International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Research Paper on the Fundamentals of Artificial Neural Networks

Savan Patel

Jain Deemed-to-be University

ABSTRACT—

An Artificial Neural Network (ANN) is an information processing model inspired by the way biological nervous systems, like the brain, handle information[1]. The core feature of this model is its unique structure, consisting of numerous highly interconnected processing elements (neurons) that work together to address specific problems. ANNs learn by example, much like humans do. They are configured for specific tasks, such as pattern recognition or data classification, through a learning process. In biological systems, learning involves changes to the synaptic connections between neurons, and the same principle applies to ANNs. This paper provides an overview of Artificial Neural Networks, including their functioning and training[3]. It also discusses the applications and benefits of ANNs.

INTRODUCTION

The study of the human brain dates back thousands of years. With the advent of modern electronics, it became natural to attempt to replicate this cognitive process [4]. The first step toward artificial neural networks occurred in 1943 when Warren McCulloch, a neurophysiologist, and Walter Pitts, a young mathematician, wrote a paper on the functioning of neurons. They simulated a simple neural network using electrical circuits. Neural networks have a remarkable ability to extract patterns and detect trends from complex or imprecise data, which are often too intricate for humans or other computer techniques to discern [1]. A trained neural network can be regarded as an "expert" in analyzing the specific category of information it has been trained on. Other advantages include:

1. Adaptive Learning: The ability to learn tasks based on the training data or initial experience provided.

2. Self-Organization: An ANN can develop its own organization or representation of the information it receives during the learning phase[6].

3. Real-Time Operation: ANN computations can be executed in parallel, and specialized hardware is being designed to leverage this capability[9].

4. Fault Tolerance via Redundant Information Coding: Partial destruction of a network results in a corresponding degradation of performance, but some network capabilities may be retained even with significant damage[5].

Neural networks approach problem-solving differently from conventional computers. Conventional computers use an algorithmic approach, following a specific set of instructions to solve a problem. If the precise steps needed are unknown, the computer cannot solve the problem, limiting its problem-solving capability to issues we already understand and know how to address[1]. However, computers would be far more valuable if they could handle tasks for which we lack explicit instructions.

Neural networks process information similarly to the human brain. They consist of a large number of highly interconnected processing elements (neurons) working in parallel to solve specific problems. Neural networks learn by example rather than being explicitly programmed for a task. The examples must be chosen carefully to avoid wasting time or, worse, leading the network to function incorrectly[10]. A disadvantage of neural networks is their potential unpredictability since they determine how to solve problems independently. In contrast, conventional computers use a cognitive approach to problem-solving, requiring the problem-solving method to be known and expressed in clear, unambiguous instructions. These instructions are then converted into high-level language programs and then into machine code that the computer can understand. These systems are completely predictable in case anything goes wrong due to software or hardware glitches.[8]

Neural networks and conventional algorithmic computers are not in competition but complement each other. Certain tasks are more suited to an algorithmic approach, such as arithmetic operations, while others are better handled by neural networks[13]. Many tasks require systems that combine both approaches (typically, a conventional computer supervises the neural network) to achieve maximum efficiency.

What is an Artificial Neural Network?

Artificial neural networks (ANN) are nonlinear electronic models based on the neural structure of the brain. The brain learns from experience, demonstrating that problems beyond the scope of current computers can indeed be solved by small, energy-efficient packages[4]. This brain modeling also promises a less technical way to develop machine solutions, offering more graceful degradation during system overload than traditional methods.

These biologically inspired methods of computing are considered the next major advancement in the computing industry. Even simple animal brains can perform functions that are currently impossible for computers, which excel at rote tasks like keeping ledgers or performing complex math but struggle with recognizing simple patterns and generalizing them into future actions[6]. Advances in biological research suggest an initial understanding of the natural thinking mechanism, showing that brains store information as patterns. These patterns enable recognition of individual faces from various angles, encompassing a new field in computing that does not rely on traditional programming but involves creating massively parallel networks trained to solve specific problems.

How ANNs Work

ANNs are computers whose architecture is modeled after the brain. They typically consist of hundreds of simple processing units wired together in a complex communication network. Each unit or node is a simplified model of a real neuron, which sends off a new signal or fires if it receives a sufficiently strong input signal from connected nodes[3]. Traditionally, the term "neural network" referred to networks or circuits of biological neurons, but modern usage often refers to ANNs. ANNs are mathematical or computational models inspired by the biological nervous system, composed of interconnecting artificial neurons programmed to mimic the properties of biological neurons. These neurons work together to solve specific problems and are configured to address artificial intelligence problems without replicating a real biological system.

ANNs are used for speech recognition, image analysis, adaptive control, and more, achieved through a learning process similar to biological systems, involving adjustments between neurons through synaptic connections.[1]

Working of ANNs

The other aspect of using neural networks involves the myriad ways individual neurons can be clustered. This clustering occurs in the human mind dynamically, interactively, and self-organizingly. Biologically, neural networks are constructed in a three-dimensional world from microscopic components, allowing nearly unrestricted interconnections, unlike any existing man-made network. Integrated circuits, using current technology, are two-dimensional devices with limited interconnection layers, restraining the types and scope of artificial neural networks that can be implemented in silicon[9].

Currently, neural networks involve simple clustering of primitive artificial neurons by creating interconnected layers. How these layers connect is another aspect of the "art" of engineering networks to solve real-world problems[12].

Basic Structure of Artificial Neural Networks

All artificial neural networks share a similar structure or topology, as depicted in Figure 1. In this structure, some neurons interface with the real world to receive inputs, while others provide outputs. These outputs might represent the character the network has recognized or the image it thinks it is viewing. The remaining neurons are hidden.

However, a neural network is more than just a collection of neurons. Early researchers discovered that randomly connecting neurons without a structured approach was unsuccessful. Even simple brains, such as those of snails, are structured devices. One effective method of structuring is by creating layers of elements. The grouping of neurons into layers, the connections between these layers, and the summation and transfer functions are key components of a functioning neural network. These characteristics are common across all networks.

Although some networks can function with only one layer or even a single element, most applications require networks with at least three types of layers: input, hidden, and output. The input layer receives data from input files or electronic sensors in real-time applications. The output layer sends information directly to the external world, to secondary computer processes, or to devices such as mechanical control systems. Between these layers, many hidden layers can exist, containing numerous neurons in various interconnected structures. The inputs and outputs of each hidden neuron are directed to other neurons.

In most networks, each neuron in a hidden layer receives signals from all neurons in the layer above it, typically the input layer. After performing its function, a neuron passes its output to all neurons in the layer below, creating a feedforward path to the output. These communication lines between neurons are crucial as they provide variable strength to an input. Connections can either cause the summing mechanism of the next neuron to add or subtract, functioning as excitation or inhibition, respectively.

Lateral inhibition is a common feature where a neuron inhibits other neurons in the same layer. This is often used in the output layer to select the highest probability outcome, such as distinguishing between similar characters in text recognition. Another connection type is feedback, where the output of one layer routes back to a previous layer, influencing earlier stages of processing.

Training an Artificial Neural Network

Once structured for a specific application, an artificial neural network is ready for training. Initially, weights are chosen randomly, and then training begins. There are two primary training approaches: supervised and unsupervised.

Supervised Training

In supervised training, both inputs and desired outputs are provided. The network processes inputs and compares its outputs to the desired ones. Errors are propagated back through the network, adjusting the weights controlling the network. This iterative process continues until the weights are refined enough for the network to perform accurately. The dataset enabling this training is called the "training set." Commercial network development packages offer tools to monitor the network's convergence toward accurate predictions. Training can take days, stopping when the system reaches a statistically desired accuracy. However, some networks may never learn, possibly due to insufficient data or inadequate input-output relationships.

If a network fails to solve a problem, the designer must review the inputs, outputs, layer configurations, connections, summation, transfer, and training functions, and initial weights. Adjustments in these areas reflect the "art" of neural network design. Training rules are also crucial, with many algorithms available to implement adaptive feedback for weight adjustments. The most common technique is backward-error propagation, or backpropagation[11]. Effective training requires a balance to prevent overtraining, where the network memorizes data without generalizing well to new inputs. When training is complete, weights can be "frozen" for stability, or the system may continue to learn during production use.

Unsupervised Training

Unsupervised training provides inputs without desired outputs, requiring the network to self-organize and identify features to group the input data. This self-organization or adaptation remains a promising but not fully realized aspect of neural networks[2]. Though research continues, unsupervised learning is not as well understood or effective as supervised learning, which dominates practical applications.

Applications

Real-time applications of artificial neural networks include:

1. Function Approximation and Regression Analysis: Including time series prediction and modeling.

2. Call Control: Answering calls with hand gestures while driving.

3. Classification: Pattern and sequence recognition, novelty detection, and sequential decision-making[1].

4. Media Control: Using hand gestures to control media playback or volume.

5. Data Processing: Filtering, clustering, blind signal separation, and compression.

6. Web Page and eBook Navigation: Using hand gestures to scroll, ideal when hands are wet or dirty[12].

7. System Identification and Control: Applications in vehicle control, process control, game-playing, decision-making, pattern recognition, sequence recognition, medical diagnosis, financial applications, and data mining.

8. Smartphone as Media Hub: Controlling TV content with touch-free gestures.

9. Touch-Free Controls: Beneficial for maintaining cleanliness or for users who dislike smudges.

Advantages

1. Adaptive Learning: Ability to learn tasks based on training data.

2. Self-Organization: Can develop its own data representation during learning.

3. Real-Time Operation: Can perform computations in parallel.

4. Pattern Recognition: Effectively harnesses data information and generalizes patterns.

5. Development through Learning: Frees analysts from programming by teaching itself patterns.

6. Flexibility: Adapts well to changing environments.

7. Modeling Complex Interactions: Capable of handling complex interactions that are difficult for traditional approaches.

8. Performance: Often outperforms classical statistical modeling in less time.

Conclusion

This paper discussed artificial neural networks, their working, and training phases. ANNs have several advantages over conventional methods, particularly for dynamic or non-linear problems. They offer an analytical alternative to conventional techniques, capturing complex relationships and modeling phenomena that might otherwise be challenging to explain. The future of neural networks, however, relies heavily on advancements in hardware development, as current efforts primarily demonstrate that the principle works.

References

1. Bradshaw, J.A., Carden, K.J., and Riordan, D. (1991) Eco-app uses a new professional shell. Computational Applications in Biosciences, 7(1), 79–83.
2. Lippmann, R.P. (1987). An Introduction to Computational Neural Networks. IEEE Acoustics, Speech, and Signal Processing Magazine, 4(2), 4-22.
3. Murata, N., Yoshizawa, S., & Amari, S. (1993). Learning curves, model selection, and complexity of neural networks. In S. Jose Hanson, J.D. Cowan, & C.L. Giles (Eds.), From Advances in Neural Information Processing Systems 5, pages 607-614 San Mateo, CA: Morgan Kaufmann.
4. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.
5. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning was used for data recognition. Proceedings of the IEEE, 86(11), 2278-2324.
6. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
7. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.
8. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.
9. Zhang, W., & Du, J. (2019). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.
10. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
11. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.
12. Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout networks. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 1319-1327).
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).