The Manifold Hypothesis for Gradient-Based Explanations

Sebastian Bordt[1] Uddeshya Upadhya[1] Zeynep Akata[1,2] Ulrike von Luxburg[1,2]
[1]University of Tübingen, Germany [2]Tübingen AI Center

Abstract

When do gradient-based explanation algorithms provide perceptually-aligned explanations? We propose a criterion: the feature attributions need to be aligned with the tangent space of the data manifold. To provide evidence for this hypothesis, we introduce a framework based on variational autoencoders that allows to estimate and generate image manifolds. Through experiments across a range of different datasets – MNIST, EMNIST, CIFAR10, X-ray pneumonia and Diabetic Retinopathy detection – we demonstrate that the more a feature attribution is aligned with the tangent space of the data, the more perceptually-aligned it tends to be. We then show that the attributions provided by popular post-hoc methods such as Integrated Gradients and SmoothGrad are more strongly aligned with the data manifold than the raw gradient. Adversarial training also improves the alignment of model gradients with the data manifold. As a consequence, we suggest that explanation algorithms should actively strive to align their explanations with the data manifold. An extended version of this paper is available at https://arxiv.org/abs/2206.07387.

1. Introduction

Post-hoc explanation algorithms for image classification often rely on the gradient with respect to the input [4,38,41]. In many cases, however, model gradients and post-hoc explanations [6, 9, 28, 36, 37] posses little visual structure that can be interpreted by humans [21]. This makes image classification with neural networks one of the most challenging applications of explainable machine learning.

Recently, a number of different papers have observed conditions that lead to *perceptually aligned gradients* (PAGs) [16]. In particular, it has been shown that adversarial training, as well as other forms of robust training, lead to PAGs [21, 23, 34, 42]. However, it remains unclear what exactly makes a feature attribution perceptually-aligned.

In this work, we try to understand when a feature attribution is perceptually-aligned. We propose and investigate the following hypothesis:

Hypothesis: Feature attributions are more perceptually-aligned the more they are aligned with the tangent space of the image manifold.

To understand the intuition behind the hypothesis, note that it is widely believed that natural image data concentrates around a low-dimensional image manifold [17]. This image manifold captures the geometric structure of the data. In particular, the tangent space of an image captures all components of the image that can be slightly changed while still staying within the realm of natural images. If an attribution approximately lies in this tangent space, this means that it highlights visually meaningful components of the image that contribute to the prediction. If a an attribution lies orthogonal to the tangent space, this means that it points in some direction that would not lead to realistic images, and a human would have a hard time understanding its meaning. Random noise, in particular, lies with high probability orthogonal to the image manifold.

To provide empirical evidence for the hypothesis, we employ autoencoders to estimate the image manifolds of five different datasets: MNIST, EMNIST, CIFAR10, X-ray pneumonia and diabetic retinopathy detection. By projecting different feature attributions into the tangent space, we then provide qualitative evidence that tangent-space components are perceptually-aligned, whereas orthogonal components visually resemble random noise (Sec. 4.1). As depicted in Figure 1, we also use variational autoencoders as generative models. This allows us to generate image datasets with a completely known manifold structure.

We then show that popular post-hoc methods such as SmoothGrad, Integrated Gradients and Input $\times$ Gradient improve the alignment of attributions with the data manifold (Sec 4.2). The same is true for l_2-adversarial training, which significantly aligns model gradients with the data manifold (Sec. 4.3). These results holds consistently across all the different datasets.

Apart from the intuitive and empirical plausibility of the hypothesis, its main appeal is that it provides a clear perspective on why explaining image classifiers is difficult: While our empirical investigations show that post-hoc methods and adversarial training improve the alignment of attributions with the data manifold, in many cases there re-

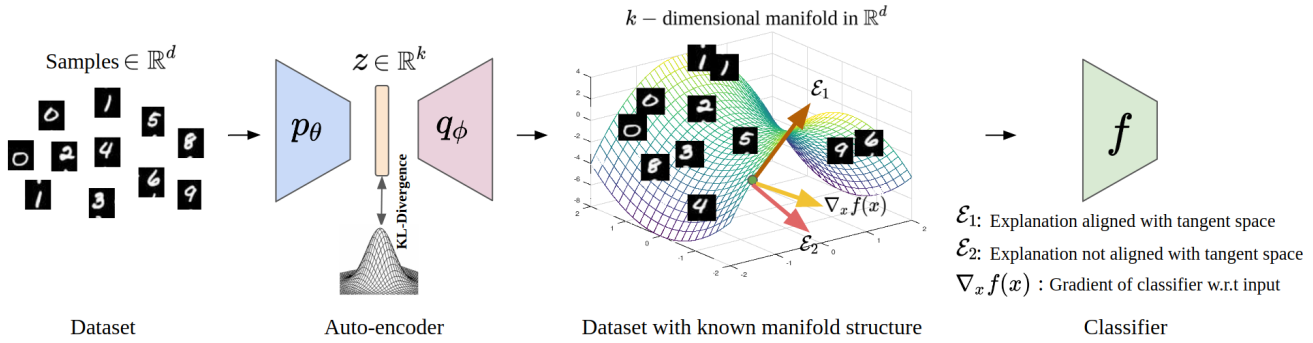

Figure 1. Conceptual overview of our approach. We first estimate the data manifold of an existing dataset with a variational autoencoder, then use the decoder as a generative model. On the generated data, we train a classifier f. For this classifier, we evaluate whether different gradient-based explanations $\mathcal{E}_i$ align with the tangent space of the data manifold. Moving along an explanation aligned with the tangent space keeps us in the manifold, whereas moving along an orthogonal explanation takes us out of the manifold. Our hypothesis is that the latter does not lead to perceptually-aligned explanations because it describes changes that lead to unnatural images.

mains much room for improvement. Overall, the manifold hypothesis is an important step toward understanding when feature attributions are explanations.

2. Related Work

Projections on the data manifold. Many different papers employ techniques where data points or model gradients are being projected on the data manifold [14, 40]. In explainable machine learning, it has been shown that explanations can be manipulated by modifying the model outside of the image manifold, and that one can defend against such attacks by projecting the explanations back on the manifold [13]. The hypothesis that natural image data concentrates around a low-dimensional image manifold is supported by a number of empirical studies [1, 15, 45].

Evaluating explanations. The unavailability of ground-truth explanations and the fact that explanations may be susceptible to adversarial attacks [13, 18] makes it difficult to evaluate them [20, 32–34]. A literature on *sanity checks* has shown that these principal difficulties non-withstanding, many explanations fail even the most basic tests such as model parameter randomization [2, 3, 8, 24].

Alignment of the implicit density model with the ground truth class-conditional density model. Srinivas et al. [39] have proposed that gradient-based explanations are more interpretable the more the density model that is implicit in the classifier f is aligned with the ground truth class-conditional density model. This criterion can be shown to be compatible with the manifold hypothesis, given assumptions on how the data centers around the manifold.

3. Overview of our approach

In order to evaluate our hypothesis, we we need to measure the alignment of an attribution $E \in \mathbb{R}^d$ at a point $x \in \mathbb{R}^d$ with the tangent space of the data manifold at x.

3.1. Background

Data manifolds and tangent spaces. A k-dimensional differentiable manifold $\mathcal{M} \subset \mathbb{R}^d$ is a subset of a d-dimensional space that locally resembles $\mathbb{R}^k$. At every point $x \in \mathcal{M}$, the tangent space $\mathcal{T}_x$ is a k-dimensional subspace of $\mathbb{R}^d$. The tangent space $\mathcal{T}_x$ consists of all directions v such that $x + v$, for $\|v\|$ small, is close to the manifold.

Model gradients and explanation algorithms. We consider DNNs that learn functions $f : \mathbb{R}^d \to \mathbb{R}^C$. Here C is the number of classes and the model prediction is given by $\text{argmax}_i f(x)_i$. The gradient of class i at point x with respect to the input is given by $\text{grad}_i(x) = \frac{\partial (f(x)_i)}{\partial x}$. Post-hoc explanation algorithms [6, 37, 38, 41] provide explanations as vectors in $E \in \mathbb{R}^d$.

3.2. How do we know the data manifold?

In the *generative approach*, we first train a variational autoencoder [19, 25] on some existing datasets. After training, we pass the entire dataset through the autoencoder. Then we train an auxiliary classifier to reproduce the original labels from latent codes and reconstructed images. Equipped with this labeling function, we sample from the prior and use decoder and labeling function to generate a new dataset with *completely known manifold structure*: the tangent space at each datapoint can be computed from the decoder via backpropagation [7, 35]. The main limitation of the generative approach is that we might not be able to obtain high-quality samples with reasonably small latent spaces.[1] To evaluate our hypothesis on real-world high-dimensional image data where it is difficult to obtain realistic samples with not-too-large latent spaces, we must rely on *estimating* the tangent space. In this *reconstructive approach*, we pass the original dataset through an autoencoder and take the reconstructed images with the original labels as our new dataset.

3.3. Measuring alignment with the data manifold

To measure how well an explanation $E \in \mathbb{R}^n$ is aligned with $\mathcal{T}_x$, we first project it into the tangent space – denoted by $\texttt{proj}_{\mathcal{T}_x} E$ – and then compute the fraction of the attri-

[1] While there have been great advances in generative modeling, state-of-the-art models like hierarchical variational autoencoders [43] require large latent spaces, i.e., $k \approx d$. For our analysis, $\sqrt{k/d}$ must be small – with $k = d$, the fraction of even a random vector in tangent space is always 1.

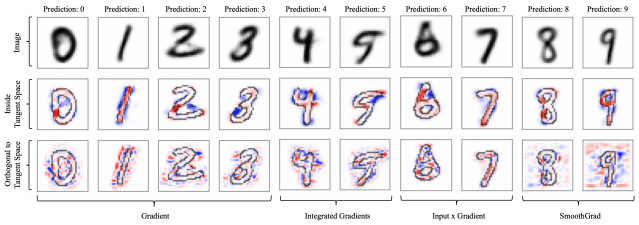

Figure 2. The part of an attribution that lies in the tangent space is perceptually-aligned, whereas the part that is orthogonal to the tangent space is not. (First row) Images from the test set of MNIST32. (Second row) The part of the attribution that lies in tangent space. (Third row) The part of attribution that is orthogonal to the tangent space. Red corresponds to positive, and blue to negative attribution (best viewed in digital format).

bution in tangent space

$$\text{Fraction of } E \text{ in } \mathcal{T}_x = \frac{\|\texttt{proj}_{\mathcal{T}_x} E\|_2}{\|E\|_2} \in [0, 1]. \qquad (1)$$

If the attribution completely lies in tangent space, we have $\texttt{proj}_{\mathcal{T}_x} E = E$. If the attribution is completely orthogonal to the tangent space, we have $\texttt{proj}_{\mathcal{T}_x} E = 0$. When we quantitatively evaluate (1), we account for the fact that even a random vector has a non-zero fraction in tangent space. The expected fraction of a random vector that lies in any k-dimensional subspace is $\approx \sqrt{k/d}$. In our MNIST32 task, for example, $d = 1024$, $k = 10$ and $\sqrt{10/1024} \approx 0.1$. Thus, we could only say that an explanation is systematically related to the data manifold if, on average, its fraction in tangent space is significantly larger than 0.1.

3.4. Datasets

We evaluate the hypothesis on six datasets. This includes (i) MNIST32 and (ii) MNIST256, two variants of the MNIST dataset [27] with 60000 grayscale training images and 10000 grayscale test images of size 32×32 and 256×256, respectively. The MNIST32 dataset was obtained from MNIST with the generative approach, using a β-TCVAE [10]. It lies on a completely known 10-dimensional image manifold in a 1024-dimensional space. The MNIST256 dataset is an up-scaled version of the MNIST32 dataset. The (iii) EMNIST128 dataset is a variant of the EMNIST dataset [12]. EMNIST128 and MNIST256 serves as examples of high-dimensional problems. The (iv) CIFAR10 dataset was created from CIFAR10 [26] with the reconstructive approach, using a convolutional autoencoder with a latent dimension of $k = 144$. We also evaluate the hypothesis on two high-dimensional medical imaging datasets: (v) X-ray Pneumonia [22] and (vi) Diabetic Retinopathy Detection[2]. These two datasets have been used before to study the properties of post-hoc explanation methods [5, 8, 11, 29, 31, 44].

[2] https://www.kaggle.com/c/diabetic-retinopathy-detection

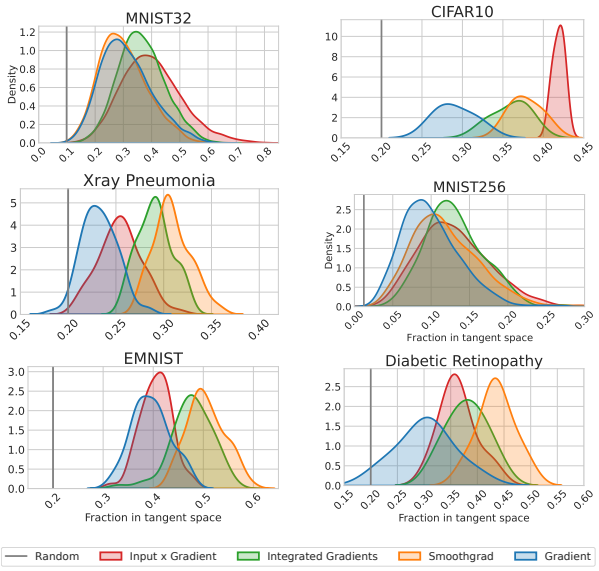

Figure 3. Post-hoc methods align attributions with the data manifold. The figure depicts the fraction of four different methods in tangent space on six different datasets. The gray line indicates the random baseline $\sqrt{k/d}$ (compare Sec. 3.3).

4. Experimental Results

Given a dataset, we train a neural network to minimize the test error. We then apply explanation algorithms and evaluate how feature attributions relate to the data manifold.

4.1. Qualitative evidence: The part of an attribution in tangent space is perceptually-aligned

We now demonstrate on MNIST32 that the part of an attribution that lies in tangent space is perceptually-aligned, whereas the part of the attribution that is orthogonal to the tangent space is not. Figure 2 depicts the gradient [37] Integrated Gradients [41], Input $\times$ Gradient [6], and SmoothGrad [38] attributions for a variant of a LeNet [27] that achieves a test accuracy $> 99\%$. The attributions are decomposed into the part that lies in tangent space (second row) and the part that is orthogonal to the tangent space (third row). We see from Figure 2 that the part of an attribution that lies in tangent space is perceptually-aligned, whereas the part that is orthogonal is not. In fact, the parts that are orthogonal to the tangent space consist of seemingly unrelated spots of positive and negative attribution. Figure 2 also provides qualitative evidence that the part of an attribution that lies in the tangent space is explanatory: The attributions in the second row of Figure 2 often highlight parts of the image that are relevant for the predicted class.

4.2. Post-hoc methods align attributions with the data manifold

We now demonstrate that the attributions provided by post-hoc methods are more aligned with the tangent space

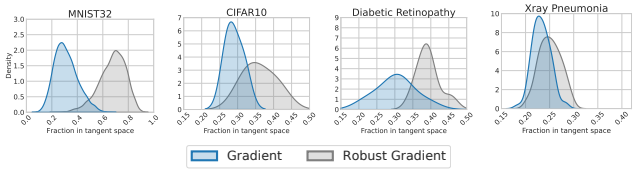

Figure 4. Adversarial training aligns model gradients with the data manifold. Figure shows the fraction of standard-and l_2-robust gradients in tangent space for four datasets.

than the gradient. Figure 3 depicts the fraction in tangent space (1) of model gradients, SmoothGrad, Integrated Gradients and Input $\times$ Gradient on six different datasets. All attributions have a fraction in tangent space that is considerably larger than random. In particular, the mean fraction of the raw gradient in tangent space is significantly larger than random on all datastes. Moreover, the gradient is the method with the weakest connection to the data manifold. Integrated Gradients, Input $\times$ Gradient and SmoothGrad improve upon the gradient on all datasets.

4.3. Adversarial training aligns model gradients with the data manifold

Previous work has observed that model gradients of adversarially trained models are perceptually-aligned [42]. According to our hypothesis, this should imply that model gradients of adversarially trained models are aligned with the tangent space of the data manifold. Figure 4 shows that this is indeed the case. Across four different datasets, the gradient of a model trained with projected gradient descent (PGD) against an l_2-adversary [30] is consistently more aligned with the tangent space than the gradient of a standard model. The alignment effect of adversarial training is substantial. On MNIST32, the mean fraction of robust gradients in tangent space is 0.68, compared with 0.31 for the standard model, and 0.40 for Input $\times$ Gradient (Figure 3).

4.4. A user study on the perceptual-alignment of attributions supports the hypothesis

To assess whether attributions that are more aligned with the data manifold are indeed more perceptually-aligned, we conducted a user study. The study consisted of three different tasks on our MNIST32 and CIFAR10 datasets. Each task took the form of an A/B-test where the participants were repeatedly shown images from two different groups of images (group A and group B). In the first task on MNIST32, the participants decided that the components of an attribution in tangent space are more perceptually-aligned than the corresponding orthogonal components ($N_A = 0$, $N_B = 580$, t-test $p < 0.01$). In the second task on MNIST32, the participants decided that among different attributions for the same image, those with a larger fraction in tangent space are more perceptually-aligned ($N_A = 143$, $N_B = 315$, t-test $p < 0.01$). In the third task on CIFAR10,

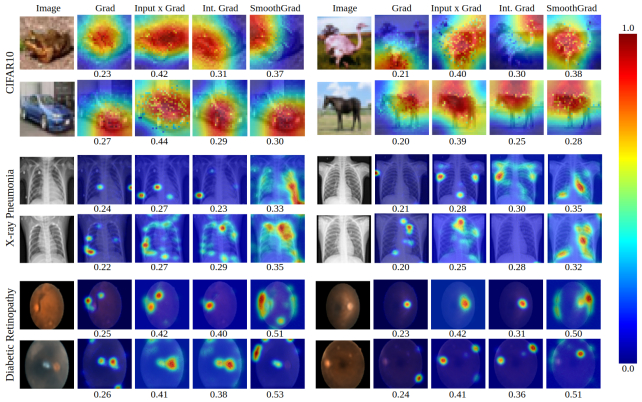

Figure 5. There is evidence that feature attributions that are more aligned with the data manifold are more explanatory. In a user study on CIFAR10, participants found that attributions that were more aligned with the data manifold better highlighted the object in the image. Figure depicts images and attributions from our CIFAR10 (Top row), X-Ray Pneumonia (Middle row) and Diabetic Retinopathy datasets (Bottom row). The number below an image depicts the fraction of the attribution in tangent space.

the participants decided that Input x Gradient attributions, which have on average the larger fraction in tangent space, better highlighted the object in the image than the gradient ($N_A = 36$, $N_B = 217$, t-test $p < 0.01$). Interestingly, the third task provides evidence that attributions that are more aligned with the tangent space are also more explanatory. For a visual comparison, see Figure 5 which depicts different attributions with various fractions in tangent space.

5. Conclusion

In this work, we focus on a particular aspect of feature attributions: whether they aligned with the tangent space of the data manifold. The main claim of this paper is that alignment with the data manifold makes attributions perceptually-aligned. While current models and algorithms provide only imperfect alignment, it is an open question whether this is due to the fact that we have not yet found the right model architecture or algorithm, or because the problem is more difficult than classification alone.

The objective of this paper is not to claim that the gradients of existing models provide good explanations, or that any particular post-hoc method works especially well. Instead, we would like to contribute to a line of work that, independently of particular algorithms, develops criteria by which explanations can be judged.

A main appeal of the manifold hypothesis is its broad potential for the analysis and improvement of different explanation algorithms. We believe that it will be interesting to explore the connections between the manifold hypothesis and other criteria for the evaluation of explanations, such as model sanity checks and the ROAR benchmark [2, 3, 20].

Acknowledgements

This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645), the BMBF Tübingen AI Center (FKZ: 01IS18039A), and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

[1] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 2019. 2

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 2018. 2, 4

[3] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *Advances in neural information processing systems*, 2020. 2, 4

[4] Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Steven Wu, and Himabindu Lakkaraju. Towards the unification and robustness of perturbation and gradient based explanations. In *International Conference on Machine Learning*, 2021. 1

[5] Amine Amyar, Romain Modzelewski, Hua Li, and Su Ruan. Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation. *Computers in Biology and Medicine*, 2020. 3

[6] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *International Conference on Learning Representations*, 2018. 1, 2, 3

[7] 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*, 2020. 2

[8] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv preprint arXiv:2008.02766*, 2020. 2, 3

[9] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 2015. 1

[10] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 2018. 3

[11] Mohamed Chetoui and Moulay A Akhloufi. Explainable diabetic retinopathy using efficientnet. In *IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020. 3

[12] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017. 3

[13] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 2019. 2

[14] Ann-Kathrin Dombrowski, Jan E Gerken, and Pan Kessel. Diffeomorphic explanations with normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021. 2

[15] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 2016. 2

[16] Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply robustness? *arXiv preprint*, 2023. 1

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 1

[18] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in neural information processing systems*, 2019. 2

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 2

[20] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 2019. 2, 4

[21] Simran Kaur, Jeremy Cohen, and Zachary C Lipton. Are perceptually-aligned gradients a general property of robust classifiers? *arXiv preprint arXiv:1910.08640*, 2019. 1

[22] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 3

[23] Beomsu Kim, Junghoon Seo, and Taegyun Jeon. Bridging adversarial robustness and gradient interpretability. *ICLR Workshop on Safe Machine Learning*, 2019. 1

[24] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. 2019. 2

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 3

[28] Dohun Lim, Hyeonseok Lee, and Sungchan Kim. Building reliable explanations of unreliable neural networks: Locally smoothing perspective of model interpretation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[29] Juan Eduardo Luján-García, Cornelio Yáñez-Márquez, Yenny Villuendas-Rey, and Oscar Camacho-Nieto. A transfer learning method for pneumonia classification and visualization. *Applied Sciences*, 2020. 3

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. 4

[31] Sivaramakrishnan Rajaraman, Sema Candemir, George Thoma, and Sameer Antani. Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs. In *Medical Imaging 2019: Computer-Aided Diagnosis*, 2019. 3

[32] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 2016. 2

[33] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 2021. 2

[34] Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems*, 2021. 1, 2

[35] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2

[36] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 2017. 1

[37] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. 1, 2, 3

[38] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1, 2, 3

[39] Suraj Srinivas and François Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. *International Conference on Learning Representations*, 2021. 2

[40] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019. 2

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017. 1, 2, 3

[42] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *International Conference on Learning Representations*, 2019. 1, 4

[43] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 2020. 2

[44] Toon Van Craenendonck, Bart Elen, Nele Gerrits, and Patrick De Boever. Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection. *Translational vision science & technology*, 2020. 3

[45] Kilian Q Weinberger and Lawrence K Saul. Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision*, 2006. 2