Title: Mechanistic interpretability

URL: https://en.wikipedia.org/wiki/Mechanistic_interpretability

PageID: 79868032

Categories: Category:Artificial intelligence, Category:Machine learning

---

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

Mechanistic interpretability (often abbreviated as mech interp , mechinterp , or MI) is a subfield of research within explainable artificial intelligence that aims to understand the internal workings of neural networks by analyzing the mechanisms present in their computations. The approach seeks to analyze neural networks in a manner similar to how binary computer programs can be reverse-engineered to understand their functions. [1]

History

The term mechanistic interpretability was coined by Chris Olah. [2] Early work combined various techniques such as feature visualization, dimensionality reduction, and attribution with human-computer interaction methods to analyze models like the vision model Inception v1 . [3] Later developments include the 2020 paper Zoom In: An Introduction to Circuits , which proposed

an analogy between neural network components and biological neural circuits. [4]

In recent years, mechanistic interpretability has gained prominence with the study of large language models (LLMs) and transformer architectures. The field is expanding rapidly, with multiple dedicated workshops such as the ICML 2024 Mechanistic Interpretability Workshop being hosted. [5]

Key concepts

Mechanistic interpretability aims to identify structures, circuits or algorithms encoded in the weights of machine learning models. [6] This contrasts with earlier interpretability methods that focused primarily on input-output explanations. [7]

Multiple definitions of the term exist, from narrow technical definitions (the study of causal mechanisms inside neural networks) to broader cultural definitions encompassing various AI interpretability research. [2]

Linear representation hypothesis

This hypothesis suggests that high-level concepts are represented as linear directions in the activation space of neural networks. Empirical evidence from word embeddings and more recent studies supports this view, although it does not hold up universally. [8] [9]

Superposition

Superposition describes how neural networks may represent many unrelated features within the same neurons or subspaces, leading to densely packed and overlapping feature representations. [10]

Methods

Probing

Probing involves training simple classifiers on neural network activations to test whether certain features are encoded. [1]

Causal interventions

Mechanistic interpretability employs causal methods to understand how internal model components influence outputs, often using formal tools from causality theory. [11]

Sparse decomposition

Methods such as sparse dictionary learning and sparse autoencoders help disentangle complex overlapping features by learning interpretable, sparse representations. [12]

Applications and significance

Mechanistic interpretability is crucial in the field of AI safety to understand and verify the behavior of increasingly complex AI systems. It helps identify potential risks and improves transparency. [13]

References

Further reading

Nanda, Neel (2023). "Emergent Linear Representations in World Models of Self-Supervised Sequence Models" . BlackNLP Workshop : 16– 30. doi : 10.18653/v1/2023.blackboxnlp-1.2 .

Transformer Circuits Thread : a series of articles from Anthropic on mechanistic interpretability in transformers