Machine Learning Explainability from an Information-theoretic Perspective

Anonymous Author(s)
Affiliation
Address
email

Abstract

The primary challenge for practitioners with multiple *post-hoc gradient-based* interpretability methods is to benchmark them and select the best. Using information theory, we represent finding the optimal explainer as a rate-distortion optimization problem. Therefore :

- We propose an information-theoretic test `InfoExplain` to resolve the benchmarking ambiguity in a model agnostic manner without additional user data (apart from the input features, model, and explanations).
- We show that `InfoExplain` is extendable to utilise human interpretable concepts, deliver performance guarantees, and filter out erroneous explanations.

The adjoining experiments, code and data will be released soon.

1 Introduction

Multiple methods have been proposed to perform *gradient-based feature attribution*. Given a particular model input instance, these attribution methods rank its features in order of importance towards the model's decision. These *explanations* allow humans to debug model behaviour [3] and comprehend model biases - especially in sensitive environments such as healthcare and law enforcement [26, 31].

Gradient-based Feature Attribution In a classification paradigm, with a model θ that outputs $\hat{y}$ when x is input, gradient-based attribution schemes compute the input gradient $\nabla_x \text{logit}_\theta(x, \hat{y})$ and sort the features inside x based on gradient magnitude. This concept or a variant of it underlies methods such as Saliency maps [21], Integrated Gradients [27], Guided Backprop [25] and SmoothGrad [24].

Information theoretic modelling Information is a common language across domains, including machine learning - since every communication must abide by its laws of encoding, decoding, transmitting and manipulating signals. Our `InfoExplain` method describes machine and human cognition as a system of compression and transmission of Information given constrained resources.

2 Related Work

Recent work in explainable AI has focused on designing and improving techniques to make machine learning more robust and reliable.

Evaluating explanation fidelity Feature attributions by explainability methods are usually evaluated via subjective visual estimation by humans.[11, 15, 24] Perturbation-based evaluation paradigms [17, 6, 5] gauge fidelity by measuring the change in model performance after changing input features considered important by the model. However, distribution shifts induced by such perturbations mean

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

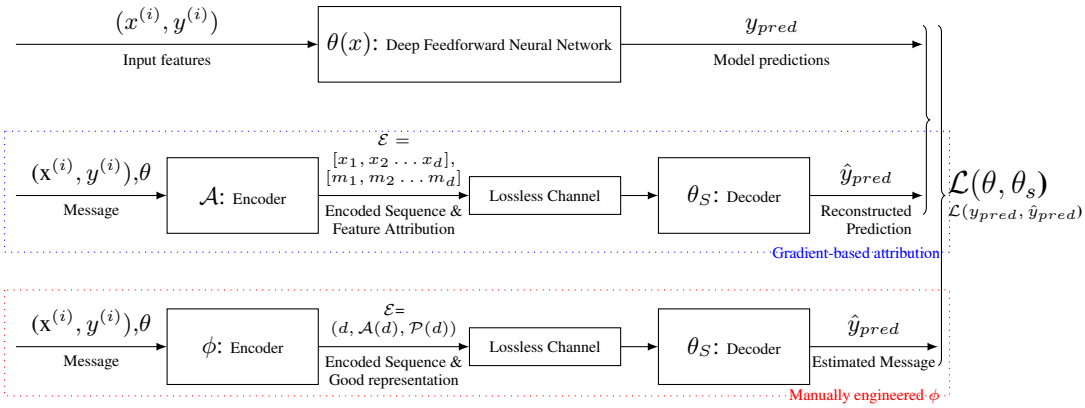

Figure 1: Problem construction as compression and constrained transmission of information. (1) shows the inference of the model. Then, (2) demonstrates the setup for evaluating attribution methods. Finally, (3) shows the setup to assess all generalised explainability methods.

these results cannot be regarded as conclusive[2]. As a result, many recent analyses have relied on custom image datasets to find false positive explanations[29], and "sanity checks" to demonstrate that attributions can often inaccurately reflect model behaviour[23, 4, 30].

Information theoretic paradigms Jung et al. (2020) established the information-theoretic underpinnings of deep learning [12]. Ziv et al. (2017) [20] showed the loss of information as it gets propagated through neural networks.

3 Our evaluation framework

In this section, we present our evaluation framework `InfoExplain` to estimate the amount of information about the model's decision process inside feature attributions.

Problem Setting Data points are considered to be of the format $(x^{(i)}, y^{(i)})$, where $x^{(i)} \in \mathbb{R}^d$ and label $y^{(i)} \in \mathcal{Y}$, which is drawn from the distribution $\mathcal{D}_o$ on $\mathbb{R}^d \times \mathcal{Y}$. Let us consider the deep-feedforward *neural network* θ, which is trained to perform this task. Deep learning strategy relies on learning $\phi(x)$ where $y_{pred} = f(x; \lambda, \mu) = \phi(x, \lambda)^T \mu$, similar to how the kernel trick is used to extend linear models to non-linear decision boundaries [10]. The parameters λ are used to learn ϕ from a broad range of functions, and the parameters μ map $\phi(x)$ to the desired output space $\mathcal{Y}$. This parameterised representation, $\phi(x; \lambda)$, uses optimisation algorithms to learn the λ that leads to *a good representation*. We can *manually engineer* ϕ to generate a good representation if we are not using deep learning.

An observer $\mathcal{O}$ wants to understand the decision process. Since neural networks learn ϕ, λ, and μ to create an optimal compression [20], these parameters of θ are not human-interpretable - but exist in a disentangled representation[7].

The *feature attribution scheme* $\mathcal{A} : \mathbb{R}^d \to \{\sigma\}$ maps this d-dimensional input x to the model, to σ : an ordering that ranks features in decreasing order of their importance, based on a magnitude m_d.Therefore inside the sequence σ, the respective $[\nabla_x \text{logit}_\theta(x, \hat{y})]_{\sigma_i} > [\nabla_x \text{logit}_\theta(x, \hat{y})]_{(\sigma_i+1)}$.

Explainability as Lossy Compression Deep neural networks spend a majority of their training time learning a compressed version (i.e. a good representation) of the input[20]. We assume explanations to be compressed representations of the model and input. Rate distortion theory, using an information-theoretic lens, addresses the problem of determining the minimal bits per symbol (R) that should be communicated over a channel so that the input signal can be approximately reconstructed at the receiver with a bounded maximum distortion D. In our setting, since R is fixed, we use the distortion to discover how much of the original signal about the model's decision process is transmitted through $\mathcal{D}_x$. Therefore, we are representing the problem of finding the optimal explainability method as a *rate-distortion optimisation*, where we analytically quantify information preserved by the "compression" performed by explainability methods.

The general mathematical model for a communication system contains a message W is to be transmitted, after processing via an encoding function f_{enc}, in blocks of length n. A channel, whether noisy or noiseless - can be modelled as a conditional probability distribution $p(y|x) = p_{Y|X}(y|x)$. The channel output is passed through a decoding function f_{dec}, which yields the estimate of the transmitted message. In our setting, based on the Information provided by the explanations, we try to reconstruct the model's decision from the input data.

The encoder can be either assumed to be a gradient-based attribution method or an arbitrary hand-engineered explanation representation. By using Shannon's source coding theory, all information is of the form $\mathcal{E}$ that is represented using the triplet $(d, \mathcal{A}_d, \mathcal{P}_d)$ [13] - where d are the different features inside $\mathcal{E}$, $\mathcal{A}_d$ are the different ways in which d manifests i.e. m_d, and $\mathcal{P}_d$ is the probability of $\mathcal{A}_d$. We refer to this distribution as $\mathcal{D}_x$

In our metric, we seek to learn a *decoder* θ_s, belonging to the family of interpretable models [22] $\theta_S \in G_i$ to map between the distribution $\mathcal{D}_x$ and the deep neural network's output, y_{pred}. The magnitudes in σ are vectorised to make $(x^{(i)})$ while the model's outputs are $(y^{(i)})$. Therefore, the optimal decoder can be learned by optimising for hamming distortion.

$$\text{Decoder}^*(x) = \underset{\theta_s \in G_i}{\text{argmin}} \mathcal{L}\left(\theta_s, \theta\right); \mathcal{L} \rightarrow d(y_{pred}, \hat{y}_{pred}) = \begin{cases} 0 & \text{if } y_{pred} = \hat{y}_{pred} \\ 1 & \text{if } y_{pred} \neq \hat{y}_{pred} \end{cases}$$

The InfoExplain consistency metric, $\mathcal{I}_c = 1 - \mathcal{L}(\theta, \theta_s)$, essentially measures how accurately we can predict the model's decision, based on information present in explanations. When running over $\mathcal{T}_i$ techniques, the score $\mathcal{I}_c^i$ allows the ranking for (A). Since the surrogate decoder, θ_s is interpretable, such as a decision tree, we can parse it to find counterfactuals. In susceptible domains, a regulatory body may verify the automated decision process by looking at the decision process. Furthermore, a human can intervene appropriately when the verified surrogate model and the deep neural network provide different decisions in production deployments.

Experimentally, we know that interpretable models are poor at fitting complex decision boundaries compared to neural networks. Therefore, we define the reference InfoExplain ability metric, $\mathcal{I}_a = 1 - \mathcal{L}(\theta, \theta_{s*})$, where θ_{s*} is the same model type as θ_s but is trained to map between x. This measures how well an interpretable decoder can fit a very complex decision boundary of the original neural network based on the training data.

4 Results

Benchmarking *feature attributions* For the sake of demonstrability, we choose to illustrate a simple task. On the titanic dataset, a deep feedforward neural network θ was trained to predict survival y, based on factors such as class, age etc (x). For the same decision, and the same model - we observed very different feature attributions $\mathcal{E}$ from other explainer methods $\mathcal{T}$. To benchmark, we generated explanations for every decision, and tried to reconstruct the model's output(y_{pred}) based on these attributions $\mathcal{E}$. All metrics for θ_s were calculated on randomly chosen, unseen test data.

Explainer method $\mathcal{T}$ i.e. Encoder $\mathcal{A}$	Rule-Fit [9]	Skope-Rules[1]	Boosted Rules[8]	C4.5 [16]	Greedy Rule[22]	Decision Tree[14]
$\mathcal{I}_a$ (Ability-score)	68.70	69.21	63.86	67.93	96.43	**97.45**
Saliency ($\mathcal{I}_c$)[21]	67.17	61.06	69.21	59.28	76.08	**78.62**
IG ($\mathcal{I}_c$) [28]	64.88	69.21	51.65	64.12	89.05	**91.34**
DeepLift [18, 4]($\mathcal{I}_c$)	66.15	69.21	60.05	63.35	91.60	**92.36**
GuidedBackprop ($\mathcal{I}_c$)[25]	65.64	68.95	38.67	58.77	77.60	**80.66**
InputXGradient[19] ($\mathcal{I}_c$)	64.37	69.21	60.30	63.86	**88.80**	88.29

Table 1: Ability and consistency scores, benchmarking different explainability techniques and decoders.

Using hand-engineered explanations For this example, we consider the Olivetti dataset a standard benchmark for the facial recognition task that contains ten images of 40 people with varying lighting conditions and expressions. We use the pre-trained Facenet model, with a downstream gradient-boosting model as the black box neural network. The encoder can be hand-engineered to use

human-understandable explanatory variables. We extract facial features, such as the nose, eyes, jaw, etc. - as features of the explanation space. Each feature is classified without supervision, and the yielded cluster labels are used in the encoded explanation. The decoder has to map between explanation space to the y_{pred}. Using InfoExplain, we obtain a $\mathcal{I}_c$= 8.92% on the testing data, which is just a little better than random guessing. However, on further analysis, the optimal decoder had an accuracy of 98.2% on its training set. This indicates too much noise in $\mathcal{E}$.

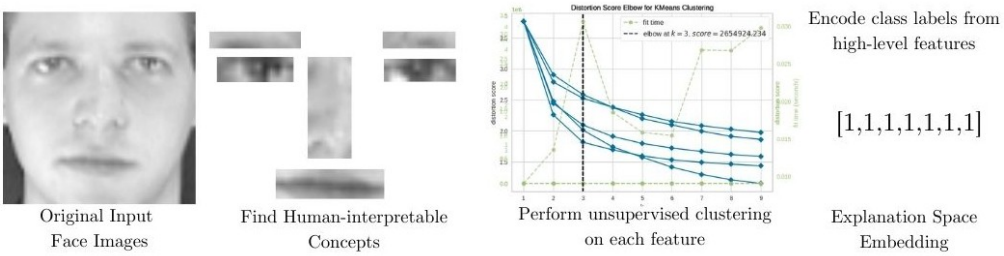

Figure 2: Using human-interpretable explanatory variables for Facial Recognition.

5 Discussion

The crux of our method relies on an encoder $\mathcal{A}$ mapping from the input space $\mathcal{X}$ to an explainer space $\mathcal{E}$, with a decoder θ_s that maps from explainer space $\mathcal{E}$ to model prediction space y_{pred}. Explainer methods can also be handcrafted so humans and machines can speak the same language. We should, however note that the dimensionality of $\mathcal{E}$ is very important. If $\mathcal{E}$ is in a high enough dimension, θ_s can always have enough capacity to fit the training set, but generalisation on the test set remains very poor. Moreover, these explanations do not encode enough prior information. An AI system must use the same way of thinking to make decisions, i.e. has fidelity. The interpretable nature of θ_s provides performance guarantees and can trigger human intervention when the system cannot offer the guarantee.

Is θ_s faithful to θ? A complete rule-based breakdown θ_s explaining the model θ cannot exist. If these rules ultimately captured the behaviour of a black box neural network, then there would be no need for the black box - as it could be discarded in favour of our interpretable model. It has been argued that an explanation of a black box model is, by definition inaccurate. Our work aims to find the best possible θ_s, based on the best $\mathcal{E}$. Since θs is a surrogate model, it may make decisions in a very different manner when compared to θ. We are looking for the θ_s that closely matches the overall behaviour of θ.

Distortion scores The notion of *distortion* is a part of ongoing discussion. Our work defines distortion mathematically as hamming distortion (squared error distortion could also be used). However, the data processed by lossy compression algorithms (music encoders, explainers) are consumed by humans - for whom this may not necessarily hold. Therefore, arguments are often made that distortion measures should be modelled on human perception and aesthetics. In that case, distortion measures can be redefined with perceptual distortion measures, such as those used in mp3 music. This paradigm, however, does not fit inside our information-theoretic problem setup.

6 Conclusion

Different techniques may give different explanations for the same decision and model. To tackle this issue, we propose a novel information theoretic test to quantify the amount of information inside explanations. This provides us with a framework for making more trustable systems and that allows artificial intelligence to emulate human-like cognition. We want to explore more nuanced and complex methods of building human-aligned artificial intelligence in future work.

References

[1] Skope rules 0.1.0 documentation. https://skope-rules.readthedocs.io/en/latest/.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps, November 2020. Number: arXiv:1810.03292 arXiv:1810.03292 [cs, stat].

[3] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging Tests for Model Explanations, November 2020. Number: arXiv:2011.05429 arXiv:2011.05429 [cs].

[4] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for Deep Neural Networks, March 2018. Number: arXiv:1711.06104 arXiv:1711.06104 [cs, stat].

[5] Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating Recurrent Neural Network Explanations, June 2019. Number: arXiv:1904.11829 arXiv:1904.11829 [cs, stat].

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. Publisher: Public Library of Science.

[7] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations, April 2017. Number: arXiv:1704.05796 arXiv:1704.05796 [cs].

[8] Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[9] Jerome H. Friedman and Bogdan E. Popescu. Predictive Learning via Rule Ensembles on JSTOR. https://www.jstor.org/stable/30245114.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[11] Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz*, 34(2):193–198, June 2020.

[12] Alexander Jung and Pedro H. J. Nardelli. An Information-Theoretic Approach to Personalized Explainable Machine Learning, March 2020. arXiv:2003.00484 [cs, stat].

[13] David MacKay. *Information Theory, Inference, and Learning Algorithms*.

[14] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, March 1986.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016. Number: arXiv:1602.04938 arXiv:1602.04938 [cs, stat].

[16] Steven L. Salzberg. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, 16(3):235–240, September 1994.

[17] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. Evaluating the visualization of what a Deep Neural Network has learned, September 2015. Number: arXiv:1509.06321 arXiv:1509.06321 [cs].

[18] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, October 2019. Number: arXiv:1704.02685 arXiv:1704.02685 [cs].

[19] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, April 2017. Number: arXiv:1605.01713 arXiv:1605.01713 [cs].

[20] Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information, April 2017. Number: arXiv:1703.00810 arXiv:1703.00810 [cs].

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. Number: arXiv:1312.6034 arXiv:1312.6034 [cs].

[22] Chandan Singh, Keyan Nasseri, Yan Shuo Tan, Tiffany Tang, and Bin Yu. imodels: a python package for fitting interpretable models. *Journal of Open Source Software*, 6(61):3192, May 2021.

[23] Leon Sixt, Maximilian Granz, and Tim Landgraf. When Explanations Lie: Why Many Modified BP Attributions Fail, August 2020. Number: arXiv:1912.09818 arXiv:1912.09818 [cs, stat].

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise, June 2017. Number: arXiv:1706.03825 arXiv:1706.03825 [cs, stat].

[25] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, April 2015. Number: arXiv:1412.6806 arXiv:1412.6806 [cs].

[26] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1379, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1379.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017. Number: arXiv:1703.01365 arXiv:1703.01365 [cs].

[28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017. Number: arXiv:1703.01365 arXiv:1703.01365 [cs].

[29] Mengjiao Yang and Been Kim. Benchmarking Attribution Methods with Relative Feature Importance, November 2019. Number: arXiv:1907.09701 arXiv:1907.09701 [cs, stat].

[30] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (In)fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[31] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, November 2018. Publisher: Public Library of Science.