Seminars In AI

Topic 2: Explainable AI (XAI)

Farooq Ahmad Wani

Matricola: 1946707

(Seminar delivered by Biagio La Rosa)

1. Abstract

The recent success of machine learning from big data, especially deep learning has made AI very popular. The widespread use in industrialization and its super human performance in a significant number of tasks has made it reach every human being. However, unfortunately, this surge in performance has been achieved by using complex models which are not interpretable to human beings and turning these models into the black box. As a result of this many fields are abstaining from the use of these techniques where they could have resulted in and contributed to enormous growth. The research field of explainable AI (XAI) provides the necessary foundations and methods to explain the decisions and the internal mechanism of complex AI systems. This article will be focused to explain the present techniques and methods available to explain the interpretability of the machine learning and deep learning models. Some links and references to the programming implementation of these techniques will also be provided.

2. Introduction

In the past decade, we have seen a tremendous development in machine learning and related fields. Due to the increase in data and availability of cheap computer hardware deep learning has re-invigorated machine learning. The super human performance of deep learning and AI to solve complex problems has made AI extremely popular. However, its power is also its jeopardy, the deep learning models are composed of millions of complex internal calculations among millions of parameters which make it quite tough for humans to interpret them and thus act practically as the black box. There are many domains where the chances of its failure also increase. Over the last decade, the demand for the toolbox which can explain these mockups to common people, researchers, and domain experts has increased. The field of explainable AI focuses on the development of such tools. Although this field is mainly associated with the development of the methods and foundations for the transparency and tractability of AI and deep learning models, in the future we may need to devise the measuring metric to express the causatives and quality of the explanation. This will not only improve the legal handling of models but also increase the interpretability and human expert augmentation of the AI models. Trusted AI requires both robustness and explainability. The AI solutions should conform to human values, ethical principles, and legal requirements and ensure the privacy, security, and safety of the user.

The narrow definition of the XAI is “the techniques and the methods that make AI and deep learning model decisions understandable by people”. The broader definition is “everything that makes AI understandable including data, functions, and performance, etc.”

Let us get the pictural understanding from the given figures. **Figure 1** shows the simple task of the classification using the features of the data. **Figure 2** shows what are the different questions the XAI is answering.

3. Developmental Overview

Over the years the number of the methods have been introduced to describe the working of deep neural networks, mostly the methods have been developed over the two axes:

1. **Local and Global methods:** as suggested by the name these are the methods that are used to explain the local and global behavior of the parameters respectively. In some cases, the global explanation arises from the local explanations but it is not necessarily true for all AI models.
2. **Post-hoc and Ante-hoc methods:** post-hoc "after this event" are the models which provide the explanation for the specific solution of the black box approach, examples are LIME, BETA, LRP, etc. Ante-hoc are the models which are interpreted immanently in the system, i, e they are transparent by nature "white box" like interact machine learning models.

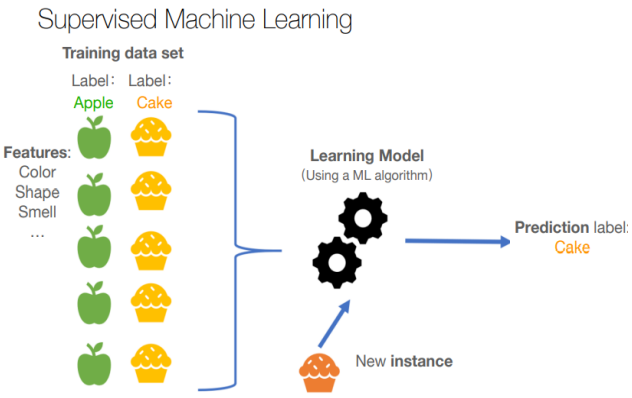


Figure 1

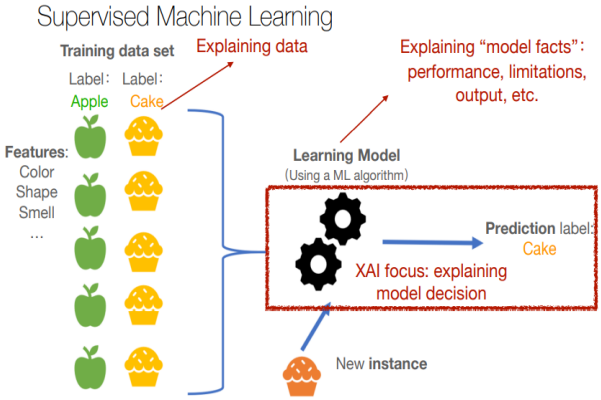


Figure 2

There have been several attempts to explain the prediction of the models, the most active among them has been on the problem of **feature attribution.** The feature attribution aims to explain which part of the input is most responsible for the output of the model. For example, in computer vision tasks we have **heatmaps** to show the region of the image that affects mostly the output. Other techniques include **feature visualization, interpretability by design** etc. With increased research in this field there have been a number of the gatherings and conferences which have set the benchmark for the explainability and interpretability.

4. Current Methods

In this section, I will provide a short overview of the current models used in explainable AI and their drawbacks or challenges. **Quantus** is the toolbox that provides an exhaustive collection of the evaluation methods and metrics for the explanations. The **CLEVR-XAI** is the benchmark dataset for the ground truth evaluation of the neural networks. Before diving into the details let us provide a chronological overview of the explanatory methods:

- **LRP (2015):** layer-wise relevance propagation heatmap method based on the deep Taylor decomposition.
- **LIME (2016):** use interpretable feature space and local approximation with sparse K-LASSO.
- **SHAP (2017):** additive, use Shapley values (game theory) unifies Deep LIFT, LRP, LIME.
- **Integrated gradients (2017):** aggregates gradients through the path between baseline and observation of interest.

- **TCAV (2018):** gradient based models, use concepts learned from annotated samples.
- **Anchors (2018):** model agnostic and rule based, sparse, with interactions.
- **ASV (2020):** Asymmetric Shapley values use variable distributions and dependency structure.
- **Graph LIME (2020):** Interpretable model for graph networks from N-hop neighbourhood.
- **XGNN (2020):** post-hoc global-level explanations for graph neural networks.
- **Shap Flow (2021):** use graph-like dependency structure between variables.

i. LRP

Layer-wise relevance propagation had been initially developed to explain the convolution neural network models but now it has been extended to many other networks by different researchers. For example, we can use it to explain the output of the LSTM, GNN, and transformers. The main idea of the LRP is to explain the models using the decomposition. The principal approach of the method is to access the different internals of the model like topology, weights, activations, etc. This additional information about the model makes the explanation problem easy for the LRP. It exploits the network structure and redistributes the explanatory factors from layer to layer starting from the output. These explanatory factors are called the relevance factors.

ii. LIME

The Local Interpretable Model agnostic explanation does not have direct access to the model parameters of the actual or the target model, it only has the access to the input and the output of the target model. It surrogates the explanation of the target model by generating input samples close in the vicinity of the actual inputs and then evaluates them using the target model and subsequently approximates the target model by the simple linear function. The main limitation with the use of the LIME is that it heavily relies on the quality of the surrogate fit which can be only done by dense sampling.

iii. SHAP

In this method, the importance of the different variables or features is analyzed by adding the variable “i” to set “S” and then analyzing how the value of the function “es” changes. Here the function “es” can be defined as a value function that explains the model “f” at some individual point “x*”. Typically, the function is defined as the expected value for the conditional distribution in which the conditioning applies to all variables in a subset. The contribution of the variable is calculated as the weighted average over all the possible subset “S”. In short, the analysis of single ordering shows how adding the consecutive variables changes the value of the value function. This is the adaptation of Shapley values which will be discussed later in this report. SHAP draws from a rich theoretical underpinning in game theory and fulfills desirable axioms, for example, those features that do not contribute to the prediction, get an attribution of zero.

iv. Integrated Gradients

Fundamentally the method is based on the two axioms', **sensitivity and the implementation invariance**. The term sensitivity is calculated by assigning a non-zero attribute to every input and baseline the one which differs in different predictions. The invariance is calculated by checking the difference in two models which are functionally identical. The approach proposed by the integrated gradients method for the model aggregates the gradients computed along the path connecting the pointed interest to the highlight observation. The integral can be replaced by the sum over the set of alpha values in the limit from {0,1}. The disadvantage of this method is the gradient shattering in the deep neural models. There are some extensions on this model which are used to tackle the shattering problem but they are out of scope for this report.

v. TCAV

This method is based on the Concept Activation Vectors, which describes how the neural activations are influenced by the presence or absence of a particular concept like color etc. In this method, we combine the two datasets one with the concept features and the other without them, then we send it to the regressor for classifying these two datasets. The regressor parameters are stored and passed as the activation parameters in the target network to check the effect of the concept on output. Although it is related to a single concept local, we can combine them across and give a global explanation. It calculates the ratio of the images with positive conceptual sensitivity. This technique is used to check whether the network has learned the flawed concepts and is very powerful in this regard.

vi. ANCHOR

The anchors are model-independent and can be applied to any domain of data although the present implementations prefer tabular and text data. This technique mainly works by finding the decision rules which can explain the model output. The rules are primarily in the form of IF_THEN statements. The explanation remains the same no matter how much other features change which are not part of the anchors or decision rules. The features chosen for the anchoring should have very high precision and coverage. Although this method is easy to explain it needs a lot of the calculations to select the anchors.

vii. ASV

This is the extension of the SHAP technique in a way that if two values have the same effect on the model's behavior it will not take the individual effects but check whether there is any causal effect on one another. The relation of the causal effects is expressed in the form of the causal graphs which allows the redistribution of the source variables. This method also calculates the effect of the variable as the average effect on a coalition of the other variables, the order of the variables is followed by obeying the casual graph links. A particular application of the ASV value is the model fairness analysis.

viii. XGNN

It is the post-hoc method. It starts by finding the adequate graph starting randomly from the chosen node as defined by the prior knowledge. It tries to increase the performance of the Graph neural network and keeps generating valid graphs depending on the domain requirements. This technique grows the graph only by the edge addition to the existing nodes and if the node added has a non-desirable contribution, then it is updated by the negative reward. This method was mainly invented to describe the graph classification, it is worth mentioning here that this is the only technique that gives the model-level explanation for the GNN.

ix. GraphLIME

Although it takes most of the concepts from its parent LIME it is non-linear in estimation. This technique is mainly applied to GNN (graph neural networks). These methods work on the non-Euclidean datasets and try to find the interpretable method to explain the working of the GNN. In this case, it is the Hilbert Schmidt Independence criterion that is used to explain the node in the graph Neural Network. This method believes that during the node classification by GNN the nodes are represented by several nonlinear aggregations and combinations. Finally, it is important to note that GraphLIME is successfully used for the investigation of backdoor attacks on GNNs by uncovering the relevant features of the graph's nodes.

x. SHAP FLOW

Shap flow or the Shapley flow like the ASV allows the use of the dependency structure between the variables in the explanation process. In this case, contrary to other methods of this type the contribution is not to the node but to the edges of the casual graph if the removal of the edge has a role in changing the value of the prediction of the model. The edge attribution has the property of the addition. The Shapley flow determines the contribution of each edge of the causal graph rather than the nodes of the graph. Shapley Flow attribution analysis carries a lot of information about both the structure of the relationship between variables and the effect of groups of variables (explanation boundaries) on the predictions.

5. General Drawbacks

a) The vast number of the methods makes it difficult for the researcher to choose the interpretation model that can answer all the questions and on other hand fitting the single interpretability technique for all can lead to dangerous misinterpretation.
b) The behaviour of the techniques for the under and overfit can lead to misleading interpretations concerning the true features. Formally the models are designed to explain the methods not to draw the inferences.
c) Sometimes there is the unnecessary use of the complex models, a common mistake of using the complex ML models when an interpretable model would have been sufficient.
d) When features are dependent, perturbation-based IML methods such as PFI, PDP, LIME, and Shapley values extrapolate in areas where the model was trained with little or no training data, which can cause misleading interpretations.
e) Features with a Pearson correlation coefficient (PCC) close to zero can still be dependent and cause misleading model interpretations. While independence between two features implies that the PCC is zero, the converse is generally false.
f) Conditional variants of these interpretation methods do not replace feature values independently of other features, but in such a way that they conform to the conditional distribution.
g) Some methods are true to the model but not true to data if the dependencies change in data.
h) These models mostly ignore the model approximation and uncertainty.
i) There are some misleading interpretations due to feature integrations like aggregation etc.
j) Sometimes different models explain the data-generating process equally well but contradict each other. This phenomenon is called the Rashomon effect, named after the movie “Rashomon” from the year 1950.
k) These methods mostly fail if the dimensionality is increased, and needs very high computational effort.
l) Only a few methods among them have causal links and thus lack the interpretability of the causal links.

6. Predictions for Unsupervised Learning Models

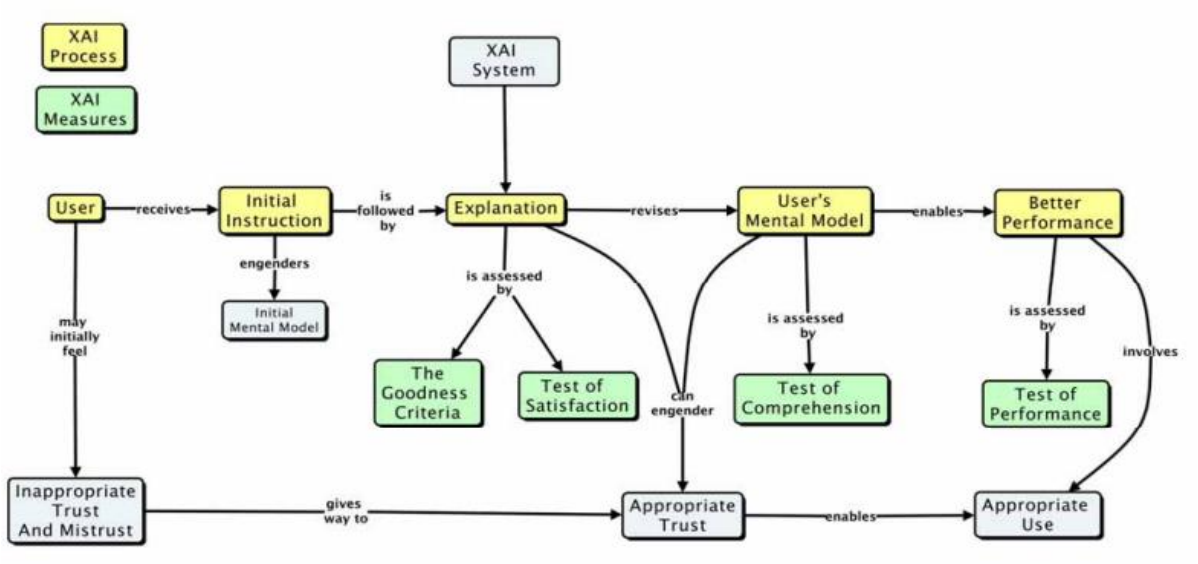
The **kernel density estimation method** is one of the most widely used methods for the explainability of unsupervised learning. The KDE method was originally used to detect the anomaly. it assumes the data set is completely unlabeled and typically the kernel of gaussian type is used. The KDE explains the prediction for the data point by computing the function which can be treated as the unnormalized probability density function. This score is used to determine the inliers and outliers of different data points. This method is fully described in the paper attached in reference.

7. Conclusion

Although the XAI topic is very hot and there are number of techniques which are used to explain different AI models, but while explaining these models some core properties get lost like speed,

efficiency, accuracy, and reliability, thus we need some newer and global methods which will be compact as well as useful.

8. Metrics for XAI



Robert R. Hoffman, Shane T. Mueller, Gary Klein & Jordan Litman 2018. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608.

9. References

1. XXAI - Beyond Explainable AI, Andreas Holzinger · Randy Goebel · Ruth Fong · Taesup Moon · Klaus-Robert Müller · Wojciech Samek (Eds.), International Workshop Held in Conjunction with ICML 2020 July 18, 2020, Vienna, Austria, Revised and Extended Papers.
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS (2018)
3. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. In: NeurIPS (2020)
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7), e0130140 (2015)
5. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: CVPR (2017)
6. Bengio, Y., Lecun, Y., Hinton, G.: Deep learning for AI. Commun. ACM 64(7), 58–65 (2021). https://doi.org/10.1145/3448250
7. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In: ICLR (2019)
8. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. In: NeurIPS (2019)
9. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: ICCV (2019)

10. Fong, R., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV (2017)
11. Agarwal, C., Nguyen, A.: Explaining image classifiers by removing input features using generative models. In: Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J. (eds.) ACCV 2020. LNCS, vol. 12627, pp. 101–118. Springer, Cham (2021). https://doi. org/10.1007/978-3-030-69544-6 7
12. Alber, M., et al.: investigate neural networks! J. Mach. Learn. Res. (JMLR) 20(93), 1–8 (2019) 3. Ali, A., Schnake, T., Eberle, O., Montavon, G., Muller, K.R., Wolf, L.: XAI for transformers: better explanations through conservative propagation. arXiv preprint arXiv:2202.07304 (2022)
13. Anders, C.J., Neumann, D., Samek, W., Muller, K.R., Lapuschkin, S.: Software for dataset-wide XAI: from local explanations to global insights with Zennit, CoRelAy, and ViRelAy. arXiv preprint arXiv:2106.13200 (2021)
14. Anders, C.J., Weber, L., Neumann, D., Samek, W., Muller, K.R., Lapuschkin, S.: Finding and removing clever HANs: using explanation methods to debug and improve deep models. Inf. Fusion 77, 261–295 (2022) 6. Arras, L., et al.: Explaining and interpreting LSTMs