ICML
International Conference On Machine Learning

Position: A Theory of Deep Learning Must Include Compositional Sparsity

David A. Danhofer   Davide D'Ascenzo   Rafael Dubach   Tomaso Poggio

CENTER FOR Brains Minds+ Machines

Funded by the European Union
NextGenerationEU

The Deep Learning Puzzle

- DNNs excel in vision, language, reasoning.
- Classical theory: **curse of dimensionality** makes high-dimensional learning intractable.

The Deep Learning Puzzle

- DNNs excel in vision, language, reasoning.

- Classical theory: **curse of dimensionality** makes high-dimensional learning intractable.

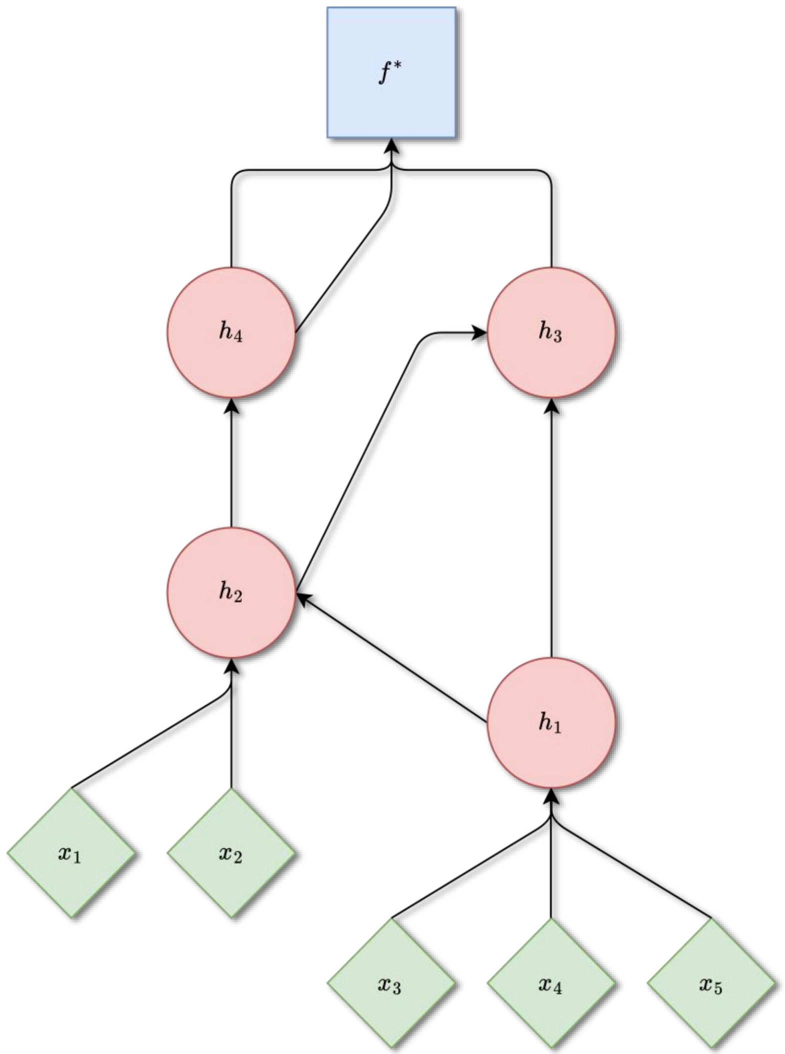
Which property of real-world target functions enables DNNs to overcome this?

What Is Compositional Sparsity?

- **Definition**
  A function is compositionally sparse if it can be expressed as a composition of a polynomial number of subfunctions, where each subfunction depends only on a constant number of inputs.

- **Example**
  Hierarchical compositions in vision, language, and reasoning tasks.

- **Visualization**
  DAG structure: inputs as leaves, subfunctions as nodes, output at the root.



Total input dimension **d = 5**
Local input dimension **c = 3**

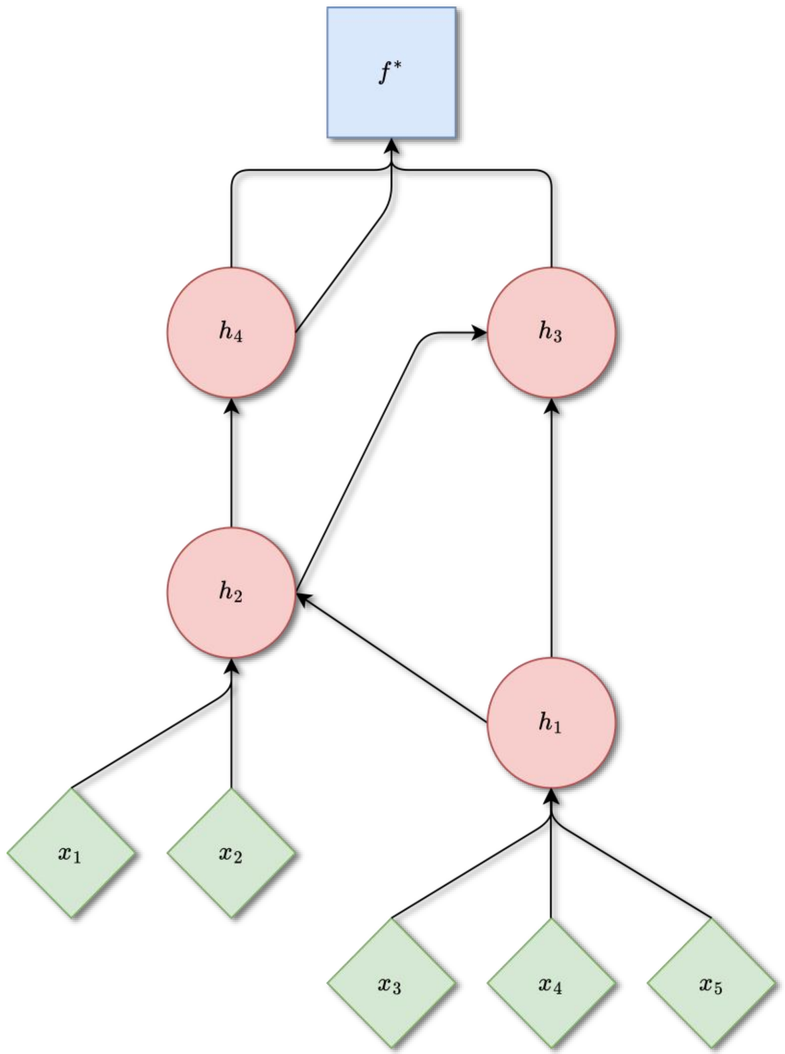
What Is Compositional Sparsity?

- **Definition**
  A function is compositionally sparse if it can be expressed as a composition of a polynomial number of subfunctions, where each subfunction depends only on a constant number of inputs.

- **Example**
  Hierarchical compositions in vision, language, and reasoning tasks.

- **Visualization**
  DAG structure: inputs as leaves, subfunctions as nodes, output at the root.

All efficiently (polynomial-time) Turing-computable functions are compositionally sparse*



Total input dimension **d = 5**
Local input dimension **c = 3**

*Conjectured in *Poggio & Fraser, Bull. Amer. Math. Soc. 61 (2024), 438-456.*

Approximation (Poggio et al. 2017)

- **Shallow** networks need **exponentially** many parameters to approximate a compositionally sparse function.

- **Deep** neural networks can approximate the same function efficiently, using only a **polynomial** number of parameters.

Generalization (Xu et al. 2023)

- For CNNs, accounting for **weight sparsity** yields much **tighter generalization bounds** than naive Rademacher complexity.

- In fact, the **local filters** in CNNs act as subfunctions that each depend on only a **constant** number of inputs.

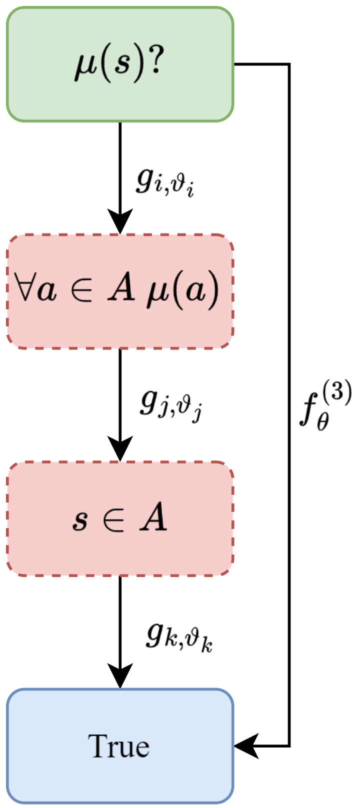
Optimization

- Learning **arbitrary** polynomial-time computable functions is **exponentially** hard under standard cryptographic assumptions. (Goldreich et al. 1986)

- In practice, real-world tasks and architectures (CNNs, Transformers, **Chain-of-Thought**) provide structure, making optimization tractable.

Chain-of-Thought

Conjecture: Chain-of-Thought explicitly decomposes a compositionally sparse learning problem into sparse subproblems, each one of which can be learned. As such, it overcomes the complexity of one-shot learning.



Is Socrates mortal? CoT-style intermediate solving steps can simplify this famous question to a sequence of general reasoning steps of less complexity than the specific question at hand.

Open Question(s)

Which functions are efficiently learnable?

Thanks