VIEW POINT

EXPLAINABLE AI (XAI)

Abstract

Artificial Intelligence is making inroads in every industry, from doctors diagnosing cancer to advanced surveillance systems. However, it is also attracting a lot of negative sentiments because of failing systems and unforgivable errors. This has called for official bodies and general users seeking more and more transparency into every decision made by AI based systems, paving way for Explainable AI. This PoV discusses the emergence of XAI and its underlying principles.

Infosys®
Navigate your next

The proliferation of AI based systems is fast evolving from being just augmenting human decision making to an independent sole decision system. Courts are using AI to sentence criminals. Banks are using AI to grant loans and predict NPAs while doctors use AI to identify cancer from scans. Accident claims arising from autonomous vehicles, profits and losses from Algorithmic trading are also in the gambit of AI. However, it is attracting lot of negative press due to failing systems, resulting lawsuits and other societal implications to the point that today Regulators, official bodies and general users are seeking more and more transparency into every decision made by AI based systems. In the United States, insurance companies need to explain their rates and coverage decisions while the European Union introduced right to explanation in GDPR.

Geoffrey Hinton (University of Toronto), often called the godfather of deep learning, explains: "A deep-learning system doesn't have any explanatory power. The more powerful the deep-learning system becomes, the more opaque it can become." [1]

For instance, consider a husky versus wolf classifier [5] that misclassifies some huskies as wolves. Using the training data set, the classifier learned to use snow as a feature for classifying images as "wolf", which might make sense in terms of separating wolves from huskies in the training dataset, but not in real-world use.

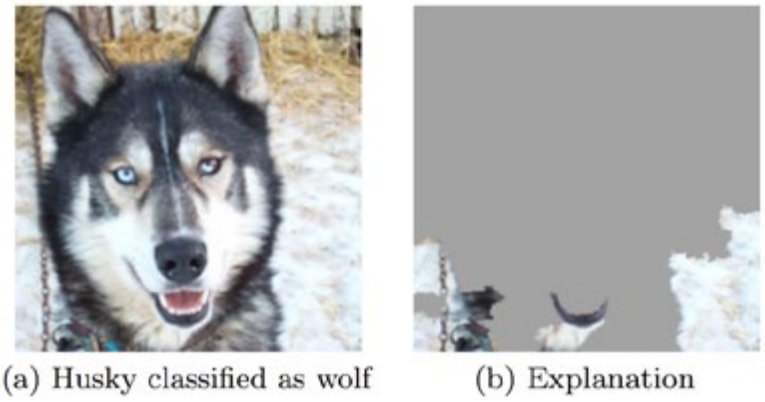


Figure 1: In the figure it can be seen that the features on which the model is classifying a Husky(dog) as a Wolf are only the areas where there is snow [5]

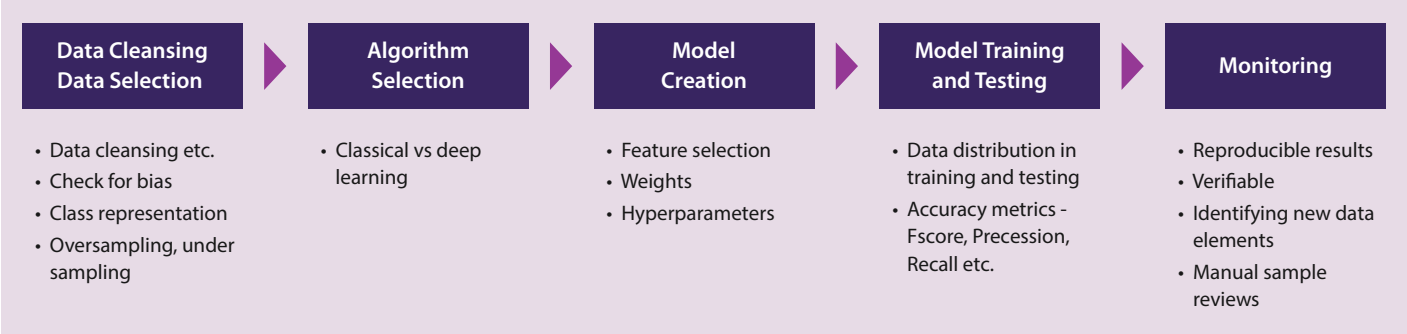
In a more recent incident, ImageNet is going to remove 600,000 images of people from its database after finding racial bias in it. On uploading an image of a white woman, the model trained on it, labels it as "stunner, looker, mantrap". Whereas, if uploading image of colored people it labels them as "Black, Black African, Negroid or Negro." The same bias can be seen in Asian and Non-Asian people. [2]

Use of Machine Learning and AI in medical field requires even more care. ML models to detect Pneumonia by examining the Chest X-rays failed when tested on real world data or data outside of the train-test dataset. Near about 1.5L chest X-rays from three medical institutions viz. the National Institutes of Health, The Mount Sinai Hospital and Indiana University Hospital, were collected and tested. The performance of Convolution Neural Network in diagnosing diseases was considerably lower when tested on different datasets. [3]

All the above scenarios call for need of tools and techniques to make ML and AI models more transparent and interpretable. The ML model should not only classify an object but also should be able to explain logic behind its classification. For example, in case of simple Cat Vs Dog classifier, the classifier model should be able to highlight the specific parts of a dog or a cat because of which it could classify an image in a particular category.

When should XAI come in?

ExplainableAI is not just to be introduced at the output explanation stage, but during complete AI life cycle. The key stages where it has an important role is as follows



- Data selection – equal representation for all classes, consideration for data collection
- Algorithm selection – selection of right algorithm, classical ones (logistic regression, Naïve Bayes, etc.) or neural network (CNN, LSTM, etc.)
- Model creation – right feature selection, assigning appropriate weights or hyper parameters
- Model training and testing – right distribution and magnitude of training, validation and testing data
  a. Selection of right frameworks such as LIME, SHAP for verification
  b. Accuracy metrics – selection of right metrics such as F1-Score, Precision, Recall, etc. instead of just plain emphasis on accuracy
- Monitoring – in production watching and verifying results, and keeping tab on incoming data varieties, spotting them in case if they are not used during model training
  a. Selection of tools such as LIME, SHAP, etc.

Traits of a Good Explainable Model

All the new Machine Learning algorithms help ease manual work and increase productivity. However, overall there are only few requirements that a Machine Learning model should satisfy.

- Unbiased Dataset
- Dataset should be non-discriminatory in nature
- Human understandable Output
- Justifiable Predictions
- Accountable ML Models

Principles

To ensure trust and reliability in Machine Learning Models, it is essential to adhere to some basic principles as below:

- Human Involvement: Though Machine Learning Models are built to operate independently without human interference, in some cases human dependency is a necessity. For example, in fraud detection or cases where law enforcement is involved, we need some human supervision in the loop to check or review decisions made by ML models from time to time.
- Bias Detection: An unbiased dataset is an important prerequisite for Machine Learning Model to perform reliable and non-discriminating predictions. ML models are being used by banks for credit scoring, resume shortlisting, and also in some judicial system, however, it has been noticed that, in some cases the dataset had some inherent bias in them on basis of color, age and sex.

For example, to detect bias, German credit dataset is used [12]. This dataset is used to classify people into good or bad credit risks based on 20 attributes. Some of the attributes used are credit history, credit amount, employment status, gender, and property. Ideally, the machine learning model should classify people into good or bad credit risk based on features such as credit history, employment status, salary and so on. However, presence of other attributes like age, sex and address can also affect the prediction of the models that leads to undesirable classification. For instance, if the model gives more weightage to features like age and sex, this may lead to unethical practices.

To detect such biases in the dataset, AIF 360 library is used [13]. The attribute 'Age' is tested for bias detection i.e. to check if there is any bias in the dataset on basis of age. The age attribute is changed into a binary label i.e. if age is greater than 25 (privileged group) then it is set to 1 and if the age is less than 25(unprivileged group) then it is set to 0. Then a mean difference is calculated between favorable results for privileged groups and unprivileged groups.

In the above figure, the mean difference is " 0.16410" which states that the privileged group has almost 17 percent more positive outcomes. This AIF 360 library also provides various methods for removing such biases. One of those is re-weighing, where weights of the individual samples are changed to balance the dataset before feeding it into the machine learning model.

- Explainabilty: Explainable AI comes into picture when we talk about justifiable predictions and Feature Importance. Explainable AI helps in understanding how the model is thinking or at which features of the given input it is emphasizing while making predictions.
- Reproducibility: The ML model should be consistent when giving predictions and should not go haywire when testing with new data.

Explainability through Feature Importance

One of the basic principles of Machine learning is Explainability by Justification [6]. One way to explain the predictions of a Machine Learning Model is to highlight features in the input that are contributing towards prediction of a class and not rely on some random feature as in the case of Husky Vs Wolf classifier model.

Working in the same direction, two explainable AI libraries are used viz. SHAP and LIME on our image classification and text classification models. LIME highlights those parts of the image (Super pixel) which are dominant in prediction of a class whereas SHAP gives us insights on how a layer in the model is impacting the output probabilities.

LIME (Local Interpretable Model-agnostic Explanations)

LIME stands for Local Interpretable Model-Agnostic Explanations [5]. In LIME, a temporary model is trained to mimic the black box model prediction. Given a sample input to the model, LIME generates new dataset by creating various permutations of the given sample and their corresponding output by training a simple and more interpretable local model on this dataset.

For example, we trained a model to classify cars based on their visual features. We used transfer learning for model training. A ResNet Architecture and its pre-trained weights were used [15]. Then the last layer was trained on Stanford Car dataset [14] which had 196 categories of cars. Each category contained 40-50 images for training. This model currently gives an accuracy of 90 percent on test set.

We then passed the car classifier model and images to LIME library to see whether the model is looking at the right regions in the images for classifications. The LIME created 1000 different samples out of the image by trying various permutations of super pixels based on segmentations. Super pixels are group of similar colored pixels.

In the below Figure 2, The input to the model is an Audi car and output of the explainer is the region because of which it is classifying it as an Audi Car. The model is focusing on the logo in the image to correctly classify the image.

Figure 2: The input to the model is an Audi car and output of the explainer is the region because of which it is classifying it as an Audi Car. The model is focusing at logo in the image to correctly classify the image

In Figure 3 below, input to the Car model is a Lamborghini. We then observe that the model is taking shape of the car as an important feature for classification.

Figure 3: In this input to the model is a Lamborghini. We see model is taking shape of the car as an important feature for classification.an important feature for classification.

SHAP (SHapley Additive exPlanations)

SHAP stands for SHapley Additive exPlanations [7]. SHAP library is primarily based on Game theory in which we calculate Shapley values. In SHAP, contribution of each feature of the input sample towards prediction is calculated. These features act as players when the contribution is calculated. In case of images, the features can be pixels or group of pixels (super pixel), then contribution of each features is calculated towards the prediction. The contribution can be positive or negative. We also provide a sample dataset to calculate average prediction of the model. Then we calculate the contribution of individual features by giving different permutations of them to the model and calculating if it is increasing the value of average prediction or decreasing it.

The same car classifier model is used for explanations using SHAP. We pass the trained Model, an image and a background dataset. SHAP gives us layer wise explanation of the Deep Learning Model.

In Figure 4, input image on the left is an Audi car and we see the explanation on the right. The red pixels are contributing more towards prediction of the class whereas blue pixels are reducing the probability. We can see the model is able pick-up the symbolic Audi rings.

Figure 4: Input image on the left is an Audi car and we see the explanation on the right. The red pixels are contributing more prediction of the class whereas blue pixels are reducing the probability. We see the model is able pick-up the symbolic Audi rings.

In Figure 5, here we see the model is picking-up the shape of car's window pane as prominent feature for classification.

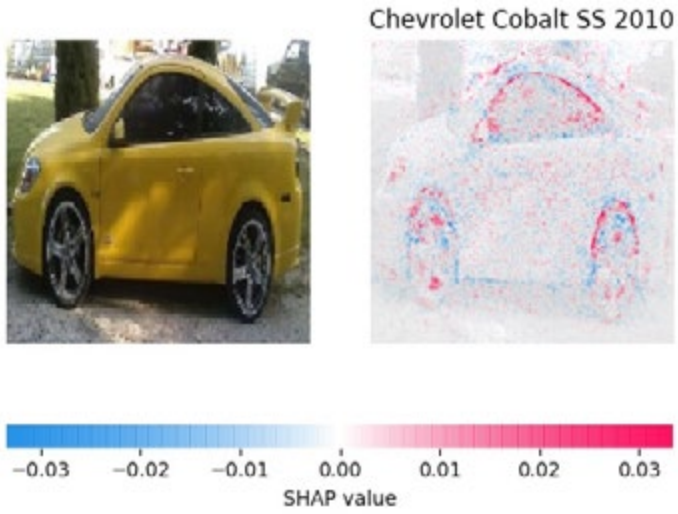


Figure 5: Here we see the model is picking-up the shape of window pane as prominent feature for classification.

The model is tested on two Explainable AI algorithms and we see that the results are consistent and can capture relevant features for classification, hence helping us trust the prediction.

Thanks to the Explainable AI techniques, we not only know what is being predicted but also know why it is being predicted. We can gain insight into the behavior of the model for a prediction and rectify the model if it is not picking-up the correct features for prediction. This will make complex ML algorithms and models more transparent and trustworthy.

Explainable AI for Text

Text based use cases, be it identifying sentiment or establishing toxic content, is important to understand which words or sentences are contributing to positive or negative sentiment or toxicity. Explainable AI applied in text scenarios can help understand the root of the cause.

For text based usecases, we train two Text classification models using BERT [16], on top of which few dense layers to classify text sentences are added.

We train the first model to classify a given sentence into two classes viz. Question and Non-Question. Before passing it to LIME library we need to convert raw text to vectorized representations, for this we make use of a pipeline of pre-processing steps right from taking raw texts from users to passing the text to the trained model. LIME then creates various samples from the given text and produces an output image explaining the contribution of a word to the predicted class.

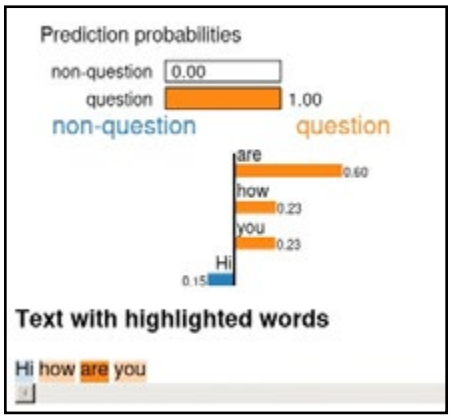


Figure 6: 'are', 'how', 'you' are contributing towards the class Question

The second model is trained for classifying a sentence into a Toxic sentence and Non-toxic sentence. Same preprocessing steps are used as mentioned in the above classifier. As one can see, in the sentence, "Are you stupid?", it has correctly highlighted "stupid" keyword as toxic.

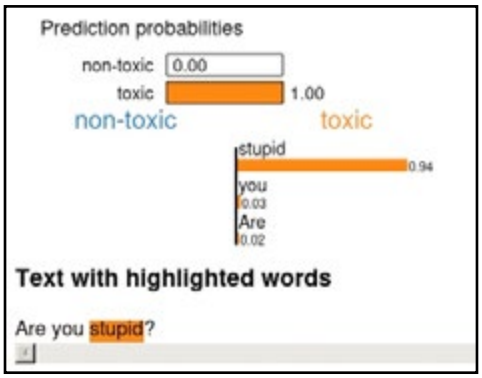


Figure 7. Here we see 'stupid' is Highlighted which is contributing towards class Toxic.

Here is an example as to how Explainable AI shows outputs in toxic classification on an excerpt taken from a customer service conversation.

More recent trends and future work

Domain experts are coming together to lay some basic principles that ML and AI models should follow. More emphasis is now being given to ensure transparency in the working of complex ML models. Google has started this with integrating 'What if tool' in their Tensorflow framework [8] in which one can inspect ML models, compare two models , visualize results, visualize feature attributions, confusion matrices and much more with minimum coding.

The Institute for Ethical AI & Machine Learning [6] is currently laying a framework that ensures ethical and conscious development of AI projects across all industries. In their work towards this, they have published ethical AI principles and also have developed an open source git hub toolbox for Explainability [11] .

References

1. http://open-shelf.ca/180201-ocula-explainable-artificial-intelligence/
2. https://hyperallergic.com/518822/600000-images-removed-from-ai-database-after-art-project-exposes-racist-bias/
3. https://www.hindustantimes.com/tech/ai-tools-may-fail-during-key-medical-diagnosis-researchers/story-t4ql1v70j19P8AXbq7ziVO.html
4. https://christophm.github.io/interpretable-ml-book/
5. Tulio Ribeiro, M., Singh, S., Guestrin, C.\ 2016.\ ``Why Should I Trust You?'': Explaining the Predictions of Any Classifier.\ arXiv e-prints arXiv:1602.04938.
6. https://ethical.institute/principles.html/
7. Lundberg and Lee(2017)]{2017arXiv170507874L} Lundberg, S., Lee, S.-I.\ 2017.\ A Unified Approach to Interpreting Model Predictions.\ arXiv e-prints arXiv:1705.07874.
8. https://towardsdatascience.com/googles-new-explainable-ai-xai-service-83a7bc823773
9. https://pair-code.github.io/what-if-tool/
10. https://thenextweb.com/artificial-intelligence/2019/12/17/8-biggest-ai-trends-of-2020-according-to-experts/
11. https://github.com/EthicalML/XAI
12. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
13. https://github.com/IBM/AIF360/blob/master/examples/tutorial_credit_scoring.ipynb
14. Jonathan Krause, Michael Stark, Jia Deng, Li Fei-Fei 3D Object Representations for Fine-Grained Categorization at 4th IEEE Workshop on 3D Representation and Recognition, at ICCV 2013 (3dRR-13). Sydney, Australia. Dec. 8, 2013.
15. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.
16. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.\ 2018.\ BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.\ arXiv e-prints arXiv:1810.04805.
17. H3 Trends in AI Algorithms - https://www.infosys.com/services/incubating-emerging-technologies/offerings/Documents/ai-algorithms.pdf
18. Transfer Learning - https://www.linkedin.com/pulse/ai-trends-transfer-learning-sudhanshu-hate/
19. AI Ethics and Biases - https://www.linkedin.com/pulse/ai-biases-ethics-sudhanshu-hate/
20. Artificial General Intelligence (AGI) - https://www.linkedin.com/pulse/artificial-general-intelligence-agi-sudhanshu-hate/

Authors

Sudhanshu Hate

Sudhanshu Hate, Senior Principal Technology Architect with iCETS, is an inventor and architect of Infosys Enterprise Cognitive Platform(iECP), a microservices API based Artificial Intelligence platform. He has over 22 years of experience in creating products, solutions and working with clients on industry problems. His current areas of interests are Computer Vision, Speech and Unstructured Text based AI possibilities.

Ram Swaroop Mishra

Ram Swaroop Mishra is Senior System Engineer with Infosys Center for Emerging Technology Solutions with over 3 years of experience. He is a Computer Vision and Machine Learning enthusiast.

To know more about our work on the H3 trends in AI, write to icets@infosys.com.