Leveraging Explainability Methods For Class Discovery Within Under-labeled Datasets

Informatics AIAI PhD Research Proposal

Patrick Kage

July 14, 2021

1 Abstract

My proposed area of research is in utilizing explanatory methods for deep neural classifiers to determine properties of the dataset over which the classifier operates—specifically latent structure. Existing methods in this field are limited by the fact that they do not handle high-dimensional data, or data such as images or audio. However, by generating instance explanations of every point in a dataset a projection may be formed which is potentially easier to interpret using low-power statistical methods (e.g. principal component analysis) than the original dataset; effectively, this projection allows these low-power methods to inherit the statistical power of a deep neural classifier. I plan to initially start this investigation by utilizing this explanatory space to find finer delineations of classes within datasets than the labels of that dataset imply (i.e. latent structure), and using these methods to automatically create deeper label hierarchies than already exist in the dataset. This is building off the foundations of my undergraduate honours project, which proposed a novel technique for detecting unlabeled subclasses within a dataset by analyzing the outputs of explanatory methods over pre-trained classifiers, clustering instances with similar explanations to discover latent subclasses [9]. This project was a very limited exploration of the topic, and it is distinct from discovery of general latent structure as it was only concerned with detecting latent subclasses within a singular class and not detecting general structure across class boundaries. Additionally, the project's interpretation of the explanation data was severely limited by the short scope of the project, and further study into this area is warranted.

2 Introduction

Explainability of machine learning models is a nascent but rapidly expanding field; as more and more decisions are made by computers rather than humans the need for auditing models became apparent [5]. Several techniques were developed for determining the reasoning behind automated decision making, such as guided backpropagation methods (e.g. DeepLIFT) and local decision modeling (e.g. LIME), and these methods determine the reasoning steps behind a single instance classification of a model.

The central hypothesis of this research proposal is that these explainability methods can be used in aggregate to produce an "explanation space" representation of a dataset, and this explanation space can be used to guide dataset discovery and auditing in ways that would be difficult to impossible to do with traditional analytic methods. As a specific focus, latent class detection is a difficult problem to solve for complex inputs (e.g. image/audio/video data, high dimensional data) with traditional techniques, however with an explanation space

representation latent classes are detectable. Interpreting the explanation space has significant benefits over simply representing the input with a finite mixture model (FMM), as FMMs are limited in several respects; this is discussed further in Section 3.

Utilizing explanatory methods is not without basis—in fact, my honours project explored this topic and showed that by using basic methods a simple classifier could accurately show intra-class latent structure segmentation in image data with low power methods (using principal component analysis)[9]. This was achieved by clustering over the explanation space, as seen in Figure 1. With this technique, latent structure was detectable in explanation space with simple PCA and clustering where the baseline was not able to detect this structure. This positive result from this exploration shows that this technique has merit, and is deserving of further study as this shows that explanation-guided analysis can outperform both benchmark methods and other state-of-the-art methods.

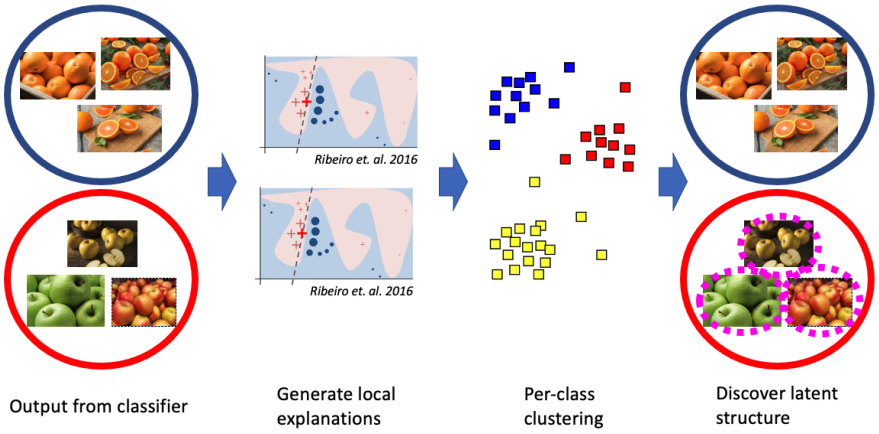

Figure 1: An example of a explanation-guided pipeline to find latent class structure via hierarchical clustering over an explanation space. Here, latent classes are identified strictly within an existing label structure.

3 Background

This proposal puts forward a set of techniques for discovery of features within a dataset through the lens of an aggregate set of instance explanations of a classifier (thereby inheriting the neural classifier's statistical power). This is in contrast to a set of traditional methods for extracting information from datasets, which are limited in comparison. Traditionally, solutions in this space involved either manual subgroup analysis, latent class analysis (LCA), or mixture-model-based methods [2]. Manual analysis is fraught with complications as it is easy to miss latent classes, due to the human-driven nature of the analysis. Latent class analysis is another common technique in this field, and is a form of finite mixture model (FMM) [2]. Finite mixture models (FMMs) assume that the dataset is comprised of a linear combination of basis functions—which is not necessarily the case, especially for complex data such as images or speech [1]. FMMs also require that the full distribution of the data be re-learned (adding a costly training step), and they are only capable of representing the data as a linear combination of basis functions. Additionally, FMMs do not take advantage of the existing labeling on the dataset—instead requiring a from-scratch allocation of basis functions[1].

This proposal builds off of the nascent field of explainable artificial intelligence (XAI). Briefly, this field is concerned with determining the decision making process of machine learning

models in order to improve their auditability, and most methods focus around generating explanations for individual instances, typically in the form of saliency maps over the input features to determine which features were relevant to a classification (so called "outcome explanations") [5]. Several techniques exist for generating saliency maps from neural networks, mainly based around backpropagation methods such as gradient ascent [3] and DeepLIFT [7]. These methods calculate the gradient of the network outputs with respect to each input feature, effectively creating a saliency map over input image. Additionally, other techniques exist such as leave-one-out encoding [8] where a section of the image is blanked out and inference is re-run to generate a saliency map; this provides an effective saliency map even in black-box networks (i.e.agnostic of the exact classifier architecture used) at the expense of greatly increased processing time.

The technique of using explanatory or saliency methods to guide other algorithms has some backing in existing literature, though it is a nascent area of research. For example, a technique in this space is the detection of rare subclasses via *commonality metrics* [6]. This analyzes the input training classes for a given dataset, and for each class determines the average activation in the penultimate neural layer. Each instance activation is then compared to the average activation and is scored on how similar it is, with the hypothesis that activations far from the mean represent instances which are in some way different from the average. Effectively, this technique is extracting additional information from the trained classifier than was originally contained within the class labels—which is similar in concept to this proposal. However, this proposal aims to use more advanced explainability methods to generate the explanatory space. Another paper in the same theme is bottom-up (saliency-guided) multiple-class learning (bMCL) [4]. bMCL focuses on fully unlabeled data, and proposes a method of finding objects within images (multiclass object detection) by using saliency methods to zero in on interesting regions of an image. This is an example of using saliency to guide other algorithms, which is thematically similar to this proposal—in fact, bMCL's paper itself notes that the use of saliency as a guide for other algorithms is little-exploited [4]. In contrast to this proposal, however, bMCL is focused on using saliency maps to guide new object detection within an image before using classical techniques, and not on finding whole classes within an existing dataset.

4 Methodology

As stated previously, the central aim of this proposal is to investigate the hypothesis that the explanation space contains useful information about the dataset and classifier. To be explicit, this is proposing to find an algorithm that does the following:

Input: A trained deep neural classifier, and a dataset (with class labels).
Output: A new set of class labels, ideally capturing latent structure in the dataset.

Using explainability methods to achieve this end is the goal of this proposal, and to benchmark the performance of the algorithms baseline techniques will need to be used. These include finite mixture model based methods and simply clustering over the original space.

To generate an explanation space, a series of explanatory methods may be used. These are discussed more fully in Section 3, however in general methods which produce saliency maps are convenient to work with as they have the same dimensionality as the original data. A further area of research would be into more advanced explanation methods which produce other kinds of explanations, such as directly tracking hidden layer activations. These individual instance explanations can then be aggregated to produce a projection of the original dataset into the explanation space.

Once this space is created, the next step is to find a robust clustering method through that space. Methods for interpreting the saliency maps can be borrowed from machine vision, such as a histogram of gradients (HoG) or a scale-invariant feature transform (SIFT). To

find latent classes, a clustering over this space can be performed—the exact method of this clustering is an area of research. Additional uses of this spaces can be envisioned; proposed research areas are proposing automatic class labels for under-labeled datasets, detecting anomalies in training data or in target datasets, and auditing the performance of a classifier. These are areas for which low-statistical-power methods exist, but which are poorly-suited to complex data.

The datasets over which these methods will be tested is an additional concern—in order to detect latent structure, we need a dataset in which latent structure exists. Unfortunately, it is difficult to find a dataset with such structure in it as most publicly-available test datasets are well-labeled. To score the methods on how well they perform, we also need the true labels that capture the latent structure as well as the original labels which do not. A possible solution to this is artificial latent structure[9], which "bridges" class labels together to create a superclass containing instances from two or more classes into one, effectively re-labeling the dataset. This has the benefit of both providing a dataset with latent structure and providing labels for that structure.

This methodology is appropriate to answer the research question as these are the current leading methods of generating explanations [5], and in a review of currently published research this is a novel method of dataset discovery. Additionally, this approach has significant benefits over more traditional methods of dataset discovery as this method inherits the statistical power of a full neural classifier; this is in contrast to a more traditional methods (e.g. FMMs for latent structure analysis). An additional area of research will be in baseline methods for generating mixture models, e.g. using generative adversarial networks as basis functions.

Success of this methodology will be determined by its ability to outperform the baseline methods (discussed above) in determining the latent structure in a series of different model architectures and datasets.

5 Considerations

5.1 Communicability

This area of research has the potential to meaningfully contribute to the field of dataset discovery, as well as being broadly applicable to a wide range of other machine learning topics. For example, latent structure detection is an important component of many medical studies, and automating the discovery process (and therefore augmenting/replacing traditional subgroup analyses) is an important step towards ensuring that these studies are correct.

5.2 Ethics

As an entirely computation-based proposal, there are no major ethical considerations for this project. Testing of the latent structure detection methods will occur on already-publicly-available datasets, and the need for additional datasets is unlikely to arise. In the event that it does, additional datasets will be created in full compliance with ethical guidelines.

The use of this technology has ethical ramifications as it is a machine learning proposal. Here, we are using ML to increase our understanding of the datasets that our algorithms are operating over which allows for a better understanding of the assumptions inherent in the data. This has a tremendous potential for social good, as it enables us to carefully examine both our data and the classifiers which operate over it.

6 Summary

In summary, this proposal puts forth a new methodology for finding finer class delineations than exist currently in a dataset through leveraging a classifier's explanations of the dataset itself. This methodology has some precedent in literature (see Zhu et al. [4] and Paterson & Calinescu [6]), but is an open area of research. Further study into this area would potentially uncover a new set of methods for dataset discovery, and also allow for the porting of low-statistical-power methods (such as simple clustering algorithms) to high-dimensional/complex data (such as images). This in turn has wide-reaching effects through the field of dataset discovery and classifier auditability, and through those benefit fields as diverse as medical artificial intelligence and autonomous vehicles.

7 Bibliography

[1] Wilfried Siedel. "Mixture Models". en. In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer, 2011, pp. 827–829. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_455.

[2] Stephanie T. Lanza and Brittany L. Rhoades. "Latent Class Analysis: An Alternative Perspective on Subgroup Analysis in Prevention and Treatment". In: *Prevention science : the official journal of the Society for Prevention Research* 14.2 (Apr. 2013), pp. 157–168. ISSN: 1389-4986. DOI: 10.1007/s11121-011-0201-1.

[3] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv:1312.6034 [cs]* (Apr. 2014). arXiv: 1312.6034 [cs].

[4] Jun-Yan Zhu et al. "Unsupervised Object Class Discovery via Saliency-Guided Multiple Class Learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.4 (Apr. 2015), pp. 862–875. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2014.2353617.

[5] F. K. Došilović, M. Brčić, and N. Hlupić. "Explainable Artificial Intelligence: A Survey". In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. May 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040.

[6] Colin Paterson and Radu Calinescu. "Detection and Mitigation of Rare Subclasses in Neural Network Classifiers". en. In: *arXiv:1911.12780 [cs, stat]* (Nov. 2019). arXiv: 1911.12780 [cs, stat].

[7] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features Through Propagating Activation Differences". In: *arXiv:1704.02685 [cs]* (Oct. 2019). arXiv: 1704.02685 [cs].

[8] Ali Abdalla. "A Visual History of Interpretation for Image Recognition". In: *The Gradient* (2021).

[9] Patrick Kage and Pavlos Andreadis. "Class Introspection: A Novel Technique for Detecting Unlabeled Subclasses by Leveraging Classifier Explainability Methods". In: *arXiv:2107.01657 [cs]* (July 2021). arXiv: 2107.01657 [cs].