DEEP LEARNING

What is deep learning?

Deep learning is a subset of machine learning driven by multilayered neural networks whose design is inspired by the structure of the human brain. Deep learning models power most state-of-the-art artificial intelligence (AI) today, from computer vision and generative AI to self-driving cars and robotics.

Unlike the explicitly defined mathematical logic of traditional machine learning algorithms, the artificial neural networks of deep learning models comprise many interconnected layers of "neurons" that each perform a mathematical operation. By using machine learning to adjust the strength of the connections between individual neurons in adjacent layers—in other words, the varying model weights and biases—the network can be optimized to yield more accurate outputs. While neural networks and deep learning have become inextricably associated with one another, they are not strictly synonymous: "deep learning" refers to the training of models with at least 4 layers (though modern neural network architectures are often much "deeper" than that).

It's this distributed, highly flexible and adjustable structure that explains deep learning's incredible power and versatility. Imagine training data as data points scattered on a 2-dimensional graph, and the goal of model training to be finding a line that runs through each of those data points. Essentially, traditional machine learning aims to accomplish this using a single mathematical function that yields a single line (or curve); deep learning, on the other hand, can piece together an arbitrary number of smaller, individually adjustable lines to form the desired shape. Deep neural networks are universal approximators: it has been proven theoretically that for any function, there exists a neural network arrangement that can reproduce it.1

Deep learning models are most commonly trained through supervised learning on labeled data to perform regression and classification tasks. But because large-scale neural networks usually require a massive amount of training data to reach optimal performance, the cost and labor of acquiring sufficiently large datasets of annotated training examples can be prohibitive. This has led to development of techniques to replicate supervised learning tasks using unlabeled data. The term self-supervised learning was coined by Yann LeCun in the late 2010s to disambiguate such methods from traditional unsupervised learning. Self-supervised learning has since emerged as a prominent mode of training neural networks, particularly for the foundation models underpinning generative AI.

HOW ITS WORKS

How deep learning works

Artificial neural networks are, broadly speaking, inspired by the workings of the human brain's neural circuits, whose functioning is driven by the complex transmission of chemical and electrical signals across distributed networks of nerve cells (neurons). In deep learning, the analogous "signals" are the weighted outputs of many nested mathematical operations, each performed by an artificial "neuron" (or node), that collectively comprise the neural network.

In short, a deep learning model can be understood as an intricate series of nested equations that maps an input to an output. Adjusting the relative influence of individual equations within that network using specialized machine learning processes can, in turn, alter the way the network maps inputs to outputs.

While that framework is very powerful and versatile, it's comes at the expense of interpretability. There's often little, if any, intuitive explanation—beyond a raw mathematical one—for how the values of individual model parameters learned by a neural network reflect real-world characteristics of data. For that reason, deep learning models are often referred to as "black boxes," especially when compared to traditional types of machine learning models informed by manual feature engineering.

Relative to classic machine learning techniques, deep learning requires an exceedingly large amount of data and computational resources for training. Given the cost and complexity of the enterprise-level hardware needed to develop and implement sophisticated deep learning applications, cloud computing services have become an increasingly integral part of the deep learning ecosystem.

Deep neural network structure

Artificial neural networks comprise interconnected layers of artificial “neurons” (or nodes), each of which performs its own mathematical operation (called an “activation function”). There exist many different activation functions; a neural network will often incorporate multiple activation functions within its structure, but typically all of the neurons in a given layer of the network will be set to perform the same activation function. In most neural networks, each neuron in the input layer is connected to each of the neurons in the following layer, which are themselves each connected to the neurons in layer after that, and so on.

The output of each node’s activation function contributes part of the input provided to each of the nodes of the following layer. Crucially, the activation functions performed at each node are nonlinear, enabling neural networks to model complex patterns and dependencies. It’s the use of nonlinear activation functions that distinguishes a deep neural network from a (very complex) linear regression model.

While some specialized neural network architectures, such as mixture of expert models or convolutional neural networks, entail variations, additions or exceptions to this arrangement, all neural networks employ some version of this core structure. The specific number of layers, numerical

Types of deep learning models

Convolutional neural networks (CNNs)

Convolutional neural networks (CNNs) are primarily (but not exclusively) associated with computer vision tasks such as object detection, image recognition, image classification and image segmentation, as they excel at "local" pattern recognition (such as relationships between adjacent pixels in an image).

Recurrent neural networks (RNNs)

Recurrent neural networks (RNNs) are used for tasks involving sequential data, such as time-series forecasting, speech recognition or natural language processing (NLP).

Whereas conventional feedforward neural networks map a single input to a single output, RNNs map a sequence of inputs to an output by operating in a recurrent loop in which the output for a given step in the input sequence serves as input to the computation for the following step.

Autoencoders

Autoencoders are designed to compress (or encode) input data, then reconstruct (decode) the original input using this compressed representation. In training, they're optimized to minimize reconstruction loss: the divergence between the reconstructed data point and the original input data. Though this type of deep learning uses unlabeled, unstructured data, autoencoders are generally considered to be an archetypal example of self-supervised learning.

Generative adversarial networks (GANs)

Like VAEs, generative adversarial networks (GANs) are neural networks are used to create new data resembling the original training data. GANs are a joint architecture combining two deep learning networks trained adversarially in a zero-sum game.

A standard feedforward neural network with 3 hidden layers.