SFU SIMON FRASER UNIVERSITY

McMaster University

Microsoft Research

Part II:
Expressive Capacity of Deep Learning Models

Presenter: Lingyang Chu

siam 2021 International Conference on Data Mining

Outline

- **Introduction**
- Depth Efficiency
- Width Efficiency
- Expressible Functional Space
- VC Dimension and Rademacher Complexity

<u>Expressive Capacity</u> reflects how well a deep learning model can approximate complex problems.

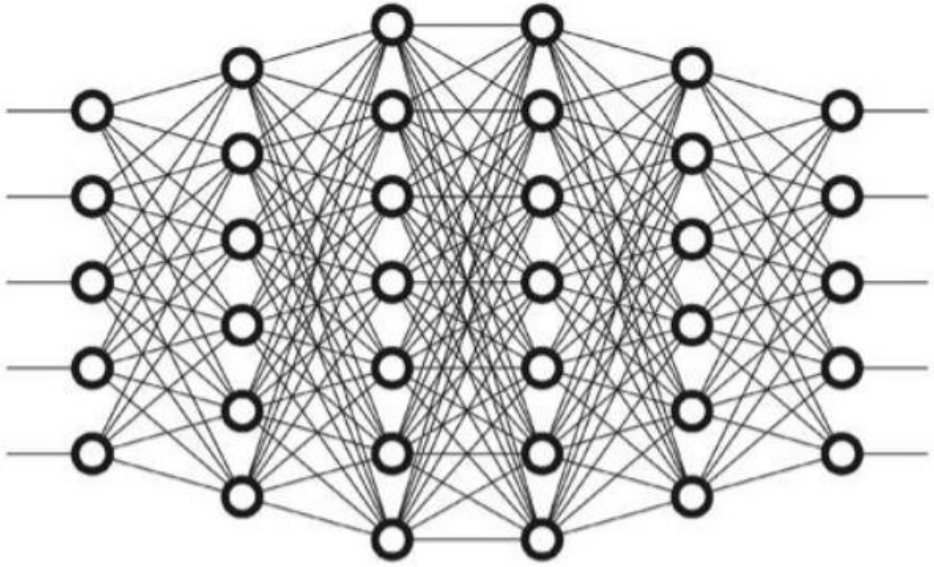
If we regard a deep learning model as a "container".

Expressive Capacity is affected by …

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

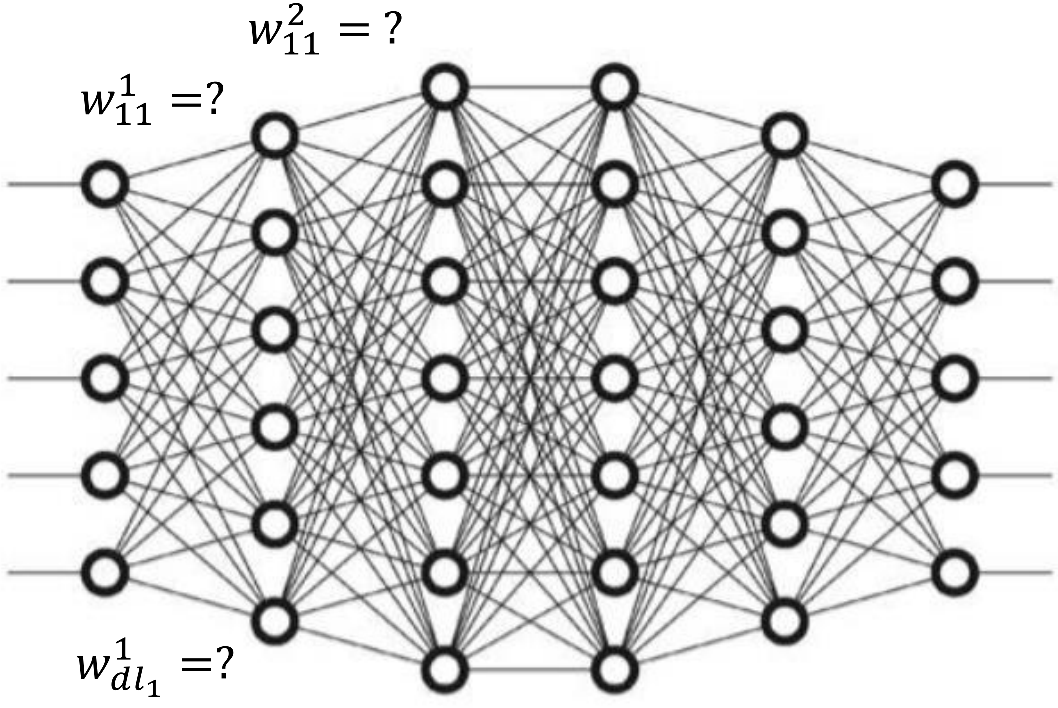
Expressive Capacity is affected by …

- Model framework
  - Model architecture? 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 = ?
  - …



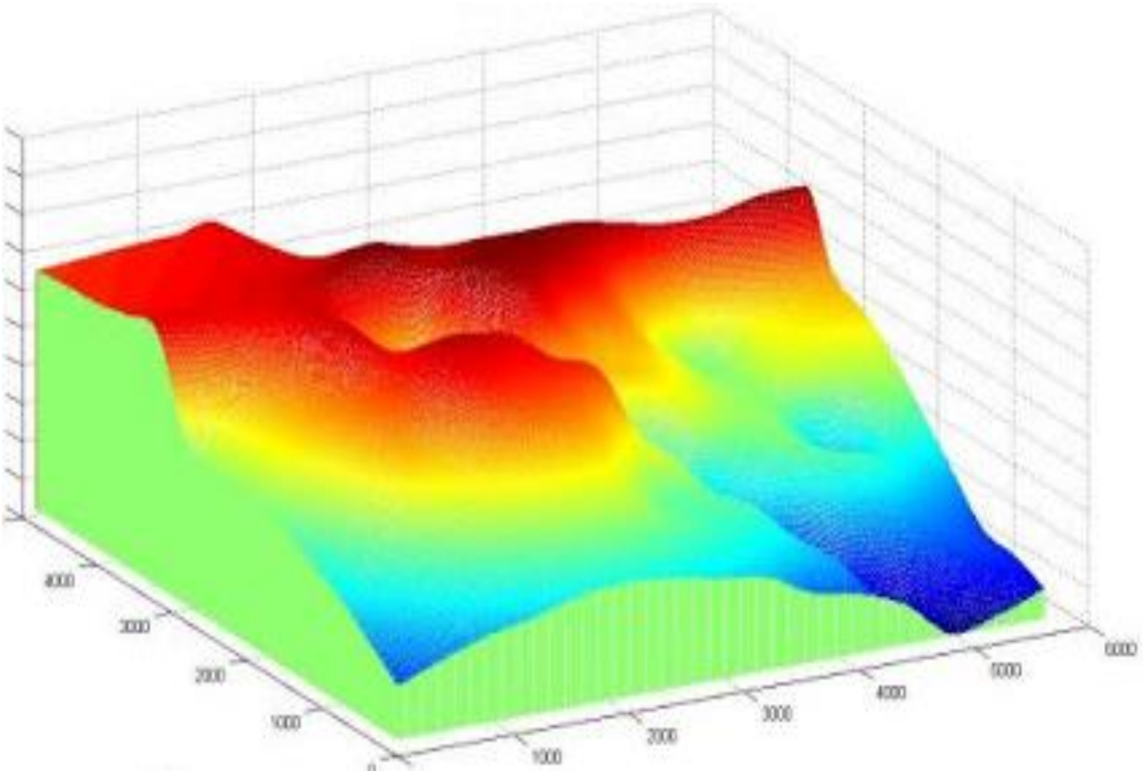
Expressive Capacity is affected by …

- 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
  - …



Expressive Capacity is affected by …

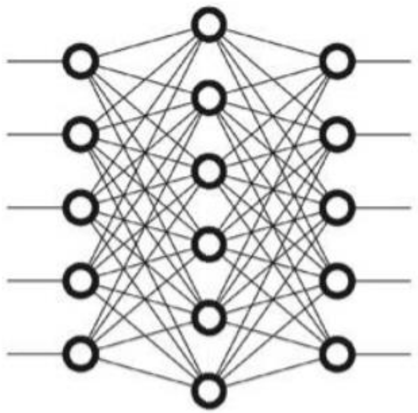
- Model framework and size fixed
  - Model N
  - Corresponding hypothesis space H
- Optimization and data complexity
  - A smaller hypothesis space $H' \subset H$



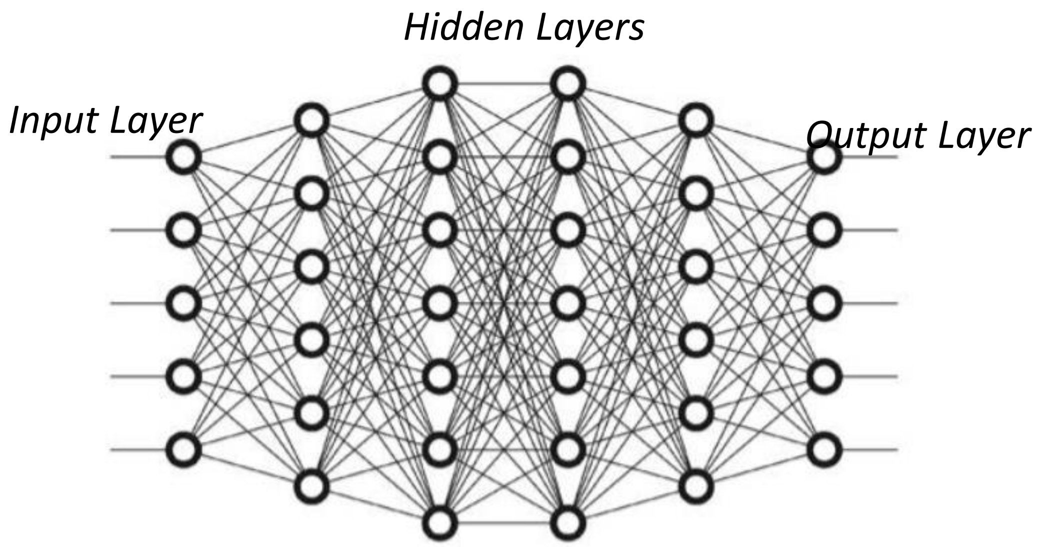
Outline

- Introduction
- **<u>Depth Efficiency</u>**
- Width Efficiency
- Expressible Functional Space
- VC Dimension and Rademacher Complexity

Depth Efficiency analyzes why deep architectures can obtain good performance and measures the effects of depth on expressive capacity.



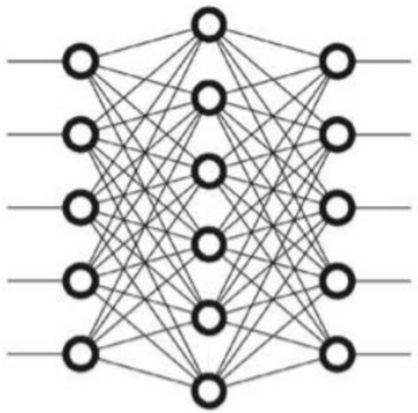
"Shallow" feedforward neural network



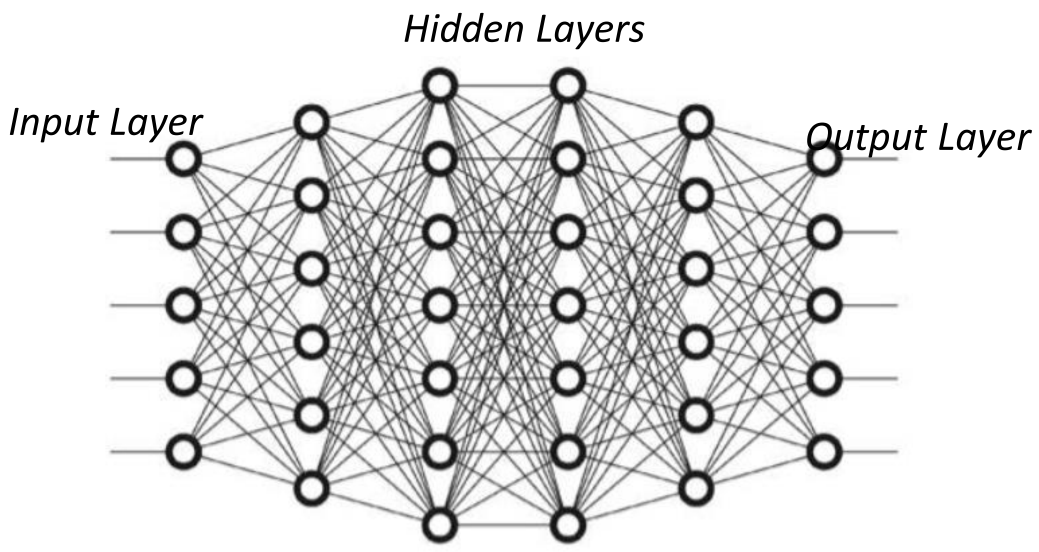
"Deep" feedforward neural network

Depth Efficiency analyzes why deep architectures can obtain good performance and measures the effects of depth on expressive capacity.

Why being <u>DEEP</u> performs so good?



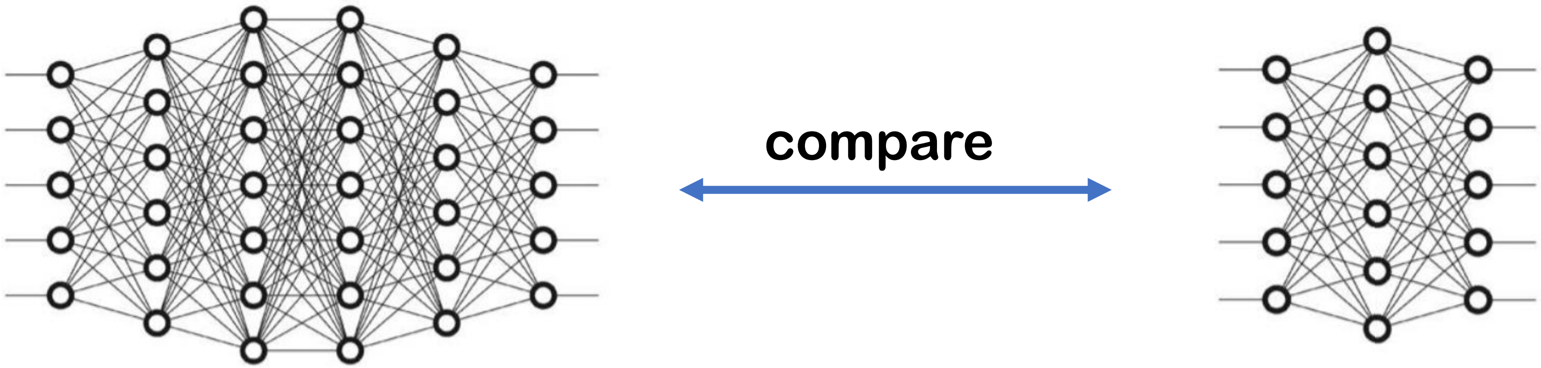
"Shallow" feedforward neural network



"Deep" feedforward neural network

Model Reduction Approaches

- Reduce deep learning models to some understandable problems and functions for analysis.
- To compare the representation efficiency between deep models with shallow ones.



Model Reduction Approaches

- A family of functions $R \to R$

$$\bigcup_{M>0} \underbrace{(\Delta_M^{w-1} \times \Delta_M^{w-1} \times \ldots \times \Delta_M^{w-1})}_{k \text{ times}}$$

- Representable by a $(k+1)$-layer ReLU DNN of width w
- Representable by a $(k'+1)$-layer ReLU DNN $(k' < k)$ of

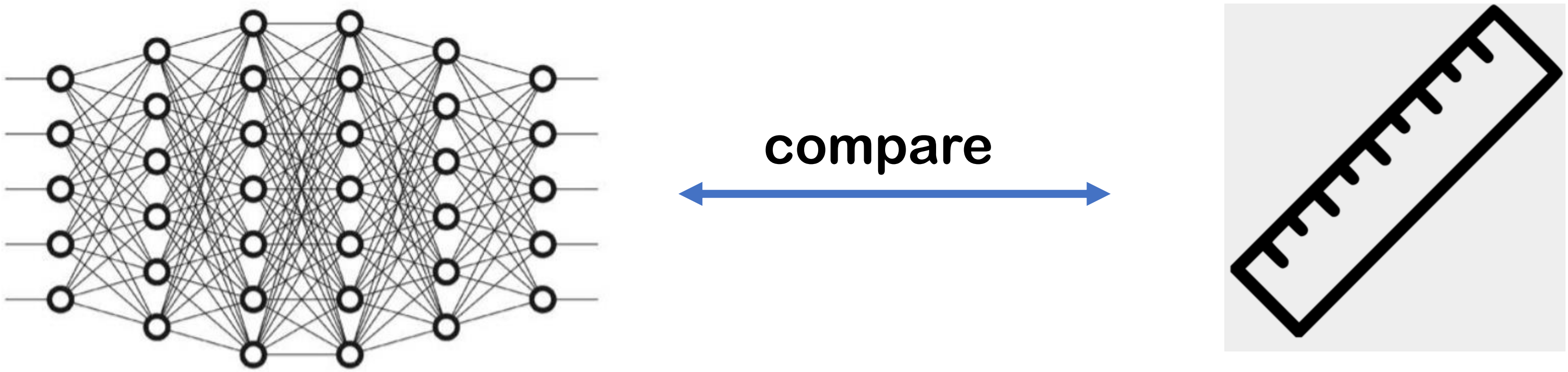
$$\frac{1}{2} k' w^{k/k'} - 1$$

hidden neurons

[Arora et al., 2018]

Measure-based Approaches

- Develop an appropriate measure of expressive capacity
- Study how the expressive capacity changes when the depth and layer width of a model increase



Measure-based Approaches

- DNNs with piecewise linear activation functions (e.g., ReLU)
- The number of linear regions as a model complexity measure
- The maximum number of linear regions

$$\geq \left(\prod_{i=1}^{l-1} \left\lfloor \frac{m_i}{m_0} \right\rfloor^{m_0} \right) \sum_{j=0}^{m_0} \binom{m_i}{j}$$

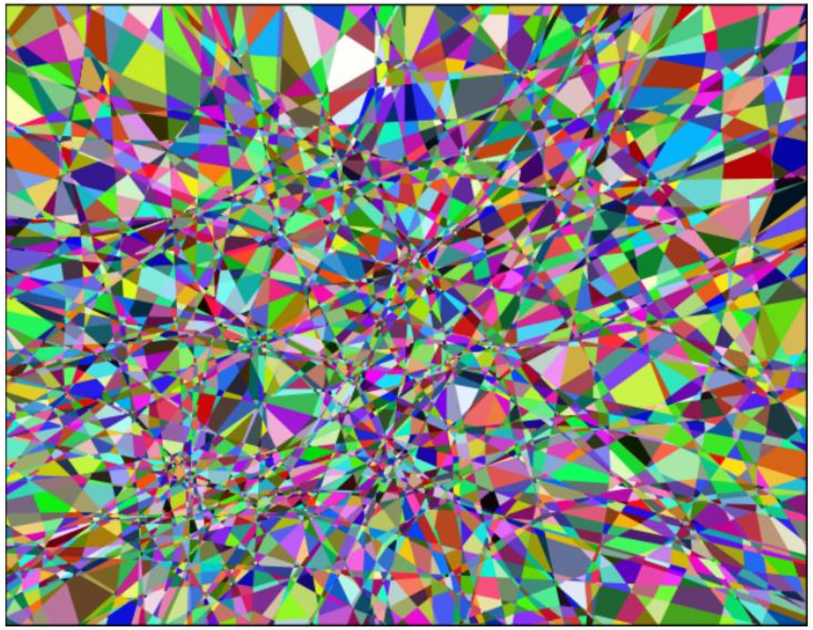


Figure from [Hanin and Rolnick, 2019]

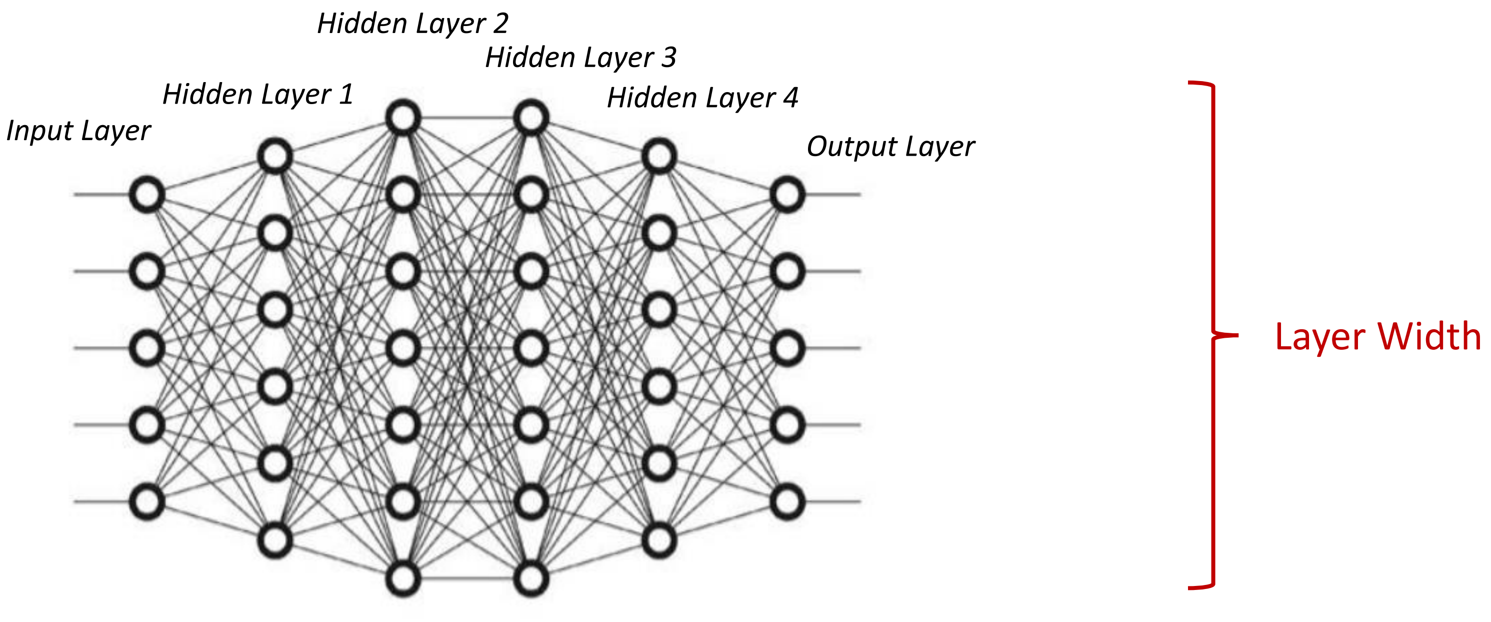
Linear Regions

[Montufar et al., 2014]

Outline

- Introduction
- Depth Efficiency
- **Width Efficiency**
- Expressible Functional Space
- VC Dimension and Rademacher Complexity

Width Efficiency analyzes how width affects the expressive capacity of deep learning models.



- Important for fully understanding expressive capacity
- Validate the insight obtained from depth efficiency

Width Efficiency

- Universal approximation theorem of width-bounded ReLU neural networks.
  - A Lebesgue-integrable function can be approximated to any desired performance by a ReLU network whose width

$$\leq d + 4$$

- Width efficiency:
  - F_A: ReLU DNN, depth $= 3$, width $= 2k^2$, $k \geq d + 4$
  - F_B: ReLU DNN, depth $\leq k + 2$, width $\leq k^{3/2}$, parameters $\in [-b, b]$
  - $\forall b, \exists \epsilon\ s.t.$

$$\int_{R^d} |F_A(x) - F_B(x)| dx \geq \epsilon$$

[Lu et al., 2017]

Depth Efficiency v.s. Width Efficiency

- Exponential lower bound of Depth Efficiency:
  - To approximate a deep model, a shallow model requires at least an exponential increase in width.

$$\left(\boxed{\prod_{i=1}^{l-1}}\left\lfloor\frac{m_i}{m_0}\right\rfloor^{m_0}\right)\sum_{j=0}^{m_0}\binom{m_i}{j}$$

[Montufar et al., 2014]

Depth Efficiency v.s. Width Efficiency

- Exponential lower bound of Depth Efficiency:
  - To approximate a deep model, a shallow model requires at least an exponential increase in width.
- Polynomial lower bound for Width efficiency
  - To approximate a shallow, wide model whose width increases linearly, a deep, narrow model requires at least a polynomial increase in depth.

$$\left(\prod_{i=1}^{l-1}\left\lfloor\frac{\boxed{m_i}}{m_0}\right\rfloor^{m_0}\right)\sum_{j=0}^{m_0}\binom{\boxed{m_i}}{j}$$

[Montufar et al., 2014]

Depth Efficiency vs. Width Efficiency

- Exponential lower bound of Depth Efficiency:
  - To approximate a deep model, a shallow model requires at least an exponential increase in width.
- Polynomial lower bound for Width efficiency
  - To approximate a shallow, wide model whose width increases linearly, a deep, narrow model requires at least a polynomial increase in depth.

$$\left(\prod_{i=1}^{l-1}\left\lfloor\frac{m_i}{m_0}\right\rfloor^{m_0}\right)\sum_{j=0}^{m_0}\binom{m_i}{j}$$

[Montufar et al., 2014]

Requires a polynomial upper bound of width to strictly prove that depth is more effective than width. [Lu et al., 2017]

Outline

- Depth Efficiency
- Width Efficiency
- **Expressible Functional Space**
- VC Dimension and Rademacher Complexity

Expressible Functional Space explores the class of functions that can be expressed by deep learning models with specific frameworks and specified size.

Model-Specific Approaches

- Focus on specific types of deep learning models
- ReLU networks can express every piecewise linear function with

$$\{limited\ depth\}$$

[Arora et al., 2018]

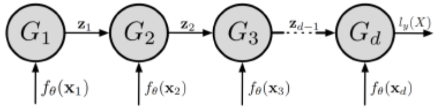
- ReLU networks can express a function f from Sobolev space with

$$\{limited\ depth\ and\ \#\ neurons\}$$

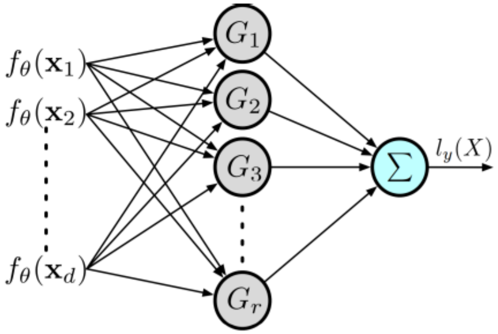
[Gühring et al., 2019]

Cross-Model Approaches

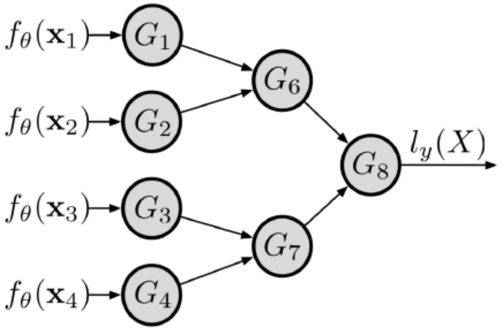
- Several network architectures correspond to various tensor decompositions



(a) TT-decomposition



(b) CP-decomposition



(c) HT-decomposition

- Compare the cross-model expressive capacity

	TT-Network	HT-Network	CP-Network
TT-Network	r	$r^{\log_2(d)/2}$	r
HT-Network	r^2	r	r
CP-Network	$\geq r^{d/2}$	$\geq r^{d/2}$	r

[Khrulkov et al., 2017]

Outline

- Depth Efficiency
- Width Efficiency
- Expressible Functional Space
- **VC Dimension and Rademacher Complexity**

Two typical measure metrics of expressive capacity (i.e., the complexity of hypothesis space):

- VC Dimension
- Rademacher Complexity

Vapnik-Chervonenkis Dimension

- The VC dimension of a hypothesis class H is the size of the largest set that can be shattered by H:

$$VCdim(H) = \max\{m : \Pi_H(m) = 2^m\}$$

- Feedforward neural network with linear threshold gates:

$$VCdim(H) = \Theta(\underbrace{W}_{\text{The number of parameters}} log W)$$

[Khrulkov et al., 2017]

Vapnik-Chervonenkis Dimension

- Feedforward neural network with piecewise polynomial activation functions

$$C_1 WL\log\left(\frac{W}{L}\right) \leq VCdim(H) \leq C_2(WL^2 + WL\log WL)$$

Network depth

- ReLU DNN

$$C_1 WL\log\left(\frac{W}{L}\right) \leq VCdim(H) \leq C_2(WL\log W)$$

[Bartlett et al., 1998&2019]

Rademacher Complexity

- The Rademacher Complexity of a hypothesis class H on a data distribution D is defined as

$$\mathfrak{R}_n(H) = E[\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\epsilon_i}_{\text{Rademacher variables}} h(x_i)]$$

- Two-layer ReLU neural network:

$$\mathfrak{R}_n(H) = \Omega(\frac{s_1 s_2 \sqrt{m} \, ||X||_F}{n})$$

[Neyshabur et al., 2018]

VC dimension and Rademacher Complexity

- Deep learning models are always over-parameterized

$$W \gg n$$

- VC bound can be very high

$$O(WL\log\left(\frac{W}{L}\right))$$

Conclusion

		Model Framework	Model Size	Learning Process	Data Complexity
Expressive Capacity	Depth Efficiency		√		
	Width Efficiency		√		
	Expressible Functional Space	√	√	√	√

Table: Summarize the aspects affecting every categories of the expressive capacity.