V-ML: VISUALIZATIONS FOR MACHINE LEARNING

Viraj Chogle
Department of Information Technology
Vidyalanakar Institue of Technology
Mumbai, India
virajchogle@gmail.com

Ankit Katre
Department of Information Technology
Vidyalanakar Institue of Technology
Mumbai,India
ankit_katre@yahoo.com

Pratyusha Parashar
Department of Information Technology
Vidyalanakar Institue of Technology
Mumbai,India
pratyushparashar1999@gmail.com

***Abstract*— Data visualization is a quite new and promising field in computer science. It uses computer graphic effects to reveal the patterns, trends and relationships out of datasets. However, there is a dearth of visualization tools for deep learning algorithms along with the details buried in the algorithms that are properly accounted for. Gaining an understanding of this technology is somewhat difficult. While the theory is important, it is also helpful for novices and users not familiar with machine learning to develop an intuitive feel for the effect of different hyperparameters and structural variations. We propose a visualization tool for some of these algorithms along with the description of "intricate details" that often result in novices describing them as 'black boxes'.**

Keywords—data visualization, deep learning, machine learning, hyperparameters.

I. INTRODUCTION

In many areas, such as artificial intelligence, bioinformatics and finance, learning data representation is a critical step in performing follow-up planning, retrieval and recommendations. Often, in large applications, how you can learn the structure of the data and get important information from the data becomes more urgent, important and challenging.

The recent success of in-depth learning and machine learning has led to a wave of interest for non-professionals. Gaining an understanding of this technology, however, is difficult. While perspective is important, it is also helpful for novice to develop an accurate sense of the effect of various hyperparameters and structural variations.

Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with graphicsThis is important because it allows styles and patterns to be easily recognizable. With the rise of big data over us, we need to be able to translate large data sets. Machine learning makes it easy to perform analyzes similar to forecasting analysis, which can serve as a useful detection to present. But data visibility is not only important for data scientists and data analysts, it is important to understand data perception in any field. Whether you work in finance, marketing, technology, design, or anything else, you need to visualize data. That fact illustrates the value of recognizing data..

II. AIM AND OBJECTIVES

Provide a visual summary of information that makes it easier to identify patterns and trends than to look at thousands of lines in a spreadsheet.. It's the way the human brain works. As the purpose of data analysis is to gain understanding, data is very important when viewed. Although a data analyst can draw comprehension from data without being seen, it will be very difficult to communicate the meaning without seeing it. Charts and graphs make data acquisition easier even if you can identify patterns without

them. Visualization tool will help to communicate information clearly and efficiently to users, learners and professionals. With the extensive scope of machine learning and big data, more and more people want to educate themselves about these fields. Visual graphics and models will make it simpler for even laymen to understand patterns, working and other intricacies of data.

III. RELATED WORKS

A. *Decision Tree ML Model*

In paper [1] S. Yee and T. Chu developed and hosted a website which gives a visual introduction to Machine Learning. Using a dataset about homes, they lay out the basic steps involved in a building a simple machine learning model (decision tree) which can be used to distinguish homes in New York and San Francisco from each other with a high accuracy (=89%). It is essentially an experiment in expressing statistical thinking with interactive design.

B. *Direct Manipulation Visualization of Deep networks*

In paper [2] Daniel Smilkov, Shan Carter and others proposed direct manipulation visualization of Deep networks such as Convolutional neural nets. They Described TensorFlow Playground(http://playground.tensorflow.org) which is an interactive, open-sourced visualization that allows users to experiment via direct manipulation rather than coding, enabling such users to quickly build an intuition about neural nets.

Playground view structure is a standard network diagram. The view shows a network designed to solve division or reversal problems based on two values with real values, x1 and x2, which vary between -1 and 1. Input units, representing these features and various mathematical combinations, are on the left. Hidden layer units are displayed as small boxes, with interactions between the units drawn as curves their color and width indicate the values of the weight. Finally, to the right, a network output display is displayed: a square with a heat map showing the output value of one unit up the final layer of the network. When the user presses the "play" button, the network begins to train Units.

C. *Machine Learning models for Diagnostic Purposes*

In paper [3] Dong Chen and Rachel K. E. Bellamy Proposed visualizing State-of-the-art Machine Learning models for Diagnostic Purposes. This study had a 2-phase design. They started with a "static" probe study with twelve ML practitioners, in which they probed for user's reaction to

ten visualizations published in previous literature. They then summarized their findings and prototyped an interactive visualization embodying several of the design themes. Their contributions include:

- an investigation of user needs in ML diagnosis through both static and interactive design probe studies; and
- a set of design themes to help inform the design of future supporting tools for ML diagnosis

D. Data Visualization using Machine Learning

ML practitioners are hoping on summary statistics to assess the standard of a model. Metrics like accuracy or error rate, precision and recall, and AUC (Area Under Curve) give people a short overview of model performance. The confusion matrix visualization offers a more detailed view by laying out data during a table with columns because the predicted value and rows because the actual value specified a cell contains the count of information items where the anticipated value matches that actual value. The modeler's goal is to possess high counts for cells where the anticipated and actual value are the identical. Sometimes the matrix cells are color coded to spotlight the problematic predictions [10]. the matter with these conventional techniques, however, is that they only convey the performance of a model and don't inform users of error severity or causes of the errors. Users need to depend on separate tools to focus on data instances that were incorrectly predicted. They often don't have any clue what causes these errors and the way to resolve them. In recent years the research community has been calling for empowering the role of user in improving ML models [7], [1], [6]. Fails and Olson [11] proposed interactive ML, geared toward lowering the expertise barrier in ML and enabling users to iteratively evaluate and improve a model. variety of advanced visualization tools are developed in recent years. SmartStripes [12] helps users interactively select features. EnsembleMatrix [10] converts the obscure activity, model ensembling, to an interesting visual interaction, and enables people to experiment with various combinations of models to optimize the result. ManiMatrix [8] empowers users to point their error tolerance, and lets user interactively steer a model to their preference. ModelTracker [4] provides an intuitive interface for performance analysis and debugging. Like us the designers of ModelTracker seek to facilitate ML diagnosis. The tool itself is proscribed to binary classification problems. In our study we include these visualizations as probe materials. By lecture participants about how they'd use these visualizations, and having them act out a diagnosis scenario, we are able to better understand what works and what doesn't. They also help us seek inspirations for brand spanking new visualizations to support ML diagnosis.

IV. PROBLEM sTATEMENT

To develop an application for data visualization that bridges the gap between machine learning concepts and their practical implementation through visuals and graphics so on make them easier to understand. Development of an interactive platform that helps not only professionals but also laymen and learners to raised understand the intricacies and functioning of machine learning models and algorithms. rather than applying algorithms to the various data sets individually and separately, development of a system that applies such algorithms in an exceedingly more generalized way and together. Development of a visually appealing platform for data representation site that caters to both learning and entertainment simultaneously.

V. IMPLEMENTATION

The system tries to supply an interactive way for the user/learner to visualise and understand some machine learning models. While the idea is vital, it's also helpful for novices to develop an intuitive condole with the effect of various hyperparameters and structural variations. The working essentially involves a user providing a dataset and putting in place the specification of the neural network(say) such variety of layers, learning rate, regularization parameter etc and also the system outputs pertinent visualizations supported the information together with the model stats regarding its accuracy and other such metrics. The system would preprocess the information at backend, create and train the model with parameters set by the user via a graphical computer program. together with the visualizations we might also help the user in understanding the code behind it with help of various sources. The proposed system would enable a novice to achieve basic understanding of the concepts and its implementation/ best practices.

Visual story-telling of information is that the foremost approach to present not just a table of numbers, but importantly the statistical results and interpretations of the algorithm results to make prescriptive actions. A standardised approach to information design with a user-centric approach, and by designing the correct navigation workflow, pertinent representations and relevant visual design is that the right place to urge started on this journey.

A) Algorithms used

1) Random Forest Algorithm

Random forest could be a supervised learning algorithm which is employed for both classification additionally as regression. Random forest algorithm creates decision trees on data samples so gets the prediction from each of them and at last selects the most effective solution by means of voting. it's an ensemble method which is healthier than one decision tree because it reduces the over-fitting by averaging the result.

Parameters utilized in Random Forest Algorithm: Number of trees (n_estimators), minimum samples per leaf, max features, Out of Bag Score (oob_score)

2) Neural Networks

Neural networks are a group of algorithms, modeled loosely after the human brain, that are designed to acknowledge patterns. They interpret sensory data through a sort of machine perception, labeling or clustering raw input.

The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or statistic, must be translated.

Parameters used for Neural Network: Learning rate, dropout, weight decay, batch normalization, number of epochs, etc.

3) K-Nearest Neighbors Algorithm

The k-nearest neighbors (KNN) algorithm could be a simple, easy-to-implement supervised machine learning algorithm that may be wont to solve both classification and regression

problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are around one another Parameters used for K-Nearest Neighbors: Number of neighbors, weights and leaf size.

4) Naïve Bayes Classifier

It is a classification technique supported Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a specific feature in an exceedingly class is unrelated to the presence of the other feature.

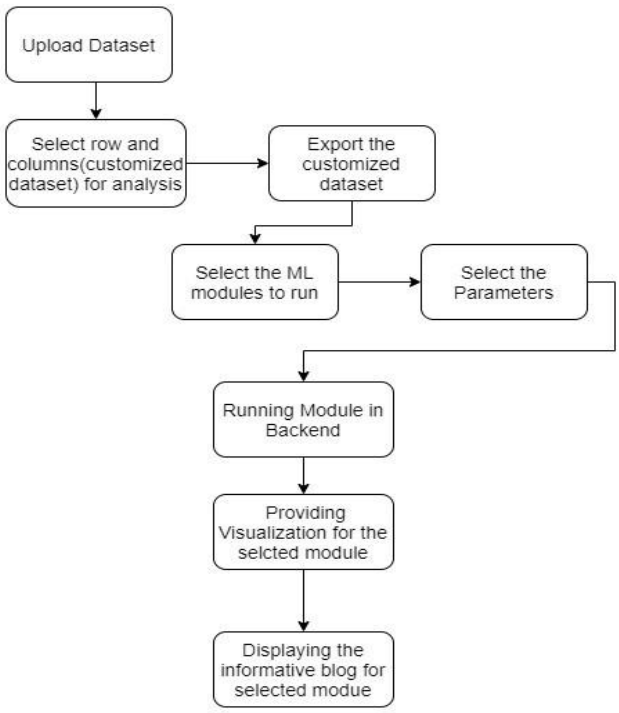

Fig 1:Flowchart for proposed system

ACKNOWLEDGMENT

We are pleased to present "V-ML: Visualizations for Machine Learning" as our project and take this opportunity to express our profound gratitude to all those people who helped us in completion of this project. We thank our college Vidyalankar Institute of Technology for providing us with excellent facilities that helped us to complete and present this project. We express our deepest gratitude towards our project guide Prof. Indu Anoop and Prof. Rasika Ransing for her valuable and timely advice during the various phases in our project. We would also like to thank her for providing us with all proper facilities and support as the project co-coordinator. We would like to thank her for support, patience and faith in our capabilities and for giving us flexibility in terms of working and reporting schedules.

REFERENCES

[1] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer, "Guiding feature subset selection with an interactive visualization," in Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, 2011, pp. 111–120. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6102448 [

[2] B. Gaver, T. Dunne, and E. Pacenti, "Design: Cultural probes," interactions, vol. 6, no. 1, pp. 21–29, Jan. 1999. [Online]. Available: http://doi.acm.org/10.1145/291224.291235

[3] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.

[4] UCI, "UCI Machine Learning Repository," 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance [

[5] J. Corbin and A. Strauss, Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage publications, 2014.

[6] S. Yee and T. Chu, "A Visual Introduction to Machine Learning," 2015. [Online]. Available: http://www.r2d3.us/visual-intro-to-machinelearning-part-1/

[7] D. A. Norman, The design of everyday things: Revised and expanded edition. Basic books, 2013.

[8] [1] S. Yee and T. Chu, "A Visual Introduction to Machine Learning," 2015. [Online]. Available: http://www.r2d3.us/visual-intro-to-machinelearning-part-1/

[9] Daniel Smilkov and Shan Carter, Direct-Manipulation Visualization of Deep NetworksAvailable: http://playground.tensorflow.org

[10] Dong Chen and Rachel K. E. Bellamy, "Diagnostic Visualization for Non-expert Machine Learning Practitioners: A Design Study"

[11] K. Patel, "Lowering the barrier to applying machine learning," Ph.D. dissertation, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=1866222

[12] N. Elmqvist, T. N. Do, H. Goodell, N. Henry, and J. D. Fekete, "ZAME: Interactive large-scale graph visualization," IEEE Pacific Visualisation Symposium 2008, PacificVis - Proceedings, pp. 215–222,2008.