Explainability in Machine Learning

Gaurav Goyal (17CS30013), Aadarsh Sahoo (17CS30041), Rohit Pathak (17CS30029)

Introduction

Artificial intelligence (AI) had for many years mostly been a field focused heavily on theory, without many applications of real-world impact. This has radically changed over the past decade as a combination of more powerful machines improved learning algorithms, as well as easier access to vast amounts of data that enabled advances in Machine Learning (ML) led to its widespread industrial adoption. Around 2012 Deep Learning methods started to dominate accuracy benchmarks, achieving superhuman results and further improved in the subsequent years. As a result, today, a lot of real-world problems in different domains, stretching from retail and banking to medicine and healthcare are tackled using machine learning models. Systems whose decisions cannot be well-interpreted are difficult to be trusted, especially in sectors such as healthcare and self-driving cars, where moral and fairness issues have also naturally arisen. This need for trustworthy, fair, robust and high performing models for real-world applications led to the revival of the field of eXplainable Artificial Intelligence (XAI) —a field focused on the understanding and interpretation of the behaviour of AI systems, which in the years prior to its revival, had lost the attention of the scientific community, as most research focused on the predictive power of algorithms rather than the understanding behind these predictions.

Fundamental Concepts and Background

The terms interpretability and explainability are usually used by researchers interchangeably; however, while these terms are very closely related, some works identify their differences and distinguish these two concepts. There is not a concrete mathematical definition for interpretability or explainability, nor have they been measured by some metric; however, a number of attempts have been made in order to clarify not only these two terms but also related concepts such as comprehensibility. However, all these definitions lack mathematical formality and rigorousness. One of the most popular definitions of interpretability is the one of Doshi-Velez and Kim, who, in their work, define it as “the ability to explain or to present in understandable terms to a human”. Another popular definition came from Miller in his work, where he defines interpretability as “the degree to which a human can understand the cause of a decision”. Although intuitive, these definitions lack mathematical formality and rigorousness.

Doshi-Velez and Kim proposed the following classification of evaluation methods for interpretability: application-grounded, human-grounded, and functionally-grounded, subsequently discussing the potential trade-offs among them. Application-grounded evaluation concerns itself with how the results of the interpretation process affect the human, domain expert, end-user in terms of a specific and well-defined task or application. Concrete examples under this type of evaluation include whether an interpretability method results in better identification of errors or less discrimination. Human-grounded evaluation is similar to application-grounded evaluation; however, there are two main differences: first, the tester, in this case, does not have to be a domain expert but can be any human end-user and secondly, the end goal is not to evaluate a produced interpretation with respect to its fitness for a specific application, but rather to test the quality of produced interpretation in a more general setting and measure how well the general notions are captured. An example of measuring how well an interpretation captures the abstract notion of input would be for humans to be presented with different interpretations of the input and then select the one that they believe best encapsulates the essence of it. Functionally grounded evaluation does not require any experiments that involve humans but instead uses formal, well-defined mathematical definitions of interpretability to evaluate the quality of an interpretability method. This type of evaluation usually follows the other two types of evaluation: once a class of models has already passed some interpretability criteria via human-grounded or application-grounded experiments, then mathematical definitions can be used to further rank the quality of the interpretability models. Functionally-grounded evaluation is also appropriate when experiments that involve humans cannot be applied for some reason (e.g ethical considerations) or when the proposed method has not reached a mature enough stage to be evaluated by human users. That said, determining the right measurement criteria and metric for each case is challenging and remains an open problem.

Different Scopes of Machine Learning Interpretability: A Taxonomy of Methods

Different viewpoints exist when it comes to looking at the emerging landscape of interpretability methods, such as the type of data these methods deal with or whether they refer to global or local properties. The classification of machine learning interpretability techniques should not be one-sided. There exist different points of view, which distinguish and could further divide these methods. Hence, in order for a practitioner to identify the ideal method for the specific criteria of each problem encountered, all aspects of each method should be taken into consideration. This taxonomy focuses on the purpose that these methods were created to serve and the ways through which they accomplish this purpose. As a result, according to the presented taxonomy, four major categories for interpretability methods are identified:

1. Methods for explaining complex black-box models,
2. Methods for creating white-box models,
3. Methods that promote fairness and restrict the existence of discrimination, and, lastly,
4. Methods for analysing the sensitivity of model predictions.

This first category encompasses methods that are concerned with black-box pre-trained machine learning models. More specifically, such methods do not try to create interpretable models, but, instead, try to interpret already trained, often complex models, such as deep neural networks.

The second category encompasses methods that create interpretability and are easily understandable from humans models. The models in this category are often called intrinsic, transparent, or white-box models. Such models include the linear, decision tree, and rule-based models and some other more complex and sophisticated models that are equally transparent and, therefore, promising for the interpretability field.

Because machine learning systems are increasingly adopted in real-life applications, any inequities or discrimination that are promoted by those systems have the potential to directly affect human lives. Machine Learning Fairness is a sub-domain of machine learning interpretability that focuses solely on the social and ethical impact of machine learning algorithms by evaluating them in terms of impartiality and discrimination. The study of fairness in machine learning is becoming more broad and diverse, and it is progressing rapidly.

The fourth category includes interpretability methods that attempt to assess and challenge the machine learning models in order to ensure that their predictions are trustworthy and reliable. These methods apply some form of sensitivity analysis, as models are tested with respect to the stability of their learnt functions and how sensitive their output predictions are with respect to subtle yet intentional changes in the corresponding inputs.

Explainability in Computer Vision

Convolutional neural networks have achieved tremendous success in solving tasks of Computer Vision. With deep learning and AI based systems becoming an increasing part of our daily lives, from the image and facial recognition systems, autonomous machines, etc, there comes a need to explain the decisions taken by these systems in order to trust them.

Explainable AI in computer vision tries to address how black box decisions are taken by different vision models. Defining explainability for computer vision can be challenging. We will discuss some of the popular methods which are used in order to interpret/explain various deep vision cnn models.

Saliency Maps

Saliency maps are a very popular visualization technique for gaining insight into “why” a deep learning vision model made an individual decision, such as classifying an image of a dog. It can be useful in explaining correct as well as wrong decisions. For example, if we have an image of a dog and a vision model classifies it as a cat, then saliency maps can help us identify why the vision model thinks that the image is a cat’s image instead of a dog’s image. They are rendered usually as a heat map, where the hotness corresponds to regions that have a big impact on the model’s final decision. Some of the famous techniques used for generating saliency maps are: GradCAM, RISE.

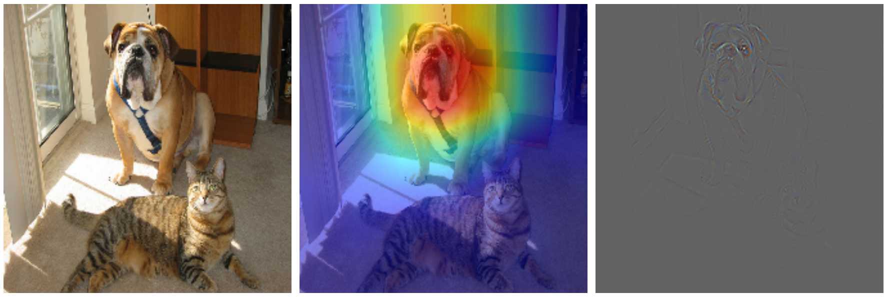


Fig. The figure shows a sample original image (left), saliency map using a method called Grad-CAM (center), and another using Guided Backpropagation (right).

Local Interpretable Model-Agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME)[1] is a technique that explains how the input features of a machine learning model affect its predictions. For instance, for image classification tasks, LIME finds the region of an image (set of super-pixels) with the strongest association with a prediction label. LIME creates explanations by generating a new dataset of random perturbations (with their respective predictions) around the instance being explained and then fitting a weighted local surrogate model. This local model is usually a simpler model with intrinsic interpretability such as a linear regression model. For the case of images, LIME generates perturbations by turning on and off some of the super-pixels in the image, it then predicts the class for each of the perturbations and then computes the weights of importance for each of the perturbations. After this, it shows the part of the image with the highest importance weight for the corresponding decision taken by the model.

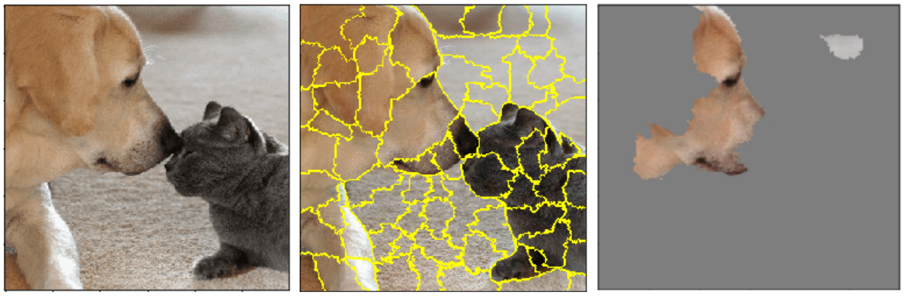


Fig. The figure demonstrates the steps involved in producing LIME explanations. The original image (left) is perturbed into various super-pixels (center) and the region with highest importance is selected as an explanation. Note: The image was predicted as ‘labrador’ by the vision model.

Neural-Backed Decision Trees (NBDT)

Before deep learning, decision trees were the gold standard for accuracy and interpretability. Decision Trees use a tree like model where each node is a ”test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The predicted class depends on the outcome of the intermediate tests thus making this model interpretable. For accuracy, however, decision trees lag behind neural networks by up to 40% accuracy on image classification datasets. Therefore, a recent work has combined neural networks with decision trees, to create highly accurate explainable models. NBDTs combine neural networks with decision trees, preserving high-level interpretability while using neural networks for low-level decisions. NBDTs are as interpretable as decision trees and achieve neural network accuracy.

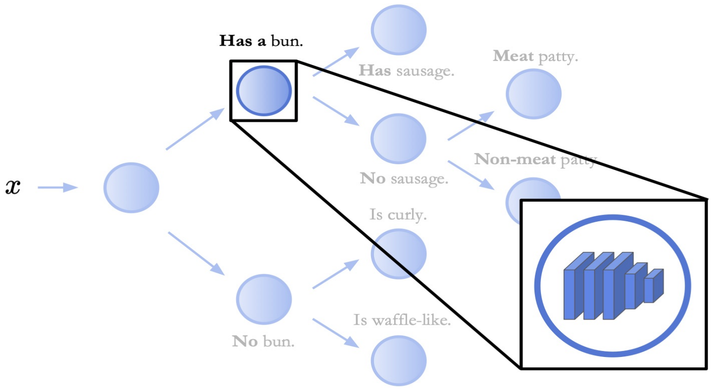
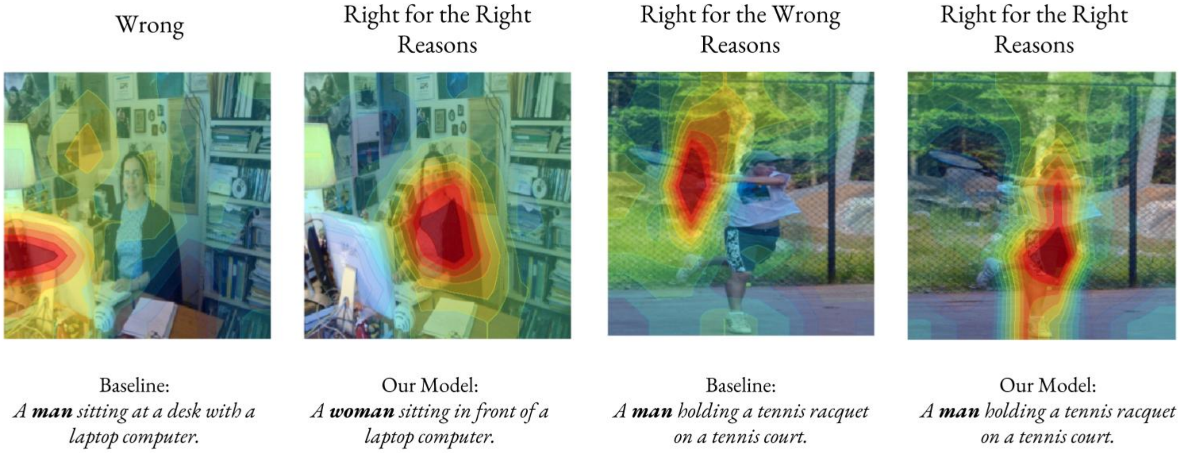


Fig. In this figure, each node contains a neural network. The figure only highlights one such node and the neural network inside. In a neural-backed decision tree, predictions are made via a decision tree, preserving high-level interpretability. However, each node in the decision tree is a neural network making low-level decisions. The "low-level" decision made by the neural network above is "Has sausage" or "no sausage".

Importance of Explaining Decisions and Bias in Computer Vision

The above discussed methods can be used to assess different aspects of deep vision models. One of the important applications is detecting bias in visual models. Bias in AI as a whole has become a very sensitive, and debatable topic in the current era of growth in deep learning applications. Involving AI systems in various critical applications makes it essential for the system to be unbiased and safe. Let us discuss one of the works which studied the bias in image captioning models. The work - "**Women also Snowboard: Overcoming Bias in Captioning Models**" discusses how machine learning methods capture and exploit biases of the training data and how image captioning models tend to exaggerate biases present in training data.



Wrong

Right for the Right Reasons

Right for the Wrong Reasons

Right for the Right Reasons

Baseline:
*A **man** sitting at a desk with a laptop computer.*

Our Model:
*A **woman** sitting in front of a laptop computer.*

Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Our Model:
*A **man** holding a tennis racquet on a tennis court.*

The above image is a very interesting visualisation on how the underlying captioning model is gender biased. Starting from the left, the captioning model tells that the person in the image is a 'man', and when asked to explain it's decision, it focuses on the desktop instead of the person present. This says that in the training data there are a lot of images with a 'man' present alongside/in front of a desktop computer, so whenever a model 'sees' a desktop computer, it blindly predicts it as a 'man' instead of a 'woman'. A similar scenario is explained in the second pair of images, where the model just focuses on the tennis racket to come up with a prediction that the person is a 'man'.

This scenario highlights the importance of explainable and interpretable models in computer vision and how models can be biased to training data.

Explainability in Natural Language Processing (NLP)

Introduction and Motivation for research in NLP Explainability

Traditionally, Natural Language Processing (NLP) systems have been mostly based on techniques such as white box techniques that are inherently explainable. However, in recent years, with the rise in popularity in various black box methods/techniques such as deep learning models, model quality has advanced at the expense of being less explanable. To maintain and improve the trust in various AI, NLP, etc.. systems that people interact with, this trend of becoming less explainable is very problematic. Hence, in the broader AI community, day by day, the importance of explanabilty is increasing in perception. A new field known as Explainable AI (XAI) has emerged. Since tasks are more amenable to particular approaches, here we will focus on works done in the domain of NLP in the last few years.

Our focus on explainability is from the perspective of an end user whose goal is to understand how any model arrives at its result. This is also known as the outcome explanation problem. This helps in building trust between developers of these NLP-based AI systems and the users using them. Also, it becomes easier for the users to provide feedback for the models.

Explanation categorization

Explanations are categorized along two main aspects. The first way to do it is to distinguish explanations done for individual predictions or global prediction (done for a model's prediction process as a whole). This is known as local and global explanations. Another way to do it is to distinguish whether the explanation is directly from the prediction process or requires some post-processing.

- Local vs Global - Justification/information or reasoning is provided for the prediction of models on a specific input in case of local explanation. For global explanations, justification/reasoning is provided for the model predictions in a general way, independently of any specific input. Most of the work done till now has been done with local category.
- Self explaining vs post prediction - Different explanations are different from one another on the basis whether they are a part of the prediction process, or if some post-processing is required after the model makes a prediction. In case of self-explaining, explanation is generated at the same time when the prediction happens. It uses the information from the model that is acquired as a result from the model making that prediction. In case of post-processing, we need to perform more additional operations after the prediction takes place to get to the explanation.

There are three main aspects of explanations:

- The methods and techniques used for deriving the explanations.
- Various sets of operations used to enable explainability.
- Visualization and presentation of the result to the user.

Techniques for deriving explanations

Derivation of explanations is mainly focused on finding mathematically motivated justifications and reasons for the generated output of a model and using different available explainability techniques to produce explanations. Broadly, there are five major techniques used for explaining that are different from each other in their mechanisms to find the explanation that will be provided to the user.

- Feature importance - Importance scores of different features used for generating output are found out and explanation is derived from them. Examples of features on which this can be done: lexical features such as words, tokens and n-grams, etc. manual features from feature engineering or Neural Network features. Two different ways to get feature-importance based explanations are attention and first derivative saliency.
- Surrogate/substitute model - A surrogate model is a second model that is learned instead of the main model to generate explanations. It is usually much more explainable than the main model and is used as a proxy. Used for both local and global approaches. One problem is that the surrogate model can theoretically have a completely different approach of getting predictions which raises questions about its use.
- Example-driven - Similar to nearest neighbour based approaches like KNN, prediction of an instance is explained with the help of other semantically similar instances gathered from the labelled data. Text classification and Question-answering use this kind of methods.
- Provenance - If the final prediction is due to a series of intermediate reasoning steps, illustrating some or all prediction derivation process steps would be a good, effective and intuitive way to explain predictions. Question-answering based problems are observed to use this method.
- Declarative induction - Representations that are easily human readable and understandable such as rules, trees and programs are induced as explanations.

Operations used for enabling explanations

Here are a few operations that we regularly encounter in many different literatures which are based on NLP explainability.

- First-derivative saliency - Contribution of inputs towards output is measured by gradient based methods. If there is input i and with output o, value is calculated by computing partial derivative of o with respect to i. Since Neural Network models generally provide methods such as auto - differentiation, which can calculate gradients for any intermediate layer when simply called. It is commonly used with feature importance technique, especially when features are word/tokens.
- Layerwise relevance propagation - Another type of operation used for finding attribute relevance of the intermediate layer of a Neural Network. Commonly used with feature importance technique and example - driven technique.
- Input perturbations - Input perturbation is a method in which output for some input i is explained by randomly generated i-perturbations and then training an explainable model. Used commonly in conjunction with linear models which are a substitute model in the surrogate model technique.
- Attention - Attention layers can be added to different types of Neural Network models / architectures to help find on which layers or where exactly the models are "focusing". This makes it more of a plan or strategy than an operation. Used previously very commonly with feature importance technique.
- LSTMs- Since languages are inherently sequential, recurrent network layers, especially LSTMs are used commonly. Generally, output of LSTM cells are mined for information regarding output explanation. It is also possible to get information from outputs of gates produced within the cells.

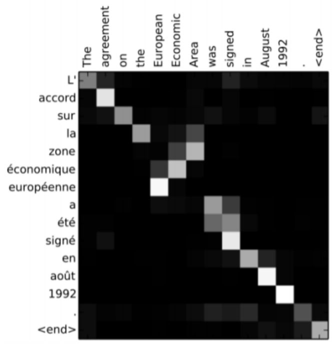
Combination of LSTMs with other operations to interpret the gating signals is also a way to find information. Used commonly with feature explainability techniques.

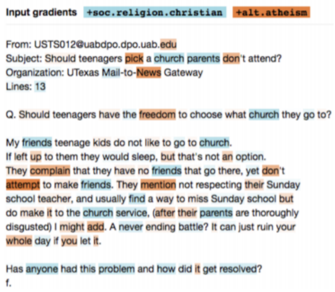
- Explainability - aware architecture design - Architectures are constructed that can somewhat mimic the process humans use to arrive at a solution. Since deep learning is quite flexible, it is very commonly used while making such architectures. Since the architecture contains human recognizable components, the learned model becomes explainable. Used commonly in solving and explaining math problems and sentence simplification problems. Commonly used or applied with surrogate model technique.

Techniques of Visualization

It is absolutely imperative to use an appropriate technique for visualization or presentation of explanation to a user. Failure to do so can make an otherwise successful explainable model useless. One major example is that to visualize attention based explanation models, two types of methods can be used - raw attention scores or saliency heatmaps. Even though viewing raw attention scores directly gives us more data related to a model, since saliency heatmap is more visually user friendly, it is much more preferred and has become a standard way of visualizing attention approaches. Major visualization techniques used in different famous literatures given below.

- Saliency - The most widespread and famous way used to visualize the importance scores of different features in explainable learning models. Ex: to show input-output word alignment. It has been observed that almost all the feature importance technique based explainable literature use saliency based visualizations. They are popular as they present visually perceptible explanations that can be easily understood and observed by users.




Input gradients +soc.religion.christian +alt.atheism

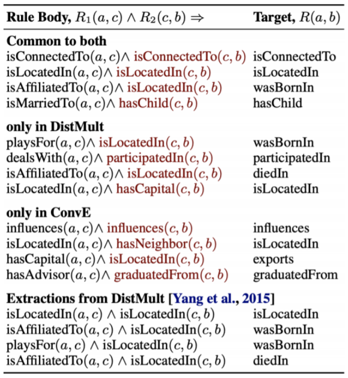
From: USTS012@uabdpo.dpo.uab.edu
Subject: Should teenagers pick a church parents don't attend?
Organization: UTexas Mail-to-News Gateway
Lines: 13

Q. Should teenagers have the freedom to choose what church they go to?

My friends teenage kids do not like to go to church.
If left up to them they would sleep, but that's not an option.
They complain that they have no friends that go there, yet don't
attempt to make friends. They mention not respecting their Sunday
school teacher, and usually find a way to miss Sunday school but
do make it to the church service, (after their parents are thoroughly
disgusted) I might add. A never ending battle? It can just ruin your
whole day if you let it.

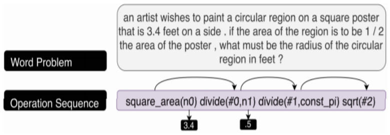
Has anyone had this problem and how did it get resolved?
f.

- Raw declarative representations - Declarative representations that are learned are presented directly. Ex: logic rules, decision and reasoning trees, first order logic rules and programs. It is used if the target users can understand such representations.



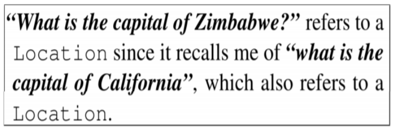
Rule Body, $R_1(a,c) \wedge R_2(c,b) \Rightarrow$	**Target,** $R(a,b)$
Common to both	
isConnectedTo$(a,c)\wedge$ isConnectedTo(c,b)	isConnectedTo
isLocatedIn$(a,c)\wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c)\wedge$ isLocatedIn(c,b)	wasBornIn
isMarriedTo$(a,c)\wedge$ hasChild(c,b)	hasChild
only in DistMult	
playsFor$(a,c)\wedge$ isLocatedIn(c,b)	wasBornIn
dealsWith$(a,c)\wedge$ participatedIn(c,b)	participatedIn
isAffiliatedTo$(a,c)\wedge$ isLocatedIn(c,b)	diedIn
isLocatedIn$(a,c)\wedge$ hasCapital(c,b)	isLocatedIn
only in ConvE	
influences$(a,c)\wedge$ influences(c,b)	influences
isLocatedIn$(a,c)\wedge$ hasNeighbor(c,b)	isLocatedIn
hasCapital$(a,c)\wedge$ isLocatedIn(c,b)	exports
hasAdvisor$(a,c)\wedge$ graduatedFrom(c,b)	graduatedFrom
Extractions from DistMult [Yang et al., 2015]	
isLocatedIn$(a,c) \wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
playsFor$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	diedIn

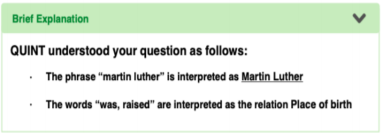
- Natural language explanation - The explanation is textualized or verbalized in human comprehensible natural language. Many declarative representations use natural language generation models to turn rules and first order logics into human comprehensible language, making them much more accessible.



an artist wishes to paint a circular region on a square poster that is 3.4 feet on a side . if the area of the region is to be 1 / 2 the area of the poster , what must be the radius of the circular region in feet ?

square_area(n0) divide(#0,n1) divide(#1,const_pi) sqrt(#2)


"What is the capital of Zimbabwe?" refers to a `Location` since it recalls me of *"what is the capital of California"*, which also refers to a `Location`.


Brief Explanation

QUINT understood your question as follows:

- The phrase "martin luther" is interpreted as Martin Luther
- The words "was, raised" are interpreted as the relation Place of birth

Some other visualization techniques are also used frequently such as examples driven approaches like raw examples, dependency parse trees, etc.

Evaluation

A model's quality is evaluated not only by its accuracy or performance, but also by how good an explanation is provided for the result. However, since this field is relatively very young, there is no general consensus regarding how evaluation of explanations should be done. Even major works done related to this field lack a standardized method of evaluation. Following are some of the main evaluation categories that have been used in various literatures related to NLP-explanations.

- Informal examination - It involves examination of explanations done informally such as high-level discussions of how generated explanations are related to human values and intuition. Examples - output of a single explainability approach is reviewed in isolation, comparing to reference approaches, etc.
- Ground truth comparison - Comparing ground truth data to the explanations that are generated to quantify the performance of these explainability techniques. Related approaches typically involve multiple annotators that report mean human performance or inter annotator agreement by evaluating the explanations at different granularities to account for cases where alternative valid explanations could be there.
- Human evaluation - Most direct method to evaluate the explanation quality. Humans are asked to evaluate the explanations that are generated. One advantage is that the assumption that there is only one or a few main explanations is avoided. Similar to the ground truth method, multiple annotators, reporting inter annotator agreement and subjectivity are characteristics of this evaluation method. Number of humans involved can vary a lot.

Another important aspect that is generally ignored is the part of the prediction process that is being covered by the explanation. Most explainability methods explain only parts of the process. Users are left to fill the gaps by themselves. Intuitively, higher coverage seems to be a positive aspect always, but care should be taken considering the target users and audience of the explanations as a lower coverage may be more palatable for certain users.

Summary, Conclusion and Future

The main contribution of this study is a taxonomy of the existing machine learning and deep learning interpretability methods that allow for a multi-perspective comparison among them. Under this taxonomy, four major categories for interpretability methods were identified: methods for explaining complex black-box models, methods for creating white-box models, methods that promote fairness and restrict the existence of discrimination, and, lastly, methods for analysing the sensitivity of model predictions.

As a result of the high attention that is paid by the research community to deep learning, the literature around interpretability methods has been largely dominated by neural networks and their applications to computer vision and natural language processing. Most interpretability methods for explaining deep learning models refer to image classification and produce saliency maps, highlighting the impact of the different image regions. In many cases, this is achieved through exploiting the gradient information flowing through the layers of the network, Grad-CAM, a direct extension of, being a prime and most influential example in terms of citations per year. Another way of creating saliency maps, and the most influential overall while using the same metric, is through the adoption of deconvolutional neural networks. In terms of explaining any black-box model, the LIME and RISE methods are, by far, the most comprehensive and dominant across the literature methods for visualising feature interactions. White-box highly performing models are very hard to create, especially in computer vision and natural language processing, where the gap in performance against deep learning models is unbridgeable. Furthermore, because models are more than ever expected to be competitive on more than one tasks and knowledge transfer from one domain to another is becoming a recurring theme, white-box models, being able to perform well only in a single given task, are losing traction within the literature and are quickly falling further behind in terms of interest.

A great deal of effort and progress has been made towards tackling discrimination and supporting fairness in machine learning that sensitive domains, like banking, healthcare, or law, could benefit from. However, these methods are neither commonly found, nor well promoted within the dominant machine learning frameworks. That being said, only a few studies deal with fairness in non-tabular data, such as images and text, which leaves plenty of room for improvements and innovation in these unexplored areas in the coming years.

Sensitivity analysis, which is the last category of interpretability methods under this taxonomy has seen tremendous growth over the past several years following the breakthrough works on adversarial examples and the weaknesses of deep learning models against adversarial attacks. Numerous methods for producing adversarial examples have been developed, with some of them focusing on a more general setting, while others being tailored to specific data types, such as image, text, or even graph data, and to specific learning tasks, such as reading comprehension or text generation. Despite its rapid growth, explainable artificial intelligence is still not a mature and well established field, often suffering from a lack of formality and not well agreed upon definitions. Consequently, although a great number of machine learning interpretability techniques and studies have been developed in academia, they rarely form a substantial part of machine learning workflows and pipelines.

Acknowledgements

We would like to thank Prof. Aritra Hazra, Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur for his constant support and amazing way of teaching this course. It helped us connect to various concepts with clarity and connect to real life applications while preparing this report.

References:

https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/?sh=3ff7144b7c9e
https://bair.berkeley.edu/blog/2020/04/23/decisions/
https://medium.com/@thelastalias/saliency-maps-for-deep-learning-part-1-vanilla-gradient-1d0665de3284
https://kayburns.github.io/projects.html
https://github.com/eclique/RISE
https://medium.com/@ODSC/visualizing-your-convolutional-neural-network-predictions-with-saliency-maps-9604eb03d766
https://towardsdatascience.com/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13
[2010.00711] A Survey of the State of Explainable AI for Natural Language Processing
https://towardsdatascience.com/rise-of-modern-nlp-and-the-need-of-interpretability-97dd4a655ac3
[1702.08608] Towards A Rigorous Science of Interpretable Machine Learning by Doshi-Velez and Kim