On Interpretability of Artificial Neural Networks: A Survey

Feng-Lei Fan, *Student Member, IEEE*, Jinjun Xiong, *Senior Member, IEEE*,
Mengzhou Li, *Graduate Student Member, IEEE*, and Ge Wang, *Fellow, IEEE*

***Abstract*—Deep learning as performed by artificial deep neural networks (DNNs) has achieved great successes recently in many important areas that deal with text, images, videos, graphs, and so on. However, the black-box nature of DNNs has become one of the primary obstacles for their wide adoption in mission-critical applications such as medical diagnosis and therapy. Because of the huge potentials of deep learning, the interpretability of DNNs has recently attracted much research attention. In this article, we propose a simple but comprehensive taxonomy for interpretability, systematically review recent studies on interpretability of neural networks, describe applications of interpretability in medicine, and discuss future research directions, such as in relation to fuzzy logic and brain science.**

***Index Terms*—Deep learning, interpretability, neural networks, survey.**

I. Introduction

DEEP learning [71] has become the mainstream approach in many important domains targeting common objects, such as text [40], images [181], videos [132], and graphs [88]. However, deep learning works as a black box model in the sense that although deep learning performs quite well in practice, it is difficult to explain its underlying mechanism and behaviors. Questions are often asked, such as how deep learning makes such a prediction, why some features are favored over others by a model, and what changes are needed to improve model performance, etc. Unfortunately, only modest success has been made to answer these questions.

Interpretability of deep neural networks (DNNs) is essential to many fields, and to healthcare [67], [68], [173] in particular for the following reasons. First, model robustness is a vital issue in medical applications. Recent studies suggest that model interpretability and robustness are closely connected [131]. On the one hand, the improvements in model robustness prompt model interpretability. For example, a deep model trained via adversarial training, a training method that augments training data with adversarial examples, shows better interpretability (with more accurate saliency maps) than the same model trained without adversarial examples [131]. On the other hand, when we understand a model deeply, we can thoroughly examine its weaknesses because the interpretability can help identify potential vulnerabilities of a complicated model, thereby improving its accuracy and reliability. Also, interpretability plays an important role in ethic use of deep learning techniques [57]. To build patients' trust in deep learning, interpretability is needed to hold a deep learning system accountable [57]. If a model builder can explain why a model makes a particular decision under certain conditions, users would know whether such a model contributes to an adverse event or not. It is then possible to establish standards and protocols to use the deep learning system optimally.

However, the lack of interpretability has become a main barrier of deep learning in its wide acceptance in mission-critical applications. For example, regulations were proposed by European Union in 2016 that individuals affected by algorithms have the right to obtain an explanation [61]. Despite great research efforts made on interpretability of deep learning and availability of several reviews on this topic, we believe that an up-to-date review is still needed, especially considering the rapid development of this area. The review of Zhang and Zhu [201] is mainly on the visual interpretability. The representative publications from their review fall under the *feature analysis*, *saliency*, and *proxy* taxonomy in our review. The review of Chakraborty *et al.* [28] took opinions of Lipton [112] on levels of interpretability, and accordingly structured their review to provide in-depth perspectives but with limited scope. For example, only 49 references are cited there. The review of Du *et al.* [43] has a similar weakness, only covering 40 papers, which are divided into *post-hoc* and *ad-hoc* explanations, as well as global and local interpretations. Their taxonomy is coarse-grained and neglects a number of important publications, such as publications on *explaining-by-text*, *explaining-by-case*, etc. In contrast, our review is much detailed and comprehensive, with the latest results included. While publications in [58] are classified into understanding the workflow of a neural network, understanding the representation of a neural network, and explanation producing, we cover all these aspects and also discuss the studies on how to protype an interpretable neural network. Reviews by Guidotti *et al.* [65] and Adadi and Berrada [2] cover existing

Manuscript received October 26, 2020; revised January 9, 2021; accepted March 10, 2021. Date of publication March 17, 2021; date of current version November 3, 2021. This work was supported in part by the Rensselaer-IBM AI Research Collaboration Program (http://airc.rpi.edu); in part by the IBM AI Horizons Network (http://ibm.biz/AIHorizons); and in part by NIH under Grant R01 EB026646, Grant R01 CA233888, Grant R01 CA237267, and Grant R01 HL151561. *(Corresponding author: Ge Wang.)*

Feng-Lei Fan, Mengzhou Li, and Ge Wang are with the Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: wangg6@rpi.edu).

Jinjun Xiong is with the AI and Hybrid Clouds Systems, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: jinjun@us.ibm.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TRPMS.2021.3066428.

Digital Object Identifier 10.1109/TRPMS.2021.3066428

black-box machine learning models instead of focusing on neural networks. As a result, several hallmark papers on explaining neural networks are missing in their survey, such as the interpretation from the perspectives of mathematics and physics.

Arrieta *et al.* [10] provided an extensive review on explainable AI (XAI), where concepts and taxonomies are clarified, and challenges are identified. While that review covers interpretability of AI/ML in general, our review is specific to DNNs and offers unique perspectives and insights. Specifically, our review is novel in the following senses: 1) We treat *post-hoc* and *ad-hoc* interpretability separately, because the former explains the existing models, while the latter constructs interpretable ones; 2) we include widely studied generative models, advanced mathematical/physical methods that summarize advances in deep learning theory, and the applications of interpretability in medicine; 3) important methods are illustrated with customized examples and publicly available codes through GitHub; and 4) interpretability research is a rapidly evolving field, and many research articles are published every year. Hence, our review should be a valuable and up-to-date addition to the literature.

Before we start our survey, let us first state three essential questions regarding interpretability: 1) What does interpretability mean? 2) Why is interpretability difficult? and 3) How to build a good interpretation method? The first question has been well addressed in [112], and we include their statements here for completeness. The second question was partially touched in [112] and [145], and we incorporate those comments and complement them with our own views. We provide our own perspectives on the third question.

A. What Does Interpretability Mean?

Although the word "interpretability" is frequently used, people do not reach a consensus on the exact meanings of interpretability, which partially accounts for why current interpretation methods are so diverse. For example, some researchers explore *post-hoc* explanations for models, while some focus on the interplay mechanism between machineries of a model. Generally speaking, interpretability refers to the extent of human's ability to understand and reason a model. Based on the categorization of Lipton [112], we summarize the implications of interpretability in different levels.

Simulatability is considered as the understanding over the entire model. In a good sense, we can understand the mechanism of a model at the top level in a unified theoretical framework, one example is what was reported in [140]: a class of radial basis function (RBF) networks can be expressed by a solution to the interpolation problem with a regularization term, where an RBF network is an artificial neural network with RBFs as activation functions. In view of simulatability, the simpler the model, the higher simulatability the model has. For example, a linear classifier or regressor is totally understandable. To enhance simulatability, we can change some facilities of models or use crafted regularization terms.

Decomposability is to understand a model in terms of its components, such as neurons, layers, blocks, and so on. Such a modularized analysis is quite popular in engineering fields. For instance, the inner working of a complicated system is factorized as a combination of functionalized modules. A myriad of engineering examples, such as software development and optical system design have justified that a modularized analysis is effective. In machine learning, a decision tree is a kind of modularized methods, where each node has an explicit utility to judge if a discriminative condition is satisfied or not, each branch delivers an output of a judgement, and each leaf node represents the final decision after computing all attributes. Modularizing a neural network is advantageous to the optimization of the network design since we know the role of each and every component of the entire model.

Algorithmic transparency is to understand the training process and dynamics of a model. The landscape of the objective function of a neural network is highly nonconvex. The fact that deep models do not have a unique solution hurts the model transparency. Nevertheless, it is intriguing that the current stochastic gradient descent (SGD)-based learning algorithms still perform efficiently and effectively. If we can understand why learning algorithms work, deep learning research and applications will be accelerated.

B. Why Is Interpretability Difficult?

After we learn the meanings of interpretability, a question is what obstructs practitioners to obtain interpretability. This question was partially addressed in [145] in terms of *commercial barrier* and *data wildness*. Here, we complement their opinion with additional aspects on *human limitation* and *algorithmic complexity*. We believe that the hurdles to interpretable neural networks come from the following four aspects.

Human Limitation: Expertise is often insufficient in many applications. Nowadays, deep learning has been extensively used in tackling intricate problems, which even professionals are unable to comprehend adequately. What is worse is that these problems are not uncommon. For example, in a recent study [46], we proposed to use an artificial neural network to predict pseudo-random events. Specifically, we fed 100 000 binary sequential digits into the network to predict the 100 001 th digit in the sequence. In our prediction, the highly sophisticated hidden relationship was learned to beat a purely random guess with a 3σ precision. Furthermore, it was conjectured that high sensitivity and efficiency of neural networks may help discriminate the fundamental differences between pseudorandomness and real quantum randomness. In this case, it is no wonder that interpretability for neural networks will be missing, because even most talented physicists know little about the essence of this problem, let alone fully understand predictions of the neural network.

Commercial Barrier: In the commercial world, there are strong motives for corporations to hide their models. First and foremost, companies profit from black-box models. It is not a common practice that a company makes capital out of totally transparent models [145]. Second, model opacity helps protect hard work from being reverse engineered. An effective black box is ideal in the sense that customers being served can obtain satisfactory results while competitors are not able to steal their intellectual properties easily [145]. Third, prototyping an interpretable model may cost too much in terms of financial, computational, and other resources. Existing open-sourced superior models are accessible to easily construct a well-performed algorithm for a specific task. However, generating reliable and consistent understanding to the behavior of the resultant model demands much more endeavors.

Data Wildness: On the one hand, although it is a big data era, high-quality data are often not accessible in many domains. For example, in the project of predicting electricity grid failure [145], the data base involves text documents, accounting data about electricity dating back to the 1890s, and data from new manhole inspections. Highly heterogenous and inconsistent data hamper not only the accuracy of deep learning models but also the construction of interpretability. On the other hand, real-world data have the character of high dimensionality, which suppresses reasoning. For example, given an MNIST image classification problem, the input image is of size $28 \times 28 = 784$. Hence, the deep learning model tackling this problem has to learn an effective mapping of 784 variables to one of the ten digits. If we consider the ImageNet dataset, the number of input variables goes up to $512 \times 512 \times 3 = 768\,432$.

Algorithmic Complexity: Deep learning is a kind of large-scale and highly nonlinear algorithms. Convolution, pooling, nonlinear activation, shortcut, and so on contribute to the variability of neural networks. The number of trainable parameters of a deep model can be on the order of hundreds million or even more. Despite that nonlinearity may not necessarily result in opacity (for example, a decision tree model is not linear but interpretable), deep learning's series of nonlinear operations indeed prevent us from understanding its inner working. In addition, recursiveness is another source of difficulty. A typical example is the chaos behavior resultant from nonlinear recursiveness. It is well-known that even a simple recursive mathematical model can lead to an intractable dynamics [107]. In [174], it was proved that there are chaotic behaviors such as bifurcations even in simple neural networks. In chaotic systems, tiny changes of initial inputs may lead to huge outcome differences, adding to the complexity of interpretation methods.

C. How to Build Good Interpretation Method?

The third major issue is the criteria for assessing the quality of a proposed interpretation method. Because existing evaluation methods are still premature, we propose five general and well-defined rules-of-thumb: 1) *exactness*; 2) *consistency*; 3) *completeness*; 4) *universality*; and 5) *reward*. Our rules-of-thumb are fine-grained and focus on the characteristics of interpretation methods, compared to that described in [42]: application-grounded, human-grounded, and function-grounded.

Exactness: Exactness means how accurate an interpretation method is. Is it just limited to a qualitative description or with a quantitative analysis? Generally, quantitative interpretation methods are more desirable than qualitative counterparts.

Consistency: Consistency suggests that there is no contradiction in an explanation. For multiple similar samples, a fair interpretation should produce consistent answers. In addition, an interpretation method should conform to the predictions of the authentic model. For example, the proxy-based methods are evaluated based on how closely they replicate the original golden model.

Completeness: Mathematically, a neural network is to learn a mapping that best fits data. A good interpretation method should show effectiveness in support of the maximal number of data instances and data types.

Universality: With the rapid development of deep learning, the deep learning armory has been substantially enriched. Such diverse deep learning models play important roles in a wide spectrum of applications. A driving question is whether we can develop a universal interpreter that deciphers as many models as possible so as to save labor and time. But, this is technically challenging due to the high variability among models.

Reward: What are gains from the improved understanding of neural networks? In addition to the trust from practitioners and users, fruits of interpretability can be insights into network design, training, etc. Due to its black-box nature, using neural networks is largely a trial-and-error process with sometimes contradictive intuitions. A thorough understanding of deep learning will be instrumental to the research and applications of neural networks.

Briefly, our contributions in this review are threefold: 1) we propose a comprehensive taxonomy for interpretability of neural networks and describe key methods with our insights; 2) we systematically illustrate interpretability methods as educational aids, as shown in Figs. 3, 5, 6, 7, 9, 10, 16, and 17; and 3) we shed light on future directions of interpretability research in terms of the convergence of neural networks and rule systems, the synergy between neural networks and brain science, and interpretability in medicine.

II. Survey on Interpretation Methods

In this section, we first present our taxonomy and then review interpretability results under each category of our taxonomy. We enter the search terms "deep learning interpretability," "neural network interpretability," "explainable neural network," and "explainable deep learning" into the Web of Science on Sep 22, 2020, with the time range from 2000 to 2019. The number of articles with respect to years is plotted in Fig. 1, which clearly shows an exponential trend in this field. With the survey, our motive is to cover as many important papers as possible. Therefore, we do not limit ourselves within Web of Science. We also search related articles using Google Scholar, PubMed, IEEE Xplore, and so on.

A. Taxonomy Definition

As shown in Fig. 2, our taxonomy is based on our surveyed papers and existing taxonomies. We first classify the surveyed papers into *Post-hoc* interpretability analysis and *ad-hoc* interpretable modeling. *Post-hoc* interpretability analysis explains

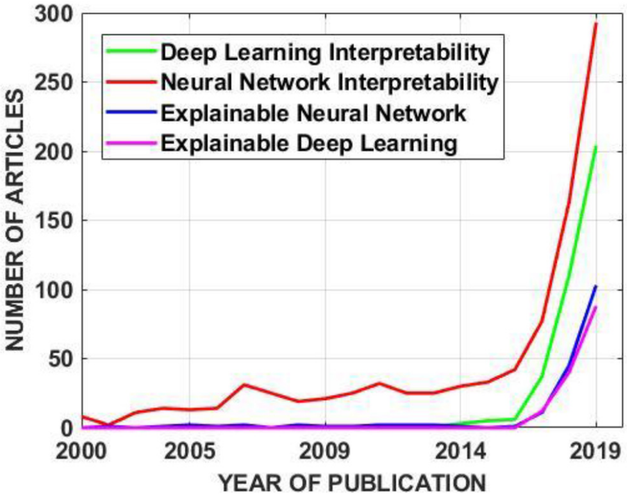


Fig. 1. Exponential growth of the number of articles on interpretability.

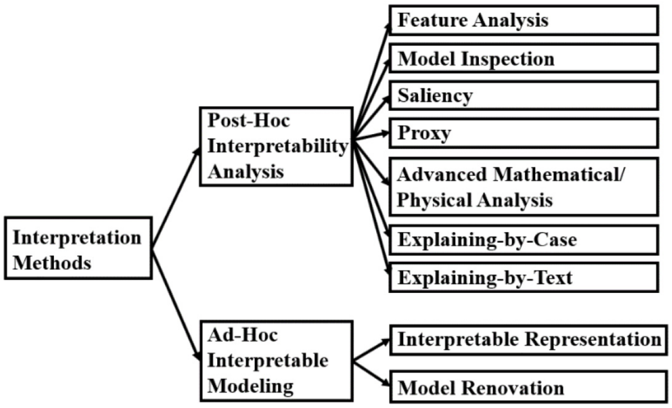

Interpretation Methods
Post-Hoc Interpretability Analysis
Ad-Hoc Interpretable Modeling
Feature Analysis
Model Inspection
Saliency
Proxy
Advanced Mathematical/ Physical Analysis
Explaining-by-Case
Explaining-by-Text
Interpretable Representation
Model Renovation

Fig. 2. Taxonomy used for this interpretability review.

the existing models and can be further classified into *feature analysis*, *model inspection*, *saliency*, *proxy*, *advanced mathematical/physical analysis*, *explaining-by-case*, and *explaining-by-text*, respectively. *Ad-hoc* interpretable modeling builds interpretable models and can be further categorized into *interpretable representation* and *model renovation*. In our proposed taxonomy, the class *advanced mathematical/physical analysis* is novel, but it is unfortunately missing in the previous reviews. We argue that this class is rather essential, because the incorporation of math/physics is critical in placing deep learning on a solid foundation. In the following, we clarify the taxonomy definition and its illustration. We would like to underscore that one method may fall into different classes, depending on how one views it.

Post-Hoc Interpretability Analysis: Post-hoc interpretability is conducted after a model is well learned. A main advantage of *post-hoc* methods is that one does not need to compromise interpretability with the predictive performance since prediction and interpretation are two separate processes without mutual interference. However, a *post-hoc* interpretation is usually not completely faithful to the original model. If an interpretation is 100% accurate compared to the original model, it becomes the original model. Therefore, any interpretation method in this category is more or less inaccurate. What is worse is that we often do not know the nuance [145]. Such a nuance makes it hard for practitioners to have a full trust to an interpretation method, because the correctness of the interpretation method is not guaranteed.

Feature analysis techniques are centered in comparing, analyzing, and visualizing features of neurons and layers. Through feature analysis, sensitive features and ways to process them are identified such that the rationale of the model can be explained to some extent.

Feature analysis techniques can be applied to any neural networks and provide qualitative insights on what kinds of features are learned by a network. However, these techniques lack an in-depth, rigorous, and unified understanding, and therefore cannot be used to revise a model toward a higher interpretability.

Model inspection methods use external algorithms to delve into neural networks by systematically extracting important structural and parametric information on inner working mechanisms of neural networks.

Methods in this class are more technically accountable than those in *feature analysis* because analytical tools such as statistics are directly involved in the performance analysis. Therefore, the information gained by a model inspection method is more trustworthy and rewarding. In an exemplary study [183], finding important data routing paths is used as a way to understand the model. With such data routing paths, the model can be faithfully compressed to a compact one. In other words, interpretability improves the trustworthiness of model compression.

Saliency methods identify which attributes of input data are most relevant to a prediction or a latent representation of a model. In this category, human inspection is involved to decide if a saliency map is plausible. A saliency map is useful, i.e., if a polar bear always appears in a picture coupled with snow or ice, the model may have misused the information of snow or ice to detect the polar bear rather than real features of polar bears for detection. With a saliency map, this issue can be found and hence avoided.

Saliency methods are popular in interpretability research, however, extensive random tests reported that some saliency methods might be model independent and data independent [3]; i.e., saliency maps offered by some methods are highly similar to results produced with edge detectors. This is problematic because it means that those saliency methods fail to find the true attributes of the input that account for the prediction of the model. Consequently, a model-relevant and data-relevant saliency method should be developed in these cases.

Proxy methods construct a simpler and more interpretable proxy that closely resembles a trained, large, complex, and black-box model. Proxy methods can be either local in a partial space or global in a whole solution space. The exemplary proxy models include decision trees, rule systems, and so on. The weakness of proxy methods is the extra cost needed to construct a proxy model.

Advanced mathematical/physical analysis methods put a neural network into a theoretical mathematics/physics framework, in which the mechanism of a neural network is understood with advanced mathematics/physics tools. This class covers theoretical advances of deep learning including nonconvex optimization, representational power, and generalization ability.

A concern in this class is that to establish a reasonable interpretation, unrealistic assumptions are sometimes made to facilitate a theoretical analysis, which may compromise the practical validity of the explanation.

Explaining-by-case methods are along the line of case-based reasoning [90]. People favor examples. One may not be engaged by boring statistic numbers of a product but could be amazed while listening to other users' experience of using such a product. This philosophy wins the heart of many practitioners and intrigues the case-based interpretation for deep learning. Explaining-by-case methods provide representative examples that capture the essence of a model.

Methods in this class are interesting and inspiring. However, this practice is more like a sanity check instead of a general interpretation because not much information regarding the inner working of a neural network is understood from selected query cases.

Explaining-by-text methods generate text descriptions in image-language joint tasks that are conducive to understanding the behavior of a model. This class can also include methods that generate symbols for explanation.

Methods in this class are particularly useful in image-language joint tasks such as generating a diagnostic report from an X-ray radiograph. However, explaining-by-text is not a general technique for any deep learning model because it can only work when a language module exists in a model.

Ad-Hoc Interpretable Modeling: Ad-hoc interpretable modeling eliminates the biases that more or less exist in the *post-hoc* interpretability analysis. Although it is generally believed that there is a tradeoff between interpretability and model expressibility [123], it is still possible to find a model that is both powerful and interpretable. One notable example is the work reported in [30], where an interpretable two-layer additive risk model has won the first place in FICO Recognition Contest.

Interpretable representation methods employ regularization techniques to steer the optimization of a neural network toward a more interpretable representation. Properties, such as decomposability, sparsity, and monotonicity can enhance interpretability. As a result, regularizing features becomes a way to allow more interpretable models. Correspondingly, the loss function must contain a regularization term for the purpose of interpretability, which restricts the original model to perform its full learning task.

Model renovation methods seek interpretability by the means of designing and deploying more interpretable machineries into a network. Those machineries include a neuron with purposely designed activation function, an inserted layer with a special functionality, a modularized architecture, and so on. The future direction is to use more and more explainable components that can at the same time achieve similar state-of-the-art performance for diverse tasks.

B. Post-Hoc Interpretability Analysis

Feature Analysis: Inverting-based methods [41], [117], [163], [200] crack the representation of a neural network by inverting feature maps into a synthesized image. For example, Mahendran and Vedaldi [117] assumed that a representation of a neural network Ω_0 for an input image x_0 was modeled as $\Omega_0 = \Omega(x_0)$, where Ω is the neural network mapping, usually not invertible. Then, the inverting problem was formulated as finding an image x^*, whose neural network representation best matches Ω_0; i.e., arg min $_x \|\Omega(x) - \Omega_0\|^2 + \lambda R(x)$, where $R(x)$ is a regularization term representing prior knowledge about the input image. The goal is to reveal the lost information by comparing differences between the inverted image and the original one. Dosovitskiy and Brox [41] directly trained a new network with features generated by the model of interest as the input and images as the label, to invert features of intermediate layers to images. It was found that contours and colors could still be reconstructed even from deeper layer features. Zeiler and Fergus [200] designed a deconvolution network consisting of unpooling, rectification, deconvolution operations, to pair with the original convolutional network so that features could be inverted without training. In the deconvolution network, an unpooling layer is realized by using locations of maxima; rectification is realized by setting negative values to 0; and deconvolution layers use transposed convolutional filters.

Activation maximization methods [45], [128], [129], [168] devote to synthesizing images that maximize the output of a neural network or neurons of interest. The resulting images are referred as "deep dreams" as these can be regarded as dream images of a neural network or a neuron.

In [16], [85], [108], [196], and [210], it was pointed out that information about a deep model could be extracted from each neuron. Yosinski *et al.* [196] straightforwardly inspected the activation values of neurons in each layer with respect to different images or videos. They found that live activation values changes for different inputs are helpful to understand how a model work. Li *et al.* [108] contrasted features generated by different initializations to investigate if a neural network learns a similar representation when randomly initialized. The receptive field (RF) is a spatial extent over which a neuron connects with an input volume [111]. To investigate the size and shape of RF of a given input for a neuron, Zhou *et al.* [210] presented a network dissection method that first selected K images with high activation values for neurons of interest and constructed 5000 occluded images for each of K images, and then fed them into a neural network to observe the changes in activation values for a given unit. A large discrepancy signals an important patch. Finally, the occluded images that have large discrepancy were recentered and averaged to generate an RF. This network dissection method has been scaled to generative networks [17]. In addition, Bau *et al.* [16] scaled up a low-resolution activation map of a given layer to the same size as the input, thresholded the map into a binary activation map, and then computed the overlapping area between the binary activation map and the ground-truth binary segmentation map as an interpretability measure. Karpathy *et al.* [85] defined the gate in

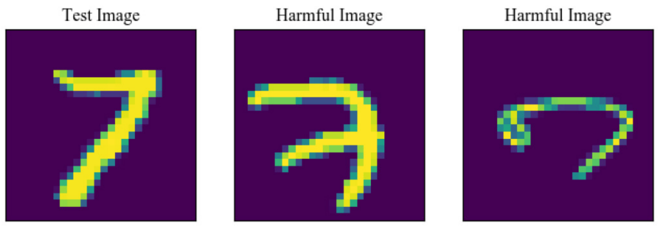


Fig. 3. Based on the influence function, two harmful images that have the same label as the test image are identified.

LSTM [73] to be either left or right saturated depending on its activation value being either less than 0.1 or more than 0.9. In this regard, neurons that were often right saturated are interesting because this means that these neurons can remember their values over a long period. Zhang *et al.* [202] dissected feature relations in a network with the premise that the feature map of a filter in each layer can be activated by part patterns in the earlier layer. They mined part patterns layer by layer, discovered activation peaks of part patterns from the feature map of each layer, and constructed an explanatory graph to describe the relations of hierarchical features, with each node representing a part pattern and the edge between neighboring layers representing a co-activation relation.

Model Inspection: The empirical influence function is to measure the dependence of an estimator on a sample [99]. Koh and Liang [89] applied the concept of the influence function to address the following question: Given a prediction for one sample, do other samples in the dataset have positive effects or negative effects on that prediction? This analysis could also help identify misannotated labels and outliers existing in the data. As Fig. 3 shows, given a LeNet-5 like network, two harmful images for a given image are identified by the influence function.

Bansal *et al.* [12], Lakkaraju *et al.* [97], and Zhang *et al.* [203] worked on the detection of failures or biases in a neural network. For example, Bansal *et al.* [12] developed a model-agnostic algorithm to identify which instances a neural network is likely to fail to provide any prediction for. In such a scenario, the model would instead give a warning like "Do not trust these predictions" as an alert. Specifically, they annotated all failed images with a collection of binary attributes and clustered these images in the attribute space. As a result, each cluster indicates a failure mode. To recognize those mislabeled instances with high predictive scores in the dataset efficiently, Lakkaraju *et al.* [97] introduced two basic speculations: the first is that mislabeling an instance with high confidence is due to the systematic biases instead of random perturbation, while the second is that each failed example is representative and informative enough. Then, they clustered the images into several groups and designed a multiarmed bandit search strategy by taking each group as a bandit that plans which group should be queried and sampled in each step. To discover representation biases, Zhang *et al.* [203] utilized ground-truth relationships among attributes according to human's common knowledge (fire-hot versus ice-cold) to examine if a mined attribute relationship by a neural network well fits the ground truth.

Wang *et al.* [183] demystified a network by identifying critical data routes. Specifically, a gate control binary vector $\boldsymbol{\lambda}_k \in \{0, 1\}^{n_k}$, where n_k is the number of neurons in the kth layer, was multiplied to the output of the kth layer, and the problem of finding control gate values is formulated as searching $\boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda}_K$ by the formula

$$\arg\min_{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K} \; d(f_\theta(x), f_\theta(x; \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)) + \gamma \sum_k \|\boldsymbol{\lambda}_k\|_1$$

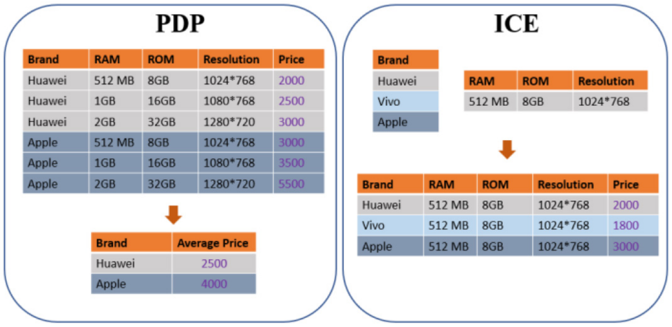
where f_θ is the mapping represented by a neural network parameterized by θ, $f_\theta(x; \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda_K})$ is the mapping when control gates $\boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda}_K$ are enforced, $d(\cdot, \cdot)$ is a distance measure, γ is a constant controlling the tradeoff between the loss and regularization, and $\| \cdot \|_1$ is the l_1 norm such that $\boldsymbol{\lambda}_k$ is sparse. The learned control gates could expose the important data processing paths of a model. Kim *et al.* [86] developed the concept activated vector (CAV) that can quantitatively measure the sensitivity of the concept C with respect to any layer of a model. First, a binary linear classifier h was trained to distinguish between layer activations stimulated by two sets of samples: $\{f_l(x) : x \in P_C\}$ and $\{f_l(x) : x \notin P_C\}$, where $f_l(x)$ is the layer activation at the lth layer, and P_C denotes data embodying the concept C. Then, the CAV was defined as the normal unit vector $\boldsymbol{v}_C^l$ to a hyperplane of the linear classifier that separated samples with and without the defined concept. Finally, $\boldsymbol{v}_C^l$ was used to calculate the sensitivity for a concept C in the lth layer as the directional derivatives

$$S_{C,k,l} = \lim_{\epsilon \to 0} \frac{h_{l,k}\big(f_l(x) + \epsilon \boldsymbol{v}_C^l\big) - h_{l,k}(f_l(x))}{\epsilon} = \nabla h_{l,k}(f_l(x)) \boldsymbol{v}_C^l$$

where $h_{l,k}$ denotes the logits of the trained binary linear classifier for the output class k. You *et al.* [195] mapped a neural network into a relational graph and then studied the relationship between the graph structures of neural networks and their predictive performance through massive experiments (transcribed a graph into a network and implemented the network on a dataset). They discovered that the predictive performance of a network was correlated with two graph measures: 1) the clustering coefficient and 2) the average path length.

Saliency: There is a plethora of methods to obtain a saliency map. Partial dependence plot (PDP) and individual condition expectation (ICE) [53], [59], [74] are model-agnostic statistical tools to visualize the dependence between the responsible variables and the predictive variables. To compute the PDP, assume that there are p input dimensions and let $S, C \subseteq \{1, 2, \ldots, p\}$ be two complementary sets, where S is the set one will fix, and C is the set one will change. Then, the PDP for $\boldsymbol{x}_S$ is defined by $f_S = \int f(\boldsymbol{x}_S, \boldsymbol{x}_C) d\boldsymbol{x}_C$, where f is the model. Compared with PDP, the definition of ICE is straightforward. The ICE curve at $\boldsymbol{x}_S$ is obtained by fixing $\boldsymbol{x}_C$ and varying $\boldsymbol{x}_S$. Fig. 4 shows a simple example illustrating how to compute PDP and ICE, respectively.

A simple approach is to study the change of prediction after removing one feature, also known as leave-one-out attribution [4], [83], [105], [143], [211]. For example, Kádár *et al.* [83] utilized this idea to define an omission score:



Brand	RAM	ROM	Resolution	Price
Huawei	512 MB	8GB	1024*768	2000
Huawei	1GB	16GB	1080*768	2500
Huawei	2GB	32GB	1280*720	3000
Apple	512 MB	8GB	1024*768	3000
Apple	1GB	16GB	1080*768	3500
Apple	2GB	32GB	1280*720	5500

Brand	Average Price
Huawei	2500
Apple	4000

Brand
Huawei
Vivo
Apple

RAM	ROM	Resolution
512 MB	8GB	1024*768

Brand	RAM	ROM	Resolution	Price
Huawei	512 MB	8GB	1024*768	2000
Vivo	512 MB	8GB	1024*768	1800
Apple	512 MB	8GB	1024*768	3000

Fig. 4. Toy examples illustrating the definitions of PDP and ICE, respectively. On the left, to measure the impact of the brand on the price with the PDP method, we fix the brand and compute the average of prices as other factors change, obtaining that the PDP of "Huawei" is 2500 and the PDP of "Apple" is 4000. On the right, ICE scores regarding brands "Huawei," "Vivo," and Apple are computed by varying brands and fixing other factors.

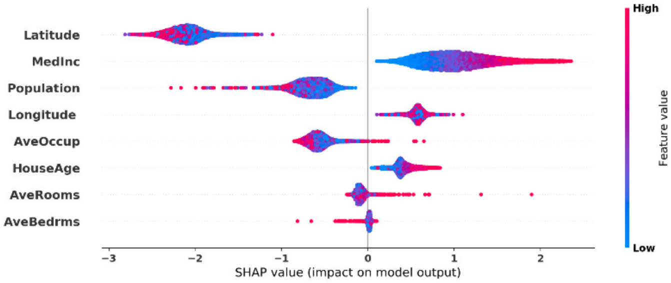


Fig. 5. Positive Shapley value indicates a positive impact on the model output, and vice versa. Shapley value analysis shows that the model is biased because the house age has the positive Shapley value on the house price, which goes against our real experience.

$1-\text{cosine}(\boldsymbol{h}(S), \boldsymbol{h}(S_{\backslash i}))$, where $\text{cosine}(\cdot,\cdot)$ is the cosine distance, $\boldsymbol{h}$ is the representation for a sentence, S is the full sentence, and $S_{\backslash i}$ is the sentence without the ith word, and analyzed the importance of each word. Adler *et al.* [4] proposed to measure an indirect influence for correlated inputs. For example, in a house loan decision system, race should not be a factor for decision making. However, solely removing the race factor is not sufficient to rule out the effect of race because some remaining factors, such as "zipcode" are highly concerned with race.

Furthermore, the Shapley value from cooperative game theory was used in [6], [27], [39], [113], and [115]. Mathematically, the Shapley value of a set function $\hat{f}$ with respect to the feature i is defined as

$$\text{Shapley}_i\Big(\hat{f}\Big) = \sum_{S\subseteq P\backslash\{i\}} \frac{(N-|S|-1)!|S|!}{N!}\Big(\hat{f}(S\cup\{i\}) - \hat{f}(S)\Big)$$

where $|\cdot|$ is the size of a set, P is a total player set of N players, and the set function $\hat{f}$ maps each subset $S \subseteq P$ to a real number. Furthermore, the definition of the Shapley value can be twisted to the neural network function f by replacing the features in the input that are not in S with the zero value. Motivated by reducing the prohibitive computational cost incurred by combinatorial explosion, Ancona *et al.* [6] proposed a novel and polynomial-time approximation of the Shapley values, which basically computed the expectation of a random coalition rather than enumerated each and every coalition. Fig. 5 shows a simple example of how the Shapley values can be computed for a fully connected layer network trained on the California Housing dataset, which includes eight attributes, such as house age and room number as the inputs and the house price as the label.

Instead of removing one or more features, researchers also resort to gradients. Simonyan *et al.* [156], Smilkov *et al.* [160], Sundararajan *et al.* [167], and Singla *et al.* [159] utilized the idea of gradients to probe the saliency of an input. Simonyan *et al.* [156] calculated the first-order Taylor expansion of the class score with respect to image pixels, by which the first-order coefficients produce a saliency map for a class. Smilkov *et al.* [160] demonstrated that gradients as a saliency map show a correlation between attributes and labels, however, typically gradients are rather noisy. To remove noise, they proposed SmoothGrad that adds noise into the input image multiple times and averages the resultant gradient maps $\widehat{M_c}(x) = (1/N)\sum_{n=1}^{N} M_c^{(n)}(x+N(0,\sigma^2))$, where $M_c^{(n)}$ is a gradient map for a class c, and $N(0,\sigma^2)$ is the Gaussian noise with σ as the standard variance. Basically, $\widehat{M_c}(x)$ is a smoothened version of a salient map. Sundararajan *et al.* [167] set two fundamental requirements for saliency methods: 1) (sensitivity) if only one feature is different between the input and the baseline, the outputs of the input and the baseline are different, then this very feature should be credited by a nonzero attribution; and 2) (implementation invariance) the attributions for the same feature in two functionally equivalent networks should be identical. Noting that earlier gradient-based saliency methods did not meet the above two requirements, they put forth integrated gradients, which is formulated as $(x_i - x_i') \int_0^1 [\partial F(\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}'))/\partial x_i] d\alpha$, where $F(\cdot)$ is a neural network mapping, $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ is an input, and $\boldsymbol{x}' = (x_1', x_2', \ldots, x_N')$ is the baseline satisfying $(\partial/\partial\boldsymbol{x})F(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}'} = 0$. In practice, the integral can be transformed into a discrete summation $[(x_i - x_i')/M] \times \sum_{m=1}^{M} [\partial F(\boldsymbol{x}' + (m/M)(\boldsymbol{x} - \boldsymbol{x}'))/\partial x_i]$, where M is the number of steps in the approximation of the integral. Singla *et al.* [159] proposed to use the second-order approximations of a Taylor expansion to produce a saliency map so as to consider feature dependencies.

Bach *et al.* [11] proposed layerwise relevance propagation (LRP) to compute the relevance of one attribute to a prediction by assuming that a model representation $f(\boldsymbol{x})$ can be expressed as the sum of pixelwise relevance R_p^l, where $\boldsymbol{x}$ is an input image, l is the index of the layer, and p is the index of the pixel of $\boldsymbol{x}$. Thus, $f(\boldsymbol{x}) = \sum_p R_p^L$, where L is the final layer and $R_p^L = [(w_p x_p^{L-1})/(\sum_p w_p x_p^{L-1})] f(\boldsymbol{x})$, where w_p is the weight between pixel p of the $(L-1)$th layer and the final layer. Given a feedforward neural network, the pixelwise relevance score R_p^1 of an input is derived by calculating $R_p^l = \sum_j [z_{pj}/(\sum_{p'} z_{p'j})] R_j^{l+1}$ backwards with $z_{pj} = x_p^l w_{pj}^{(l,l+1)}$, where $w_{pj}^{(l,l+1)}$ is the weight between pixel p of layer l and the pixel j of the $(l+1)$th layer. Furthermore, Arras *et al.* [9] extended

LRP to recurrent neural networks (RNNs) for sentiment analysis. Montavon *et al.* [125] employed the whole first-order term of deep Taylor decomposition to produce a saliency map instead of just gradients. Suppose $\widehat{\boldsymbol{x}}$ is a well-chosen root for the function by a model $f(\boldsymbol{x})$: $f(\widehat{\boldsymbol{x}}) = 0$, because $f(\boldsymbol{x})$ can be decomposed as $f(\boldsymbol{x}) = f(\widehat{\boldsymbol{x}}) + ([\partial f/\partial \boldsymbol{x}]|_{\boldsymbol{x}=\widehat{\boldsymbol{x}}})^T \cdot (\boldsymbol{x} - \widehat{\boldsymbol{x}}) + \epsilon = 0 + \sum_i (\partial f/\partial x_i)|_{\boldsymbol{x}=\widehat{\boldsymbol{x}}}(x_i - \hat{x}_i) + \epsilon$, where ϵ is the high-order terms, the pixel relevance for the pixel i is expressed as $R_i = [\partial f/\partial x_i]|_{\boldsymbol{x}=\widehat{\boldsymbol{x}}}(x_i - \hat{x}_i)$. Inspired by the fact that even though a neuron is not fired, it is still likely to reveal useful information, Shrikumar *et al.* [155] proposed DeepLIFT to compute the difference between the activation of each neuron and its reference, where the reference is the activation of that neuron when the network is provided a reference input, and then backpropagate the difference to the image space layer by layer as LRP does. Singh *et al.* [158] introduced contextual decomposition, whose layer propagation formula is $\beta_i = \boldsymbol{W}\beta_{i-1} + [|\boldsymbol{W}\beta_{i-1}|/(|\boldsymbol{W}\beta_{i-1}| + |\boldsymbol{W}\gamma_{i-1}|)] \cdot b$ and $\gamma_i = \boldsymbol{W}\gamma_{i-1} + [|\boldsymbol{W}\gamma_{i-1}|/(|\boldsymbol{W}\beta_{i-1}| + |\boldsymbol{W}\gamma_{i-1}|)] \cdot b$, where $\boldsymbol{W}$ is the weight matrix between the ith and $(i-1)$th layers and b is the bias vector. The restricting condition is $g_i(x) = \beta_i(x) + \gamma_i(x)$, where $g_i(x)$ is the output of ith layer. $\beta_i(x)$ is considered as the contextual contribution of the input and $\gamma_i(x)$ implies contribution of the input to $g_i(x)$ that is not included in $\beta_i(x)$.

Fig. 6 showcases the evaluation of raw gradients, SmoothGrad, IntegratedGrad, and Deep Taylor methods with a LeNet-5-like network. Among them, IntegratedGrad and Deep Taylor methods perform superbly on five digits.

A mutual-information measure to quantify the association between inputs and latent representations of a deep model can also similarly work as saliency [63], [148], [193]. In addition, there are other methods to obtain saliency maps as well. Ross *et al.* [144] defined a new loss term $\sum_i (\boldsymbol{A}_i(\partial/\partial x_i)\sum_{k=1}^{K} \log(\hat{y}_k))^2$ for training, where i is an index of a pixel, $\mathbf{A}_i$ is the binary mask to be optimized, $\hat{y}_k$ is the k^{th} digit of the label, and K is the number of class. This loss is to penalize the sharpness of gradients toward a clearer interpretation boundary. Fong and Vedaldi [52] explored to learn the smallest region to delete, which is to find the optimal $\boldsymbol{m}^*$

$$\boldsymbol{m}^* = \mathop{\arg\min}_{\boldsymbol{m}\in[0,1]^n} \lambda\|\mathbf{1} - \boldsymbol{m}\|_1 + f_c(x_0; \boldsymbol{m})$$

where $\boldsymbol{m}$ is the soft mask, $f_c(x_0; \boldsymbol{m})$ represents the loss of the network for an image x_0 with the soft mask, and n is the number of pixels. Lei *et al.* [102] utilized a generator to specify segments of an original text as the so-called rationales, which fulfill two conditions: 1) rationales should be sufficient as a replacement for the initial text, and 2) rationales should be short and coherent. Deriving rationales is actually equivalent to deriving a binary mask, which can be regarded as a saliency map. Based on the above two constraints, the penalty term for a mask is formulated as

$$\Omega(z) = \lambda_1\|z\|_1 + \lambda_2 \sum_t |z_t - z_{t-1}|$$

where $z = [z_1, z_2, \ldots]$ is a mask, the first term penalizes the number of rationales, and the second term is for smoothness.

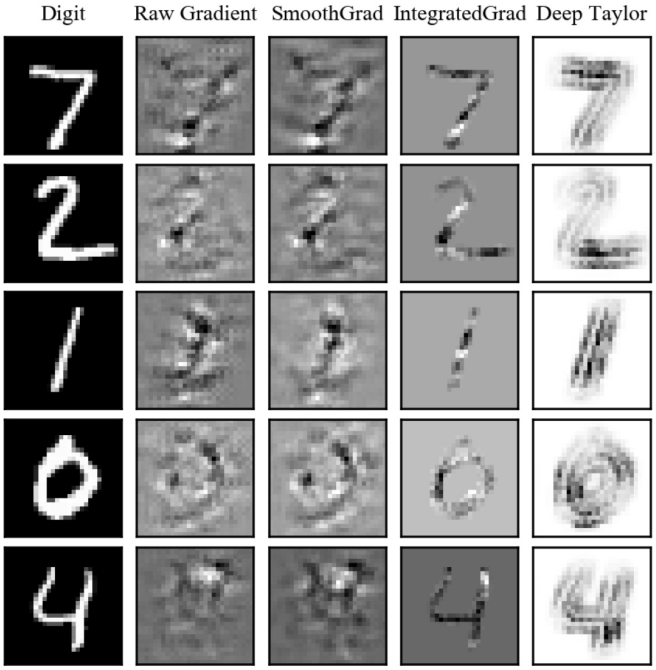


Fig. 6. Interpreting a LeNet-5-like network by raw gradient, SmoothGrad, integrated gradient, and deep Taylor methods, respectively. It is seen that integrated gradient and deep Taylor methods have sharper and less noisy saliency map.

The class activation map method (CAM [209]) and its variant [150] utilized global average pooling before a fully connected layer to derive the discriminative area. Specifically, let $f_k(x, y)$ represent the kth feature map, for a given class c, the input to the softmax layer is $\sum_k w_k^c \sum_{x,y} f_k(x, y)$,where w_k^c is the weight vector connecting the kth feature map and the class c. The discriminative area is obtained as $\sum_k w_k^c f_k(x, y)$, which directly implies the importance of the pixel at (x, y) for class c. What is more, some weakly supervised learning methods such as [135] can obtain discriminative areas as well. Specifically, they trained a network only with object labels; however, when they rescaled the feature maps produced by the max-pooling layer, it was surprisingly found that these feature maps were consistent with the locations of objects in the input.

Proxy: There are about three ways to prototype a proxy. The first one is direct extraction. The gist of direct extraction is to construct a new interpretable model, such as a decision tree [92], [191] or a rule-based system directly from the trained model. As far as the rule extraction is concerned, both decompositional [151] and pedagogical methods [146], [172] can be used. Pedagogical approaches extract rules that enjoy a similar input–output relationship with that of a neural network. These rules do not correspond to the weights and structure of the network. For example, the validity interval analysis (VIA) [118] extracts rules in the following form:

IF (input ∈ a hypercube), THEN input belongs to a certain class.

Setiono and Liu [151] clustered hidden unit activation values based on the proximity of activation values. Then, the activation values of each cluster were denoted by their average activation values, at the same time kept the accuracy of the neural network as intact as possible. Next, the input data with

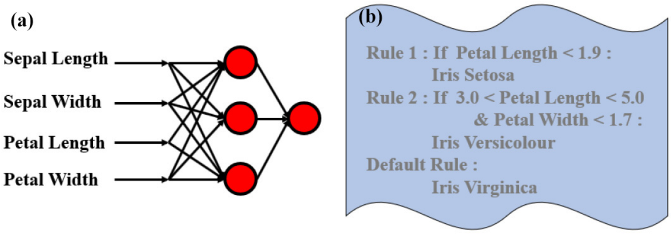


Fig. 7. Rule extraction process as proposed by Setiono and Liu [151]. (a) One-hidden-layer network with three hidden neurons is constructed to classify the Iris dataset. (b) Rules are extracted via discretizing activation values of hidden units and clustering of inputs, where Petal length and Petal width are dominating attributes for classification of Iris samples. The extracted rules have the same classification performance as that of the original neural network.

the same average hidden unit activation value were clustered together to obtain a complete set of rules. In Fig. 7, we illustrate obtained rules from a one-hidden-layer network using Setiono and Liu's method over the Iris dataset. In a neural network for a binary classification problem, decision boundaries divide the input space into two parts, corresponding to two classes, respectively. The explanation system HYPINV developed in [146] computed for each and every decision boundary hyperplane a tangent vector. The sign of an inner product between an input instance and a tangent vector will imply the position of the input instance relative to the decision boundary. Based on such a fact, a rule system can be established.

Finally, some specialized networks, such as ANFIS [80] and RBF networks [126], straightforwardly correspond to fuzzy logic systems. For example, an RBF network is equivalent to the Takagi–Sugeno rule system [171] that comprises rules, such as "if $x \in$ set A and $y \in$ set B, then $z = f(x, y)$" [136]. Fuzzy logic interpretation in [48] considers each neuron/filter in a network as a generalized fuzzy logic gate. In this view, a neural network is nothing but a deep fuzzy logic system. Specifically, they analyzed a new type of neural networks, called quadratic networks, in which all the neurons are quadratic neurons that replace the inner product with the quadratic operation [47]. Their interpretation generalized fuzzy logic gates implemented by quadratic neurons and then computed the entropy based on spectral information of fuzzy operations in a network. It was suggested that such an entropy could have deep connections with properties of minima and the complexity of neural networks.

The second one is called knowledge distillation [23], as shown Fig. 8. Although knowledge distillation techniques are mostly used for model compression, their principles can also be used for interpretability. The motif of knowledge distillation is that cumbersome models can generate relatively accurate predictions, assigning probabilities to all the possible classes, known as soft labels, that are more informative than one-hot labels. For example, a horse is more likely to be classified as a dog instead of a mountain. But, with one-hot labeling, both the dog class and mountain class have zero probability. It was shown in [23] that by the means of matching the logits of the original model, the generalization ability of the original cumbersome model could be transferred into a simpler model.

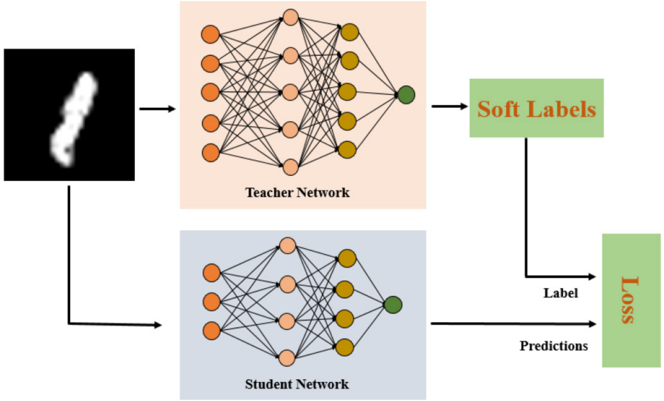


Fig. 8. Knowledge distillation is to construct an interpretable proxy by the soft labels from the original complex models.

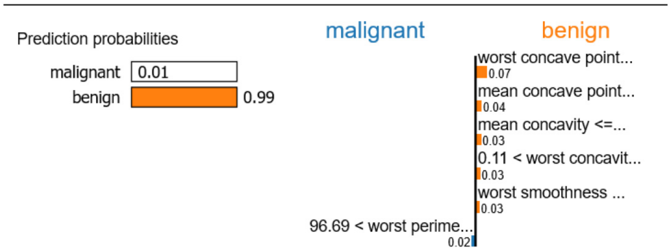


Fig. 9. Breast cancer classification task model dissected by LIME. In this case, the sample is classified as benign, where the worst concave point, mean concave point, and so on are contributing forces, while the worst perimeter is the contributing force to drive the model to predict "malignant."

Along this direction, an interpretable proxy model, such as a decision tree [38], [185], a decision set [98], a global additive model [170], and a simpler network [75] were developed. For example, Tan *et al.* [170] used soft labels to train a global additive model in the form of $h_0+\sum_i h_i(x_i)+\sum_{i\neq j} h_{ij}(x_i, x_j)+\sum_{i\neq j}\sum_{j\neq k} h_{ijk}(x_i, x_j, x_k)+\cdots$, where $\{h_i\}_{i\geq 1}$ could also work as a feature saliency map directly.

The last one is to provide a local explainer as a proxy. Local explainer methods locally mimic the predictive behaviors of a neural network. The basic rationale is that when a neural network is inspected globally, it looks complex. However, if we tackle it locally, the picture becomes clearer.

One typical local explainer is local interpretable model-agnostic explanation (LIME) [141], which synthesizes a number of neighbor instances by randomly setting elements of that sample to 0 and computing the corresponding outcomes. Then, a linear regressor is used to fit synthesized instances, where the coefficients of the linear model signify the contributions of features. As Fig. 9 shows, the LIME method is applied to a breast cancer classification model to identify which attributes are contributing forces for the model's benign or malignant prediction.

Zhang *et al.* [206] pointed out the lack of robustness in the LIME explanation, which originates from sampling variance, sensitivity to the choice of parameters, and variation across different data points. *Anchor* [142] is an improved extension of LIME, which is to find the most important

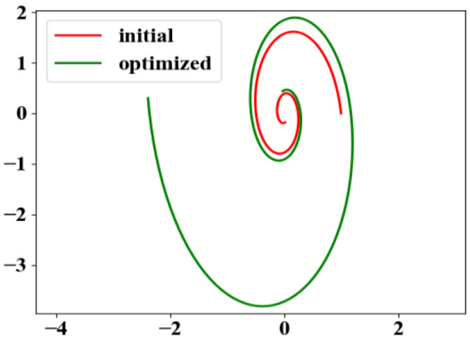


Fig. 10. ODE-Net optimizes the start point and the dynamics to fit the spiral shape.

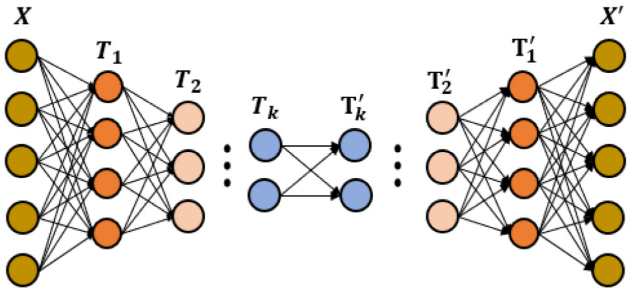


Fig. 11. Application of the information bottleneck theory to compare mutual information between symmetric layers in an autoencoder.

segments of an input such that the variability of the rest segments does not matter. Mathematically, *Anchor* searches a set: $A = \{z|f(z) = f(x), z \in x\}$, where $f(\cdot)$ is a black-box model, x is the input and z is the part of x. Another proposal local rule-based explanation (LORE) is from [64]. The LORE takes advantage of the genetic algorithm to generate the balanced neighbors instead of random neighbors, thereby yielding high-quality training data that alleviates sampling variance of LIME.

Advanced Mathematical/Physical Analysis: Lu *et al.* [114] showed that many residual networks can be explained as discretized numerical solutions of ordinary differential equations; i.e., the inner working of a residual block in ResNet [69] can be modeled as $u_{n+1} = u_n + f(u_n)$, where u_n is the output of the nth block, and $f(u_n)$ is the block operation. It was noted that $u_{n+1} = u_n + f(u_n)$ is a one-step finite difference approximation of an ordinary differential equation $du/dt = f(u)$. This idea inspired the invention of ODE-Net [32]. As Fig. 10 shows, the starting point and the dynamics are tuned by an ODE-Net to fit a spiral.

Lei *et al.* [101] constructed an elegant connection between the Wasserstein generative adversarial network (WGAN [8]) and the optimal transportation theory. They concluded that with low dimensionality hypothesis and the intentionally designed distance function, a generator and a discriminator can exactly represent each other in a closed form. Therefore, the competition between a discriminator and a generator in WGAN in the training is unnecessary.

In [153], it was proposed that the learning of a neural network is to extract the most relevant information in the input random variable X that pertains to an output random variable Y. Naively, for a feedforward neural network, the following inequality of mutual information holds:

$$I(Y; X) \geq I\left(Y; h_j\right) \geq I(Y; h_i) \geq I\left(Y; \hat{Y}\right)$$

where $I(\cdot;\cdot)$ denotes the mutual information, h_i and h_j are outputs of hidden layers ($i > j$ means that the ith layer is deeper), and $\hat{Y}$ is a final prediction. Furthermore, Yu and Principe [197] employed an information bottleneck theory to gauge the mutual information states of symmetric layers in a stacked autoencoder as shown in Fig. 11

$$I\left(X; X'\right) \geq I\left(T_1; T'_1\right) \geq \cdots \geq I\left(T_K; T'_K\right).$$

However, it is tricky to estimate the mutual information since the probabilistic distribution of data is usually unknown as a priori.

Kolouri *et al.* [91] built an integral geometric explanation for neural networks with a generalized Radon transform. Let X be a random variable for the input, which conforms to the distribution p_X, then we can derive a probability distribution function for the output of a neural network $f_\theta(X)$ parametrized with θ: $p_{f_\theta}(z) = \int_X p_X(x)\delta(z - f_\theta(x))dx$, which is the generalized Radon transform, and the hypersurface is $H(t, \theta) = \{x \in X|f_\theta(x) = t\}$. In this regard, the transform by a neural network is characterized by the twisted hypersurfaces. Huang [77] used the mean-field theory to characterize the mechanism of dimensionality reduction by a deep network that assumes weights in each layer and input data following a Gaussian distribution. In his study, the self-covariance matrix of the output of the lth layer was computed as C^l, then the intrinsic dimensionality was defined as $D = [((\sum_{i=1}^N \lambda_i)^2)/(\sum_{i=1}^N \lambda_i^2)]$, where λ_i is the eigenvalue of C^l, and N is the number of eigenvalues. The quantity D/N was investigated across layers to analyze how compact representation are learned across layers. Ye *et al.* [192] utilized a framelet theory and a low-rank Hankel matrix to represent signals in terms of their local and nonlocal bases, corresponding to convolution and generalized pooling operations. However, in their study, the network structure was simplified in concatenating two ReLU units into a linear unit such that the nonlinearity from ReLU units could be circumvented. As far as advanced physics models are concerned, Mehta and Schwab [121] built an exact mapping from the Kadanoff variational renormalized group [82] to the restricted Boltzmann Machine (RBM) [147]. This mapping is independent of forms of the energy functions and can be scaled to any RBM.

Theoretical neural network studies are essential to interpretability as well. Currently, theoretical foundations of deep learning are primarily from three perspectives: 1) representation; 2) optimization; and 3) generalization.

Representation: Let us include two examples here. The first example is to explain why deep networks are superior to the shallow ones. Recognizing successes of deep networks, Cohen *et al.* [37], Eldan and Shamir [44], Liang and Srikant [109], Mhaskar and Poggio [124], and Szymanski and McCane [169] justified that a deep network is more expressive than a shallow one. The basic idea is to construct a special class of functions that can be efficiently represented by a deep network but hard to be approximated by a shallow one. The second example is to understand utilities of shortcut connections of deep networks. Veit *et al.* [177] showed that residual connections can render a neural network to manifest an ensemble-like behavior. Along this direction, it was reported in [110] that with shortcuts, a network can be super slim to allow for universal approximation.

Optimization: Generally, optimizing a deep network is an NP-hard nonconvex problem. The pervasive existence of saddle points [56] leads to that even finding a local minimum is also NP-hard [5]. Of particular interest to us is why an over-parametrized network can still be optimized well because a deep network is a kind of over-parametrized networks. The character of an over-parameterized network is that the number of parameters in a network much exceeds the number of data instances. Soltanolkotabi *et al.* [162] showed that when data are Gaussian distributed and activation functions of neurons are quadratic, the landscape of an over-parameterized one-hidden-layer network allows global optimum to be searched efficiently. Nguyen and Hein [130] demonstrated that with respect to linearly separable data, under assumptions on the rank of weight matrices of a feedforward neural network, every critical point of a loss function is a global minimum. Furthermore, Jacot *et al.* [78] showed that when the number of neurons in each layer of a neural network goes infinitely large, the training only renders small changes for the network function. As a result, the training of the network turns into the kernel ridge regression.

Generalization: The conventional generalization theory is incompetent to explain why a deep network can generalize well despite that the number of parameters of a deep network is many more than the number of samples. Recently proposed generalization bounds [127] that rely on the norm of weight matrices partially solved this problem. However, these bounds have an abnormal dependence on data that more data lead to a larger generalization bound, which apparently contradicts the common sense. We prospect that more efforts are needed to resolve the generalization puzzle satisfactorily [18], [122].

Explaining-by-Case: Basically, case-based explanations present a case that is believed by a neural network to be the most similar to the query case needing an explanation. Finding a similar case for explanation and selecting a representative case from data as the prototype [19] are basically the same thing and just use different metrics for similarity. While prototype selection is to find a minimal subset of instances that can represent the whole dataset, case-based explanations use the similarity metric based on the closeness of representations of a neural network, thereby exposing the hidden representation information. In this light, case-based explanations are also related to deep metric learning [149].

As shown in Fig. 12, Wallace *et al.* [180] employed the k-nearest neighbor algorithm to obtain the most similar cases for the query case in the feature space and then computed the percentage of the nearest neighbors belonging to the expected class as a measure for interpretability, suggesting how much a prediction is supported by data. Chen *et al.* [31] constructed a model that could dissect images by finding prototypical parts.

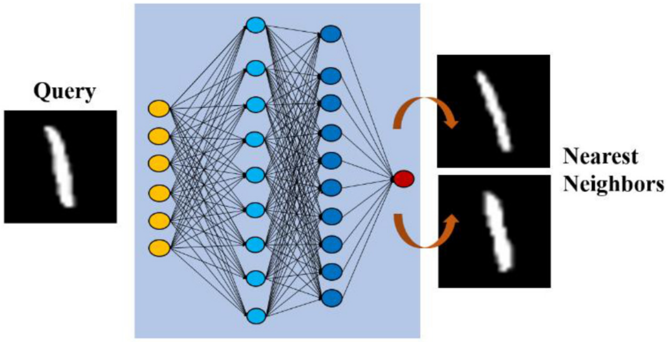


Fig. 12. Explaining-by-case presents the nearest neighbors in response to a query.

Specifically, the pipeline of the model splits into multiple channels after convolutional layers, in which the function of each channel is expected to learn a prototypical part of the input, such as the head or body of a bird. The decision for an input image is made based on the similarity of features of channels.

Wachter *et al.* [179] offered a novel case-based explanation method by providing a counterfactual case, which is an imaginary case that is close to the query but has a different output from that of the query. Counterfactual explanation provides the so-called "closest possible case" or the smallest change to yield a different outcome. For example, counterfactual explanations may produce the following statement: "If you have a good striker, your team would have won this soccer game." Coincidentally, techniques to generate a counterfactual explanation have been developed for the purpose of "adversarial perturbation"; i.e., structural attack [190]. Essentially, finding the closest possible case x' to the input x is equivalent to finding the smallest perturbation to x such that the classification result changes. For example, the following optimization can be built:

$$\underset{x'}{\operatorname{argmin}} \; \lambda\big(f\big(x'\big) - y'\big)^2 + d(x, x')$$

where λ is a constant, y' is a different label, and $d(\cdot, \cdot)$ is chosen to be the Manhattan distance in hope that the input be minimally perturbed. Goyal *et al.* [62] explored an alternative way to derive a counterfactual visual explanation. Given an image I with a label c, since the counterfactual visual explanation represents the change for the input that can force the model to yield a different prediction class c', they selected an image I' with a label c' and managed to recognize the spatial region in I and I' such that the replacement of the recognized region would alter the model prediction from c to c'.

Explaining-by-Text: Neural image captioning uses a neural network to produce a natural language description for an image. Despite that neural image captioning is initially not for network interpretability, descriptive language about images can tell the information about how a neural network analyzes an image. One representative method is from [84] that combines a convolutional neural network and a bidirectional RNN to obtain a bimodal embedding. Due to the hypothesis that the two embeddings representing similar semantics across two modalities should share the nearby locations of two spaces, the

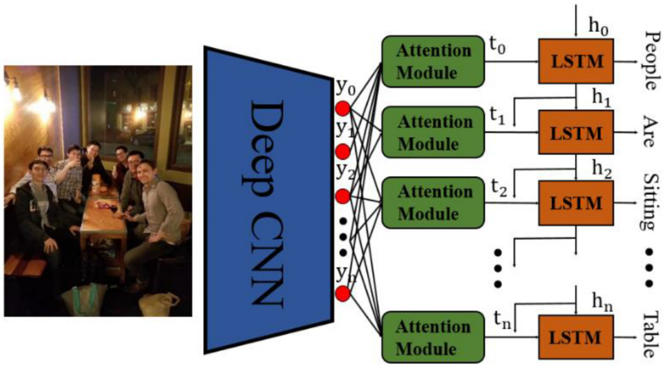


Fig. 13. Image captioning with attention modules provides an explanation to the features mined by a deep convolutional network.

objective function is defined as

$$S_{IT} = \sum_{t \in g_T} \max_{i \in g_I} v_i^T s_t$$

where v_i is the ith image fragment in the set g_I, and s_t is the tth word in a sentence g_T. Another representative method is the attention mechanism [137], [178], [188], [189], where deep features are to align the corresponding text descriptions by a recursive neural network such as LSTM [73]. An explanation for deep features is provided by the corresponding words in the text and attention maps, which reflect which parts of an image attract the attention of the neural network.

As shown in Fig. 13, in the kth attention module that takes $y_0, y_1, \ldots, y_n$ as input, suppose its output is $t_k = \sum_i y_i s_{ki}$. $s_{k0}, s_{k1}, \ldots, s_{kn}$ together form an attention map for t_k with respect to the associated word. However, Jain and Wallace [79] argued that an attention map is not qualified as an explanation because they observed that the attention map was not correlated with other importance measures of features such as gradient-based measures, and the change of attention weights yielded no changes in prediction.

C. Ad-Hoc Interpretable Modeling

Interpretable Representation: Traditionally, regularization techniques for deep learning are primarily designed to avoid overfitting. However, it is also feasible to devise regularization techniques to enhance an interpretable representation in terms of decomposability [33], [164], [181], [204], monotonicity [194], nonnegativity [34], sparsity [166], human-in-the-loop prior [96], and so on.

For example, Chen *et al.* [33] invented *Info*GAN, which is a simple but effective way to learn an interpretable representation. Traditionally, a generative adversarial network (GAN) [60] imposes no restrictions on how a generator utilizes the noise. In contrast, *Info*GAN maximizes the mutual information between the latent codes and observations, forcing each dimension of noise to encode a semantic concept. Particularly, the latent codes are made of discrete categorical codes and continuous style codes. As shown in Fig. 14, two style codes control the localized part and the digit rotation, respectively.

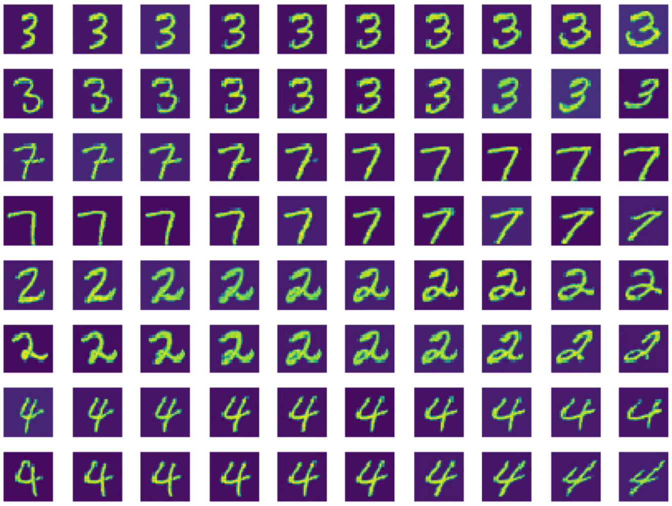


Fig. 14. In an *Info*GAN, two latent codes control the localized parts and rotation parts, respectively.

Incorporating monotonicity constraints [194] is also useful to enhance interpretability. A monotonical relationship means when the value of a specified attribute increases, the predictive value of a model either increases or decreases. Such a simplicity promotes interpretability as well. Chorowski and Zurada [34] imposed non-negativity to weights of neural networks and argued that it could improve interpretability because it eliminated the cancelation and aliasing effects among neurons. Subramanian *et al.* [166] employed a k-sparse autoencoder for word embedding to promote sparsity in the embedding and claimed that this enhanced interpretability because a sparse embedding reduced the overlap between words. Lage *et al.* [96] proposed a novel human-in-the-loop evaluation in selecting a model. Specifically, a diverse set of models were trained and sent to users for evaluation. Users were asked to predict the label of a data point that would be assigned by a model M. The shorter the response time was, the better a user understood the model. Then, the model with the lowest response time was chosen.

Model Renovation: Chu *et al.* [35] proposed to use piecewise linear functions as activations for a neural network (PLNN); thereby, the decision boundaries of PLNN could be explicitly defined and further a closed-form solution could be derived for predictions of a network. As Fig. 15 shows, Fan *et al.* [49] proposed soft-autoencoder (Soft-AE) by using adaptable soft-thresholding units in encoding layers and linear units in decoding layers. Consequently, Soft-AE can be interpreted as a learned cascaded wavelet adaptation system.

Fan [50] explained a neural network as a generalized Hamming network, whose neurons compute the generalized Hamming distance: $h(\boldsymbol{x}, \boldsymbol{w}) = \sum_{l=1}^{L} w_l + \sum_{l=1}^{L} x_l - 2\boldsymbol{x} \cdot \boldsymbol{w}$ for an input $\boldsymbol{x} = (x_1, \ldots, x_L)$ and a weight vector $\boldsymbol{w} = (w_1, \ldots, w_L)$. The bias term in each neuron is specified as $b = -(1/2)(\sum_{l=1}^{L} w_l + \sum_{l=1}^{L} x_l)$ so that each neuron is a generalized Hamming neuron. In this regard, the function of the batch normalization is demystified as making the bias suitable for computation of the generalized Hamming distance.

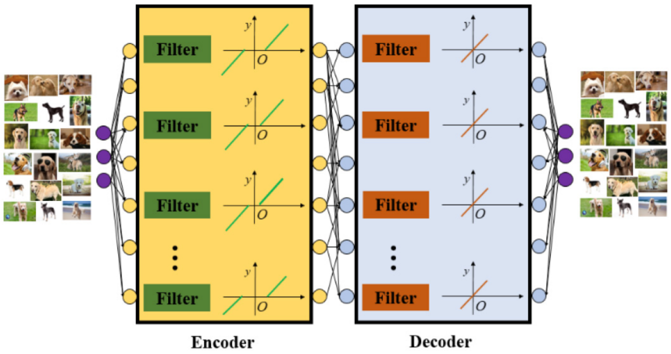


Fig. 15. Soft-AE with soft-thresholding functions as activation functions in the encoding layers and linear functions as activations in the decoding layers, thereby admitting a direct correspondence to the wavelet adaptation system.

Kuo *et al.* [95] proposed a transparent design for constructing a feedforward convolutional network without the need of backpropagation. Specifically, filters in convolutional layers were built by selecting principal components of PCA for outputs of earlier pooling layers. A fully connected layer was constructed by treating it as a linear-squared regressor.

Melis and Jaakkola [123] claimed that a neural network model f is interpretable if it has the form that $f(x) = g(\theta_1(x)h_1(x), \dots, \theta_k(x)h_k(x))$, where $h_i(x)$ is the prototypical concept from the input x and $\theta_i(x)$ is the relevance associated with that concept, g is monotonic and completely additively separable. Such a model can learn interpretable basis concepts and facilitate saliency analysis. Similarly, Vaughan *et al.* [176] designed a network structure to compatibly learn the function formulated as $f(x) = \mu + \gamma_1 h_1(\beta_1^T x) + \gamma_2 h_2(\beta_2^T x) + \cdots + \gamma_K h_K(\beta_K^T x)$, where β_k is the projection, $h_k(\cdot)$ represents the nonlinear transformation, μ is the bias, and γ_k is the weighting factor. Such a model is more interpretable than a general network, because the function of this model has simpler partial derivatives that can simplify saliency analysis, statistical analysis, and so on.

Li *et al.* [104] proposed deep supervision by using prior hierarchical tasks on features of intermediate layers. Specifically, we have a dataset $\{(x, y_1, \dots, y_m)\}$, where labels $y_1, \dots, y_m$ are hierarchical that $y_j,\ j < i$ is a strictly necessary condition for the existence of $y_i,\ i > 1$. Such a scheme introduces a modularized idea that through the supervision of a specific task for an intermediate layer, the learning of that layer is steered toward the prespecified task, thereby gaining interpretability.

Wang [182] proposed to use an interpretable and insertable substitute on a subset of data which the complex black-box model overkills. In their work, a rule set was built as an interpretable model to make a decision on the input data first. Those inputs which a rule set was handicapped to classify were passed into the black-box model for decision making. The logic of this hybrid predictive system is that an interpretable model for regular cases without compromising accuracy, a complex black-box model for complicated cases.

Jiang *et al.* [81] proposed finite automata-RNN (FA-RNN) that can be directly transformed into the regular expressions such that a good interpretability is extracted. The roadmap is that the constructed FA-RNN can be approximated into finite automata, and further transformed into regular expressions because finite automata and a regular expression are mutually convertible. In analogy, a regular expression can also be decoded into an FA-RNN as an initialization. FA-RNN is a good example to manifest the synergy between a rule system and a neural network.

III. Interpretability in Medicine

These days, reports are often seen in the news that deep learning-based algorithms outperform experts or classic algorithms in the field of medicine [152]. Indeed, given adequate computational power and well-curated datasets, a properly designed model can deliver competitive performance in most well-defined pattern recognition tasks. However, due to the high stakes of medicine-concerned applications, it is not sufficient to have a deep learning model that produces correct answers without an explanation. In this section, we focus on several exemplary papers concerning applications of interpretability methods in medicine and we organize the articles of relevance in accordance with the aforementioned taxonomy.

A. Post-Hoc Interpretability Analysis

Feature Analysis: Van Molle *et al.* [175] visualized convolutional neural networks to assist decision making for skin lesion classification. In their work, feature activations generated from the last two convolutional layers were rescaled to the size of an input image as the activation maps. Where a map has high activations were inspected. The activation strengths across different border types, skin colors, skin types, etc., were compared. The activation map exposed a risk that some unexpected regions had uncommonly high activations.

Bychkov *et al.* [24] utilized a model that combines a VGG-16 network [157] and an LSTM network [73] to predict five-year survival of colorectal cancer based on digitized tumor tissue samples. In their work, an RGB pathological image was split into many tiles. A VGG-16 network extracted a high-dimensional feature vector from each tile, which was then fed into an LSTM network to predict five-year survival. They used t-SNE [116] to map features learned by VGG-16 into a 2-D space for visualization and found that different classes of features of VGG-16 were well separated.

Saliency: Sturm *et al.* [165] applied a deep network with LRP [11] for the single-trial EEG [22] classification. The network entails two linear mean pooling layers before being activated or normalized. The feature importance score is assigned by LRP.

Zech *et al.* [199] developed a deep learning model for chest radiography to classify patients into having pneumonia or not. Through interpretability analysis by CAM [209], they reported the risk that a deep learning model could make an incorrect decision by capturing features irrelevant to diseases, such as metal tokens.

Oktay *et al.* [134] combined attention gates with the decoder part of U-Net to cope with interpatient variation in organs' shapes and sizes. The proposed model can improve model sensitivity and accuracy by inhibiting representations of irrelevant regions. Aided by attention gates, they found that the model gradually shifted its attention to regions of interest.

Ardila *et al.* [7] proposed a deep learning algorithm that considers a patient's current and previous CT volumes to predict the risk of lung cancer. They used the integrated gradient method [167] to derive saliency maps and invited experienced radiologists to examine the fidelity of these maps. It turned out that in all cases, the readers strongly agreed that the model indeed focused on the nodules.

Lee *et al.* [100] reported an attention-assisted deep learning system for detection and classification of acute intracranial haemorrhage, where an attention map identified a region relevant to the disease. They evaluated the localization accuracy of the attention maps by computing the proportion of bleeding points overlapping with the attention maps. Overall, it was found that 78.1% bleeding points were detected in the attention maps.

Caicedo-Torres and Gutierrez [25] proposed a multiscale deep convolutional neural network for the mortality prediction based on the measurement of 22 different items in ICU, such as the sodium index, urine output, etc. In their work, three temporal scales were represented by stacking convolutional kernels of dimensions 3 × 1, 6 × 1, and 12 × 1. The saliency map by DeepLIFT [155] was utilized for interpretability.

Guo *et al.* [66] introduced an effective dual-stream network that conjugates extracted features from ResNet [69] and clinical prior knowledge to predict the mortality risk of patients based on low-dose CT images. To further testify the effectiveness of the proposed model, they utilized t-SNE [116] to reduce the dimensionality of feature maps of malignant and benign samples and found that malignant and benign features were well separated. Also, they applied CAM [209] to reveal that the deceased subjects correctly classified by the model were prone to have strong activations.

Proxy: Che *et al.* [29] applied knowledge distillation into a deep model to learn a gradient boosting tree (GBT) [106], which provides not only robust prediction performance but also a good interpretability in the context of electronic health record prediction. Specifically, they trained three deep models, respectively, and then used predictions of deep models as labels to train a GBT model. Experiments on a Pediatric ICU dataset were reported that the GBT model maintained the prediction performance of deep models in terms of mortality and ventilator-free days.

Pereira *et al.* [138] combined global and local interpretation efforts for brain tumor segmentation and penumbra estimation in stroke lesions, where the global interpretability was derived from mutual information to sense the dependence between an input sample and the prediction, while the local interpretability was cast by a variant of LIME [141].

Explaining-by-Case: Codella *et al.* [36] employed saliency and explaining-by-case methods to explain a dermoscopic image analysis network, which was jointly trained by disease labels with a triple-let loss. Specifically, the interpretability was gained by the discovered neighbors and localized regions that were most relevant to the distance from queries and neighbors.

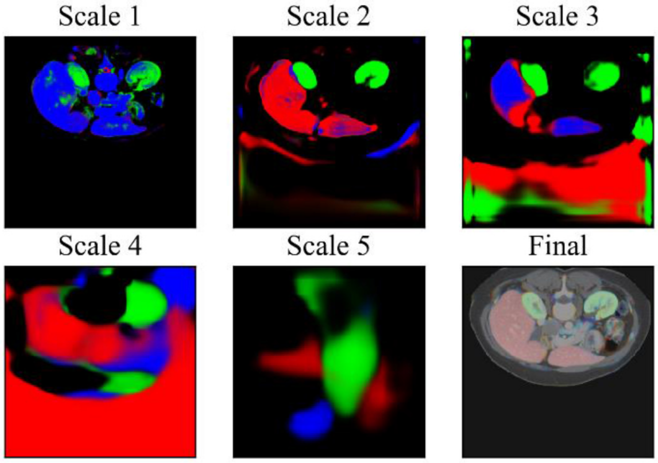


Fig. 16. Visualization of feature maps of different arms in PIPO-FAN, where low-scale subnetworks produce local structural details and high-scale subnetworks target global morphological information.

Explaining-by-Text: Zhang *et al.* [207] proposed an all-in-one network that reads pathology bladder cancer images, generated diagnostic reports, retrieved images according to symptomatic descriptions, and visualized attention maps. They designed an auxiliary attention sharpening module to improve the discriminability of attention maps. Pathologists' feedbacks suggested that the explanatory maps tended to highlight regions that concern with carcinoma-informative regions.

B. Ad-Hoc Interpretable Modeling

Interpretable Representation: Fang and Yan [51] devised the pyramid input pyramid output feature abstraction network (PIPO-FAN) with multiple arms for multiorgan segmentation. Each of the arm handles the information on one scale. The total loss is obtained by adding the segmentation loss to each of these arms such that segmentation-wise features are generated in each arm. Visualization analysis suggested that features from different arms have hierarchical semantical meanings; i.e., some are blurry but contain global classwise information, while the others contain local boundary information. As shown in Fig. 16, the segmentation loss creates semantically meaningful features, where low-scale arms produce more details and high-scale arms find global morphologies.

Model Renovation: Gale *et al.* [55] combined a DenseNet [76] model with an LSTM model [73] for the detection of hip features from pelvic X-ray radiographs. A radiologist hand-labeled standard descriptive terms to construct a semantic dataset for these radiographs. Their model consistently generated informative sentences favored by doctors over saliency maps. Also, they demonstrated that the combination of visualization and text interpretation gives an interpretation superior to either of them alone.

Biffi *et al.* [20] employed a variational autoencoder (VAE) [87]-based model for the classification of cardiac diseases as well as structurally remodeling based on cardiovascular images. In their scheme, registered left ventricular (LV) segmentations at end-diastolic (ED) and end-systolic (ES) phases were encoded in a low-dimensional latent space by VAE. The learned latent low-dimensional manifold was connected to a multilayer perceptron (MLP) for disease classification. The interpretation was given by an activation maximization technique. The “deep dream” of the MLP was derived and inverted to the image space for visualization.

Shen *et al.* [154] built an interpretable deep hierarchical semantic convolutional neural network (HSCNN) to predict the malignancy of pulmonary nodules in CT images. HSCNN consists of three modules: 1) a general feature learning module; 2) a low-level task module that predicts semantic characteristics, such as sphericity, margin, subtlety, and so on; and 3) a high-level task module absorbs information from both general features and low-level task predictions to produce an overall lung nodule malignancy. Due to the semantic meaning contained in the low-level task, HSCNN has boosted interpretability.

Zhang *et al.* [208] developed a deep convolutional network to automate the whole-slide reading of pathology images for tumors and the diagnosis process of pathologists. Specially, the network can generate a clinical pathology report along with attention-assisted features.

Lei *et al.* [103] observed that CAM [209] and Grad-CAM [150] are for interpreting localization tasks and tend to ignore fine-grained structures. Consequently, they proposed a shape-and-margin-aware soft activation map (SAM) that could probe subtle but critical features in a lung nodule classification task. The comprehensive experimental comparisons showed that compared to CAM and Grad-CAM, SAM could reveal relatively discrete and irregular features around nodules.

IV. Perspective

In this section, we suggest a few directions, in hope to advance the understanding and practice of artificial neural networks.

Synergy of Fuzzy Logic and Deep Learning: Fuzzy logic [198] was a buzz phrase in the last nighties. It extends the Boolean logic from 0–1 judgement to imprecise inference with fuzziness in the interval [0, 1]. The fuzzy theory can be divided into two branches: 1) fuzzy set theory and 2) fuzzy logic theory. The latter, with an emphasis on “IF-THEN” rules, has demonstrated effectiveness in dealing with a plethora of complicated system modeling and control problems. Nevertheless, a fuzzy rule-based system is restricted by the acquisition of a large number of fuzzy rules, a process that is tedious and computationally expensive. While a neural network is a data-driven method that extracts knowledge from data through training, with the knowledge represented by neurons in a distributed manner. However, a neural network falls short of delivering a satisfactory result in the context of small data and suffers from the lack of interpretability. In contrast, a fuzzy logic system employs experts’ knowledge and represents a system in the form of IF-THEN rules. Although a fuzzy logic system merits interpretability and accountability, it is incompetent in efficient and effective knowledge acquisition. It seems that a neural network and a fuzzy logic system are complementary to each other. Therefore, it is instrumental to combine the strengths of two worlds toward an enhanced interpretability. In fact, this roadmap is not totally new. There have been several combinations along this direction: ANFIS model [80], generic fuzzy perceptron [126], RBF networks [21], and so on.

One suggestion is to build a deep RBF network. Given the input vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, an RBF network is expressed as $f(\boldsymbol{x}) = \sum_i^n w_i \phi_i(\boldsymbol{x} - \boldsymbol{c}_i)$, where $\phi_i(\boldsymbol{x}-\boldsymbol{c}_i)$ is usually selected as $\exp(-[(||\boldsymbol{x} - \boldsymbol{c}_i|| \wedge 2)/(2\sigma^2)])$, where $\boldsymbol{c}_i$ is the cluster center of the ith neuron. It was proved the functional equivalence between an RBF network and a fuzzy inference system under mild conditions [21]. Also, an RBF network is shown to be a universal approximator [136]. Hence, an RBF network is a potentially sound vehicle that can encode fuzzy rules into its adaptive representation without loss of accuracy. Reciprocally, rule generation and fuzzy rule representation in an adaptable RBF network are more straightforward compared to an MLP. Although the current RBF networks are of one-hidden-layer structures, it is feasible to develop deep RBF networks, which can be viewed as a deep fuzzy rule system. A greedy layerwise training algorithm was developed in [71], which successfully solved the training problem for deep networks. It is possible to translate such success for training of deep RBF networks. Then, the correspondence between a deep RBF network and a deep fuzzy logic system can be applied to obtain a deep fuzzy rule system. We believe that efforts should be made to synergize fuzzy logic and deep learning techniques aided by big data along this direction.

Convergence of Neuroscience and Deep Learning: Up to date, truly intelligent systems are only human. The artificial neural networks in their earlier forms were clearly inspired by biological neural networks [120]. However, subsequent developments of neural networks were, to a much less degree, pushed by neurological and biological insights. As far as interpretability is concerned, since biological and artificial neural networks are deeply connected, advances in neuroscience should be relevant and even instrumental to the development and interpretation of deep learning techniques. We believe that the neuroscience has a great potential of deep learning interpretability in the following aspects.

Cost Function: The effective use of cost functions is a driving force for the development of deep networks in the past years; for example, the adversarial loss used in GANs [60]. In the previous sections, we have highlighted cases that an appropriate cost function can enable a model to learn an interpretable representation, such as enhance feature disentanglement. Along this direction, a myriad of cost functions can be built to reflect biologically plausible rationales. Indeed, our brain can be modeled as an optimization machine [119], which has a powerful credit assignment mechanism to form a cost function.

Optimization Algorithm: Despite the huge success by backpropagation, it is far from ideal in the view of neuroscience. Truly in many senses, backpropagation fails to manifest how a human neural system handles the synapses of a neuron. For example, in a biological neural system, synapses are updated in a local manner [94] and only depend on the activities of

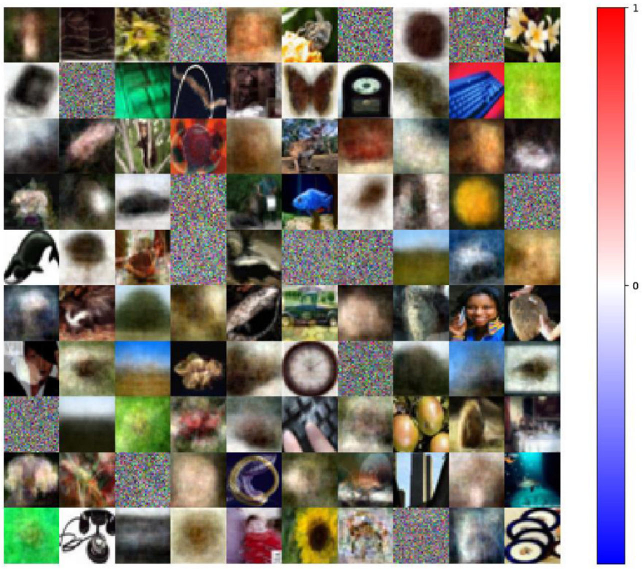


Fig. 17. Visualization of weights of a network learned by a bio-plausible algorithm, where prototypes of training images are captured [94].

presynaptic and postsynaptic neurons. However, connections in deep networks are tuned through nonlocal backpropagation. Fig. 17 shows a bio-plausible learning algorithm for a two-layer network on CIFAR-100 [93]. Additionally, a neuromodulator is missing in deep networks in contrast to the inner working of a human brain, where the state of one neuron can exhibit different input–output patterns controlled by a global neuromodulator, such as dopamine, serotonin, and so on [161]. Neuromodulators are believed to be critical due to their ability to selectively control on and off states of one neuron, which is equivalent to modifying the involved cost function [13].

Considering that there are quite few studies discussing the interpretability of training algorithms, powerful and interpretable training algorithms will be highly desirable. Just like what were proved for classic optimization methods, we wish that future nonconvex optimization algorithms will have some kinds of uniqueness, stability, and continuous dependency on data, etc.

Bio-Plausible Architectural Design: In the past decades, neural networks were designed in diverse architectures from simple feedforward networks to deep convolutional networks and other highly sophisticated networks. The structure determines functionality; i.e., a specific network architecture regulates the information flow with distinct characteristics. Therefore, specialized architectures are useful as effective solutions for intended problems. Currently, the structural differences between deep learning and biological systems are eminent. A typical network is used and tuned for most tasks based on big data, while a biological system learns from a small number of data and generalizes very well. Clearly, a huge amount of knowledge needs to be learned from biological neural networks so that more desirable and explainable neural network architectures can be designed.

Interpretability in Medicine: A majority of interpretability research efforts in medicine are currently for classification tasks, but radiological imaging covers a large variety of tasks, such as image segmentation, registration, reconstruction, and so on. Clearly, interpretability is also closely relevant to these areas and, therefore, it is in need to promote interpretability research in these domains. On one hand, more efforts should be made to extend the existing interpretation methods to other tasks that have not been explored. On the other hand, practitioners can design task-specific interpretation methods with their expertise and insights. For example, explaining why a voxel receives a class label in image segmentation is much harder than explaining which area in the input image is responsible for a prediction in image classification. Similarly, for image reconstruction, interpretability could be quite complicated. In this regard, our recently proposed ACID framework allows a synergistic integration of data-driven priors and compressed sensing (CS)-modeled priors, enforcing both of which iteratively via physics-based analytic mapping [187]. By doing so, modern CS and state-of-the-art deep networks are united to overcome the vulnerabilities of existing deep reconstruction networks, at the same time transferring the interpretability of the model-based methods to the hybrid DNNs.

In addition to the above referenced publications, gaining interpretability ultimately also relies on medical doctors, who have invaluable professional training despite some biases and errors. As a result, active collaboration among medical doctors, technical experts, and theoretical researchers will be an important avenue for future development of deep learning methods.

V. Conclusion

In conclusion, we have reviewed key ideas, implications, limitations of the existing interpretability studies, and illustrated some typical interpretation methods through examples. In doing so, we have depicted a holistic landscape of interpretability research using our proposed taxonomy and introduced applications of interpretability in medicine. Figs. 3, 5, 6, 7, 9, 10, 16, and 17 are visualization results from our own implementations of the chosen interpretation methods. We have publicly shared relevant codes in the GitHub (https://github.com/FengleiFan/IndependentEvaluation). There is no doubt that a unified and accountable interpretation framework is critical to elevate interpretability research into a new phase. In the future, more efforts are needed to reveal the essence of deep learning. Because this field is highly interdisciplinary and rapidly evolving, there are great opportunities ahead that will be both academically and practically rewarding.

Acknowledgment

The authors are grateful for Dr. Hongming Shan's suggestions (Fudan University) and anonymous reviewers' advice.

References

[1] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, no. 1, pp. 39–59, 1994.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[3] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. NeurIPS*, 2018, pp. 9525–9536.

[4] P. Adler *et al.*, "Auditing black-box models for indirect influence," *Knowl. Inf. Syst.*, vol. 54, no. 1, pp. 95–122, 2018.

[5] A. Anandkumar and R. Ge, "Efficient approaches for escaping higher order saddle points in non-convex optimization," in *Proc. COLT*, 2016, pp. 81–102.

[6] M. Ancona, C. Öztireli, and M. H. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley values approximation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, May 2019, pp. 272–281.

[7] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nat. Med.*, vol. 25, no. 6, pp. 954–961, 2019.

[8] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017. [Online]. Available: arXiv:1701.07875.

[9] L. Arras, G. Montavon, K. R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," 2017. [Online]. Available: arXiv:1706.07206.

[10] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[11] 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, 2015, Art. no. e0130140.

[12] A. Bansal, A. Farhadi, and D. Parikh, "Towards transparent systems: Semantic characterization of failure modes," in *Proc. ECCV*, 2014, pp. 366–381.

[13] C. I. Bargmann, "Beyond the connectome: How neuromodulators shape neural circuits," *Bioessays*, vol. 34, no. 6, pp. 458–465, 2012.

[14] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neuralnetworks," in *Proc. NeurIPS*, 2017, pp. 6240–6249.

[15] O. Bastani, C. Kim, and H. Bastani, "Interpretability via model extraction," 2017. [Online]. Available: arXiv:1706.09773.

[16] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. CVPR*, 2017, pp. 3319–3327.

[17] D. Bau, J. Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proc. Nat. Acad. Sci.*, vol. 117, no. 48, pp. 30071–30078, 2020.

[18] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. Nat. Acad. Sci.*, vol. 116, no. 32, pp. 15849–15854, 2019.

[19] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *Ann. Appl. Stat.*, vol. 5, no. 4, pp. 2403–2424, 2011.

[20] C. Biffi *et al.*, "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling," in *Proc. MICCAI*, 2018, pp. 464–471.

[21] C. Bishop, "Improving the generalization properties of radial basis function neural networks," *Neural Comput.*, vol. 3, no. 4, pp. 579–588, 1991.

[22] S. Bozinovski, M. Sestakov, and L. Bozinovska, "Using EEG alpha rhythm to control a mobile robot," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 1988, pp. 1515–1516.

[23] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. KDD*, 2016, pp. 535–541.

[24] D. Bychkov *et al.*, "Deep learning based tissue analysis predicts outcome in colorectal cancer," *Sci. Rep.*, vol. 8, no. 1, p. 3395, 2018.

[25] W. Caicedo-Torres and J. Gutierrez, "ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU," 2019. [Online]. Available: arXiv:1901.08201.

[26] R. Caruana, H. Kangarloo, J. D. N. Dionisio, U. Sinha, and D. B. Johnson, "Case-based explanation of non-case-based learning methods," in *Proc. AMIA Symp.*, 1999, pp. 212–215.

[27] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Disc. Databases*, 2018, pp. 655–670.

[28] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE Smart World Ubiquitous Intell. Comput. Adv. Trusted Comput. Scalable Comput. Commun. Cloud Big Data Comput. Internet People Smart City Innovat.*, 2017, pp. 1–6.

[29] Z. Che, S. Purushotham, R. G. Khemani, and Y. Liu, "Interpretable deep models for ICU outcome prediction," in *Proc. AMIA Annu. Symp.*, 2016, pp. 371–380.

[30] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "An interpretable model with globally consistent explanations for credit risk," 2018. [Online]. Available: arXiv:1811.12615.

[31] C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," 2018. [Online]. Available: arXiv:1806.10574.

[32] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Proc. NeurIPS*, 2018, pp. 6571–6583.

[33] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NeurIPS*, 2016, pp. 2172–2180.

[34] J. Chorowski and J. M. Zurada, "Learning understandable neural networks with nonnegative weight constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 62–69, Jan. 2015.

[35] L. Chu, X. Hu, J. Hu, L. Wang, and J. Pei, "Exact and consistent interpretation for piecewise linear neural networks: A closed form solution," in *Proc. KDD*, Jul. 2018, pp. 1244–1253.

[36] N. C. F. Codella, C.-C. Lin, A. Halpern, M. Hind, R. Feris, and J. R. Smith, "Collaborative human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images," in *Proc. 1st Int. Workshops Understanding Interpreting Mach. Learn. Med. Image Comput. Appl. MICCAI*, Granada, Spain, 2018, pp. 97–105.

[37] N. Cohen, O. Sharir, and A. Shashua, "The power of deeper networks for expressing natural functions," in *Proc. ICLR*, 2018, p. 14.

[38] M. W. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," in *Proc. NeurIPS*, 1995, pp. 24–30.

[39] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *Proc. IEEE Symp. Security Privacy (SP)*, May 2016, pp. 598–617.

[40] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: arXiv:1810.04805.

[41] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proc. CVPR*, 2016, pp. 4829–4837.

[42] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017. [Online]. Available: arXiv:1702.08608.

[43] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," 2018. [Online]. Available: arXiv:1808.00033.

[44] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proc. COLT*, 2016, pp. 907–940.

[45] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Dept. Informatique Recherche Opérationnelle, Univ. Montreal, Montréal, QC, Canada, Rep. 1341, 2009.

[46] F. Fan and G. Wang, "Learning from pseudo-randomness with an artificial neural network–does god play pseudo-dice?" *IEEE Access*, vol. 6, pp. 22987–22992, 2018.

[47] F. Fan, W. Cong, and G. Wang, "A new type of neurons for machine learning," *Int. J. Numer. Methods Biomed. Eng.*, vol. 34, no. 2, 2018, Art. no. e2920.

[48] F. Fan and G. Wang, "Fuzzy logic interpretation of quadratic networks," *Neurocomputing*, vol. 374, pp. 10–21, Jan. 2020.

[49] F. Fan, M. Li, Y. Teng, and G. Wang, "Soft autoencoder and its wavelet adaptation interpretation," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1245–1257, Aug. 2020.

[50] L. Fan, "Revisit fuzzy neural network: Demystifying batch normalization and ReLU with generalized hamming network," in *Proc. NeurIPS*, 2017, pp. 1923–1932.

[51] X. Fang and P. Yan, "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction," 2020. [Online]. Available: arXiv:2001.00208.

[52] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. CVPR*, 2017, pp. 3429–3437.

[53] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[54] L. Fu, "Rule generation from neural networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 24, no. 8, pp. 1114–1124, Aug. 1994.

[55] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, "Producing radiologist-quality reports for interpretable artificial intelligence," 2018. [Online]. Available: arXiv:1806.00340.

[56] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. COLT*, 2015, pp. 797–842.

[57] J. R. Geis *et al.*, "Ethics of artificial intelligence in radiology: Summary of the joint European and North American multisociety statement," *Can. Assoc. Radiol. J.*, vol. 70, no. 4, pp. 329–334, 2019.

[58] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. DSAA*, 2018, pp. 80–89.

[59] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Stat.*, vol. 24, no. 1, pp. 44–65, 2015.

[60] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[61] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.

[62] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," 2019. [Online]. Available: arXiv:1904.07451.

[63] C. Guan, X. Wang, Q. Zhang, R. Chen, D. He, and X. Xie, "Towards a deep and unified understanding of deep neural models in NLP," in *Proc. ICML*, 2019, pp. 2454–2463.

[64] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," 2018. [Online]. Available: arXiv:1805.10820.

[65] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys (CSUR)*, vol. 51, no. 5, p. 93, 2019.

[66] H. Guo, U. Kruger, G. Wang, M. K. Kalra, and P. Yan, "Knowledge-based analysis for mortality prediction from CT images," 2019. [Online]. Available: arXiv:1902.07687.

[67] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 162–169, Mar. 2019.

[68] M. Hatt, C. Parmar, J. Qi, and I. EI Naqa, "Machine (deep) learning methods for image processing and radiomics," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 104–108, Mar. 2019.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[70] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[71] G. E. Hinton, S. Osindero, and Y. W. The, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[72] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: arXiv:1503.02531.

[73] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[74] G. Hooker, "Discovering additive structure in black box functions," in *Proc. KDD*, 2004, pp. 575–580.

[75] Z. Hu, X. Ma, Z. Liu, E. Hovy, and E. Xing, "Harnessing deep neural networks with logic rules," 2016. [Online]. Available: arXiv:1603.06318.

[76] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 4700–4708.

[77] H. Huang, "Mechanisms of dimensionality reduction and decorrelation in deep neural networks," *Phys. Rev. E*, vol. 98, no. 6, 2018, Art. no. 062313.

[78] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Proc. NeurIPS*, 2018, pp. 8571–8580.

[79] S. Jain and B. C. Wallace, "Attention is not explanation," 2019. [Online]. Available: arXiv:1902.10186.

[80] J. S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 23, no. 3, pp. 665–685, May/Jun. 1993.

[81] C. Jiang, Y. Zhao, S. Chu, L. Shen, and K. Tu, "Cold-start and interpretability: Turning regular expressions into trainable recurrent neural networks," in *Proc. EMNLP*, 2020, pp. 3193–3207.

[82] L. P. Kadanoff, "Variational principles and approximate renormalization group calculations," *Phys. Rev. Lett.*, vol. 34, no. 16, p. 1005, 1975.

[83] A. Kádár, G. Chrupała, and A. Alishahi, "Representation of linguistic form and function in recurrent neural networks," *Comput. Linguistics*, vol. 43, no. 4, pp. 761–780, 2017.

[84] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. CVPR*, 2015, pp. 3128–3137.

[85] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," 2015. [Online]. Available: arXiv:1506.02078.

[86] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," 2017. [Online]. Available: arXiv:1711.11279.

[87] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. [Online]. Available: arXiv:1312.6114.

[88] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016. [Online]. Available: arXiv:1609.02907.

[89] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. ICML*, 2017, pp. 1885–1894.

[90] J. L. Kolodner, "An introduction to case-based reasoning," *Artif. Intell. Rev.*, vol. 1, no. 6, pp. 3–4, 1992.

[91] S. Kolouri, X. Yin, and G. K. Rohde, "Neural networks, hypersurfaces, and radon transforms," 2019. [Online]. Available: arXiv:1907.02220.

[92] R. Krishnan, G. Sivakumar, and P. Bhattacharya, "Extracting decision trees from trained neural networks," *Pattern Recognit.*, vol. 32, no. 12, pp. 1999–2009, 1999.

[93] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Rep. TR-2009, 2009.

[94] D. Krotov and J. J. Hopfield, "Unsupervised learning by competing hidden units," *Proc. Nat. Acad. Sci.*, vol. 116, no. 16, pp. 7723–7731, 2019.

[95] C.-C. J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen, "Interpretable convolutional neural networks via feedforward design," *J. Visual Commun. Image Represent.*, vol. 60, pp. 346–359, Apr. 2019.

[96] I. Lage, A. Ross, S. J. Gershman, B. Kim, and F. Doshi-Velez, "Human-in-the-loop interpretability prior," in *Proc. NeurIPS*, 2018, pp. 10159–10168.

[97] H. Lakkaraju, E. Kamar, R. Caruana, and E. Horvitz, "Identifying unknown unknowns in the open world: Representations and policies for guided exploration," in *Proc. AAAI*, 2017, pp. 2124–2132.

[98] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & explorable approximations of black box models," 2017. [Online]. Available: arXiv:1707.01154.

[99] R. M. Lark, "A comparison of some robust estimators of the variogram for use in soil survey," *Eur. J. Soil Sci.*, vol. 51, no. 1, pp. 137–157, 2000.

[100] H. Lee *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nat. Biomed. Eng.*, vol. 3, no. 3, pp. 173–182, 2019.

[101] N. Lei, K. Su, L. Cui, S.-T. Yau, and X. D. Gu, "A geometric view of optimal transportation and generative model," *Comput. Aided Geometric Design*, vol. 68, pp. 1–21, Jan. 2019.

[102] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," 2016. [Online]. Available: arXiv:1606.04155.

[103] Y. Lei, Y. Tian, H. Shan, J. Zhang, G. Wang, and M. K. Kalra, "Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101628.

[104] C. Li, M. Z. Zia, Q.-H. Tran, X. Yu, G. D. Hager, and M. Chandraker, "Deep supervision with intermediate concepts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1828–1843, Aug. 2019.

[105] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," 2016. [Online]. Available: arXiv:1612.08220.

[106] T. R. Li, A. Chamrajnagar, X. Fong, N. Rizik, and F. Fu, "Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model," *Front. Phys.*, vol. 7, p. 98, Jul. 2019.

[107] T.-Y. Li and J. A. Yorke, "Period three implies chaos," *Amer. Math. Monthly*, vol. 82, no. 10, pp. 985–992, 1975.

[108] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. E. Hopcroft, "Convergent Learning: Do different neural networks learn the same representations?" in *Proc. ICLR*, 2016.

[109] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" in *Proc. ICLR*, 2017.

[110] H. Lin and S. Jegelka, "ResNet with one-neuron hidden layers is a universal approximator," in *Proc. NeurIPS*, 2018, pp. 6172–6181.

[111] T. Lindeberg, "A computational theory of visual receptive fields," *Biol. Cybern.*, vol. 107, no. 6, pp. 589–635, 2013.

[112] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[113] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Appl. Stochastic Models Bus. Ind.*, vol. 17, no. 4, pp. 319–330, 2001.

[114] Y. Lu, A. Zhong, Q. Li, and B. Dong, "Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations," 2017. [Online]. Available: arXiv:1710.10121.

[115] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 4765–4774.

[116] L. V. der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[117] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. CVPR*, 2015, pp. 5188–5196.

[118] F. Maire, "On the convergence of validity interval analysis," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 802–807, May 2000.

[119] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an integration of deep learning and neuroscience," *Front. Comput. Neurosci.*, vol. 10, p. 94, Sep. 2016.

[120] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.

[121] P. Mehta and D. J. Schwab, "An exact mapping between the variational renormalization group and deep learning," 2014. [Online]. Available: arXiv:1410.3831.

[122] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and double descent curve," 2019. [Online]. Available: arXiv:1908.05355.

[123] D. A. Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. NeurIPS*, 2018, pp. 7775–7784.

[124] H. N. Mhaskar and T. Poggio, "Deep vs. shallow networks: An approximation theory perspective," *Anal. Appl.*, vol. 14, no. 6, pp. 829–848, 2016

[125] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.

[126] D. Nauck, "A fuzzy perceptron as a generic model for neuro-fuzzy approaches," in *Proc. Fuzzy Syst. 2nd GI Workshop*, 1994, pp. 91–99.

[127] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Proc. Conf. Learn. Theory*, 2015, pp. 1376–1401.

[128] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. NeurIPS*, 2016, pp. 3387–3395.

[129] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug and play generative networks: Conditional iterative generation of images in latent space," in *Proc. CVPR*, 2017, pp. 4467–4477.

[130] Q. Nguyen and M. Hein, "The loss surface of deep and wide neural networks," in *Proc. ICML*, 2017, pp. 2603–2612.

[131] A. Noack, I. Ahern, D. Dou, and B. Li, "Does interpretability of neural networks imply adversarial robustness?" 2019. [Online]. Available: arXiv:1912.03430.

[132] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. P. Singh, "Action-conditional video prediction using deep networks in Atari games," in *Proc. NeurIPS*, 2015, pp. 2863–2871.

[133] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, p. e7, 2017.

[134] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018. [Online]. Available: arXiv:1804.03999.

[135] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? Weakly-supervised learning with convolutional neural networks," in *Proc. CVPR*, 2015, pp. 685–694.

[136] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Comput.*, vol. 3, no. 2, pp. 246–257, 1991.

[137] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *Proc. CVPR*, 2018, pp. 7680–7688.

[138] S. Pereira *et al.*, "Enhancing interpretability of automatically extracted machine learning features: Application to a RBM-random forest system on brain lesion segmentation," *Med. Image Anal.*, vol. 44, pp. 228–244, Feb. 2018.

[139] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. CVPR*, 2015, pp. 1713–1721.

[140] T. Poggio and F. Girosi, "Regularization algorithms for learning that are equivalent to multilayer networks," *Science*, vol. 247, no. 4945, pp. 978–982, 1990.

[141] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. KDD*, 2016, pp. 1135–1144.

[142] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI*, 2018, pp. 1527–1535.

[143] M. Robnik-Šikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 589–600, May 2008.

[144] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," 2017. [Online]. Available: arXiv:1703.03717.

[145] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[146] E. W. Saad and D. C. Wunsch, II, "Neural network explanation using inversion," *Neural Netw.*, vol. 20, no. 1, pp. 78–93, 2007.

[147] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proc. ICML*, 2017, pp. 791–798.

[148] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *Proc. ICLR*, 2020, p. 18.

[149] T. Scott, K. Ridgeway, and M. C. Mozer, "Adapted deep embeddings: A synthesis of methods for *k*-shot inductive transfer learning," in *Proc. NeurIPS*, 2018, pp. 76–85.

[150] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017, pp. 618–626.

[151] R. Setiono and H. Liu, "Understanding neural networks via rule extraction," in *Proc. IJCAI*, vol. 1, 2017, pp. 480–485.

[152] H. Shan *et al.*, "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction," *Nat. Mach. Intell.*, vol. 1, no. 6, pp. 269–279, 2019.

[153] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017. [Online]. Available: arXiv:1703.00810.

[154] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," *Expert Syst. Appl.*, vol. 128, pp. 84–95, Aug. 2019.

[155] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Interpretable deep learning by propagating activation differences," in *Proc. ICML*, 2016, p. 6.

[156] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013. [Online]. Available: arXiv:1312.6034.

[157] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[158] C. Singh, W. J. Murdoch, and B. Yu, "Hierarchical interpretations for neural network predictions," in *Proc. ICLR*, 2019.

[159] S. Singla, E. Wallace, S. Feng, and S. Feizi, "Understanding impacts of high-order loss approximations and features in deep learning interpretation," 2019. [Online]. Available: arXiv:1902.00407.

[160] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017. [Online]. Available: arXiv:1706.03825.

[161] S. H. Snyder, "Adenosine as a neuromodulator," *Annu. Rev. Neurosci.*, vol. 8, no. 1, pp. 103–124, 1985.

[162] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 742–769, Feb. 2019.

[163] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014. [Online]. Available: arXiv:1412.6806.

[164] A. Stone, H. Wang, M. Stark, Y. Liu, D. S. Phoenix, and D. George, "Teaching compositionality to CNNs," in *Proc. CVPR*, 2017, pp. 5058–5067.

[165] I. Sturm, S. Lapuschkin, W. Samek, and K. R. Müller, "Interpretable deep neural networks for single-trial EEG classification," *J. Neurosci. Methods*, vol. 274, pp. 141–145, Dec. 2016.

[166] A. Subramanian, D. Pruthi, J. H. Jhamtani, T. Berg-Kirkpatrick, and E. H. Hovy, "SPINE: Sparse interpretable neural embeddings," in *Proc. AAAI*, 2018, pp. 4921–4928.

[167] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, 2017, pp. 3319–3328.

[168] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013. [Online]. Available: arXiv:1312.6199.

[169] L. Szymanski and B. McCane. "Deep networks are effective encoders of periodicity," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1816–1827, Oct. 2014.

[170] S. Tan, R. Caruana, G. Hooker, P. Koch, and A. Gordo, "Learning global additive explanations for neural nets using model distillation," 2018. [Online]. Available: arXiv:1801.08640.

[171] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. SMC-15, no. 1, pp. 116–132, Jan./Feb. 1985.

[172] S. Thrun, "Extracting rules from artificial neural networks with distributed representations," in *Proc. NeurIPS*, 1995, pp. 505–512.

[173] M. Torres-Velázquez, W. J. Chen, X. Li, and A. B. McMillan, "Application and construction of deep learning networks in medical imaging," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 2, pp. 137–159, Mar. 2021.

[174] H. L. Van der Maas, P. F. Verschure, and P. C. Molenaar, "A note on chaotic behavior in simple neural networks," *Neural Netw.*, vol. 3, no. 1, pp. 119–122, 1990.

[175] P. Van Molle, M. De Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, and B. Dhoedt, "Visualizing convolutional neural networks to improve decision support for skin lesion classification," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Cham, Switzerland: Springer, 2018, pp. 115–123.

[176] J. Vaughan, A. Sudjianto, E. Brahimi, J. Chen, and V. N. Nair, "Explainable neural networks based on additive index models," 2018. [Online]. Available: arXiv:1806.01933.

[177] A. Veit, M. J. Wilber, and S. J. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. NeurIPS*, 2016, pp. 550–558.

[178] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015, pp. 3156–3164.

[179] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GPDR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2017.

[180] E. Wallace, S. Feng, and J. Boyd-Graber, "Interpreting neural networks with nearest neighbors," 2018. [Online]. Available: arXiv:1809.02847.

[181] G. Wang, "A perspective on deep imaging," *IEEE Access*, vol. 4, pp. 8914–8924, 2016.

[182] T. Wang, "Gaining free or low-cost interpretability with interpretable partial substitute," in *Proc. ICML*, 2019, pp. 6505–6514.

[183] Y. Wang, H. Su, B. Zhang, and X. Hu, "Interpret neural networks by identifying critical data routing paths," in *Proc. CVPR*, 2018, pp. 8906–8914.

[184] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Interpretable transformations with encoder-decoder networks," in *Proc. ICCV*, 2017, pp. 5726–5735.

[185] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: Tree regularization of deep models for interpretability," in *Proc. AAAI*, 2018, pp. 1670–1678.

[186] T. Wu, W. Sun, X. Li, X. Song, and B. Li, "Towards interpretable R-CNN by unfolding latent structures," 2017. [Online]. Available: arXiv:1711.05226.

[187] W. Wu, D. Hu, S. Wang, H. Yu, V. Vardhanabhuti, and G. Wang, "Stabilizing deep tomographic reconstruction networks," 2020. [Online]. Available: arXiv:2008.01846.

[188] Q. Xie, X. Ma, Z. Dai, and E. Hovy, "An interpretable knowledge transfer model for knowledge base completion," 2017. [Online]. Available: arXiv:1704.05908.

[189] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[190] K. Xu *et al.*, "Interpreting adversarial examples by activation promotion and suppression," 2019. [Online]. Available: arXiv:1904.02057.

[191] C. Yang, A. Rangarajan, and S. Ranka, "Global model interpretation via recursive partitioning," in *Proc. IEEE 20th Int. Conf. High Perform. Comput. Commun. IEEE 16th Int. Conf. Smart City IEEE 4th Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, 2018, pp. 1563–1570.

[192] J. C. Ye, Y. Han, and E. Cha, "Deep convolutional framelets: A general deep learning framework for inverse problems," *SIAM J. Imag. Sci.*, vol. 11, no. 2, pp. 991–1048, 2018.

[193] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9244–9255.

[194] S. You, D. Ding, K. Canini, J. Pfeifer, and M. Gupta, "Deep lattice networks and partial monotonic functions," in *Proc. NeurIPS*, 2017, pp. 2981–2989.

[195] J. You, J. Leskovec, K. He, and S. Xie, "Graph structure of neural networks," in *Proc. ICML*, 2020, pp. 10881–10891.

[196] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015. [Online]. Available: arXiv:1506.06579.

[197] S. Yu and J. C. Principe, "Understanding autoencoders with information theoretic concepts," *Neural Netw.*, vol. 117, pp. 104–123, Sep. 2019.

[198] L. A. Zadeh, "Fuzzy logic," *Computer*, vol. 21, no. 4, pp. 83–93, 1988.

[199] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med.*, vol. 15, no. 11, 2018, Art. no. e1002683.

[200] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[201] Q. Zhang and S. C. Zhu, "Visual interpretability for deep learning: A survey," *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, 2018.

[202] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S.-C. Zhu, "Interpreting CNN knowledge via an explanatory graph," in *Proc. AAAI*, 2018, pp. 4454–4463.

[203] Q. Zhang, W. Wang, and S. C. Zhu, "Examining CNN representations with respect to dataset bias," in *Proc. AAAI*, 2018, pp. 4464–4473.

[204] Q. Zhang, Y. N. Wu, and S. C. Zhu, "Interpretable convolutional neural networks," in *Proc. CVPR*, 2018, pp. 8827–8836.

[205] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh, "Predicting failures of vision systems," in *Proc. CVPR*, 2014, pp. 3566–3573.

[206] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "Why should you trust my explanation? Understanding uncertainty in LIME explanations," 2019. [Online]. Available: arXiv:1904.12991.

[207] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. CVPR*, 2017, pp. 6428–6436.

[208] Z. Zhang *et al.*, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 236–245, 2019.

[209] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016, pp. 2921–2929.

[210] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," 2014. [Online]. Available: arXiv:1412.6856.

[211] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," 2017. [Online]. Available: arXiv:1702.04595.