

I'm not a bot



equally to suit their environments. How does a neural network learn things? Information flows through a neural network in two ways. When it's learning (being trained), patterns of information are fed into the network via the input units, which trigger the layers of hidden units, and these in turn arrive at the output units. This common design is called a feedforward network. Not all units "fire" all the time. Each unit receives inputs from the units to its left, and the inputs are multiplied by the weights of the connections they travel along. Every unit adds up the inputs it receives in this way and (in the simplest type of network) if the sum is more than a certain threshold value, the unit "fires" and triggers the units it's connected to (those on its right). Photo: Bowling: You learn how to do skillful things like this with the help of the neural network inside your brain. Every time you throw the ball wrong, you learn what corrections you need to make next time. Photo by Kenneth R. Hendrix/US Navy/published on Flickr. For a neural network to learn, there has to be an element of feedback involvedjust as children learn by being told what they're doing right or wrong. In fact, we all use feedback, all the time. Think back to when you first learned to play a game like ten-pin bowling. As you picked up the heavy ball and rolled it down the alley, your brain watched how quickly the ball moved and the line it followed, and noted how close you came to knocking down the skittles. Next time it was your turn, you remembered what you'd done wrong before, modified your movements accordingly, and hopefully threw the ball a bit better. So you used feedback to compare the outcome you wanted with what actually happened, figured out the difference between the two, and used that to change what you did next time ("I need to throw it harder," "I need to roll slightly more to the left," "I need to let go later," and so on). The bigger the difference between the intended and actual outcome, the more radically you would have altered your moves. Neural networks learn things in exactly the same way, typically by a feedback process called backpropagation (sometimes abbreviated as "backprop"). This involves comparing the output a network produces with the output it was meant to produce, and using the difference between them to modify the weights of the connections between the units in the network, working from the output units through the hidden units to the input unitsgoing backward, in other words. In time, backpropagation causes the network to learn, reducing the difference between actual and intended output to the point where the two exactly coincide, so the network figures things out exactly as it should. Artwork: A neural network can learn by backpropagation, which is a kind of feedback process that passes corrective values backward through the network. Simple neural networks use simple math: they use basic multiplication to weight the connections between different units. Some neural networks learn to recognize patterns in data using more complex and elaborate math. Known as convolutional neural networks (CNNs or, sometimes, "ConvNets") their input layers take in 2D or 3D "tables" of data (like the matrices you might remember learning about in school). Their hidden layers (sometimes several dozen of them) include some that perform a mathematical process called convolution. Simply speaking, convolutional layers recognize significant patterns hidden in data and "concentrate" them into an easier-to-use form. Essentially, they're detecting key features, which can then be classified by further layers that work like a more traditional neural network. CNNs are particularly good at classifying images or videos, recognizing handwriting, and so on. Artwork: This convolutional neural network (greatly simplified) extracts and emphasizes key features (by the mathematical process of convolutionbroadly a kind of matrix multiplication). These are fed into a more conventional neural network, which uses them to recognize an unknown object or image. How does it work in practice? Once the network has been trained with enough learning examples, it reaches a point where you can present it with an entirely new set of inputs it's never seen before and see how it responds. For example, suppose you've been teaching a network by showing it lots of pictures of chairs and tables, represented in some appropriate way it can understand, and telling it whether each one is a chair or a table. After showing it, let's say, 25 different chairs and 25 different tables, you feed it a picture of some new design it's not encountered beforelet's say a chaise longueand see what happens. Depending on how you've trained it, it'll attempt to categorize the new example as either a chair or a table, generalizing on the basis of its past experiencejust like a human. Hey presto, you've taught a computer how to recognize furniture! That doesn't mean to say a neural network can just "look" at pieces of furniture and instantly respond to them in meaningful ways; it's not behaving like a person. Consider the example we've just given: the network is not actually looking at pieces of furniture. The inputs to a network are essentially binary numbers: each input unit is either switched on or switched off. So if you had five input units, you could feed in information about five different characteristics of different chairs using binary (yes/no) answers. The questions might be 1) Does it have a back? 2) Does it have a top? 3) Does it have soft upholstery? 4) Can you sit on it comfortably for long periods of time? 5) Can you put lots of things on top of it? A typical chair would then present as Yes, No, Yes, Yes, No or 10110 in binary, while a typical table might be No, Yes, No, No, Yes or 01001. So, during the learning phase, the network is simply looking at lots of numbers like 10110 and 01001 and learning that some mean chair (which might be an output of 1) while others mean table (an output of 0). What are neural networks used for? Photo: For the last two decades, NASA has been experimenting with a self-learning neural network calledIntelligent Flight Control System (IFCS) that can help pilots land planes after suffering major failures or damage in battle. The prototype was tested on this modified NF-15B plane (a relative of the McDonnell Douglas F-15). Photo by Jim Ross courtesy of NASA. On the basis of this example, you can probably see lots of different applications for neural networks that involve recognizing patterns and making simple decisions about them. In airplanes, you might use a neural network as a basic autopilot, with input units reading signals from the various cockpit instruments and output units modifying the plane's controls appropriately to keep it safely on course. Inside a factory, you could use a neural network for quality control. Let's say you're producing clothes washing detergent in some giant, convoluted chemical process. You could measure the final detergent in various ways (its color, acidity, thickness, or whatever), feed those measurements into your neural network as inputs, and then have the network decide whether to accept or reject the batch. There are lots of applications for neural networks in security, too. Suppose you're running a bank with many thousands of credit-card transactions passing through your computer system every single minute. You need a quick automated way of identifying any transactions that might be fraudulentand that's something for which a neural network is perfectly suited. Your inputs would be things like 1) Is the cardholder actually present? 2) Has a valid PIN number been used? 3) Have five or more transactions been presented with this card in the last 10 minutes? 4) Is the card being used in a different country from which it's registered? and so on. With enough clues, a neural network can flag up any transactions that look suspicious, allowing a human operator to investigate them more closely. In a very similar way, a bank could use a neural network to help it decide whether to give loans to people on the basis of their past credit history, current earnings, and employment record. Photo: Handwriting recognition on a touchscreen, tablet computer is one of many applications perfectly suited to a neural network. Each character (letter, number, or symbol) that you write is recognized on the basis of key features it contains (vertical lines, horizontal lines, angled lines, curves, and so on) and the order in which you draw them on the screen. Neural networks get better and better at recognizing over time. Many of the things we all do everyday involve recognizing patterns and using them to make decisions, so neural networks can help us out in zillions of different ways. They can help us forecast the stockmarket or the weather, operate radar scanning systems that automatically identify enemy aircraft or ships, and even help doctors to diagnose complex diseases on the basis of their symptoms. There might be neural networks ticking away inside your computer or your cellphone right this minute. If you use cellphone apps that recognize your handwriting on a touchscreen, they might be using a simple neural network to figure out which characters you're writing by looking out for distinct features in the marks you make with your fingers (and the order in which you make them). Some kinds of voice recognition software also use neural networks. And so do some of the email programs that automatically differentiate between genuine emails and spam.Neural networks have even proved effective in translating text from one language to another. Google's automatic translation, for example, has made increasing use of this technology over the last few years to convert words into one language (the network's input) into the equivalent words in another language (the network's output). In 2016, Google announced it was using something it called Neural Machine Translation (NMT) to convert entire sentences, instantly, with a585 percent reduction in errors. This is just one example of how Google deploys neural-network technology. Google Brainis the name it's given to a massive research effort that applies neural techniques across its whole range of products, including its search engine. It also uses deep neural networks to power the recommendations you see on YouTube, with models that "learn approximatelyone billion parameters and are trained on hundreds of billions of examples." [5] All in all, neural networks have made computer systems more useful by making them more human. So next time you think you might like your brain to be as reliable as a computer, think againand be grateful you have such a superb neural network already installed in your head! Chris Woodford is the author and editor of dozens of science and technology books for adults and children, including DK's worldwide bestselling Cool Stuff series and Atoms Under the Floorboards, which won the American Institute of Physics Science Writing award in 2016. You can hire him to write books, articles, scripts, corporate copy, and more via his website chriswoodford.com. Just imagine what must go on inside the human brain when tasked with recognizing digits, given the varying ways the digits 0-9 can be hand-drawn. Human readers can instantly classify these digits, even from an early age. The goal of neural networks is for a computer algorithm to be able to classify these digits as well as a human. TheMNIST(ModifiedNational Institute of Standards and Technology)database is a largedatasetof handwritten digits that is frequently used fortrainingvariousimage processingand machine learning systems. This video presentation by Michael Garris, senior scientist, provides an overview. You may download the MNIST dataset files directly for further exploration. In this section, we introduce simple models of neural networks and discover how they work. Many technical details are deferred to later sections or are outside the scope of this text. A neural network is a structure made up of components called neurons, which are individual decision-making units that take some number of inputs $x_1, x_2, \dots, x_{n1}, x_2, \dots, x_n$ and produce an output y , as in Figure 7.2. How the output is determined by the inputs could be quite complex. Moreover, any two neurons may behave quite differently from each other on the same inputs. Figure 7.2 Single Neuron in a Neural Network The neural network itself may consist of hundreds, thousands, or even millions of neurons connected to each other in layers, or groups of neurons that all receive the same inputs from previous layers and forward signals in aggregate to the next layer. There are always at least two layers, the input layer (containing neurons that accept the initial input data) and output layer (containing the neurons that are used to interpret the answer or give classification information), together with some number of hidden layers (layers between the input and output layers). Figure 7.3 shows a simple neural network diagram for reference. Figure 7.3 Neural Network Diagram. This neural network has four layers, two of which are hidden layers. There are three input neurons and one output neuron. In this example, all nodes in adjacent layers are connected, but some neural network models may not include all such connections (see, for example, convolutional neural networks in Introduction to Deep Learning). The main purpose of a neural network is to classify complex data. Problems for which neural networks are especially well suited include the following: Image recognition, including facial recognition, identifying handwritten letters and symbols, and classifying parts of images. This is a huge area of innovation, with powerful tools such as TensorFlow developed by Google and PyTorch developed by Meta. Speech recognition, such as Googles Cloud Speech-to-Text service, offers accurate transcription of speech and translation for various languages. Recommendation systems, which are used to serve ads online based on each users browsing habits, have been developed and used by Amazon, Meta, Netflix, and many other large companies to reach target markets. Anomaly detection has been developed to aid in fraud detection, data security, and error analysis/correction by finding outliers in large, complex datasets. An example is Microsofts Azure Anomaly Detector, which can detect anomalies in time series data. Autonomous vehicles and robotics, including Teslas Autopilot technology, are becoming more and more prominent as automation alleviates some of the routine aspects of daily life and leads to increased efficiencies in business, manufacturing, and transportation. Generative art, music, video, and poetry, leverage vast stores of human creative output to produce novel variations. Predictive text, including natural language processing models such as ChatGPT (see Convolutional Neural Networks). If you want to get your feet wet with neural networks, check out this interactive web-based neural network tool, called TensorFlow Playground, which uses TensorFlow to train and update outputs in real time. There, you can choose a dataset from a list, adjust the number of hidden layers and neurons per layer by clicking the plus (+) and minus (-) buttons, and adjust the learning rate, choice of activation function, and other parameters (topics that we will learn more about in the rest of the chapter). Then click the play button to start training the model and watch as it learns how to classify the points in your chosen dataset! If you want to start over, just click reset and start from scratch! The way neurons work in a neural network is conjectured to be similar, or analogous, to how neurons work in the brain. The input of the neural network is fed into the neurons of the input layer. Each neuron then processes it and produces an output, which is in turn pushed to the neurons in the next layer. Individual neurons only send signals, or activate, if they receive the appropriate input required to activate. (Activation is the process of sending an output signal after having received appropriate input signals.) After passing through some number of hidden layers, the output of the last hidden layer feeds into the output layer. Lastly, the output of this final layer is interpreted based on the nature of the problem. If there is a single output neuron, then the interpretation could be true if that neuron is activated or false if not. For neural networks used to classify input into various classes (e.g., number recognition), there is usually one output neuron per class. The one that is most activated would indicate the classification, as shown in Figure 7.4. Figure 7.4 Output Neurons. In this figure, there are four output neurons, labeled A, B, C, and D. Since B has the highest activation level, the output of the neural network is B. Each connection from one neuron to another has two parameters associated with it. The weight is a value w that is multiplied to the incoming signal, essentially determining the strength of the connection. The bias is a value b that is added to the weighted signal, making the neuron more likely (or less likely, if b is negative) to activate on any given input. The values of w and b may be positive, negative, or zero. Once the weight and bias are applied, the result is run through an activation function f that determines whether the neuron activates and, if so, how strongly. (Youll see this later in Figure 7.6.) Consider the simplest case of all, a neuron with single input x and output y . Then the flow of signal through the neuron follows the formula $y=f(wx+by)=f(wx+b)$. For example, suppose that the input value is $w=0.87x=0.87$, with weight $w=0.53w=0.53$ and bias $b=0.12b=0.12$. Then the signal would first be combined as $wx+b=(0.53)(0.87)+(0.12)=0.3411wx+b=(0.53)(0.87)+(0.12)=0.3411$. This value would be fed as input to the activation function to produce $y=f(0.3411)y=f(0.3411)$. In the next section, we will discuss various activation functions used in neural networks. When there are multiple inputs, $x_1, x_2, \dots, x_{n1}, x_2, \dots, x_n$, each will be affected by its own weight. Typically, bias is thought of as a property of the neuron itself and so bias affects all the inputs in the same way. To be more precise, first, the inputs are multiplied by their individual weights, the result is summed, and then the bias is added. Finally, the activation function is applied to obtain the output signal, $y=f(w_1x_1+w_2x_2+\dots+w_nx_n+b)=f((i=1nwx_i)+by)=f(w_1x_1+w_2x_2+\dots+w_nx_n+b)=f((i=1nwx_i)+b)$ The weights and biases are parameters that the neural network learns during the training process, a topic we will explain in Backpropagation. A typical neural network may have hundreds or thousands of inputs for each neuron, and so the equation can be difficult to work with. It would be more convenient to regard all the inputs $x_1, x_2, \dots, x_{n1}, x_2, \dots, x_n$ as parts of a single mathematical structure, called a vector. A vector is simply an ordered list of numbers, that is, $x=(x_1, x_2, \dots, x_n)x=(x_1, x_2, \dots, x_n)$. (Note: In this text we use boldface letters to represent vectors. Also, the components of a vector are listed within a set of parentheses. Some texts vary on these notational details.) The number of components in a vector is called its dimension, so for example the vector $(5, 0, 2, 1, 2)$, $(5, 0, 2, 1, 2)$ has dimension 5. Certain arithmetic operations are defined on vectors. Vectors of the same dimension can be added: if $x=(x_1, x_2, \dots, x_n)$ and $y=(y_1, y_2, \dots, y_n)$, then $x+y=(x_1+y_1, x_2+y_2, \dots, x_n+y_n)x+y=(x_1+y_1, x_2+y_2, \dots, x_n+y_n)$. Any real number can be multiplied to a vector: $kx=k(x_1, x_2, \dots, x_n)=(kx_1, kx_2, \dots, kx_n)kx=k(x_1, x_2, \dots, x_n)=(kx_1, kx_2, \dots, kx_n)$. The dot product of two vectors of the same dimension results in a real number (not a vector) and is defined by: $xy=(x_1, x_2, \dots, x_n)(y_1, y_2, \dots, y_n)=1nx_iy_i=x_1y_1+x_2y_2+\dots+x_ny_nxy=(x_1, x_2, \dots, x_n)(y_1, y_2, \dots, y_n)=1nx_iy_i=x_1y_1+x_2y_2+\dots+x_ny_n$ If the inputs and weights are regarded as vectors, $x=(x_1, x_2, \dots, x_n)x=(x_1, x_2, \dots, x_n)$ and $w=(w_1, w_2, \dots, w_n)w=(w_1, w_2, \dots, w_n)$, respectively, then the formula may be re-expressed more concisely as: $y=f(wx+by)=f(wx+b)$ For example, if $w=(0.3, 0.1, 0.9)w=(0.3, 0.1, 0.9)$, $x=(1.2, 0.4, 0.6)x=(1.2, 0.4, 0.6)$, and $b=0.1b=0.1$, then $wx+b=(0.3)(1.2)+(0.1)(0.4)+(0.9)(0.6)+0.1=0.12wx+b=(0.3)(1.2)+(0.1)(0.4)+(0.9)(0.6)+0.1=0.12$ So in this example, the output would be $y=f(0.12)y=f(0.12)$, the exact value of which depends on which activation function $f(x)$ is chosen. Activation functions come in many types. Here are just a few of the most common activation functions. Step function, $f(x)={0, ifx$