Methods and Utility of Interpreting Artificial Neural Networks

Eliza Starr
3 April 2019

1. Introduction

The artificial neural network (ANN) is an automatic learning tool capable of supervised and unsupervised tasks such as classifying objects, predicting outcomes, clustering data, and recognizing patterns [1]. As a result, this tool has applications in a growing number of domains. However, ANNs have limited interpretability because it is difficult to decipher how a trained ANN arrives at an output. As a result, the ANN is called a black box algorithm, which means that its inner workings remain hidden while its input and output can be known and understood [2]. This is a challenge for applications requiring model transparency, which provides additional insight and understanding. In response, several mathematical approaches have been developed to interpret black box models, including ANNs. This paper will focus on a technique proposed by [4] in 2015 called layer-wise relevance propagation (LRP) and its utility in model validation. The ability to interpret black box algorithms will further increase the practicality of ANNs. For the sake of brevity, the background section will introduce ANNs and LRP, leaving complete mathematical explanation to [3] and [4]. The discussion section will lastly present the implications of ANN interpretation.

2. Background

2.1 Artificial Neural Networks

Although multiple ANN architectures exist, this paper covers the feed-forward ANN architecture called the multi-layer perceptron (MLP). Figure 1 depicts an example MLP architecture.

An MLP is a weighted directed graph whose nodes are organized into layers that are either partially or fully connected. A network has one input layer, one or more hidden layers, and one output layer. If the MLP has more than one hidden layer it is called a deep neural network.

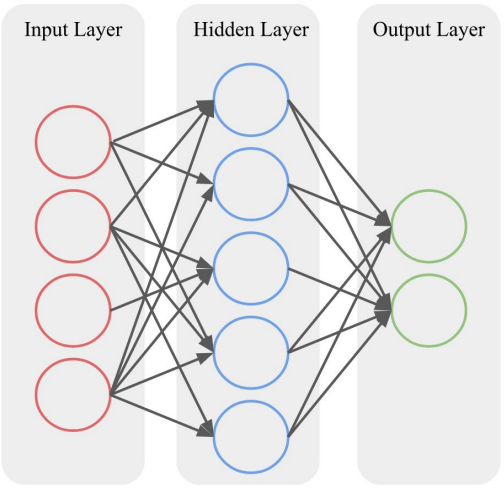

Fig. 1. A simplification of the MLP architecture with four input nodes, one hidden layer with five nodes, and two output nodes. This graph is not fully connected.

Each connection has a weight, which in the context of supervised learning is adjusted through a training method such as back error propagation [1]. The goal of training is to minimize the error between the output values and the given target values.

A node is a processing unit that accepts inputs from connected preceding nodes, performs a function, and sends the outputs of the function to connected nodes in succeeding layers. The function takes the weighted sum of the inputs and evaluates the result through an activation function, $g(n)$, to obtain the output. The output sent by a neuron i in layer l to a neuron j in layer $l+1$ is denoted $x_i^{(l+1)}$ [5]. Equation 1 shows this computation.

$$x_i^{(l+1)} = g\left(\sum_i w_{ij}^{(l \to l+1)} x_i^{(l)} + b_j^{(l+1)}\right) \tag{1}$$

The weight and bias parameters are denoted w and b, respectively. The activation function $g(n)$

can output discrete or continuous values. For example, the binary step function evaluates to 0 if $n < 0$ and to 1 if $n \geq 0$. The sigmoid function can be used to produce continuous probabilities ranging from 0 to 1 [1].

2.2 Layer-Wise Relevance Propagation

While [6] claims that layer-wise relevance propagation (LRP) has broad applications, the novel technique is primarily used for image classification. As a result, this section will explain LRP through the well-documented application of image recognition.

LRP is an interpretation technique that assigns relevance scores to individual pixels. The LRP process begins at the output layer and ends at the input layer, the first layer. Once the input layer is reached, the relevance score of pixel p from an image x is produced and is denoted R_p^1. The relevance score of pixel p is evidence against a class if $R_p^1 < 0$. Conversely, $R_p^1 > 0$ is evidence in favor of a class [7]. The relevance scores of all pixels in x can be visualized on a heatmap, which shows how each pixel contributes to the image's classification.

The scoring process relies on a conservation rule stating that relevance is neither gained nor lost between layers [8]. The rule is applied through a backward propagation process. To begin, the image's classification score is decomposed as the sum of relevance scores belonging to the output layer [9]. This sum is then conserved at each layer as LRP propagates backwards through the MLP until the input layer is reached.

Consider a trained MLP f with L layers. Each layer is a vector V indexed by $l \in \{1, 2, 3, \ldots, L\}$. Each layer vector is composed of neurons, each denoted as d. The MLP outputs a classification score $f(x)$, which will then be used to find the relevance score R_d using Equation 2 [9].

$$f(x) = \sum_{d \in V_L} R_d^{(L)} = \ldots = \sum_{d \in V_l} R_d^{(l)} = \ldots = \sum_{d \in V_1} R_d^1 \tag{2}$$

The naive propagation rule in Equation 3 shows how the relevance score from a neuron i at layer l is determined from the network weights of the layers that came before it in the LRP process [5].

$$R_i^{(l)} = \sum_j \frac{w_{ij}^{(l \to l+1)} x_i^{(l)}}{\sum_k w_{kj}^{(l \to l+1)} x_k^{(l)}} R_j^{(l+1)} \tag{3}$$

3. Discussion

The insights offered by LRP allow for knowledge extraction, ethical compliance, noise and error reduction, and model validation [6]. This section will focus on the opportunities LRP presents for model validation.

Without a way to interpret how a trained ANN uses features to produce outputs, train-test-holdout error estimation is the primary model validation procedure available. Techniques like LRP facilitate more methods of model validation including model comparison and improvement.

In the case of image classification, relevance heatmaps can be used to differentiate two trained MLP models with different architectures but similar train-test-holdout errors [8]. When plotted in a heatmap, pixel-wise relevance scores expose how the models use groups of pixels to classify an image. This visualization allows models to be compared by their inner logic. The following example from [6] describes this use case well.

Lapuschkin et al. [8] trained an MLP and a Fisher vector classifier to classify images of horses. The models performed with similar test-train-holdout accuracies, but the heatmap created from LRP reveals that the MLP relies on the outline of the horse while the Fisher vector model uses a copyright tag commonly found in the images of horses [6]. With this information, the MLP is preferred over the Fisher vector because it used more robust classification logic.

Models can also be improved using interpretation methods. This can be seen through a hypothetical extension of the previous example. The correlation between copyright marks and horse images suggests that the models could be improved if they were trained and tested using different data [6]. For example, removal of the copyright from the horse images or the inclusion of non-horse images with copyright symbols could allow the Fisher net to pick up on relevant pixel patterns having more to do with the horse.

4. Conclusion

The artificial neural network (ANN) is a popular technique with seemingly limitless applications in business and academia. Various ANN architectures, like the multi-layer perceptron (MLP), expand the types of tasks the tool can accomplish. However, the hidden and abstracted structure of the ANN limits its interpretability.

Layer-wise Relevance Propagation (LRP) is one of several methods created to explain the inner workings of ANNs. This technique was applied successfully to various image recognition tasks, but other applications of LRP have yet to be explored and documented. Nonetheless, interpretation methods like LRP make the ANN more transparent, a characteristic preferred and sometimes required by certain applications. Among other benefits, LRP supplements rudimentary train-test-holdout error analysis by providing new modes of model validation. These developments in model interpretability will likely preserve the popularity and improve the utility of ANNs for years to come.

4. References

[1] J.E. Dayhoff, J.M. DeLeo, Artificial Neural Networks Opening the Black Box, CANCER Supplement. 91 (2001) 1615-1635. doi:10.1002/1097-0142(20010415)91:8 3.0.co;2-l.

[2] Definition Of Black Box, Merriam-Webster. (n.d.). https://www.merriam-webster.com/dictionary/black%20box (accessed April 2, 2019).

[3] S. Haykin, Neural Networks, 2nd ed., Pearson Education, Delhi, 1999.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.R. Müller, W. Samek, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLoS ONE. 10 (2015). doi:10.1371/journal.pone.0130140.

[5] A. Binder, S. Bach, G. Montavon, K.-R. Müller, W. Samek, Layer-Wise Relevance Propagation for Deep Neural Network Architectures, Lecture Notes in Electrical Engineering . 376 (2016) 913-922. doi:10.1007/978-981-10-0557-2_87.

[6] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digital Signal Processing. 73 (2017) 1-15. doi:10.1016/j.dsp.2017.10.011.

[7] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, W. Samek, Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers, Lecture Notes in Computer Science. 9887 (2016) 63-71. doi:10.1007/978-3-319-44781-0_8.

[8] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, W. Samek, Analyzing Classifiers: Fisher Vectors and Deep Neural Networks, Computer Vision Foundation. (2016) 2913-2920. doi:10.1109/cvpr.2016.318.

[9] H. Li, Y. Tian, K. Mueller, X. Chen, Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation, Image and Vision Computing. (2019) 6-7. doi:10.1016/j.imavis.2019.02.005.

Honor Statement: This essay is my own work, and I have cited all ideas stemming from outside sources.