SFU SIMON FRASER UNIVERSITY

McMaster University

Microsoft Research

Part III:
Effective Complexity of Deep Learning Models

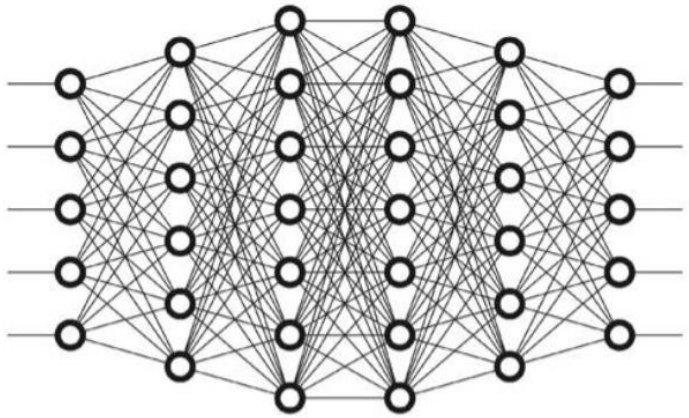
Presenter: Xia Hu

siam 2021 International Conference on Data Mining

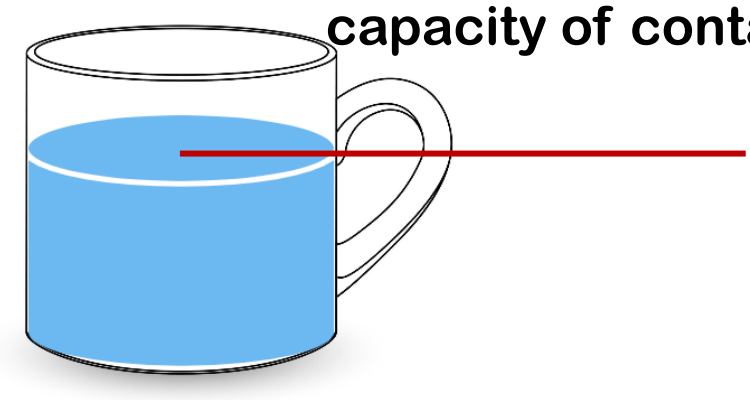
Outline

- **Introduction**
- Measure of Effective Model Complexity
- High Capacity Low Reality Phenomenon
- Discussion

Effective Model Complexity reflects the complexity of the function represented by deep models with specific parameterizations



=



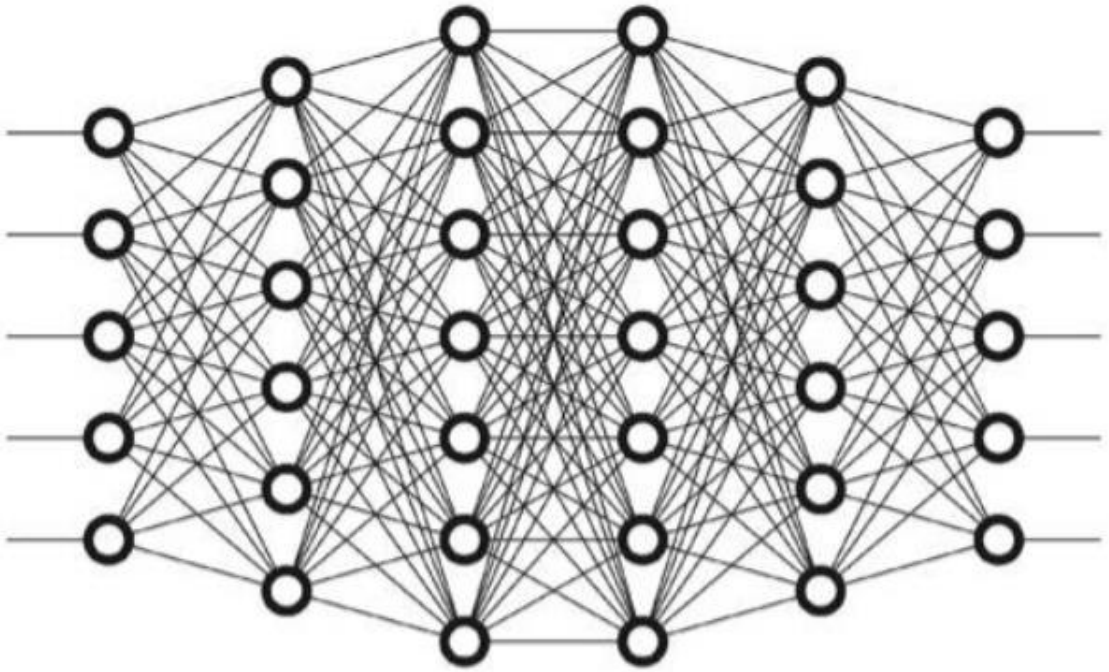
If we informally regard a deep learning model as a "container".

Effective Model Complexity is affected by …

- Model framework
- Model size
- Optimization process
- Data complexity

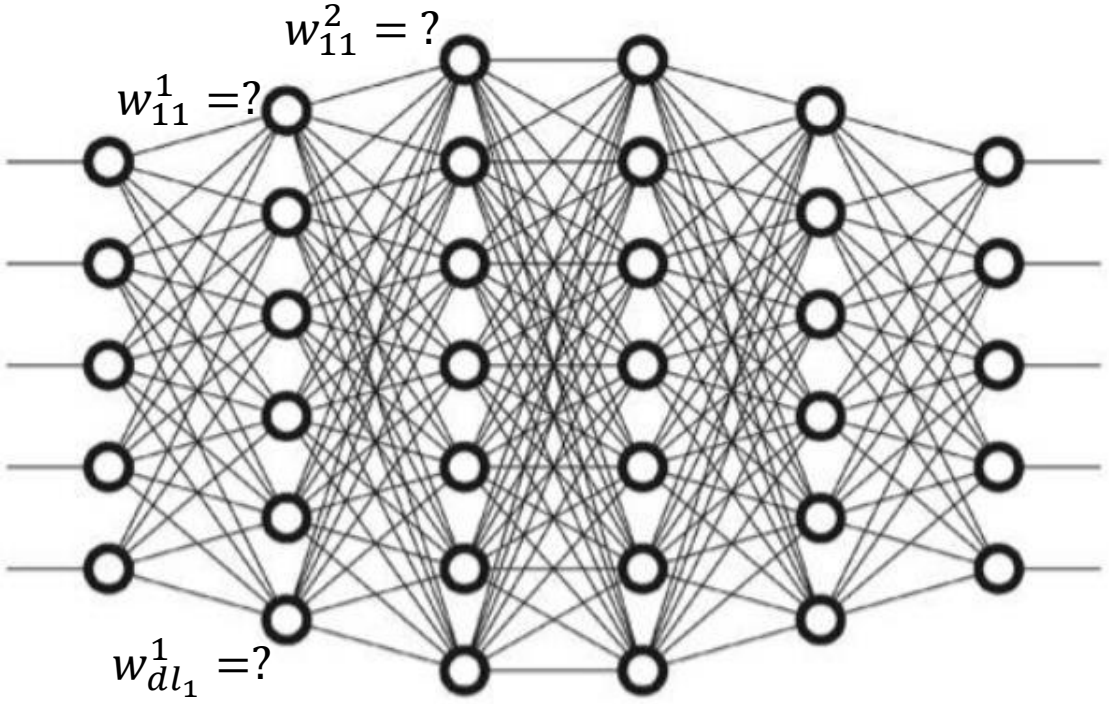
Effective Model Complexity is affected by …

- Model framework
  - Model type? FCNN, CNN, RNN, ResNet …
  - Activation function? Tanh, ReLU …
- Model size
  - Number of hidden layers = ?
  - Width of each layer = ?
  - Number of filters = ?
  - Number of trainable parameters = ?
  - …



Effective Model Complexity is affected by …

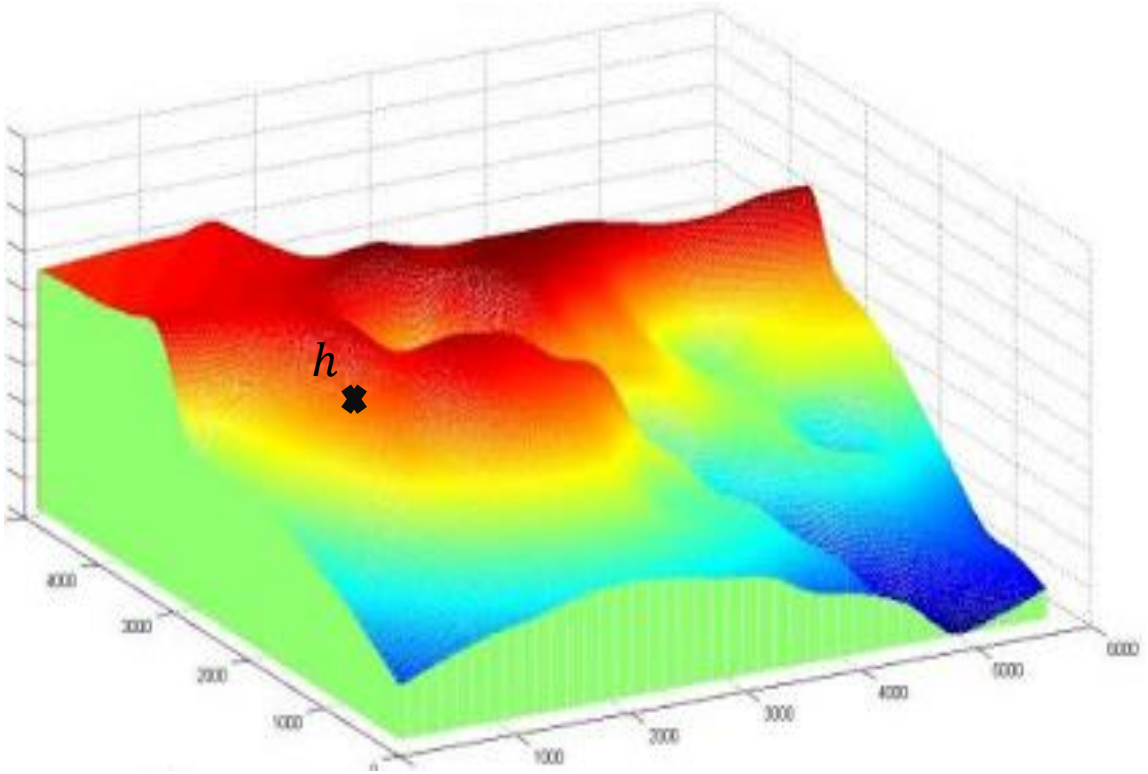
- Optimization process
  - What is the objective function?
  - What is the optimization algorithm?
  - The setting of hyper-parameters
- Data complexity
  - Data dimensionality
  - Number of class labels
  - Data distribution
  - …



Reflected in the values of model parameters

Effective Model Complexity is affected by …

- Model framework and size fixed
  - Model N
  - Corresponding hypothesis space H
- Model parameters fixed
  - A specific $\mathrm{h} \in H$

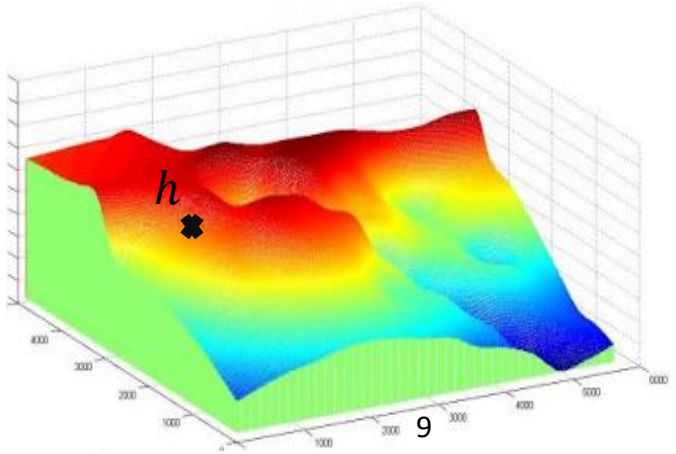


Outline

- Introduction
- **Measure of Effective Model Complexity**
- High Capacity Low Reality Phenomenon
- Discussion

Measure of Effective Complexity

- Necessity to design quantitative measures for effective complexity of deep learning models.
  - Effective complexity cannot be directly derived from the model structure alone.
  - Different parameter values lead to different effective complexity.
  - Be sensitive to distinguish between models with the same structure but different parameter values.



What is the model complexity of h ?

Measure by Piecewise Linear Property

- Piecewise linear property
  - A finite number of linear regions
  - Local linear model
- Deep neural networks with piecewise linear activation functions
  - ReLU
  - Hard Tanh
  - Maxout
  - …

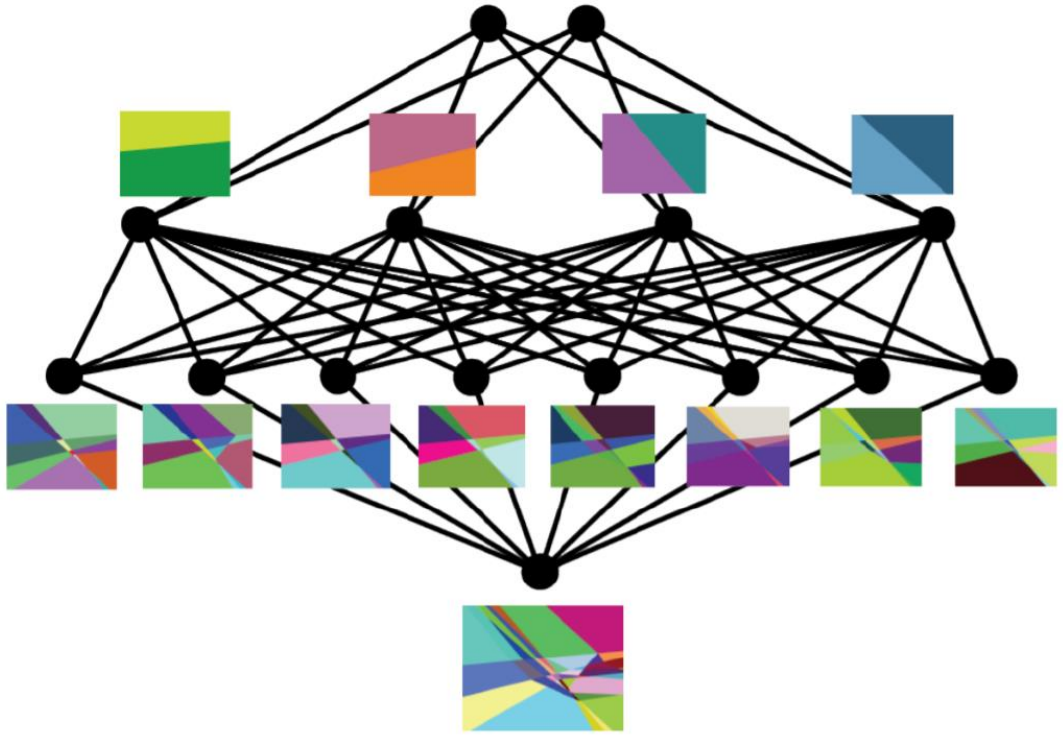


Figure from [Hanin and Rolnick, 2019]

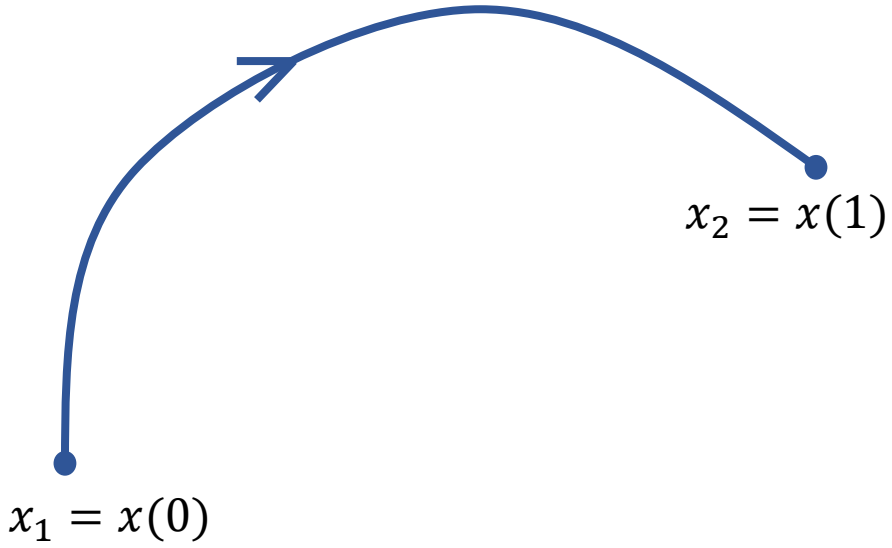
Measure by Piecewise Linear Property

- Trajectory path

$$x(t), t \in [0, 1]$$

- Number of linear regions through trajectory path

$$\Gamma(N(x(t); W))$$



[Raghu et al., 2017]

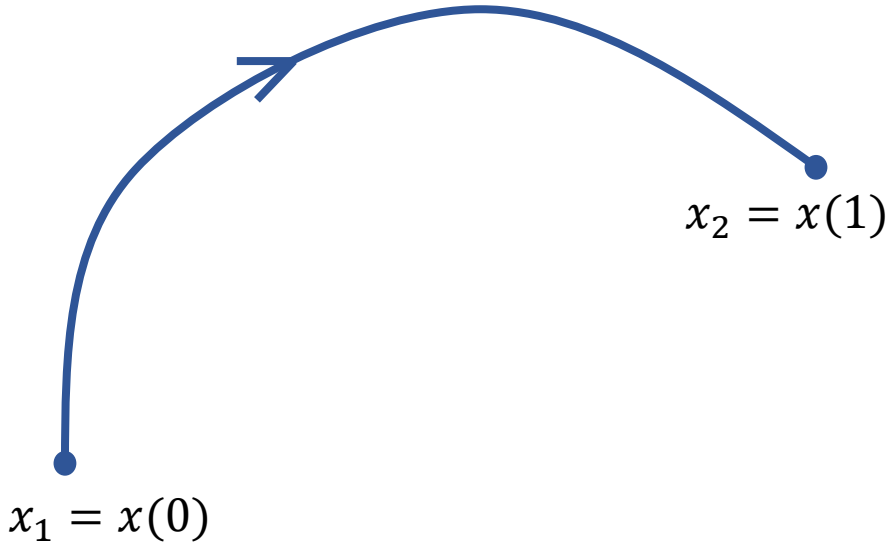
Measure by Piecewise Linear Property

- Trajectory path

$$x(t), t \in [0, 1]$$

- Length of trajectory path

$$l\big(x(t)\big) = \int || \frac{dx(t)}{t} || dt$$



[Raghu et al., 2017]

Measure by Piecewise Linear Property

- A deep ReLU neural network

$$E\left[l\left(z^{(i)}(x(t))\right)\right] \geq O\left(\frac{\sigma_w\sqrt{m}}{\sqrt{m+1}}\right)^i l(x(t))$$

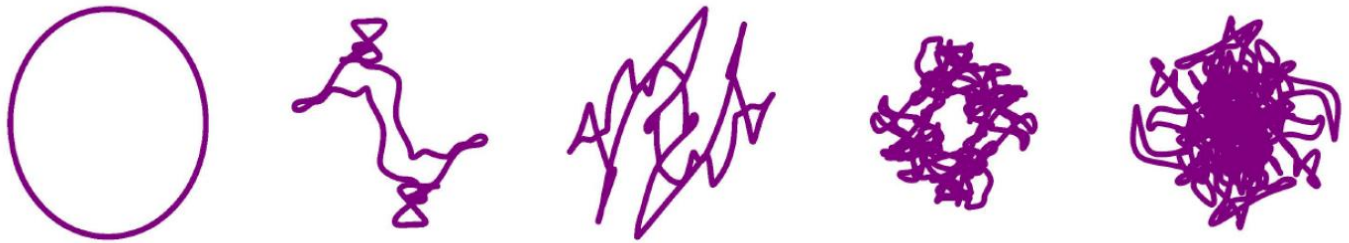


Figure from [Raghu et al., 2019]

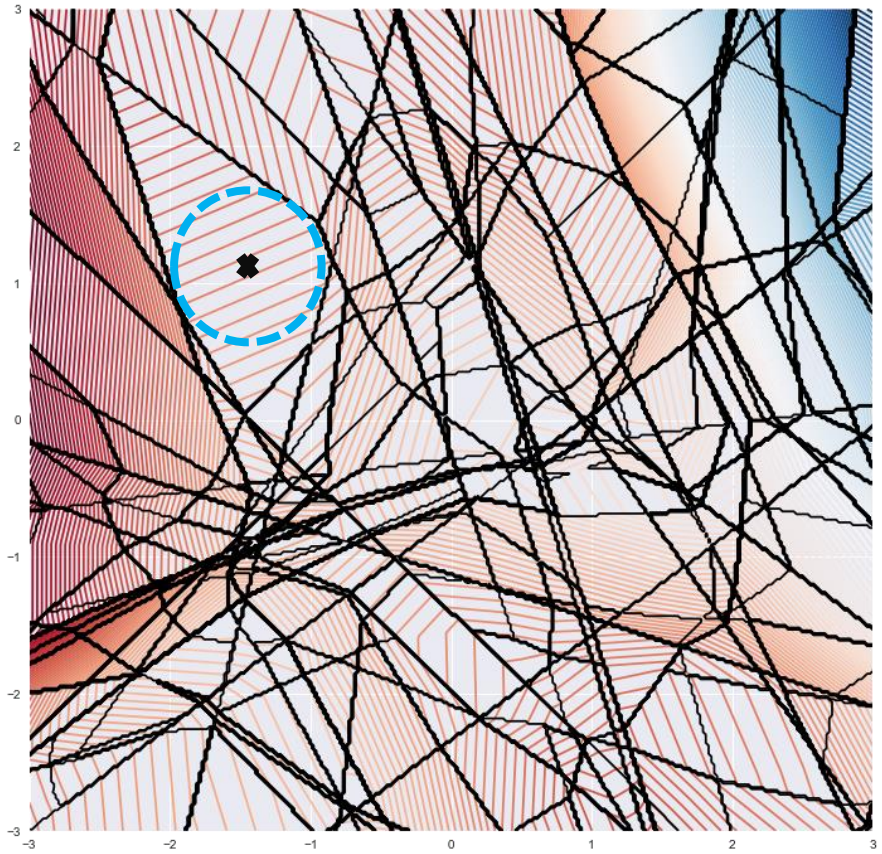
[Raghu et al., 2017]

Measure by Piecewise Linear Property

- Local sensitivity measure

$$E_{x \in D}[\left|\left|J(x)\right|\right|_F]$$

  - An input is perturbed within the same linear region



[Novak et al., 2018]

Measure by Piecewise Linear Property

- The volume of boundaries between linear regions

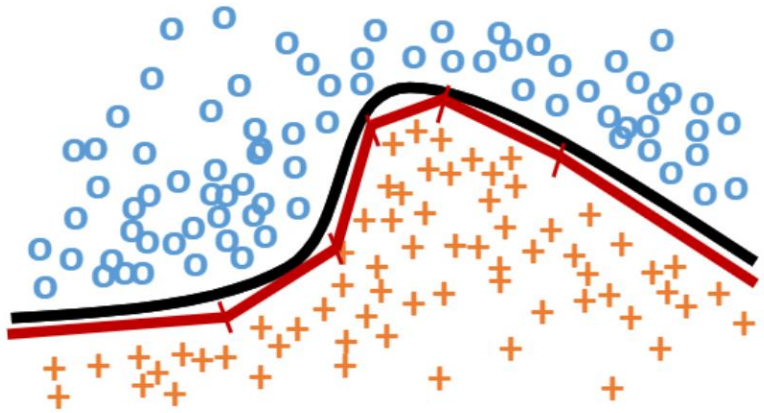
$$E\left[\frac{volume_{d-1}(B_N \cap K)}{volume_d(K)}\right] \approx T \cdot M$$

  - Given a ReLU neural network N: $\mathbb{R}^d \to \mathbb{R}, d > 1$

[Hanin and Rolnick, 2019]

Measure by Piecewise Linear Property

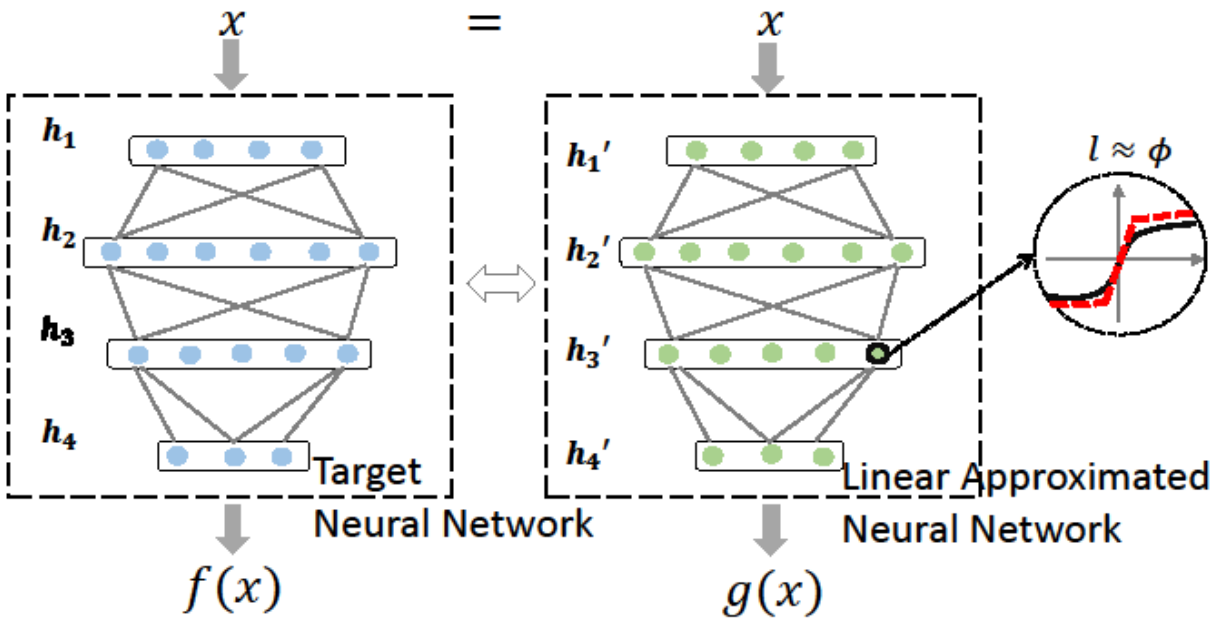
- Deep neural networks with non-piecewise linear activation functions
  - Sigmoid, Tanh, …
- Make them benefit from the piecewise linear property.



[Hu et al., 2020]

Measure by Piecewise Linear Property

- Linear approximation neural network $g(x)$
  - Piecewise linear approximation with as small number of linear regions as possible



[Hu et al., 2020]

Measure by Piecewise Linear Property

- Number of linear regions of the piecewise linear approximation

$$d \sum_{i} log \left(\sum_{j} k_{i,j} - m_i + 1 \right)$$

[Hu et al., 2020]

Other Measure Metrics

- Number of samples with non zero training error

$$\max\{n \mid E_{S\sim D}\left[Error_S\left(\Gamma(S)\right)\right] \leq \epsilon\}$$

[Nakkiran et al., 2020]

Other Measure Metrics

- Fisher-Rao Information

$$\langle \theta, I(\theta)\theta \rangle$$

The set of parameters of the model

$$I(\theta) = E[\nabla_\theta \ell(N(X), Y) \otimes \nabla_\theta \ell(N(X), Y)]$$

Fisher Information matrix. $\ell(N(X), Y)$ is the loss function.

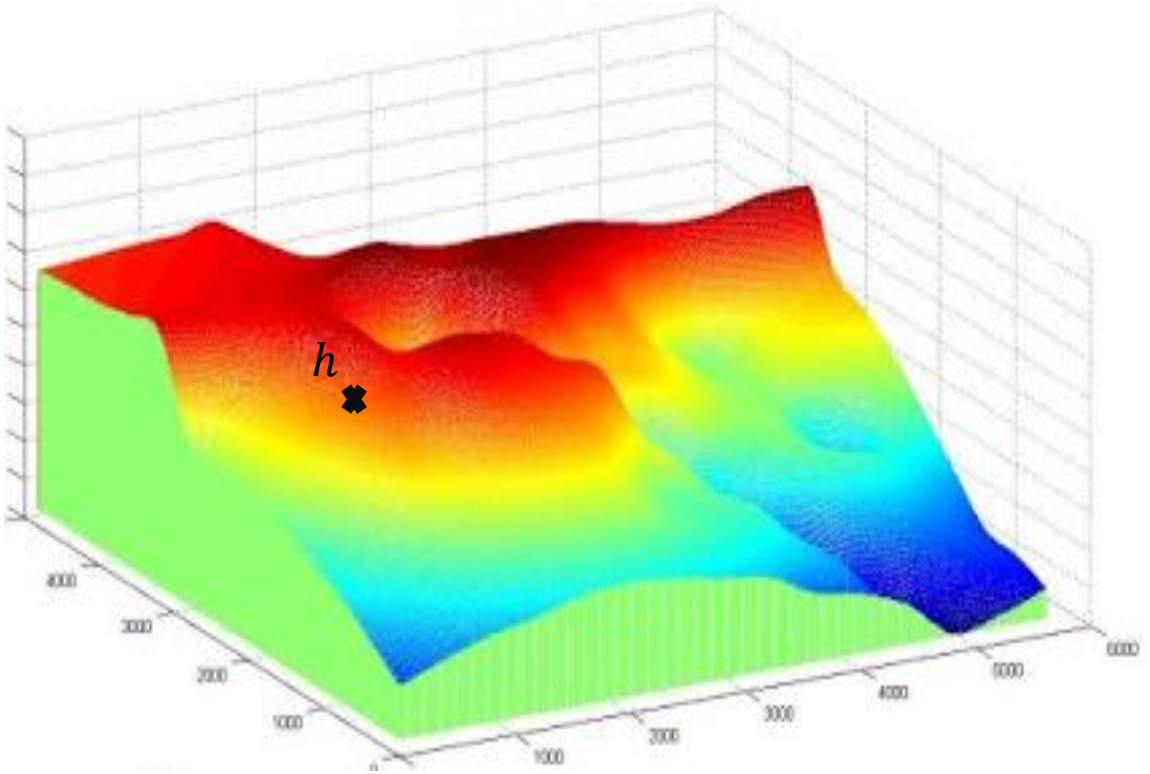
[Liang et al., 2019]

Outline

- Introduction
- Measure of Effective Model Complexity
- **High Capacity Low Reality Phenomenon**
- Discussion

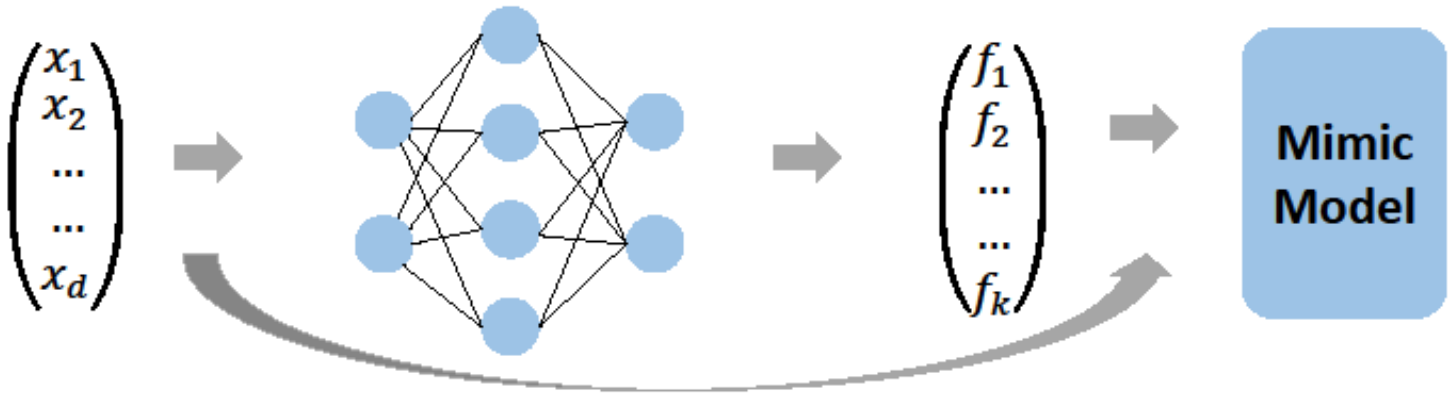
High Capacity Low Reality Phenomenon

- Expressive capacity is high
  - Deep learning models are always highly over-parameterized.
- Effective complexity of learned model is found to be low
- There might be a huge gap!



High Capacity Low Reality Phenomenon

- Given a well-trained deep model
- A shallow model
  - Can mimic the deep model to as high accuracy as the deep one
  - Sometimes only requires the same number of parameters



[Ba and Caruana, 2014]

High Capacity Low Reality Phenomenon

- The effective complexity of the trained deep model might not be very high
  - It can be mimicked by a shallow one
- The strength of deep learning may arise in part from a good match between deep architecture and current training algorithms
  - Deep architectures might be easier to train by current optimization algorithms.

[Ba and Caruana, 2014]

High Capacity Low Reality Phenomenon

- Given a ReLU neural network N: $\mathbb{R} \to \mathbb{R}$

$$E[\#\{linear\ regions\ in\ S\}] \approx |S| \cdot T \cdot M$$

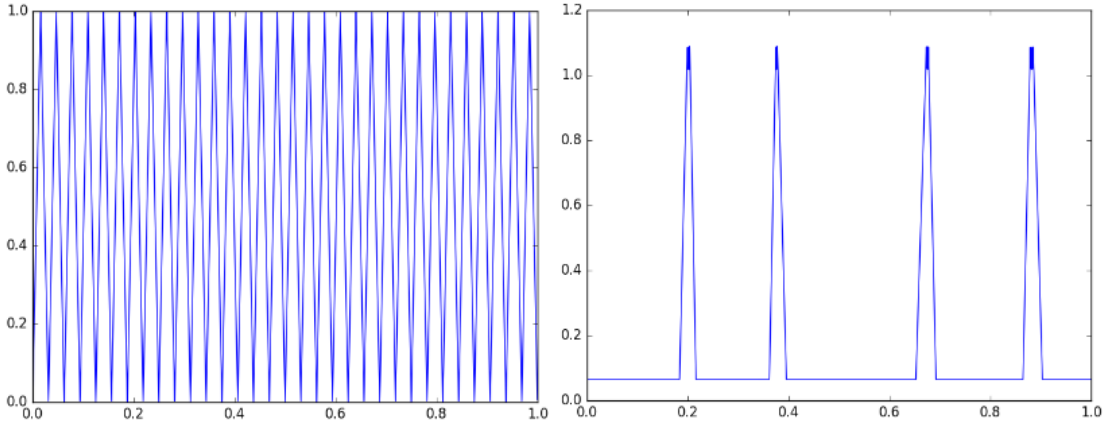


Figure from [Hanin and Rolnick, 2019]

[Hanin and Rolnick, 2019]

High Capacity Low Reality Phenomenon

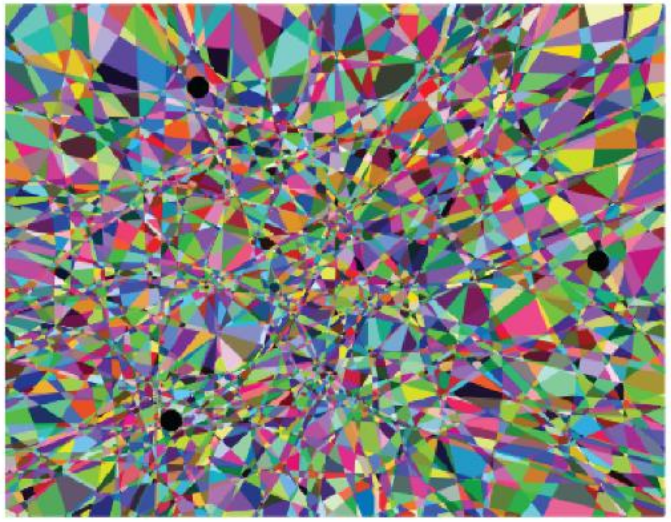
- Given a ReLU neural network N: $\mathbb{R}^d \to \mathbb{R}, d > 1$

$$E\left[\frac{volume_{d-1}(B_N \cap K)}{volume_d(K)}\right] \approx T \cdot M$$

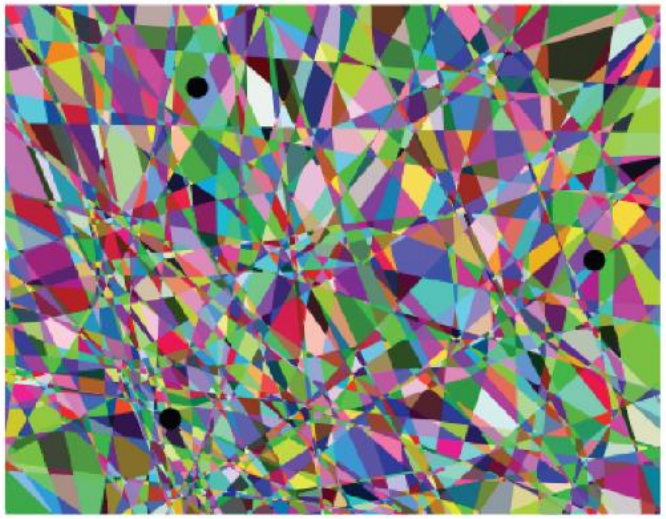
[Hanin and Rolnick, 2019]

High Capacity Low Reality Phenomenon

Epoch 0: 9744 regions



Epoch 1: 4196 regions



Epoch 20: 8541 regions

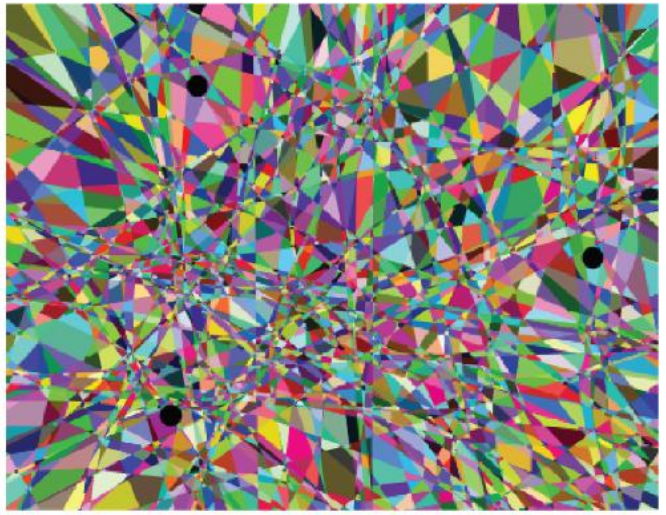


Figure from [Hanin and Rolnick, 2019]

[Hanin and Rolnick, 2019]

Outline

- Introduction
- Measure of Effective Model Complexity
- High Capacity Low Reality Phenomenon
- **Discussion**

Discussion

- Effective model complexity is a relatively new, promising and useful problem in deep learning
  - Investigate the usefulness of optimization algorithms
  - Study the rule of regularizations
  - Develop new regularization approaches
  - … …

Discussion

- Other interesting problems
  - Cross-model comparison
  - Specify granularity of effective complexity measures