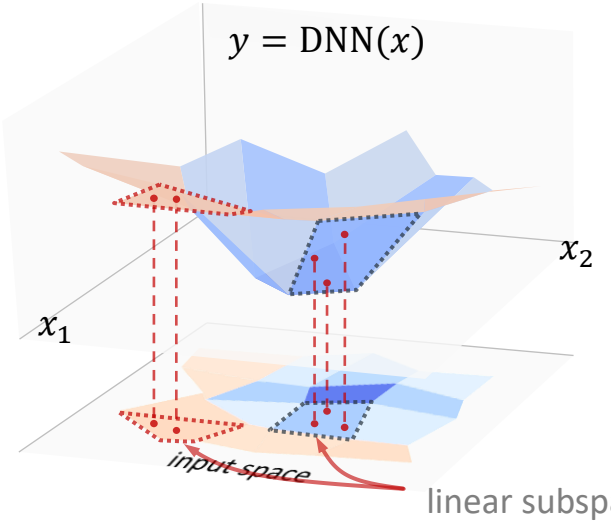
A **complex** DNN
*(**high** complexity)*



Overview

This study analyzes ***the diversity of transformations*** in piecewise linear DNNs.

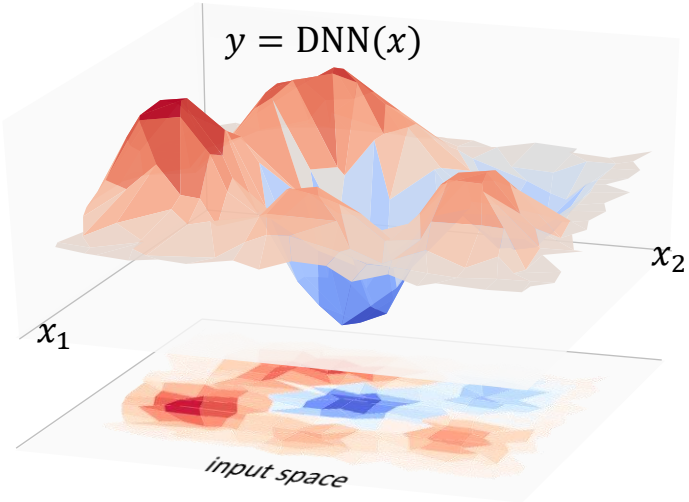
A **simple** DNN
*(**low** complexity)*



Overview

This study analyzes ***the diversity of transformations*** in piecewise linear DNNs.

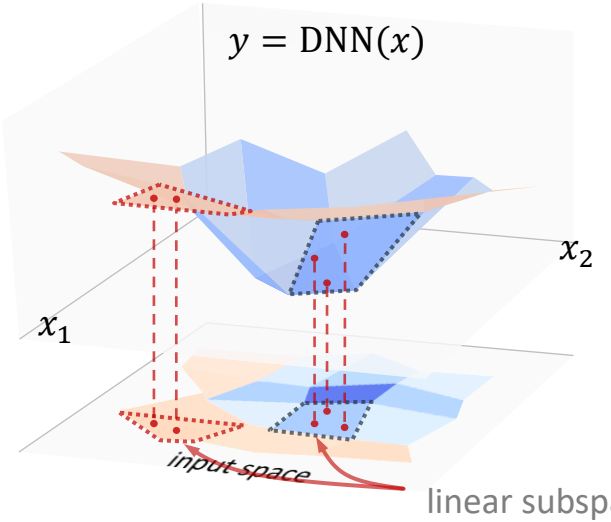
A **complex** DNN
*(**high** complexity)*



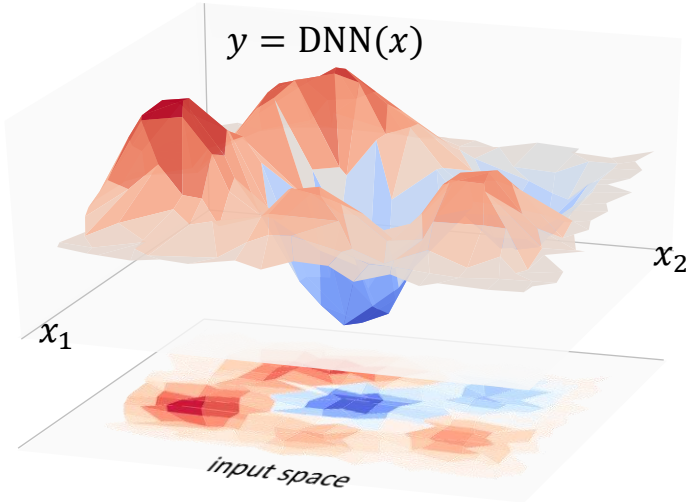
Overview

This study analyzes ***the diversity of transformations*** in piecewise linear DNNs.

A **simple** DNN
*(**low** complexity)*



A **complex** DNN
*(**high** complexity)*



Gating states in piecewise linear DNNs

The diversity of gating states determines the complexity of a DNN.

- Given a piecewise linear DNN, the mapping from x to y:

$$y = g(z_{L+1}),$$

$$z_{L+1} = W_{L+1}\boldsymbol{\sigma}_L(\dots\boldsymbol{\sigma}_2(W_2\boldsymbol{\sigma}_1(W_1x + b_1) + b_2)\dots) + b_{L+1}$$

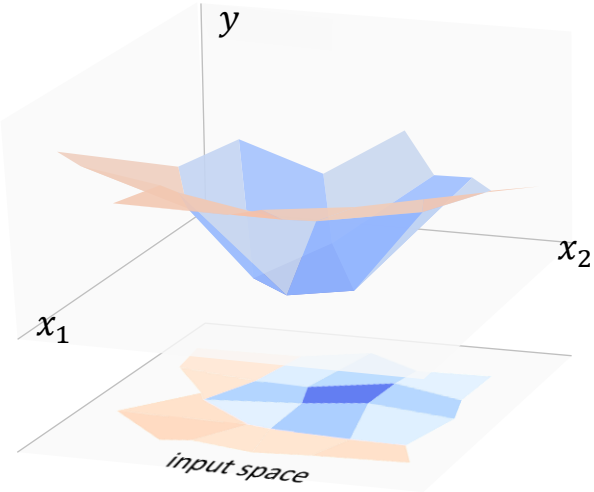
gating states

E.g., for ReLU layers, $\boldsymbol{\sigma}_l = diag(\sigma_{l,1}, \dots, \sigma_{l,d})$, $\sigma_{l,i} = \begin{cases} 1, z_{l,i} \geq 0 \\ 0, z_{l,i} < 0 \end{cases}$

Different gating states lead to different transformations.

A ***simple*** DNN

low diversity of gating states



Gating states in piecewise linear DNNs

The diversity of gating states determines the complexity of a DNN.

- Given a piecewise linear DNN, the mapping from x to y:

$$y = g(z_{L+1}),$$

$$z_{L+1} = W_{L+1}\boldsymbol{\sigma}_L(\dots\boldsymbol{\sigma}_2(W_2\boldsymbol{\sigma}_1(W_1x + b_1) + b_2)\dots) + b_{L+1}$$

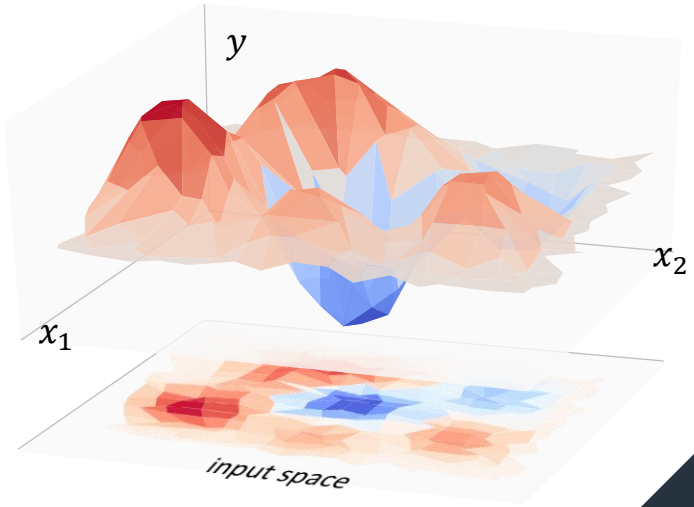
gating states

E.g., for ReLU layers, $\boldsymbol{\sigma}_l = diag(\sigma_{l,1}, \dots, \sigma_{l,d})$, $\sigma_{l,i} = \begin{cases} 1, z_{l,i} \geq 0 \\ 0, z_{l,i} < 0 \end{cases}$

Different gating states lead to different transformations.

A ***complex*** DNN

high diversity of gating states



Definitions of three complexity metrics: $H(\Sigma), \mathrm{I}(\mathrm{X};\Sigma), I(X;\Sigma;Y)$

Let $\Sigma = [\sigma_1, \sigma_2, \dots, \sigma_L]$ denote the random variable of all gating states in all layers of the DNN.

- $H(\Sigma)$: the entropy of gating states among all inputs.
- $I(X;\Sigma)$: the complexity of transformations that are caused by the input. Both the random sampling operation and the dropout operation introduce additional uncertainty that is not caused by the input.
- $I(X;\Sigma;Y)$: the complexity of transformations that are caused by inputs and used for inference.

Properties of transformation complexity metrics

Non-negativity

Property 1. If the DNN does not introduce additional information that is not contained by the input X (e.g., *there are no operations of randomly sampling or dropout throughout the DNN), then we have* $I(X;\Sigma;Y) \geq 0$.

Monotonicity 1

Property 2. If the DNN does not introduce additional complexity that is not caused by the input, then the complexity increases along with the number of gating layers.

Monotonicity 2

Property 3. If the DNN does not introduce additional complexity that is not caused by the input, then the complexity decreases when we use features of high layers for inference. This property shows that the transformation complexity decreases through the layerwise propagation.

These properties ensure the trustworthiness of the complexity metrics.

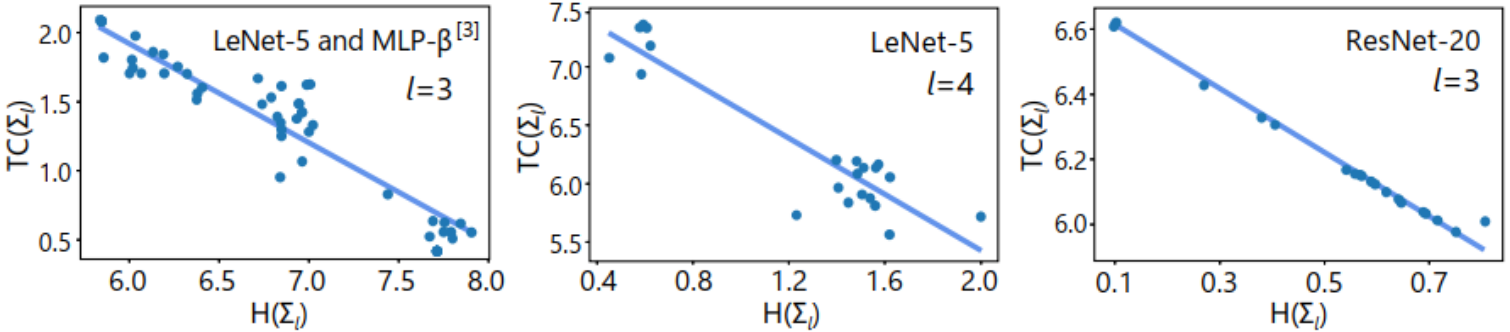
Negative correlation between complexity and entanglement

We prove the ***negative correlation*** between complexity $H(\Sigma_l)$ and entanglement $TC(\Sigma_l)$.

$$\underbrace{H(\Sigma_l)}_{\text{complexity}} + \underbrace{TC(\Sigma_l)}_{\text{entanglement}} = C_l$$

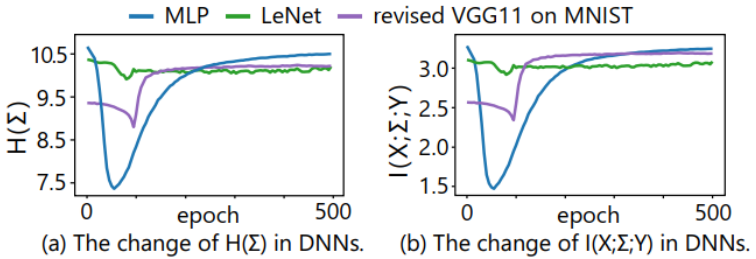
$TC(\Sigma_l)$ measures the dependence of different feature dimensions.

Verification of negative relationship between complexity $H(\Sigma_l)$ and entanglement $TC(\Sigma_l)$



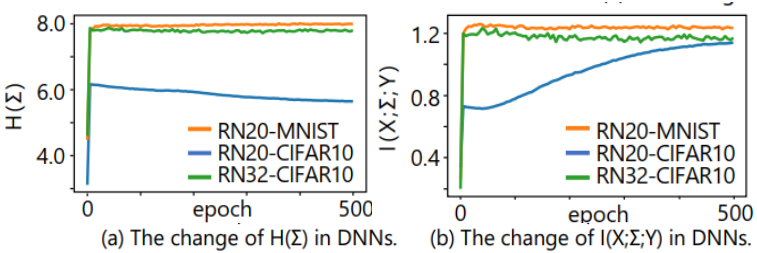
Three phenomena in the training process

- *Phenomenon 1:* For most traditional stacked DNNs, the transformation first decreased and then increased.



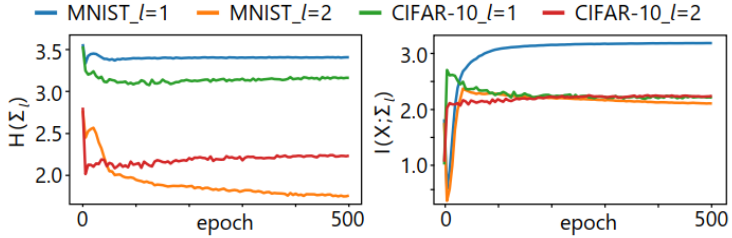
(a) The change of H(Σ) in DNNs. (b) The change of I(X;Σ;Y) in DNNs.

- *Phenomenon 2:* For residual DNNs with skip-connections, the complexity increased monotonously in the early stage and saturated later.



(a) The change of H(Σ) in DNNs. (b) The change of I(X;Σ;Y) in DNNs.

- *Phenomenon 3: For DNNs that contain additional uncertainty (e.g., VAE), the difference between $H(\Sigma_l)$ and $I(X;\Sigma_l)$ gradually decreased during the training process.*

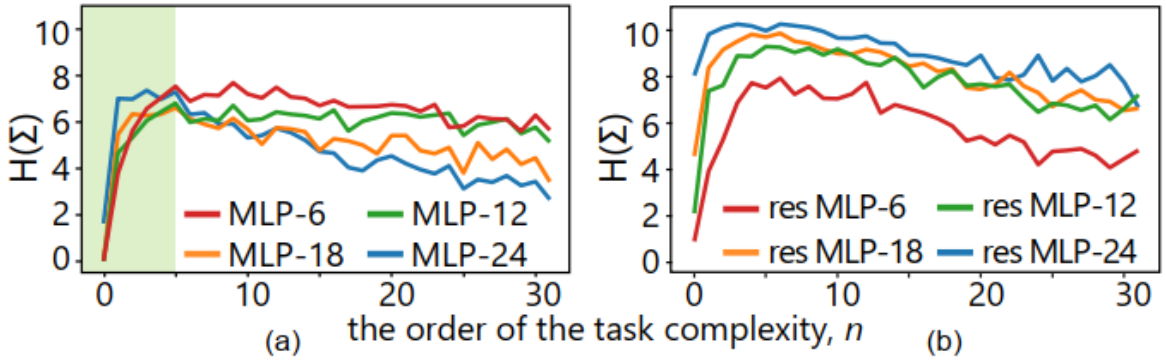


The ceiling of a DNN's complexity

We found that

(1) *the complexity of a DNN did not monotonously increase with the network depth.*

(2) *the complexity of transformations did not increase monotonously along with the increase of the complexity of tasks.*

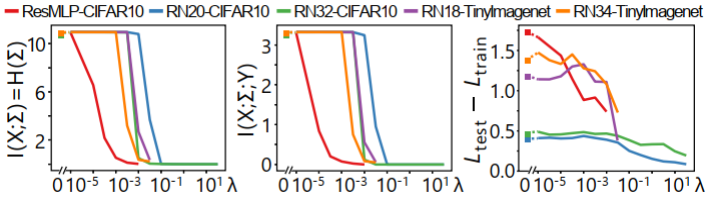


Learning a DNN with minimum complexity

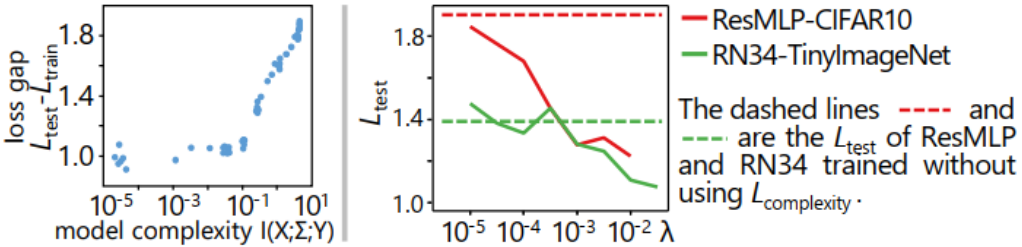
- *We propose the following loss to penalize a DNN's complexity, thereby avoiding learning an over-complex DNN.*

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{complexity}},$$

$$\mathcal{L}_{\text{complexity}} = \sum_{l=1}^{L} H(\Sigma_l) = \sum_{l=1}^{L} \{-\mathbb{E}_{\sigma_l}[\log p(\sigma_l)]\}$$



- ***Utility of minimizing the transformation complexity:*** *reducing the gap between the testing loss and the training loss, alleviating the over-fitting problem.*

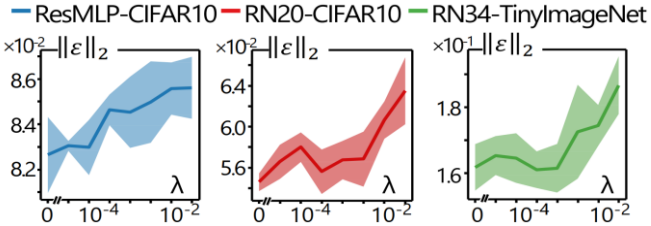


Negative correlation between transformation complexity and adversarial robustness

For ***adversarial robustness***, we found that there is a negative correlation between transformation complexity and adversarial robustness.

- DNNs with low transformation complexity usually exhibited high adversarial robustness
- DNNs with high transformation complexity were usually sensitive to adversarial perturbations.

Model	RN-20		RN-32		RN-44		LeNet	
	Normal	AT	Normal	AT	Normal	AT	Normal	AT
Layer 1	**3.845**	2.979	**2.718**	2.507	**3.938**	1.600	**7.624**	5.358
Layer 2	**6.079**	4.485	**5.660**	4.370	**5.426**	3.374	**7.417**	1.216
Layer 3	**6.671**	6.573	**6.817**	6.786	6.395	**6.828**	**10.966**	10.949

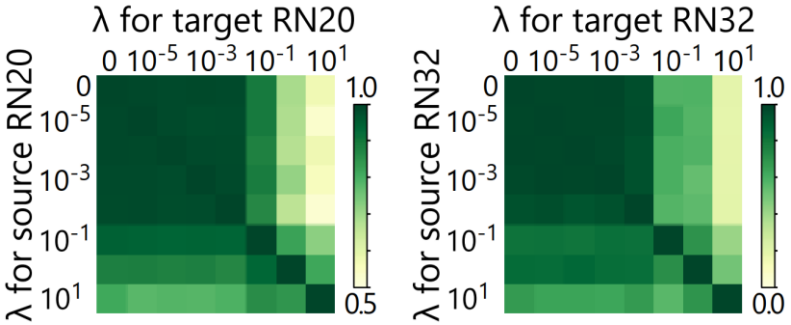


The minimum *L*2 norm of the adversarial perturbations increased along with the increase of the weight of the complexity loss λ.

Simple DNNs usually have higher adversarial transferability

For ***adversarial transferability***, we found that simple DNNs encoded common knowledge that could be transferred to DNNs learned for the same task.

- adversarial perturbations for complex DNNs could not be well transferred to simple DNNs
- adversarial perturbations for simple DNNs could be transferred to complex DNNs.



Summary

- We define three metrics to evaluate the complexity of transformations in piecewise linear DNNs, which have a great theoretical extensibility.
- We prove the negative correlation between the complexity and the entanglement of transformations.
- Comparative studies reveal the ceiling of a DNN's complexity.
- We further use the transformation complexity as a loss to learn a minimum-complexity DNN, which also reduces the gap between the training loss and the testing loss.

THANK YOU !