A SURVEY ON VISUALIZATION FOR EXPLAINABLE CLASSIFIERS

by

YAO MING

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Supervised by Prof. Huamin Qu

October 2017, Hong Kong

TABLE OF CONTENTS

A SURVEY ON VISUALIZATION FOR EXPLAINABLE CLASSIFIERS

by

YAO MING

Department of Computer Science and Engineering

The Hong Kong University of Science and Technology

ABSTRACT

Classification is a fundamental problem in machine learning, data mining and computer vision. In practice, interpretability is a desirable property of classification models (classifiers) in critical areas, such as security, medicine and finance. For instance, a quantitative trader may prefer a more interpretable model with less expected return due to its predictability and low risk. Unfortunately, the best-performing classifiers in many applications (e.g., deep neural networks) are complex machines whose predictions are difficult to explain. Thus, there is a growing interest in using visualization to understand, diagnose and explain intelligent systems in both academia and in industry. Many challenges need to be addressed in the formalization of explainability, and the design principles and evaluation of explainable intelligent systems.

The survey starts with an introduction to the concept and background of explainable classifiers. Efforts towards more explainable classifiers are categorized into two: designing classifiers with simpler structures that can be easily understood; developing methods that generate explanations for already complicated classifiers. Based on the life circle of a classifier, we discuss the pioneering work of using visualization to improve its explainability at different stages in the life circle. The survey ends with a discussion about the challenges and future research opportunities of explainable classifiers.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Classification is the problem of identifying if an observation or object belongs to a set or not, or which of several sets. It is a fundamental problem in machine learning, data mining and computer vision. With the support of the increasing capacity of computation resources and growing volume of available data, the last decades have witnessed an explosion of breakthroughs in these fields. Nowadays, classification models (classifiers) are widely adopted to solve real world tasks, including face recognition [], handwritten recognition [], sentiment analysis [] and spam filtering []. Take image classification for instance, a well-designed convolutional neural network can achieve human-level performance in a number of benchmark datasets [,].

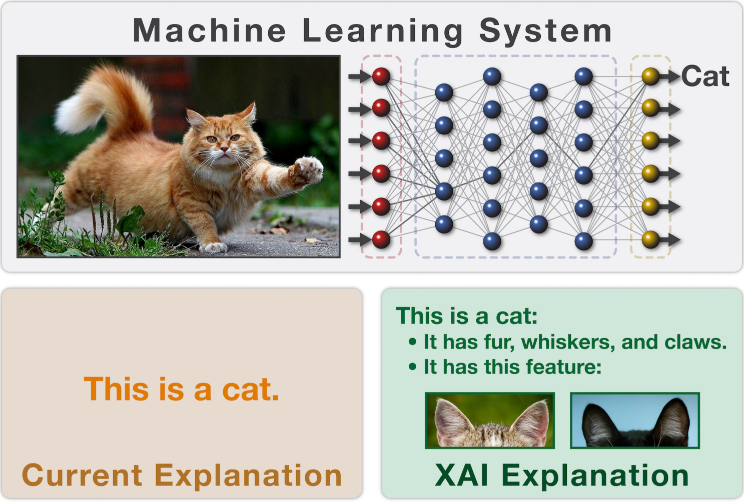

Figure 1.1. An illustration of an explainable image classifier [12].

Despite their promising capability, an often-overlooked aspect is the important role of humans [27]. When humans are to understand and collaborate with these autonomous systems, it is desirable if we have explanations of their outputs. For instance, a doctor using a machine classifier to assist identifying early signs of lung cancers would need to know why the classifier “thinks” there might be a cancer so that he/she can make a more confident diagnosis. A natural way is to provide explanations (Figure 1.1). In machine learning, the term *explainability* does not have a standard and generally accepted definition. In some literature, *interpretability* is used instead. Generally speaking, the explainability or interpretability of an intelligent system refers to the ability to explain its reasoning [8] to humans. For the sake of consistency, we use explainability as the ability to explain in this survey. Interpretability is used to refer the property of how easily a model can be understood by humans.

The research for explainable intelligent systems can be traced backed to the 1980s, when expert systems are created and proliferated [7, 23, 30]. These early works focused on reducing the difficulty of maintaining the complicated if-then rules by designing more explainable representations. A huge gap exists between todays state-of-the-art intelligent systems and the techniques that can make them explainable. The new challenges brought by the new generation of intelligent systems have attracted a growing research interest. DARPA launches the Explainable Artificial Intelligence (XAI) project [12], which aims to develop new techniques to make these systems explainable. Google Inc. initiated the People + AI Research Initiative (PAIR) [11] to advance the humanistic considerations in AI.

Visualization, is an effective and efficient technique for communicating information and understanding complex datasets for humans. The visual system is a proxy with a very large bandwidth to human brains [22]. Thus, visualization can be an ideal weapon for explaining complicated classifiers for humans. Early related research can be traced back to the software and algorithm visualization for computer science education in the 1980s and 1990s [6, 29, 25]. Visualization, especially interactive visualization, was proved to be very effective in facilitating people's understanding of complex softwares and algorithms. Few research has been done to visualize the increasingly complicated classifiers, which are actually algorithms learned from the data. It is not until recently that visualization was popularized as a media for understanding classification models, especially for image classifiers [28, 34, 2, 35]. However, these methods have limited applications to neural networks for image data. There is also a lack of a unified and convenient evaluation method for the generated visualizations.

1.2 Challenges

The need for visually explaining classifiers is actually a result of the successes and advances of AI. The major challenges of visually explaining classifiers results from the complexity of the model and data, and the limits of humans.

First, it is challenging to explain complex classification models both concisely and precisely. The best-performing classifiers (e.g., neural networks) are becoming increasingly complex, in terms of the number of parameters and operations they employed, which makes them difficult to be explained. A convolutional network typically employs thousands of neurons and millions of parameters. A random forest used for classification may employ hundreds of decision trees, each contains hundreds of nodes. Sampling a small number of parameters/neurons/nodes to explain might be easier to understand for humans, but it brings risks of misunderstanding as well. The variety of model architectures also increases the difficulty for a explanation framework for classifiers to be effective and general at the same time.

Another challenge is the volume and variety of the data used for training the classifiers. To explain a classifier, a most common strategy is to trace back to the input data. Which part of the input data contributes to the prediction? How the model behaves on this subset of data? Some explanation methods require computations over the whole training data, which may become impractical if the data set is very large. Different data types may require different forms of visual explanation. Image data are readily interpretable, but how to effectively explain classifiers on categorical, text and speech data is still a problem.

These challenges are, to some extent, due to the compromise with the limits of humans' cognition ability. If humans can make sense of the meaning of thousands of parameters and complex model structures by merely looking at the raw data or code, there is no needs struggling on how to better visualize them. There is already some studies discussing the structure, function and effectiveness of explanations in cognitive science. However, it is still unclear how we can effectively evaluate the quality of an explanation, and the load that its visual representations exert to humans.

1.3 Overview

This survey mainly focuses on how visualization techniques can be used to support explainable classifiers. In Chapter 2, we first introduce the definition of classification and classifiers, and the concept of explainable classifiers. Two major research directions towards more explainable classifiers are identified: designing classifiers that are readily interpretable, and methods that generate explanations for a classifier without modifying the model. In Chapter 3, we first articulate the life circle of a classifier into different stages, i.e., the recursive procedures of data collection and processing, model development and testing, and operations and maintenance. Then, we illustrate how visualization can be applied at different stages to provide explainability of classifiers. Based on the specified life cycle, we categorize the surveyed literature and discuss the challenges and opportunities for future research in visualization for explainable classifiers.

CHAPTER 2

EXPLAINABLE CLASSIFIERS

As discussed previously, for classification, explainability refers to the ability to explain the reasons of a classifier. To achieve explainability, existing work mainly falls into two categories. The first type of work develops more interpretable models that are easy to understand for humans. The second type of work generates explanations for a classifier without modifying the model, either by explaining the classifier locally on specific instances, or by explaining the behavior of the classifier globally.

2.1 Classification

To clarify the scope of this survey, we first briefly introduce the problem of classification, as an instance of supervised learning, and a few popular classifications models (classifiers).

2.1.1 Definition

Given an input space $\mathcal{X}$ and an output space $\mathcal{Y} = \{1, 2, ..., K\}$ with K classes, **classification** is the problem of identifying any **observation** $\mathbf{x} \in \mathcal{X}$ to a class $y \in \mathcal{Y}$. For multi-label classification, where class labels are not exclusive, we can view it as multiple related binary classification. For simplicity, we only consider the basic formulation in this survey.

A **classifier** is an algorithm f that implements classification, *i.e.*, $y = f(\mathbf{x})$. To handle ambiguity, a classifier is often used in a probabilistic setting. That is, the output of f a probabilistic distribution $p(y \mid \mathbf{x}, \mathcal{D})$ over all possible classes in $\mathcal{Y}$. $\mathcal{D}$ is the training set, which is a subset of $\mathcal{X} \times \mathcal{Y}$, that have already been observed. Thus, in practice, a classifier will often take the form of $\mathbf{y} = f(\mathbf{x})$, where $\mathbf{y} = (y_i) \in \mathbb{R}^K$ is a vector denoting the probabilistic distribution. Then the final classification will be the class i with largest probability $\arg\max_i y_i$.

Classification is now widely applied in solving many real world applications. A few examples are: face recognition [], handwritten recognition [], sentiment analysis [] and spam filtering.

2.1.2 Classifiers

Here we briefly present a few popular models for classifications, including k-nearest neighbors, support vector machines, decision trees and neural networks.

K**-nearest neighbor**.

Support vector machine.

Decision trees.

Neural networks. CNN, RNN.

Before we go into details the discussion on improving explainability of classifiers, we first present an overview of the categorization.

―――

A table goes Here

―――

2.2 Interpretable Classifiers

Interpretable classifiers are the classifiers that are commonly recognized to be more understandable than others, and hence, do not need extra explicit explanations. Summarizing existing work, we find two major strategies for creating interpretable classifiers: developing interpretable models with easy-to-understand structures, and learning simpler or sparser models.

2.2.1 Interpretable Models

To create more interpretable classifiers, a natural way is to use simple computation structures (*e.g.*, if-then rules) in classifier. Most models that falls into this category are rule-based.

A widely adopted type of models are the decision trees [5]. A decision tree classifier uses internal nodes and branches to represent its classification reasoning as conjunctions of rules. A human can trace back a specific classification from a leaf to the root to understand the prediction of the classifier. However, the difficulty of constructing a high-accuracy and interpretable decision tree has long been criticized.

Focused on balancing among performance, explainability and computation, a few recent studies introduce the Bayesian framework in rule-based classifiers. Letham *et al.*[17] develop

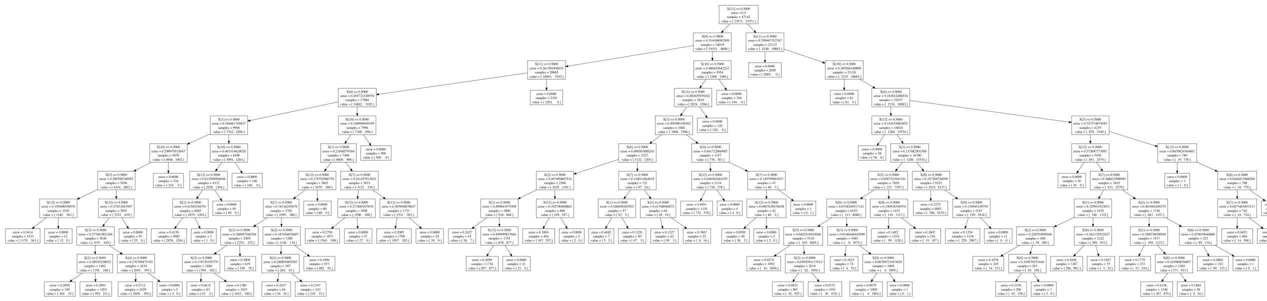

Figure 2.1. A decision tree with over one hundred nodes, which is hard to explain its reasoning.

the Bayesian Rule List which employs a prior structure that encourages sparsity in the generated decision lists with a good accuracy. Wang and Rudin [32] design the Falling Rule Lists that use an ordered if-then rule list so that the most at-risk occasion will be handled first. Wang *et al*.[33] construct rule sets based on AND and OR operations and highlight its low computation cost and on-par accuracy compared with SVM and random forest.

The most series problem of these interpretable models with easy-to-understand structures is the scalability. The performance of the rule-based models increases as the number of rules increases or the non-linearity increases. Although the rule-based models are easy to learn and understand at the first glance, it is intractable to understand the classifier as a whole when the number of nodes of rules grows up to a few hundreds. An example is shown in Figure 2.1.

Except for rule-based models, there are a few other models with more complicated models are recognized to be interpretable. One family of interpretable models worth noticing are the generalized linear models [4], which are pervasive in statistics and finance. Although these models can have highly nonlinear computations, the additive relation between nonlinear functions of features are believed to be easy-to-understand. However, the generalized linear classifiers can also be hard to understand when their non-linearity increased to a certain extent. The other non-probabilistic family of classifiers are the k-nearest neighbors (kNN) classifiers, whose prediction can be easily understood by presenting the observation's k-nearest neighbors. Numerous work has been done to boost the performance the kNN classifiers, including weighted kNN with different kernels [9] and fuzzy kNN [15]. The explainability of kNN classifiers may easily fail when there lacks near neighbors for certain observations.

2.2.2 Learning Sparser Models

As discussed above, the explainability often decreases as the complexity (*i.e.*, number of parameters or nodes) of the model increases. Thus, we can improve the explainability by learning a sparser model with the same architecture. These methods can also be regarded as model compressions, which reduce computation costs. In this category, researchers develop methods that can learn a sparser model with the similar performance, usually in a per-model manner.

For decision trees, Quinlan summarizes four techniques for model simplification[26], *i.e.*, cost-complexity pruning, reduced error pruning, pessimistic pruning and simplification to rule sets. Liu *et al.*[19] use a sparse decomposition method to zero out redundant parameters in a CNN, which achieve about 10-times speedup while lost accuracy for only 1%. Other simplification or methods include sparse linear integer models [31],

...

Although

In most cases, the efforts of developing more interpretable classifiers are tradeoffs between performance and explainability. However, for performance-critical applications, these methods cannot provide the required explainability.

2.3 Explanation of Classifiers

(4 pages)

2.3.1 Definitions of Explanation

A review of how previous work define explanations in machine learning.

General: ML:[8]. Cognitive:[20]

Data-type specific definitions: image-based, text-based, categorical data.

Some other work focuses on explaining the model itself (the mechanism, compositions of the model) to facilitate the understanding and diagnosing of the model.

To summarize, no agreed definition of explanation of classifier exists. Here we give a general definition...Next we give problem specific definitions.

Learned from surveying the literature, we separate the problem of explaining classifiers into two sub-tasks: explaining the predictions of classifiers (instance-level); explaining the classifier itself (model-level).

2.3.2 Local Explanation

Brief intro.

Model-aware methods

image data: [28, 34, 2, 35]

text data: [14, 18, 21, 1]

Model-unaware methods

- Learn to explain (Image captioning) [13]

- Model induction (Locally approximate) [27]

- Trace back to training data (influence function) [16]

2.3.3 Global Explanation

Brief intro.

Explain inner components of the model (model-aware)

RNN: [24]

CNN: [3] study interpretable units

NN: [10]

Explain the model as a whole (model-unaware)

[27] Image and text, sampling instanc-level explanations

[21] Text

2.4 Evaluation

Address the problem of evaluation, and how other work evaluates.

CHAPTER 3

VISUALIZATION FOR EXPLAINABLE CLASSIFIERS

(10-12 pages) Task driven.

3.1 Visualization for Model Development

3.1.1 Understanding

Scientific understanding. Investigate the characteristic of the model.

Existing work:

3.1.2 Diagnosing

Diagnose model and data.

Existing work:

3.1.3 Assessment and Selection

Unquantifiable assessments, Fairness (e.g., discrimination), Vulnerability

Existing work:

3.2 Visualization for Model Operation

3.2.1 Trust Establishment

3.2.2 Monitoring

3.3 Other Applications

3.3.1 Teaching and Communicating Models

Narrative, Interactive, etc. to explain your model to others.

3.3.2 Learn from the Model

Knowledge Discovery; Learn lessons from what the model learned (Alpha Go)

3.4 Evaluation

Review methods and standards of evaluating visualization.

Address the problem of the lack of evaluation standards for visualization for explaianble classifiers.

Proposed?

1. Fidelity. How visualization reflects the real model. (The relativeness and faithfulness of explanation)

2. Understandability. How easy the visualization is to be understood.

CHAPTER 4

CONCLUSION

Placeholder for Conclusion.

BIBLIOGRAPHY

[1] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 159–168. [Online]. Available: http://www.aclweb.org/anthology/W17-5221

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0130140

[3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Computer Vision and Pattern Recognition*, 2017.

[4] K. D. Bock, K. Coussement, and D. V. den Poel, "Ensemble classification based on generalized additive models," *Computational Statistics & Data Analysis*, vol. 54, no. 6, pp. 1535 – 1546, 2010.

[5] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[6] M. H. Brown, "Algorithm animation," Ph.D. dissertation, Providence, RI, USA, 1987, uMI Order No. GAX87-15461.

[7] W. Clancey, "The epistemology of a rule-based expert system: A framework for explanation," Stanford, CA, USA, Tech. Rep., 1981.

[8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[9] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 325–327, April 1976.

[10] R. Féraud and F. Clérot, "A methodology to explain neural network classification," *Neural Netw.*, vol. 15, no. 2, pp. 237–246, Mar. 2002.

[11] Google Inc. (2017) PAIR | people + ai research initiative. [Online]. Available: http://ai.google/pair

[12] D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*, 2017.

[13] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," *CoRR*, vol. abs/1603.08507, 2016. [Online]. Available: http://arxiv.org/abs/1603.08507

[14] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," in *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[15] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 4, pp. 580–585, July 1985.

[16] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 1885–1894. [Online]. Available: http://proceedings.mlr.press/v70/koh17a.html

[17] B. Letham, C. Rudin, T. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Stat.*, vol. 9, no. 3, pp. 1350–1371, 09 2015. [Online]. Available: https://doi.org/10.1214/15-AOAS848

[18] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and understanding neural models in nlp," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 681–691. [Online]. Available: http://www.aclweb.org/anthology/N16-1082

[19] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Penksy, "Sparse convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 806–814.

[20] T. Lombrozo, "The structure and function of explanations," *Trends in Cognitive Sciences*, vol. 10, no. 10, pp. 464 – 470, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661306002117

[21] D. Martens and F. Provost, "Explaining data-driven document classifications," *MIS Q.*, vol. 38, no. 1, pp. 73–100, Mar. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2600518.2600523

[22] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[23] R. Neches, W. Swartout, and J. Moore, "Enhanced maintenance and explanation of expert systems through explicit models of their development," *IEEE Transactions on Software Engineering*, vol. SE-11, no. 11, pp. 1337–1351, Nov 1985.

[24] U. H. M. of Recurrent Neural Networks, "Yao ming and shaozu cao and ruixiang zhang and zhen li and yuanzhe chen and yangqiu song and huamin qu." in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, 2017.

[25] B. Price, I. Small, and R. Baecker, "A taxonomy of software visualization," vol. ii. IEEE Publishing, 1992, pp. 597–606.

[26] J. R. Quinlan, "Simplifying decision trees," *International journal of man-machine studies*, vol. 27, no. 3, pp. 221–234, 1987.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *International Conference on Learning Representations (ICLR) Workshop*, 2014.

[29] J. T. Stasko, "Tango: a framework and system for algorithm animation," *Computer*, vol. 23, no. 9, pp. 27–39, Sept 1990.

[30] W. Swartout, C. Paris, and J. Moore, "Explanations in knowledge systems: design for explainable expert systems," *IEEE Expert*, vol. 6, no. 3, pp. 58–64, June 1991.

[31] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, pp. 349–391, Mar 2016. [Online]. Available: https://doi.org/10.1007/s10994-015-5528-6

[32] F. Wang and C. Rudin, "Falling Rule Lists," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Lebanon and S. V. N. Vishwanathan, Eds., vol. 38. San

Diego, California, USA: PMLR, 09–12 May 2015, pp. 1013–1022. [Online]. Available: http://proceedings.mlr.press/v38/wang15a.html

[33] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "A bayesian framework for learning rule sets for interpretable classification," *Journal of Machine Learning Research*, vol. 18, no. 70, pp. 1–37, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-003.html

[34] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

[35] L. Zintgraf, T. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *International Conference on Learning Representations (ICLR)*, 2017.