ISSN: 2641-2039

Global Journal of
Engineering Sciences

DOI: 10.33552/GJES.2020.05.000621

Iris Publishers

Opinion

Explainable and Interpretable Deep Learning Models

Md Shamsuzzaman[1]* and Mysore Satish[2]

[1]Engineering Division, Saint Mary's University, Halifax, Nova Scotia, Canada

[2]Civil & Resource Engineering, Dalhousie University, Halifax, Nova Scotia, Canada

***Corresponding author:** Md Shamsuzzaman, Engineering Division, Saint Mary's University, Halifax, Nova Scotia, B3H3C3, Canada.

Received Date: May 22, 2020

Published Date: June 09, 2020

Opinion

Deep learning models are becoming ubiquitous recently. However, they still suffer from many limitations. In this mini review, we will focus on one of these aspects of deep learning models.

As is well known, the objective of modeling is to capture the intended behavior of a system in such a way that it can be used for inference, prediction and/or classification. Surely, the obtained model can be used as a crucial tool for decision making in many cases. Machine learning (ML) and deep learning (DL) are branches of artificial intelligence, where the model with the associated parameters are developed using data. Here, we use DL to mean Deep Artificial Neural Network explicitly.

As the name suggests, Artificial Neural Networks (ANN) are composed of multiple simple computational blocks called artificial neurons. An artificial neuron is composed of a linear summing unit (like a linear regressor, it has weights for each input along with a bias) followed by a non-linear transfer/activation function. A single neuron is not capable of solving complex problems and therefore, multiple neurons are connected together to form a network where the output from a neuron is used as an input to other neuron(s). Conventionally, the way the neurons are connected forms a layered architecture where the first layer takes the input data known as the input layer and the last layer provides the desired prediction/classification and known as the output layer. One or more layers are sandwiched between these two layers and are known as hidden layer(s). The neurons from one layer are connected to the neurons from other layers but are not connected within a layer. The number of layers is domain/application specific. Moreover, based on the architecture, all the output from a particular layer may/may not be connected to all of the input of the following layer. Deep neural network (DNN) usually has several hidden layers. While number of input and number of output (and thereby the number of neurons in these layers) are problem specific, the number of hidden layers and the number of neurons they contain are design specific and it is beyond the scope of this review. Many popular DNN architectures have hundreds of layers with several millions of parameters. Several architectures have showed human-level performance in different problems. Deep learning models have been applied in different fields including but not limited to aerospace, automotive, banking, business, defense, engineering, entertainment, finance, image processing, insurance, medical, oil and gas, speech, securities, telecommunications, transportation (including emerging driverless vehicles) and environment. However, despite its promising performance and recent developments, DNN still suffers from many limitations, namely,

- Deep learning models need huge amount of data to train the model. In general, a deep learning model has several hidden layers (hence the name deep) and therefore it has many more parameters compared to other models. Naturally, training set should be large enough to successfully train a model (in comparison to parameters). For supervised learning cases, that means the (training) data must be labeled too. Labeling data is a labor-intensive process and it is still very difficult to obtain huge amount of labelled data.
- Even if a deep learning model is trained with large amount of data, if the distribution of the training data does not capture application domain, it might fail measurably. For example, a surface water flow model trained using Canadian data with more than 98% accuracy can't be applied in India, as the conditions are totally different. In short, it is very difficult to

generalize a model and a trained model may not always transfer well to the real world.

- Surely, one can train a model with new data set. However, it is still expensive to train a deep neural network – apart from data, it needs lots of computational power and time.
- Most importantly, it is difficult, if not impossible, to explain why a deep neural network model works or fails. In other words, a deep learning model does not convey what actually caused or contributed to the target. Therefore, it is difficult to isolate the major input factors and make decisions and/or act accordingly.

From the above discussion it is clear that deep learning models are very good at input-output mapping and the internal black-box representation is opaque to both the developers and the users. However, in many instances, for example medical, it is not enough to get a good decision alone, an explanation is often desirable and sometimes binding due to legal or ethical issues. Moreover, trust is another factor that depends on interpretability and explain ability of the model.

Unfortunately, there is no unanimous definition of these two terms and in many cases, they are used interchangeably. However, it is generally agreed that interpretable models usually come at a cost of lower predictive performance. The current goal of the researchers is to provide understandable predictions without sacrificing accuracy of the model where uninterpretable predictors has proper explanations. In this regard, different explanators are built using a number of techniques, like, decision rules, model simplification, feature importance, saliency mask, sensitivity analysis, visual analysis etc. The explanators can be either local or global.

Researchers have adopted different approaches to achieve this goal – additional implicit/explicit DL node to capture explain ability; model induction causal model to learn an explainable, causal, probabilistic programming model; pattern theory; adaptive programs; cognitive models; reinforcement and attention-based learning; question-answer system; incorporating shallow models for explanation etc.

In conclusion, if proper training data and resources are provided deep learning models could outperform other models. However, these models are not explainable and interpretable in general. This is a required and an active research area for deep learning models.

Acknowledgement

None.

Conflict of Interest

No conflict of interest.