EVALUATING FEATURE ATTRIBUTION METHODS

Benedict Gattas

June 2020

School of Information Technology and Electrical Engineering,
The University of Queensland

Submitted for the degree of
Bachelor of Engineering
in the field of Software Engineering.

Benedict Gattas
benedict.gattas@uqconnect.edu.au

June 14, 2020

Prof Amin Abbosh
Acting Head of School
School of Information Technology and Electrical Engineering
The University of Queensland
St Lucia, Q 4072

Dear Professor Abbosh,

In accordance with the requirements of the degree of Bachelor of Engineering (Honours) in the division of Software Engineering, I present the following thesis entitled "Evaluating Feature Attribution Methods". This work was performed under the supervision of Dr Alina Bialkowski.

I declare that the work submitted in this thesis is my own, except as acknowledged in the text and footnotes, and has not been previously submitted for a degree at The University of Queensland or any other institution.

Yours sincerely,

Benedict Gattas.

Acknowledgments

I wish to acknowledge the direction and support provided by Dr Alina Bialkowski over the course of this project. Feedback after each milestone was invaluable and tips on different approaches helped provide guidance at key junctions. This ranged from helping narrow the scope at the start of the project, suggestions after each deliverable to help improve the next one, and regular appraisal of the direction I was heading throughout the project. I also wish to extend my congratulations for the birth of her daughter! To my family, I would like to thank my parents for their care and support, and my siblings for the same.

Abstract

This document is a skeleton thesis for 4th-year students. The printable versions show the structure of a typical thesis with some notes on the content and purpose of each part. The notes are meant to be informative but not necessarily illustrative; for example, this paragraph is not really an abstract, because it contains information not found elsewhere in the document. The LaTeX 2_ε source file (`skel.tex`) contains some non-printing comments giving additional information for students who wish to typeset their theses in LaTeX. You can download the source, edit out the unwanted material, insert your own frontmatter and bibliographic entries, and in-line or `\include{}` your own chapter files. Of course the content of a particular thesis will influence the form to a large extent. Hence this document should not be seen as an attempt to force every thesis into the same mold. If in doubt about the structure of your thesis, seek advice from your supervisor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

"State-of-the-art" machine learning models have achieved above expert-level performance in fields as diverse as medical imaging and criminal justice. However, lack of interpretability has prevented their adoption in many of the fields that would benefit the most from them. These domains are often ones where the decisions made have a tangible impact on people's lives. Because of their requirement for trust and accountability, decision-makers in those domains are unlikely to use the predictions made by a highly accurate but opaque model.

The requirement for explainability is not in order to replace human experts, but to understand contradictions between expert and algorithm. For example, a radiologist might disagree with a diagnosis made by a model trained to predict pneumonia from chest X-rays, and attributing the error to a factor that one or the other relied upon would be helpful. Was the model relying on some unrelated part of the scan (a spurious feature) or the radiologist failing to pick up on a subtle pneumonia differential? The former was observed in practice after applying an explanation technique to a convolutional neural network (CNN) trained on such data [1].

Extrapolating this view, explanations for model predictions can be as equally important as low rates of incorrect ones. The counter-argument to the notion that a highly accurate model need not be interpretable to be effective, is that the premise of effectiveness requires a level of *trust* that a model relies on unbiased data and non-spurious features, two guarantees that are not at all provided by an objective function that seeks only to minimise prediction error. With poor visibility into the factors that a model relies upon, machine learning researchers tend to use model performance metrics as the basis for arguing a new model architecture is superior. This disconnect between "performance" in the sense of test set accuracy and performance

in the sense of accountability and reliability is not ideal.

Interpretability in a Facebook algorithm recommending product categories, for example, might not be seen as important as interpretability in a cancer diagnosis model, though the possibility of unethical model behaviour from reliance on biased data is as tangible in both domains. One study of 200 sentiment analysis classifiers found several to have significant race and gender bias [2]. The consequences of errors can certainly be higher in some domains however - a poor Netflix recommendation is not as disastrous as a naive algorithm used in government decision-making, such as the "robo-debt" scheme recently employed by the Australian Government [3].

Approaches to Interpretability

There is fortunately an active literature aimed at addressing this 'black-box' critique in machine learning. The top-level distinction among approaches is to either use inherently intepretable models to achieve explainability, or take complex, black-box models and find techniques to isolate and explain a piece of their complexity, such as an individual prediction.

The first approach includes model families with low complexity like linear/logistic regression, decision trees, k-nearest neighbours and Naive Bayes. Within these families are both parametric and non-parametric techniques, which demonstrates that lack of interpretability is more related to model complexity than a particular type of formulation - this empirically observed trade-off between accuracy and interpretability is discussed further in the next section[1]. Since many of these models are often too simplistic for obtaining competitive performance, the motivation to attack the 'black-box' critique from this angle is quite low. Instead, methods to introduce interpretability to modern, high-performance models are a more studied and popular approach to take in the literature, as in the second approach.

The second approach includes "feature attribution" or "feature importance" methods, which compute a weight score for each feature in the input space to measure its contribution to an output class. For example, a CNN classifier predicting "tree" would be expected to rely heavily on green pixels of leaves. This class includes model-specific techniques for neural network architectures, like those based on activations in a hidden layer, and model-agnostic techniques that are compatible with any model family. Within both sub-classes are a variety of techniques, with varying levels of model agnosticity and task compatibility. For example, some methods are designed solely for CNNs and are therefore mainly suited for image classification or related tasks.

[1]There is a view by some researchers ([4]) that in many domains the accuracy vs interpretability trade-off does not exist, and thus there is a responsibility to use equally effective, inherently interpretable models for high-stakes decisions where those are available.

Importantly there are both 'global' feature attribution methods, which calculate each feature's contribution to a model at large, as well as 'local' methods, which attempt to explain a single prediction. This project has focused on the latter. Local methods are dominant in the literature for modern model architectures - when the dimensionality of the data is high, such as in visual data, or when the number of parameters is too high to make conclusions about global model behaviour (as in most modern architectures) this tends to be the only effective approach to interpretability. An author of one local, model-agnostic method notes that understanding these models globally "would require understanding of the underlying prediction function at all locations of the input space" [5].

Accuracy vs Interpretability

As deep learning and other state-of-the-art model families proliferate in their typical number of parameters, global behaviour has become even less explainable. Researchers maintain some intuitions about the impact of architectural design decisions, though not on predictive behaviour. For example, filters within a CNN model have been shown to act as 'object detectors' of patterns, shapes and other connected regions [6], though these per-layer intuitions don't explain how a network of dozens of layers will determine a husky from a wolf (a recent approach in the literature, however, has looked at abstracting model behaviour into 'concept' vector encodings to capture model information across filters and layers (Net2Vec [7], TCAV [8])).

Feature attribution methods can therefore re-introduce transparency into complex, non-linear models and highlight predictive biases in the context of individual predictions. They also can reveal unexpected features involved in a prediction, such as the spurious features mentioned in the previous X-ray data example, or bugs that could lead to exploitation of adversarial examples [9]. Note these methods do not seek to add causal interpretability to the models they are applied to, only to isolate and highlight a piece of complexity in a way that might make sense to an expert reviewing the explanation. This does not make them shallow - the benefit of the 'post-training' approach is that model designers have more flexibility in their choice of models, and fewer restrictive assumptions about model complexity are made[2]. More model-agnostic methods, with the least restrictive requirements, are not well understood in context with model-specific ones in terms of this accuracy and interpretability trade-off.

[2]The counter-argument made by those who argue for the inherently interpretable model approach is that there is no guarantee these explanations are faithful to the model, and that they extend the authority of the black box instead of making it a "glass box" [4]. This is revisited in the **Discussion**.

Existing Literature

Some comparisons of feature attribution methods do exist, though typically either in a qualitative context, as a pairwise comparison, **cite** or within a single class of methods. They are not normally compared on speed/performance, *adoptability* in terms of task or model compatibility, nor explanation *quality* (a difficult criteria to design). One can reason this is because the literature for some model families (i.e. deep learning) is relatively new, but it is also because of differences in method formulation and their own hyper-parameters that make it difficult to compare different approaches fairly.

1.2 Project Overview

This project has sought to evaluate a panel of feature attribution methods representative of different approaches to the interpretability problem. The aim was to highlight their relative strengths and weaknesses and thereby increase the understanding of the benefits of one method's approach over another.

Two key contributions made over the course of the project have been a quantitative evaluation framework for explanation quality in the image classification context, and the development of a software package to collect image data explanations for multiple underlying methods at scale.

Goals

In summary, the project has aimed to:

1. **Examine and evaluate** a panel of feature attribution methods for their use cases and performance, using proxy metrics of explanation quality supported by analysis.

2. **Develop** an attribution software package for testing methods at scale with modular support for different models, making it easier for researchers to collect explanations and build more adoptable models.

Project Scope

Section 1.1 introduced a broad motivation for interpretability, though this project has focused specifically on image classification for two main reasons.

Firstly, many interpretability techniques from the pre-deep learning era were studied in this domain, and many deep learning specific methods continue to be developed and tested in this domain on modern CNN architectures. The natural 'visual' aspect of image data explanations has also made computer vision a popular venue for interpretability research, with important applications such as medical imaging.

Secondly, well-annotated datasets and pre-trained 'off the shelf' models are more easily acquired in this domain. This availability allowed for both richer evaluation metrics and the removal of model training as a project requirement.

A more detailed scope is provided at the beginning of the Methodology section, including a description of the specific datasets, models, and feature attribution methods used.

Report Overview

Chapter 2 deals with a literature review of feature attributions methods, and existing evaluation metrics for comparing them. Chapter 3 breaks down the project's methodology in terms of particular milestones, software design and the evaluation metrics that were designed. Chapter 4 lists project results from both quantitative and qualitative standpoints. Chapter 5 provides a discussion on the project's contribution and the limitations encountered, and finally Chapter 6 provides conclusions and recommendations for future work.

Chapter 2

Background Research

2.1 Scope of Research

In Chapter 1 ("Approaches to Interpretability") a brief overview of feature attribution methods was provided with reference to a distinction between model-specific and model-agnostic methods. This is a common distinction in the literature and was also used to guide research in this project. The panel of methods chosen for evaluation ultimately consisted of a balanced selection from both approaches.

A major difficulty of this project was distilling the broad literature on these methods however. For the model-specific (neural network) family, different angles are commonly taken to calculate feature "relevance" or importance. These include generally:

- **Gradient-based** methods, saliency maps and output sensitivity methods
- **Backpropagation-based** methods or 'importance signal' visualisations (such as activations in a hidden layer)
- **Perturbation-based** methods and use of occlusion masks in the input space

This categorisation is based on two recent papers that make similar categorisations of explanation approaches [10] [11]. In this chapter a broad selection of methods based on traction over time, current popularity and representativeness of approach are described, though the reader should note there are many more methods under each of those three than have been described here.

For model-agnostic methods, perturbation-based and surrogate model approaches are considered. Again the selection was based on traction and literature popularity.

First reviewed are traditional, visual approaches to interpretability to provide context to the task of feature attribution. After the exploration of feature attribution methods, a review of existing evaluation metrics and comparison studies is provided.

Terminology

A note on terminology is that all methods are variously referred to as *attribution* methods in this project for any projection on the input space that causes an explanation to take place. Goodfellow, Kim, et al. refer to the broad category of "visualisation and attribution methods aimed at interpreting trained models" as *saliency* methods instead [12]. A *saliency map* is widely-used as a 'catch-all' term to refer to input space projections (individual explanations themselves) in the context of interpreting deep neural networks for image data ([12]). However they also refer to a specific gradient-based method (**cite below**).

There seems to be little consensus around terminology. Some researchers ([11]) describe attribution methods as a subclass of explanation techniques where contribution scores are specifically calculated for each input feature (i.e. excluding higher level 'patterns' which cause neuron activations, as in Zeiler & Fergus ([13]) (Section **X**) or the back-propagation class generally).

In summary, attribution methods are used here as a general term, but can refer specifically to contribution 'calculators', and saliency methods/maps is widely used in the context of neural networks and image data.

2.2 Traditional Approaches

2.2.1 Feature Projection

Visualisation tools for high-dimensional data are a popular way to gauge insight into expected model behaviour. These pre-learning, exploratory data analysis techniques include mathematical reductions like PCA and probabilistic techniques like t-SNE, projecting high-dimensional examples that are 'similar' into a visualisable 2D or 3D space [14] (Figure 2.2).

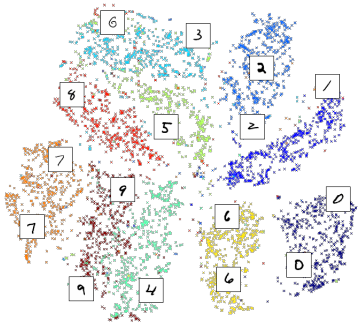

Figure 2.1: 2D embedding of 70,000 handwritten digits (0-9) from MNIST [15]

Other methods of clustering and dimensionality reduction are also widely used for interpreting data, and although useful for gaining an intuition on relationships

between features, they are not suited for explaining model behaviour as they examine only the input space itself.

2.2.2 Partial Dependence Plots

A partial dependence plot (PDP) is a tool to demonstrate the marginal effect of one or two features on a prediction outcome. It was proposed by Friedman in 2001 to interpret and visualise the features that his gradient boosting machine relied upon (though it is limited to 1 or 2 input features such that it can be displayed) [16]. A partial dependence function $\widehat{f_{xs}}$ can be calculated for some desired set of features S, by marginalising the model output over the set of 'complement' features C (all other features):

$$\widehat{f_{xs}} = \int \widehat{f}(x_s, x_c) dP(x_c) \tag{2.1}$$

It can be approximated with a Monte Carlo method. Friedman believed in 2001 that these might be used to help interpret "any black box prediction method" such as NN and SVM architectures, and that, "[...] when there are a large number of predictor variables, it is very useful to have a measure of relevance to reduce the potentially large number of variables to be considered" [16]. The mentioned relevance measure was defined only in the context of the decision trees which constituted the paper's gradient boosting machine. Certainly, PDPs are suited for the low-dimensional feature spaces that were imagined in the pre-deep learning era, and are less suitable for high-dimensional input spaces such as in image classification. They are also restricted by an unrealistic assumption of independence among features.

2.3 Model-Specific Methods

Deep learning's reputation for lack of transparency has led to many attempts to explain the predictions of complex NN architectures. This section examines representative attribution methods from the gradient-based, backpropagation-based and perturbation-based approaches overviewed in Section 2.1, with some emphasis on those developed in the context of CNNs (i.e. image data).

2.3.1 Backpropagation-Based

This class of methods try to isolate an internal model signal, such as neuron activations in a target hidden layer, and map these signals back into the input pixel space. Zeiler and Fergus (2014) introduced the motivation for signal backpropagation as, '[...] showing what input pattern originally caused a given activation in the feature maps" **zeiler**.

Feature Visualisation

First described is the feature visualisation approach, which is a technique to visualise inner model workings. This is not a true feature attribution method, as it is visualising internal 'trained features' instead of attributing the contributions of input features, though it is briefly discussed for its relevance to formulations in the backpropagation family and to gradient-based methods as well.

The success of visualisation approach is a technique to explore internal model workings, via visualisations of units within a model. These units can be layers, individual convolutional channels (or feature maps) or even an individual neuron like out an output class neuron

DeConvNet

In their creation of class-based saliency maps (Section 1.3.1) Simonyan et al. showed that DeconvNet-based reconstruction of the input to a hidden is equivalent to computing the gradient of visualised neuron activations with respect to that input [17].

Guided BackProp

Layer-wise Relevance Propagation

DeepTaylorDecomposition

2.3.2 Gradient-Based

These methods aim to explain a class output in terms of sensitivity in the input space by relying on a gradient function of the output. The goal is to find the input features that make that prediction more or less confident. For an output prediction of 'tree' for example, they seek to answer "what makes a tree more/less a tree?".

Saliency Maps

An early formulation of a local explanation method was provided by Baehrens et al. (2010) for any nonlinear classification algorithm [18] (though developed in the context of Bayesian classification). The local explanation gradient vectors that this paper devised were based on class probability gradients, characterising how much a data point has to be moved for a predicted label to change. Simonyan et al. (2014) later applied a similar idea to CNNs to create 'class saliency maps' specific to a given image and class [17].

The formulation is based on finding the derivative of an output class with respect to an input image via back-propagation. The authors also formulate a method to generate an image that maximises the output class score for a particular class, to visualise the model's 'interpretation' of a class. This paper sparked popularity for

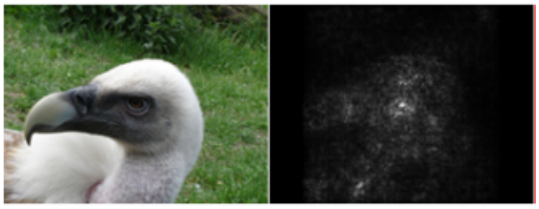

Figure 2.2: (From [17]) Example of an image-specific class saliency map.

saliency maps and further interest in creating explanations from network gradients. Note that a drawback of saliency maps is that noisy images can be produced when a model does not distinguish between objects that are being predicted and nearby objects that are associated (i.e. a tree with leaves, in an image of a bird).

Class Activation Maps

Another visual approach aimed at understanding the behaviour of CNNs was introduced by Zhou et al. (2015) and

GradCAM gradcam efficient variation https://arxiv.org/abs/1911.11293

Integrated Gradients (ref DeepLIFT)

SmoothGrad

2.3.3 Perturbation-Based

occlusion and ablation

occlusion mask, zeiler and fergus

2.4 Model-Agnostic Methods

"While we have made a case for model agnosticism, this approach is not without its challenges. For example, getting a global understanding of the model may be hard if the model is very complex, due to the trade-off between flexibility and interpretability. To make matters worse, local explanations may be inconsistent with one another, since a flexible model may use a certain feature in different ways depending on the other features. In Ribeiro et al. (2016) we explained text models by selecting a small number of representative and non-redundant individual prediction explanations obtained via submodular optimization, similar in spirit to showing prototypes (Kim et al., 2014). However, it is unclear on how to extend this approach to domains such as images or tabular data, where the data itself is not sparse. In some domains, exact explanations may be required (e.g. for legal or ethical reasons), and using a black-box may be unacceptable (or even illegal). Interpretable models may

also be more desirable when interpretability is much more important than accuracy, or when interpretable models trained on a small number of carefully engineered features are as accurate as black-box models." Model-Agnostic Interpretability of Machine Learning

2.4.1 Perturbation-Based

Both gradient and backpropagation-based methods operate under the assumption that propagating an output signal, back through a classifier model, is a means to explain how a relevant signal was originally encoded in an input [11]. An alternative approach is to treat the model as more of a black box, by perturbing the input space (such as occluding important parts) to study the effect on the output. Changes in output would reveal that occluded parts of the input are important in a prediction.

"The problem of attribution is concerned with identifying the parts of an input that are responsible for a model's output. An important family of attribution methods is based on measuring the effect of perturbations applied to the input"

2.4.2 Surrogate Models

some overlap with perturbation-based methods for how they are trained

variants of lime include KL-lime

2.5 Evaluation Metrics

2.6 Existing Explanation Frameworks

2.7 Existing Evaluation Studies

Comparisons

Chapter 3

Methodology

Procedure, design, etc. This may be one chapter or several. Again, titles should be more informative than the above.

Chapter 4

Results

Chapter 5

Discussion

KL-Lime discussion on role of itnerpretable models

Chapter 6

Conclusions

6.1 Summary and conclusions

6.2 Possible future work

Appendix A

Software Documentation

Appendix B

Software Descriptions

B.1 Attributer Class

Text

B.2 Evaluator Class

Appendix C

Software Repository Link

Bibliography

[1] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Confounding variables can degrade generalization performance of radiological deep learning models," *CoRR*, vol. abs/1807.00431, 2018. arXiv: 1807.00431. [Online]. Available: http://arxiv.org/abs/1807.00431.

[2] S. Kiritchenko and S. M. Mohammad, *Examining gender and race bias in two hundred sentiment analysis systems*, 2018. arXiv: 1805.04508 [cs.CL].

[3] D. Diviny, *How to effectively use machine learning to implement public policy*, 2019. [Online]. Available: https://www.nousgroup.com/insights/effectively-use-machine-learning-implement-public-policy/.

[4] C. Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, 2018. arXiv: 1811.10154 [stat.ML].

[5] T. Peltola, "Local interpretable model-agnostic explanations of bayesian predictive models via kullback-leibler projections," *CoRR*, vol. abs/1810.02678, 2018. arXiv: 1810.02678. [Online]. Available: http://arxiv.org/abs/1810.02678.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, *Object detectors emerge in deep scene cnns*, 2014. arXiv: 1412.6856 [cs.CV].

[7] R. Fong and A. Vedaldi, "Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks," *CoRR*, vol. abs/1801.03454, 2018. arXiv: 1801.03454. [Online]. Available: http://arxiv.org/abs/1801.03454.

[8] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)*, 2017. arXiv: 1711.11279 [stat.ML].

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial examples*, 2014. arXiv: 1412.6572 [stat.ML].

[10] A. Shrikumar, P. Greenside, and A. Kundaje, *Learning important features through propagating activation differences*, 2017. arXiv: 1704.02685 [cs.CV].

[11] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, *Learning how to explain neural networks: Patternnet and patternattribution*, 2017. arXiv: 1705.05598 [stat.ML].

[12] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *CoRR*, vol. abs/1810.03292, 2018. arXiv: 1810.03292. [Online]. Available: http://arxiv.org/abs/1810.03292.

[13] M. D. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks*, 2013. arXiv: 1311.2901 [cs.CV].

[14] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[15] A. Fabisch. (2014). "T-sne in scikit learn," [Online]. Available: https://alexanderfabisch.github.io/t-sne-in-scikit-learn.html.

[16] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, Nov. 2000. DOI: 10.1214/aos/1013203451.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2013. arXiv: 1312.6034 [cs.CV].

[18] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Mueller, *How to explain individual classification decisions*, 2009. arXiv: 0912.1128 [stat.ML].