artificial neural network (ANN)

Authors

Alexander Jung[*], Department of Computer Science, Aalto University, Espoo, Finland
[*]Corresponding author: alex.jung@aalto.fi

Abstract

An artificial neural network (ANN) is a graphical (signal-flow) representation of a hypothesis that maps the features of a data point at its input to a prediction for the label at its output. Its computational unit is the artificial neuron, which applies an activation function to the weighted sum of its inputs; the edge weights are the tunable model parameters. An ANN can be represented as a directed acyclic graph, and its connectivity structure — the architecture — is a central design choice. Deep nets arrange neurons in consecutive layers; varying the connectivity yields feedforward networks, convolutional neural networks (CNNs), recurrent networks, and transformer-based architectures.

Definition

The face recognition of a smartphone camera and the speech transcription of a voice assistant are both computed by networks of simple computing units with tuned connection strengths. Such a network is an ANN: a graphical (signal-flow) representation of a hypothesis that maps features of a data point at its input to a prediction for the corresponding label at its output.

The fundamental computational unit of an ANN is the artificial neuron, which applies an activation function $\sigma(\cdot)$ to the weighted sum of its inputs plus an offset term (Fig. 1); the edge weights of the network are its tunable model parameters.

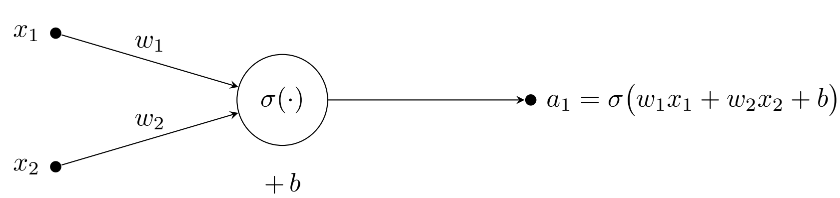

Figure 1: A single artificial neuron: it applies an activation function $\sigma(\cdot)$ to the weighted sum $w_1x_1 + w_2x_2$ of its inputs plus an offset term b, and delivers the output a_1. The weights w_1, w_2 and the offset b are tunable model parameters. This neuron appears as node a_1 of the ANN in Fig. 2, which does not draw the offset terms

The output of a neuron can be used either as the final output of the ANN or as an input to other neurons. One important design aspect of an ANN is its connectivity structure (or architecture), i.e., which outputs are connected to which downstream neuron's inputs (Goodfellow et al., 2016). As illustrated in Fig. 2, an ANN can be represented as a directed acyclic graph (DAG).

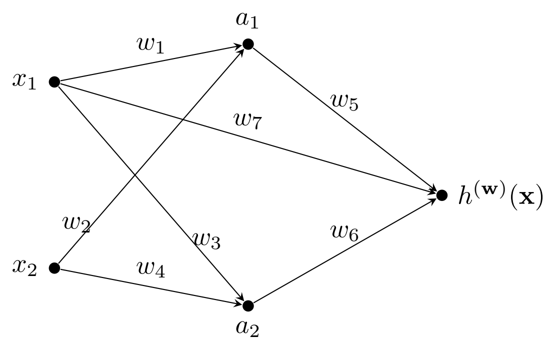

Figure 2: ANN represented as a weighted DAG with nodes corresponding to neurons or features of a data point. Feature nodes x_1, x_2 have a fixed output given by the feature value, with no input. Each of the neuron nodes a_1, a_2 applies an activation function to the weighted sum of its inputs (Fig. 1). The weighted directed edges indicate how node outputs are used as inputs to downstream neurons. The edge weights are tunable model parameters and are used to scale the inputs to the neurons. The prediction $h^{(\mathbf{w})}(\mathbf{x})$ is the output computed from the inputs x_1, a_1, and a_2

One widely used type of ANN is deep nets where neurons form consecutive layers (Fig. 3). In a deep net, the outputs of neurons in a given layer are typically only connected to the inputs of the neurons in a consecutive layer. Each layer l then applies a feature transformation $\mathbf{\Phi}^{(l)}$ whose components are computed by the neurons of that layer (each as in Fig. 1), and the deep net computes the concatenation of these layer-wise maps. The network of Fig. 3, with its dashed edge removed, delivers the prediction

$$h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{\Phi}^{(3)}\Big(\mathbf{\Phi}^{(2)}\big(\mathbf{\Phi}^{(1)}(\mathbf{x})\big)\Big),$$

starting from the input feature vector $\mathbf{x} = (x_1, x_2, x_3)^\top$: the two hidden layers compute $\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)} : \mathbb{R}^3 \to \mathbb{R}^3$ and the output layer computes $\mathbf{\Phi}^{(3)} : \mathbb{R}^3 \to \mathbb{R}$.

Sometimes it is useful to add shortcut or skip connections that directly connect the outputs of neurons in one layer to the inputs of neurons in a non-consecutive layer (Goodfellow et al., 2016; He et al., 2016).

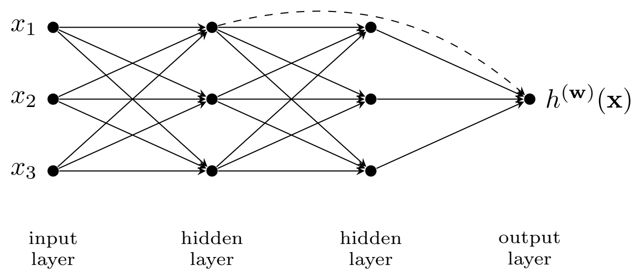

Figure 3: A deep net: the neurons form consecutive layers, and each neuron receives only the outputs of the preceding layer. Every solid edge carries a tunable edge weight (not drawn). The dashed edge is a skip connection, directly connecting two nonconsecutive layers

Varying the connectivity pattern yields different ANN architectures: feedforward networks connect layers strictly in sequence, convolutional neural networks (CNNs) share weights across spatial positions, recurrent networks include feedback connections for sequence processing, and transformer-based architectures underlie modern large language models (LLMs).

Cross References

label, artificial neuron, activation function, deep net, layer, CNN, transformer

References

1. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

2. He et al. (2016). Deep Residual Learning for Image Recognition. In: 2016 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778.