POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DIAGNOSING UNEXPECTED SPURIOUS CORRELATION

Julius Adebayo[†], **Michael Muelly**[♯], **Hal Abelson**[†],**& Been Kim**[♯]
{juliusad,hal}@mit.edu, mmuelly@stanford.edu beenkim@google.com
[†]Massachusetts Institute of Technology
[♯]Google Inc

ABSTRACT

Detecting whether an overparametrized deep network has learned a 'spurious association' from the training data is challenging. Post hoc model explanations are an increasingly promising avenue for addressing this challenge; however, it remains unclear whether they are effective.

We investigate whether three classes of post hoc explanations–feature attribution, concept ranking, and training point ranking–can detect *unknown* spurious correlation in a high-stakes medical task. Through control experiments, we assess the ability of these classes of explanations to reliably identify model reliance on spurious signals in the training set. We find that the post hoc explanation approaches tested are able to detect spurious associations *only* when the spurious signal is known *a priori*. Given that the space of all possible spurious signals that a model could rely on is large and often unknown, this finding suggests that these approaches may be ineffective for detecting spurious signals in practice.

1 INTRODUCTION

Increasingly, deep neural networks trained for use in high-stakes settings like medical images are susceptible to reliance on spurious training signals in the data that provide no meaningful information about the underlying data generating process (Geirhos et al., 2020; Sagawa et al., 2020; D'Amour et al., 2020). Models trained on medical images can rely on signals like scanner type to diagnose hip fracture (Badgeley et al., 2019). Consequently, the challenge of detecting when and what kind of spurious signal that a model has learned, ideally prior to model deployment, is important.

Post hoc explanations methods–approaches that provide ways to give insight into the associations that a model has learned–are a promising avenue for detecting a model's dependence on spurious training signals. In fact, their use for such purpose has been documented, and is often the *raison d'etre* for these approaches (Ribeiro et al., 2016; Lapuschkin et al., 2019; Rieger et al., 2020). Yet, widespread adoption of these approaches for detecting spurious signals lags. Consequently, we ask and address the following question: *are post hoc explanations effective for detecting a model's reliance on spurious training signals?*

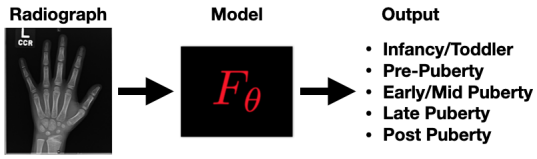

Figure 1: **Main Task.** We show the main task that we consider in this work, which is prediction bone age category from a Radiograph.

We focus on three classes of approaches: feature attributions, concept based methods, and an training point ranking via influence functions. We consider a high-stakes bone-age prediction task (see Figure 1). In our experimental setup, we manually create control settings where we induce models to learn spurious signals in the training set. Consequently, we can train models with and without reliance on such signals. We then compare post hoc explanations on models with defects to models trained to not depend on the spurious signal. We consider two kinds of spurious signal: 1) an easy one that induces dependence on a spatial location

in the image, and 2) a more difficult one that is not perceptible and is due to a low frequency noisy grid-like signal. We induce models to depend on these signals for specific classes in the dataset, and are then able to check whether the post hoc explanations approaches can detect these signals.

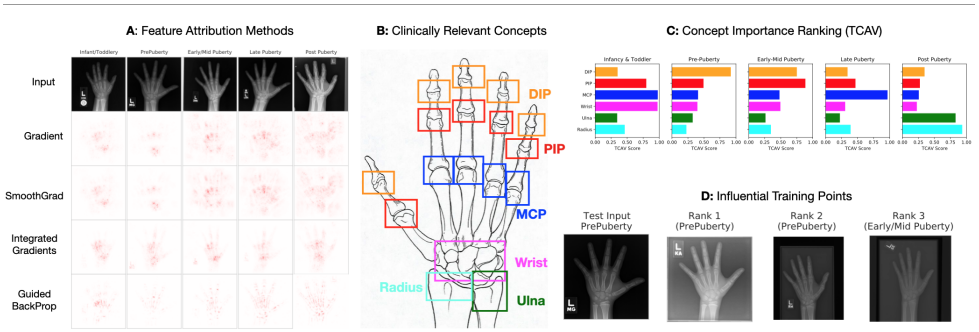

Figure 2: **Overview of the 3 post hoc explanation methods considered. A: Feature Attributions.** Here we show 5 different inputs, one for each bone age category, and the corresponding feature attribution maps for 4 different methods for a CNN model trained to predict bone age category. **B: Clinically Relevant Concepts for Bone Age Prediction.** Here we show a schematic of a hand along with the regions of the image that correspond to the clinically relevant variables for the concept importance ranking method that we consider. **C: Concept Importance Score** Here we show, for each clinically relevant concept, the concept importance score for a CNN model trained to predict bone age category from a radiograph. **D: Influential Training Points Ranking.** We show the 3 top ranked training points that most influence the test-loss for the Pre-Puberty input shown. This ranking is derived for a the CNN trained to predict bone age category.

Key Findings. We find that the 3 classes of post hoc explanations tested are only able to detect a spurious training signal when they are used to explicitly test for model dependence on these signals. However, the class of spurious signals that a model could conceivably depend on is large, so it is unlikely that one is able to exhaustively test dependence on all possible signals.

We make the following contributions:

1. We ask whether post hoc explanations are effective for detecting model reliance on spurious training signals? Through extensive control experiments and ground-truth explanations, we assess the ability of three classes of post hoc explanations–feature attribution, concept ranking, and training point ranking–to reliably identify model reliance on spurious signals in the training set.
2. We find that the 3 classes of post hoc explanations tested are only able to reliably diagnose the spurious training signal when they are used to explicitly test for model dependence on these signals. Consequently, these findings indicate that the post hoc explanation approaches tested may not be effective in practice for detecting that a model relies on an unknown spurious signals.

2 Insights into the 'Normal' Model.

We now discuss the insights that can be gleaned a normally trained model. Due to space limitations, we describe results for the feature attribution methods here and defer discussion on training point ranking and concept attribution to the appendix. Overall, we find that each of the 3 methods provides insights that suggest that the normal model is reliable. In the next section, we will contrast the insights discussed in this section with those gleaned from a model that has been spurious to manifest spurious signals.

Figure 3: **Spurious Hospital Tag on Training Image.**

Feature Attribution Methods. In Figure 2-A we show 4 feature attributions for 5 different inputs (one per each class) for a the normal model.

We observe that Gradient, SmoothGrad, and Integrated Gradients seem to highlight relevant parts of the input that a radiologist might pay attentions to. In the case of Guided Backprop, we observe that the feature attributions seem to approximate the input, which confirms insights from previous work (Nie et al., 2018).

3 Detecting Spurious Correlation.

In this section, we explain models that have been induced to rely on spurious training set signals. We show that unless one knows to test for these spurious signals explicitly, detecting that the model relies on a spurious signal for classification is challenging with the 3 explanation methods that we consider. We demonstrate this difficulty both for an 'easy' spatial spurious signal and a challenging, non-obvious signal as well.

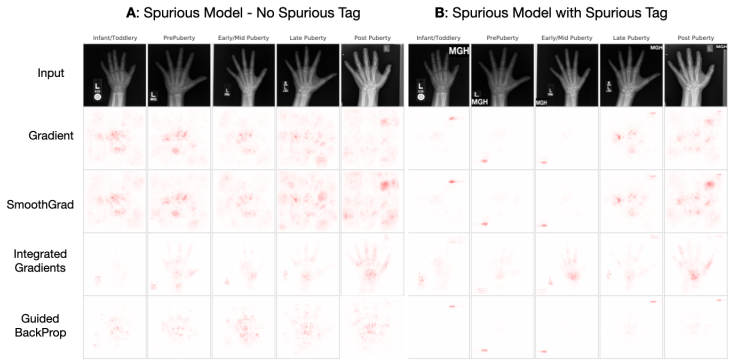

Figure 4: Feature Attribution Spurious Correlation Results. A: we show attributions for all methods on a spurious model, but these inputs do not have the hospital tag, which is the spurious signal. B: We show attributions for all methods on a spurious model, but these inputs have the hospital tag.

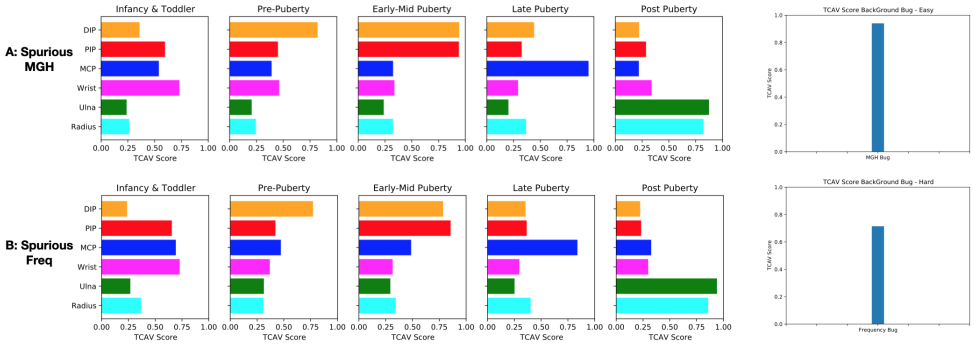

Figure 5: TCAV Spurious Correlation Results. In A, we show the TCAV rankings for a the easy spurious model setting, and B we show the TCAV rankings for the hard spurious setting. In column 6, we show the concept ranking where we have explicitly tested for the model's dependence on the spurious signal.

Experimental Setting. Spurious training signals are artifacts that encode and correlate with the input label in the training set but are not meaningfully useful for generalization. Here we consider two spurious signals as spurious signals. In the first case, which constitutes the easy settings, we 'paste' a hospital code (in this MGH for a hypothetical hospital) onto the images for all inputs for the class Pre-Puberty. See Figure 3 for an example. Here, the model trained on this task is than forced to use the spurious hospital code as a 'short-cut' for the Pre-Puberty class.

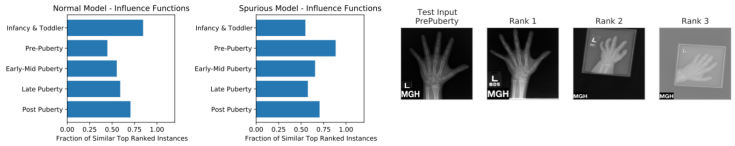

Figure 6: Influence Functions Spurious Correlation Results for the easy setting. In the first two columns, we show present the metric: for a given test-input, what is the fraction of the top 20 ranked training inputs that belong to the true class of the test-input–across all classes for the normal model and the spurious model. On the far right, we should an example for a test-input that includes the spurious hospital signal.

In the second spurious setting, we use a low frequency grid noise pattern that is not visible to the eye as the spurious training signal. We can selectively add this patterned noise to inputs in a specific class stochastically. We empirically verify that a model trained on this modified dataset learns to use the patterned noise as a short-cut. We now discuss whether the post hoc interpretations considered are able to signal that the model relies on this spurious signal.

Feature Attribution. We derive feature attributions for 5 different test inputs, one per class, for the 4 feature attribution methods tested for the easy (hospital tag) setting. In general, for both settings, we make similar observations.

In Figure 4 we show attributions for all methods on a spurious model, but for inputs that do not have the hospital tag, which is the spurious signal. We observe that the attributions obtained for this setting is qualitatively and quantitatively different from those obtained from a the normal model except in the case of Guided Backpropagation. We observe dispersed attribution across the input, but no clear signal exists to indicate that the model relies on a short-cut. If one only examines the validation and test-set metrics, we actually find that the spurious model performs within 3 percent of the test accuracy of the normal model. Yet observing Figure 4-A, one is left with an inconclusive diagnosis.

However, we compute feature attributions again, but for test images where we have artificially inserted the spurious hospital tag. Immediately, as shown in Figure 4-B, we observe that the attributions focus on the tag. To confirm the reliability of these methods, we quantitatively compare these attributions to ground-truth equivalents and observe similarity (SSIM) greater than 0.78 across all methods. The situation above demonstrates that unless one explicitly tests for a specific spurious correlation, it is unlikely to be identified by naive feature attribution inspection.

4 Conclusion

DNNs trained on image data are increasingly reliant on spurious training signals. This challenge is a non-starter for deployment of these models in high-stakes settings like medical images and records data. Ideally, we seek approaches that would allow us to detect a model reliance on a spurious training signal prior to test-time. Post hoc explanations methods are a promising direction for addressing this challenge. However, the state of the literature is currently unclear as to the effectiveness of these methods in practice, particularly in the hands of a domain expert. In this work, we have address this challenge. We investigated whether 3 classes of post hoc explanations are effective for detecting a model reliance on spurious training signals. We find that the 3 classes of post hoc explanations tested are only able to reliably diagnose the spurious training signal when they are used to explicitly test for model dependence on these signals. Consequently, our findings suggest that using the post hoc explanation approaches tested may not be effective in practice for identifying unexpected spurious signals.

Limitations. We have presented results on a single bone age dataset with three classes of post hoc explanations. Future work will need to demonstrate similar findings on additional datasets to further verify the results presented. This work is a small step towards making explicit the necessary requirements for current post hoc explanation tools to allow a user or practitioner identify whether a model is reliant on a spurious signal.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9525–9536, 2018.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 275–285, 2020.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2020.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):1–10, 2019.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *medRxiv*, 2020.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, pp. 13567–13578, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Learning explainable models using attribution priors. *arXiv preprint arXiv:1906.10670*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 3681–3688. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://doi.org/10.1609/aaai.v33i01.33013681.

Vicente Gilsanz and Osman Ratib. *Hand bone age: a digital atlas of skeletal maturity*. Springer Science & Business Media, 2005.

Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. *arXiv preprint arXiv:2012.15781*, 2020.

Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*, 2020.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, pp. 2921–2932, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 9737–9748, 2019.

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. *arXiv preprint arXiv:2012.04104*, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2673–2682, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

Himabindu Lakkaraju and Osbert Bastani. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *European Conference on Computer Vision*, pp. 120–135. Springer, 2016.

Qingjie Meng, Christian Baumgartner, Matthew Sinclair, James Housden, Martin Rajchl, Alberto Gomez, Benjamin Hou, Nicolas Toussaint, Jeremy Tan, Jacqueline Matthew, et al. Automatic shadow detection in 2d ultrasound. 2018.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *ICML*, 2018.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2018.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracking gradient descent. *arXiv preprint arXiv:2002.08484*, 2020.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pp. 1135–1144. ACM, 2016.

Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *International Conference on Machine Learning*, 2020.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In Carles Sierra (ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 2662–2670. ijcai.org, 2017. doi: 10.24963/ijcai.2017/371. URL https://doi.org/10.24963/ijcai.2017/371.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pp. 168–172, 2020.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6034.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why modified bp attribution fails. *arXiv preprint arXiv:1912.09818*, 2019.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Suraj Srinivas and Francois Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dYeAHXnpWJ4.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun D. Preece. Sanity checks for saliency metrics. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 6021–6029. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6064.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in neural information processing systems*, pp. 9291–9301, 2018.

APPENDIX A RELATED WORK.

The recent work of Adebayo et al. (2020) presents debugging tests for assessing feature attribution methods. The spurious correlation setting that we consider here fits under their framework. However, they only consider feature attribution methods. Here we extend this analysis to concept and training point ranking methods. Critically, Adebayo et al. (2020) show that feature attributions are able to identify spurious training signals. We make a similar finding in this work; however, we further demonstrate this finding for concept and training point ranking methods. Our work takes important departures from theirs: 1) we explain the source of this phenomenon, and 2) we demonstrate that naive application of these methods might be unable to detect spurious correlation in practice. Adebayo et al. (2020) assume the spurious correlation training bug is known, a priori; however, here we demonstrate that the more challenging task is identifying the spurious signal in the first place.

More recently, Han et al. (2020) demonstrate that training point ranking via influence functions is able to identify the dependence of an NLP model on dataset artifacts. In addition, they show correspondence between the insights observed with the input-gradient feature attribution and the training point ranking. Along similar lines, Guo et al. (2020) present fast approximations for computing the training point ranking for a test point. In addition, they show how to identify and correct model errors in a natural language task. Similar to the distinctions that we note with the work by Adebayo et al. (2020) above, here, they also assume that the spurious signal being identified is known a priori.

Post hoc explanations, more generally, have been shown to be able to identify a model's reliance on spurious training signals (Ribeiro et al., 2016; Meng et al., 2018; Lapuschkin et al., 2019; DeGrave et al., 2020; Ross et al., 2017). Recent work by Rieger et al. (2020) showed that regularizing model attributions during training can help lead to models that avoid spurious correlation and enable improved debugging by experts. Similarly, Erion et al. (2019) show that regularizing the expected gradient attribution during training confers similar benefits. Koh & Liang (2017) used influence functions to identify domain shift. Kim et al. (2018) also perform a user study to understand if attribution methods can be used for catch spurious correlation.

However, similar methods have also been shown to struggle in the hands of end-users for diagnosing model errors (Alqaraawi et al., 2020; Adebayo et al., 2020). This contradiction reflects the challenge that we explore in this work. Often, post hoc explanations have been shown to be effective for identifying spurious signals that were suspected or known a priori; however, these methods seem to struggle when confronted with the task of identifying an unexpected spurious signal.

Increasingly, insights into why overparametrized DNNs rely on spurious training set signals is starting to be theoretically and empirically analyzed (Sagawa et al., 2019; 2020; Khani & Liang, 2020; Nagarajan et al., 2020), yet it is still unclear how to reliably detect that a model is relying on such signals prior to model deployment.

Assessing whether a post hoc explanation approach is faithful to the underlying model being explained has been addressed in recent works, yet this challenge remains elusive (Hooker et al., 2019; Tomsett et al., 2020). Generally, the class of approaches that modify backpropagation with positive aggregation have been shown to be invariant to the higher layer parameters of a DNN (Mahendran & Vedaldi, 2016; Nie et al., 2018; Adebayo et al., 2018; Sixt et al., 2019). In an intriguing demonstration, Srinivas & Fleuret (2021) show that the input-gradient, a key feature attribution primitive, might not capture discriminative signals about input sensitivity. Instead they show that input-gradient likely capture the ability of the model to be able to generate class-conditional inputs.

User studies are typically the classic approach for evaluating the effective of an explanation (Doshi-Velez & Kim, 2017). Poursabzi-Sangdeh et al. (2018) manipulate the features of a predictive model trained to predict housing prices to assess how well end-users can identify model mistakes. Their results indicate that users found it challenging to debug these linear models with the model coefficients. Recent work by Chu et al. (2020) and Shen & Huang (2020) has shown similar results in the DNN setting as well. Alqaraawi et al. (2020) find that the LRP explanation method improves participant understanding of model behavior for an image classification task, but provides limited utility to end-users when predicting the model's output on new inputs.

Post hoc explanations have been shown to be fragile and very easily manipulated Ghorbani et al. (2019); Heo et al. (2019); Dombrowski et al. (2019); Anders et al. (2020); Slack et al. (2020);

Lakkaraju & Bastani (2020). Our work tackles a difference concern: whether they are suitable for detecting unexpected spurious training set signals.

APPENDIX B DETAILED OVERVIEW OF EXPLANATION METHODS

First, for all code used in this work we include a link to an anonymous repo in the readme.txt file. The main repo includes an demo.ipynb with additional images.

In this section, we provide additional implementation details for the explanation methods that we consider in this work. To start with the model setup: let's say we are given input-output pairs, $\{x_i, y_i\}_i^n$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$; and a classifier's goal is to learn a function, $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ that generalizes to new inputs via empirical risk minimization (ERM). In this work, we assume that f_θ is an over-parametrized deep neural network (DNN) trained on image data for classification (C classes).

Feature Attributions. An attribution functional, $E : \mathcal{F} \times \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d$, maps the input, $x_i \in \mathbb{R}^d$, the model, f_θ, output, $f_k(x)$, to an attribution map, $M_{x_i} \in \mathbb{R}^d$. The class of feature attribution methods is large, so in this work we pick: Input Gradient, SmoothGrad, Integrated Gradients, and Guided Backprop. We choose these approaches since they were the top-ranked methods tested under the spurious correlation setting of Adebayo et al. (2020).

1. **The *Input-Gradient (Gradient)*** Simonyan et al. (2014); Baehrens et al. (2010) map, $|\nabla_{x_i} F_i(x_i)|$, is a key primitive upon which several other methods are based.
2. ***SmoothGrad* Smilkov et al. (2017)** corresponds to the average of noisy input gradients: $M_{\text{sg}}(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_{x_i} F_i(x_i + n_i)$ where n_i is sampled according to a random Gaussian noise. We considered 50 noisy inputs, selected the standard deviation of the noise to be $0.15 *$ input range. Here input range refers to the difference between the maximum and minimum value in the input.
3. ***Integrated Gradients* (Sundararajan et al. (2017))** sums input gradients along an interpolation path from the "baseline input", $\bar{x}$, to x_i: $M_{\text{IntGrad}}(x_i) = (x_i - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x}+\alpha(x_i-\bar{x}))}{\partial x_i} d\alpha$. For integrated gradients we set the baseline input to be a vector containing the minimum possible values across all input dimensions. This often corresponds an all-black image. The choice of a baseline for IntGrad is not without controversy; however, we follow this setup since it is one of the more widely used baselines for image data.
4. ***Guided Backpropagation (GBP)* *Springenberg et al. (2014)*** modifies the backpropagation process at ReLU units in DNNs. Let, $a = \max(0, b)$, then for a backward pass, $\frac{\partial l}{\partial s} = 1_{s>0} \frac{\partial l}{\partial b}$, where l is a function of s. For GBP, $\frac{\partial l}{\partial s} = 1_{s>0} 1_{\frac{\partial l}{\partial s}>0} \frac{\partial l}{\partial b}$.

Feature Attributions: Implementation. We implement all of these methods from scratch in the PyTorch framework and also compare our implementations to the output of the Captum framework in PyTorch. We were able to confirm a correspondence between these outputs.

Concept-Based Approaches. We now discuss additional implementation details of our concept based approach. We select the TCAV approach to quantify the sensitivity of a DNN model's class score to user provided inputs represent a particular class. Given hidden representations, h_l, from a particular layer of a DNN for for images belonging to concept class C. We can derive the sensitivity score as: $\nabla h_{l,k}(f_l(x)).\theta_c^l$. The previous expression indicates the sensitivity of the class score (logit) for class k to inputs indicating concept, C, given hidden representations from layer l from the DNN f. The concept vector, θ_c^l, typically corresponds to a the weights of a linear classifier trained to separate the images for a particular concept class from or images.

For completeness, we show in Figure 7 an overview of the clinical concepts that we consider in this work. These are the representative clinical attributes that a radiologist would inspect to ascertain the bone age of a particular input. These concepts are: DIP, PIP, MCP, Radius, Ulna, and Wrist.

Concept Implementation. To compute the TCAV score for each concept, we collect representations from all hidden 'layers' of the model and train linear models to obtains the concept vector for the corresponding attribute. We then compute the class sensitivity score for each concept attribute.

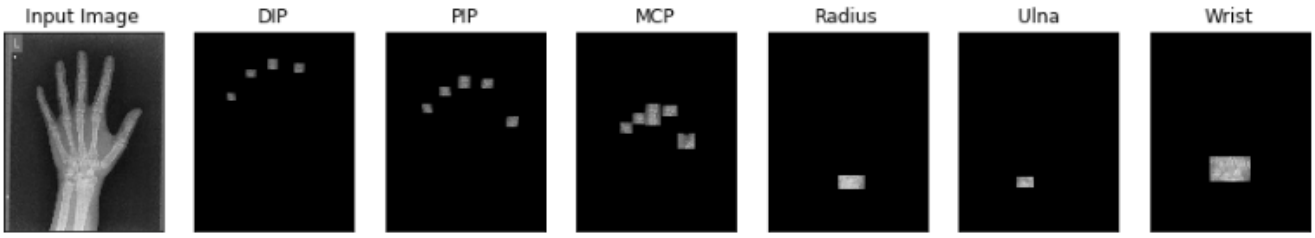

Figure 7: **Concept Partition for an Example Input.** Here we show how we partition a single instance into its constituent clinical concept components.

We train the linear model 100 times and perform statistical significance testing in order to mitigate the case where a spurious concept is selected. For each concept class, we use 325 images that part of the training, validation, or test sets. These new set of images were annotated by a board certified radiology with the clinical bone age regions (MCP, PIP, DIP etc) that we chose.

Influence Functions for Training Point Ranking. The final kind of interpretation that we consider is training point ranking via influence functions. In the case of training point ranking via influence functions, we rank the training samples, in terms of 'influence', on the loss of a test example. Specifically, if we up-weighted a training point and retrained the model, then by how much would the loss on a given test example change? Koh & Liang (2017) analytically derive the analytically formulas for computing this quantity. Given a test point, x_t, the influence of a training point, x_i, on the test loss is: $I(x_t, x_i) = -\nabla_\theta \ell(x_t, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta \ell(x_i, \hat{\theta})$, where H is the empirical Hessian of the loss.

Estimating the influence requires computing hessian-vector products, so it can be difficult to scale to model with large number of parameters, and recent work has shown that influence estimate for test points for deep networks can be inaccurate due to non-convexity (Basu et al., 2020). Consequently, we estimate influence on a linear model student network trained to mimic the original DNN. We empirically verify that the predictions of the student network seem to mimic the original DNN.

Implementation Details. There are two other training point ranking methods that we also consider in this work (Yeh et al., 2018; Pruthi et al., 2020). For 150 inputs in the test set, we compare the Spearman rank correlation of the training point due to Influence Functions to these other two approaches. We obtain mean values of 0.88, and 0.76 respectively, which suggests high similarity amongst these approaches. Ultimately, we chose to present the main results in the draft for the training point ranking due to influence functions approach.

We trained a student multi-class logistic regression model to mimic the original model for each model we want to compute influence for. Here for all training points, we collect embeddings across all layers and pass these embedding through a random projection to obtain a 1000-dimensional approximation. We then train linear models to mimic the original deep network using these features. The correlation between the output of the student models and the original teacher models was found to be 0.87.

Task and Dataset Description. We consider the high stakes task of predicting the bone age category from a radiograph to one of five classes based on age: Infancy/Toddler, Pre-Puberty, Early/MiD Puberty, Late Puberty, and Post Puberty. This task is one that is routinely performed by radiologists and as been previously studied with a variety of DNN. The dataset we use is derived from the Pediatric Bone Age Machine learning challenge conducted by the radiological society of North America in 2017 Halabi et al. (2019). The dataset consists of 12282 training, 1425 validation, and 200 test samples. We resize all images to (299 by 299) grayscale images for model training.

Models. We keep fixed the model architecture that we use throughout this work. This architecture consists of 6 layers, 5 of which are the traditional conv-relu-batchnorm-maxpool combination. The last layer is a fully-connected layer for a 5-way bone age class classification. This architecture is inspired by the CBR-Tiny architecture of Raghu et al. (2019). We train this model with the ADAM

optimizer for 40 epochs with an exponentially decaying learning rate schedule. The initial learning rate for this task was : $3e-4$, which was found via hyper-parameter optimization.

Explanation Comparison. We measure visual of feature attributions using the structural similarity index (SSIM) and feature ranking similarity using the Spearman rank correlation metrics, respectively.

Visualization Attributions and Normalization. Here and in the main text we show attributions in a single color scheme: either Gray Scale or a White-Red scheme. We do this to prevent visual clutter. For all the metrics we compute, we normalize attributions to lie between [0, 1] for SSIM and [-1, 1] for attributions that return negative relevance.

APPENDIX C ADDITIONAL METRICS

In lieu of cherry picked figures, we now present additional metrics derived for the experiments reported in the paper for the **entire test set.** We compute the mean and standard errors where necessary.

Spurious Correlation Feature Attribution To quantitatively measure whether attribution methods reflect the spurious background, we compare attributions to two ground truth masks (GT-1 & GT-2). We consider an ideal mask that apportions all relevance to the background and none to the object part. Next, we consider a relaxed version that weights the first ground truth mask by the attribution of a spurious background without the object. In Table 3, we report SSIM comparison scores across all methods for both ground-truth masks. For *GT-2*, scores range from a minimum of 0.78 to maximum of 0.97; providing evidence that the attributions identify the spurious background signal. We find similar evidence for *GT-1*.

1lightgray

Metric	Gradient	SmoothGrad	Integrated Gradient	GBP
SSIM-GT1	0.78	0.81	0.82	0.79
SSIM-GT1 (SEM)	0.012	0.013	0.077	0.089
SSIM-GT2	0.83	0.83	0.89	0.97
SSIM-GT2 (SEM)	0.013	0.013	0.02	0.0024

Table 1: **Similarity between attribution masks for inputs with spurious background and ground truth masks.** SSIM-GT1 measures the visual similarity between an ideal spurious input mask and the GT-1. SSIM-GT2 measures visual similarity for the GT-2. To calibrate this metric, the mean SSIM between a randomly sampled Gaussian attribution and the spurious attributions which is: $6e^{-05}$.

Normal Model compared to radiologist Concept Ranking For each of the 100 re-runs, we computed the rank correlation between the concept ranking for the normal model and rankings provided by a board certified radiologist.

1lightgray

Setting	Spearman Correlation w/Rad
Infancy/Toddler	0.85
Pre-Puberty	0.89
Pre-Puberty	0.94
Late Puberty	0.86
Post Puberty	0.91

Table 2: **Normal Model compared to radiologist Concept Ranking**

Spurious Correlation Concept Ranking For each of the 100 re-runs, we computed the rank correlation between the concept ranking for the normal model and the spurious-hard and the spurious-easy settings.

1lightgray

Setting	Spearman Correlation w/Normal
Spurious Easy	0.88
Spurious Hard	0.76

Table 3: **Similarity between concept rankings for model that exhibit easy spurious and hard spurious bugs compared to a normally trained model.**

APPENDIX D ADDITIONAL INSIGHTS FOR NORMAL MODEL

Concept Ranking (TCAV). We show the TCAV score for each concept, towards each bone age category (class) for a normal model, Figure 2-C.Figure 2-B provide a 'legend' for what each concept indicates. We also sought a ground-truth concept ranking for each bone age category from both the clinically recommended ATLAS (Gilsanz & Ratib, 2005), which were verified by a board certified radiologist. We then computed the similarity, using the spearman rank correlation, of the TCAV ranking derived for the normal model for each class with the ground-truth rankings. These rankings were always greater than 0.85 across all 5 bone age categories, which suggests that the normal model is relying on the expected clinical concepts.

Influence Functions (IF). In using the training point ranking for model diagnosis, we compute an influence ranking for different test inputs. Ideally, the most influential training inputs should capture semantically similar features to the test input as well. We measure this by the fraction of the top 20 ranked training inputs that belong to the true class of the test-input. Figure 2-D shows an example for a PrePuberty sample. Here we observe that the top ranked inputs for this test instance also belong to the same class, which confirms appropriate model behavior.

APPENDIX E ADDITIONAL DETAILS FOR SPURIOUS MODELS

Concept Ranking (TCAV). In Figure 5-A, we show the TCAV rankings for a the easy spurious setting, and Figure 5-B shows rankings for the hard spurious setting. On the far right of both figures (column) 6, we show the concept ranking where we have explicitly tested for the model's dependence on the spurious signal in both settings.

An immediate observation from Figure 5-A and Figure 5-B is that these rankings do not deviate from the normal rankings. To measure similarity, we compare the rank correlation across all bone age categories between each spurious setting and the normal model. In both cases, we observe relatively reasonable ranking similarity (0.67 and 0.74) respectively. In addition, we asked a board certified radiologist to inspect these rankings in order to diagnose any inconsistencies across the rankings. Here we find that it is difficult to distinguish these spurious model explanations from the the same concept rankings but for the normal model.

However, somewhat surprisingly, if we test explicitly for dependence on the spurious hospital tag and the spurious low frequency pattern, we immediately observe a high concept importance for both signals on the respective models. Again, this finding indicates that if we know the spurious signal ahead of time and explicitly test for it, then we might be able to detect the model's reliance on it. However, naive reliance on clinically relevant variables does not show explicit dependence on spurious signals.

Influence Functions (IF). We now consider the ability of influence ranking to help reveal dependence on spurious signals. Figure 6 shows results from analysis of the easy spurious setting. We defer the results of the hard spurious setting to the appendix. Here we recall the metric for assessing

a model using the influence function ranking: for a given test-input, what is the fraction of the top 20 ranked training inputs that belong to the true class of the test-input. In Figure 6, we show this metric, across all classes, for both the normal model and the spurious model.

Somewhat surprisingly, we find that the metric improves for the spurious model compared to the normal model. This indicates that in the spurious setting, the training inputs that influence the test loss typically also belong to similar classes for the test inputs. However, when we explicitly compute the training point influence ranking for a test-input where an obvious spurious signal, we find that the top ranked training points are all points that also exhibit this signal. One the one hand, this is certainly good news for the ability to use the influence ranking as a debugging signal; however, this demonstration suggests that to identify such signals one has to carefully craft inputs that exhibit the bug one would like to test.

The overriding message of our results in the section indicate that the 3 classes of post hoc explanations tested are only able to reliably diagnose the spurious training signal when they are used to explicitly test for model dependence on these signals. However, the class of spurious signals that a model could conceivably depend on is large, so it is unlikely that one is able to exhaustively test dependence on all possible signals.