Shortcomings of Interpretability Taxonomies for Deep Neural Networks

Anders Søgaard[1,2,3]

[1]Dpt. of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2200 Copenhagen
[2]Pioneer Centre for Artificial Intelligence, Lyngbyvej 2, DK-2100 Copenhagen
[3]Dpt. of Philosophy, Karen Blixens Plads 8, DK-2300 Copenhagen

Abstract

Taxonomies are vehicles for thinking about what's possible, for identifying unconsidered options, as well as for establishing formal relations between entities. We identify several shortcomings in 10 existing taxonomies for interpretability methods for explainable artificial intelligence (XAI), focusing on methods for deep neural networks. The shortcomings include redundancies, incompleteness, and inconsistencies. We design a new taxonomy based on two orthogonal dimensions and show how it can be used to derive results about entire classes of interpretability methods for deep neural networks.

Keywords

interpretability, taxonomy

1. Two Common Distinctions

Biological taxonomies provide a basis for conservation and development and are used to generate interesting questions about missing species [1, 2]. Inconsistent taxonomies can, at the same time, hinder research or lead in the wrong direction [3, 4, 5]. In engineering, taxonomies play *additional* roles: They are vehicles for thinking about what's possible, for identifying unconsidered options, as well as for establishing formal relations between methods.

Several taxonomies of interpretability methods already exist [6, 7, 8, 9, 10, 11, 12, 13]. These taxonomies provide us with technical terms for distinguishing approaches to interpretability and can be efficient tools for researchers to contextualize their work. They generate interesting research questions – e.g., *if all methods in class A but no methods in B happen to exhibit property X, it this by necessity, or can we design a method in B with property X?* – and help us see relations between methods – e.g., *two methods in class A are mathematically equivalent.* Unfortunately, the taxonomies that exist, without exception, have shortcomings and are either redundant, incomplete, or inconsistent. In §2, we show this, examining the above 10 taxonomies, one by one, also discussing between-taxonomy inconsistencies in how individual methods are classified. In §2, we present a consistent taxonomy and establish various observations and results that apply to entire classes of methods in our taxonomy. *Contributions* (a) We detect inadequacies in 10 interpretability taxonomies. (b) We establish a simple, yet superior, two-dimensional taxonomy. (c) We derive six non-trivial observations or results based on this taxonomy.

Survey	**Dimensions**	**Inconsistent**	**Incomplete**	**Redundant**	loc-glo	intr-phoc	other
[6]	4	☒			(☒)	(☒)	time, expertise
[7]	3	☒	☒		☒	☒	model-specific/model-agnostic
[8]	4	☒		☒	☒	☒	pre-model/in-model/post-model, results
[9]	4	☒		☒	☒	☒	spec./agn., results
[10]	3	☒			☒	☒	types
[11]	3	☒	☒	☒	☒	☒	technique
[12]	3		☒	☒	☒	☒	methodology
[14]	1	☒	☒				grad./pert./ simpl.
[15]	1	☒	☒				att./rule/sum.
[13]	2	☒	☒	☒	☒		inst./approx./ attr./counterf.

Table 1
10 existing taxonomies and their shortcomings: Most distinguish local from global methods, and intrinsic from posthoc methods. We argue the additional dimensions *all* lead to inconsistencies and/or redundancies, and that the intrinsic-posthoc distinction is itself problematic.

The simplest taxonomies presented are one-dimensional, i.e., simple groupings [14, 15]. Other methods introduce up to four dimensions and use these to cross-classify existing methods. The 10 taxonomies are at most a couple of years old (2019-2021) and discussed in chronological order. We first discuss two common distinctions that are largely agreed upon: *local-global* and *intrinsic-posthoc*. One of these distinctions,

AIMLAI 2022: Advances in Interpretable Machine Learning and Artificial Intelligence, October 21, 2022, Atlanta, GA
soegaard@di.ku.dk (A. Søgaard)
https://anderssoegaard.github.io/ (A. Søgaard)
0000-0001-5250-4276 (A. Søgaard)

local-global, will be useful, while the other is problematic in several respects. We define an interpretability method $\mathscr{M}(\mathbf{w}, S)$ as a complex function that takes a model $\mathbf{w}$ and a sample of token sequences $S \subseteq \mathcal{S}$ as input, and is composed of three types of functions:

Definition 1.1 (Forward functions). *Let* $\mathbf{forward}(\mathbf{w}, S)$ *return* $\mathbf{w}(f(S))$ *for all inputs* $s \in S$, *i.e.,* $\mathbf{w}_1(s), \ldots, \mathbf{w}_n(s)$ *for* n *layers, with* $f : \mathcal{S} \mapsto \mathcal{S}$ *a function from input to input, e.g., perturb, delete, identity.*

Definition 1.2 (Backward functions). *Let* $\mathbf{backward}(\mathbf{w}, S)$ *return*

$$g(\mathbf{w}^{-1}, (\mathbf{forward}(\mathbf{w}, S)))$$

where $g(\cdot, \cdot)$ *is a function that defines a backward pass of gradients, relevance scores, etc., over the inverse model* $\mathbf{w}^{-1}$.

Definition 1.3 (Inductive functions). *Let* $\mathbf{induce}(\mathbf{w}, S)$ *return a set of parameters* $\mathbf{v}$ *fitted by minimizing an objective over* $\mathbf{w}$ *and* S.

Examples of inductive functions include, for various loss functions $\ell(\cdot, \cdot)$: (a) *probing* [16], in which the objective is of the form $\ell(\mathbf{v}(S), l(\mathbf{w}_j(S)))$ where $\mathbf{w}_j(s)$ is the representation of s at the $j$the layer of $\mathbf{w}$, and l is the probe-specific re-labeling function of samples; or (c) *linear approximation* [17], in which the objective is of the form $\ell(\mathbf{v}(S), \mathbf{w}(S))$, where $\mathbf{v}$ is a linear function.

Local and global explanations The distinction between local and global interpretability methods is shared across all the taxonomies discussed in this paper, and will also be one of the two dimensions in the taxonomy we propose below. The distinction is defined slightly differently by different authors, As should be clear from the discussion below, this is not equivalent to our definition, which uses the reliance of global methods on samples, rather than the reliance of local methods on specific instances, as the distinguishing criterion. One argument against the definition in [11] is that it is not entirely clear in what sense global methods such as concept activation vectors [18], for example, are *independent of any particular input.* The function that provides us with explanations is global, but of course its output depends on the input. or not defined at all, e.g., [6], but here we present the definition that our taxonomy below relies on: A method $\mathscr{M}$ is said to be *global* if and only if it includes at least one inductive function. Otherwise $\mathscr{M}$ is said to be local. Global methods typically require access to a representative sample of data, to minimize their objectives, whereas local methods are applicable to singleton samples.[1] **Challenge** When taxonomies have tried to classify interpretation methods into local and global ones, in practice, some methods have seemed harder to classify than others. Concept activation approaches [18], for example, use joint global training to learn mappings of individual examples into local explanations. Contrastive interpretability methods [20] provide explanations in terms of pair of examples. It may also seem unclear whether a challenge dataset provides a local or global explanation. [10] discuss what they call *semi-local* approaches, and [8] introduce a category for interpretability methods that relate to groups of examples. Are there methods that are not easily categorized as global or local? **Answer** Our definition of local-global focuses on the induction of explanations from samples. This focus enables unambiguous classification and leads us to classify concept activation methods as global, since the explanatory model component is induced from a sample (and relies on the representativity of this sample). Similarly, we classify contrastive and group methods as local methods, since they do not require induction or assume representative samples; and, finally, we classify challenge datasets as local methods, since challenge datasets also do not have to be representative.[2]

Intrinsic and post-hoc explanations This distinction, also called *active-passive* in [10] and *self-explaining-ad hoc* in [11], is between *intrinsic* methods that jointly output explanations, and methods that derive these explanations *post-hoc* using auxiliary techniques. While most taxonomies introduce this distinction, we argue that it is inherently problematic. **Challenge** The distinction between intrinsic and posthoc methods can be hard to

[1] Note that our definition does not refer to *how* the methods characterize the models, e.g., whether they describe individual inferences, or derive aggregate statistics that quantify ways the models are biased. This is to avoid a common source of confusion: Local methods can be used to derive aggregate statistics that characterize global properties of models. LIME [19], for example, is mostly classified as a local method ([12] classify it as *both* local and global), but in [19], the authors explicitly discuss how LIME can be used on i.i.d. samples to derive aggregate statistics that characterize model behavior on distributions (same can be done for *all* local methods; see §3.6). Our definition makes it clear that such methods are local; local methods *can* be applied globally, whereas global methods *cannot* be applied locally. It is also clear from our definition that the two classes of interpretability methods are often motivated by different prototypical applications: Local methods are often used to explain the motivation behind critical decisions, e.g., why a customer was assessed as high-risk, why a traveling review was flagged as fraudulent, or why a newspaper article was flagged as misleading, whereas global methods are used to characterize biases in models and evaluate their robustness.

[2] Examples of *local* methods include gradients [21, 22, 23], LRP [24], deep Taylor decomposition [25], integrated gradients [26, 27], DeepLift [28], direct interpretation of gate/attention weights [29], attention roll-out and flow [30], word association norms and analogies [31], time step dynamics [32], challenge datasets [33, 34, 35, 36], local uptraining [19], and influence sketching and influence functions [37]; examples of *global* methods include unstructured pruning, lottery tickets, dynamic sparse training, binary networks, sparse coding, gate and attention head pruning, correlation of representations [38], clustering [39, 40, 41], probing classifiers [16], concept activation [18], representer point selection [42], TracIn [43], and uptraining [44].

maintain.[3] Moreover, for a method to be *posthoc* means different things to local and global methods. A post-hoc, local method is post-hoc relative to a class inference (in the case of classification); a post-hoc, global method is post-hoc relative to training, introducing a disjoint training phase for learning the interpretability functions. Strictly speaking, the fact that 'post-hoc' takes on two disjoint meanings for local and global methods, namely 'post-inference' and 'post-training', makes taxonomies that rely on both dimensions inconsistent.

2. Shortcomings of Taxonomies

We now briefly assess the 10 taxonomies, pointing out the ways in which they are inconsistent, incomplete or redundant:

Guidotti et al. (2018) [6] makes the local-global distinction, as well as two that relate to how explanations are communicated (how much time the user is expected to have to understand the model decisions, and how much domain knowledge and technical experience the user is expected to have). In addition to the terms *local* and *global*, they also refer, synonymously, to *outcome explanation* and *model explanation*. Later in their survey, [6] make a fourth distinction that is very similar to intrinsic-posthoc, namely between *transparent design* (leading to intrinsically interpretable models) and (post-hoc) *black box inspection*, but oddly, this is not seen as an orthogonal dimension, but as two additional classes *on par* with outcome and model explanation. **Challenge** How to classify methods that are *both*, say, local and post-hoc, i.e., do outcome explanation by black-box inspection? Examples would include gradients [21, 22, 23], layer-wise relevance propagation [24], deep Taylor decomposition [25], integrated gradients [26, 27], etc.

Adadi and Berrada (2019) [7] rely on the local-global and intrinsic-posthoc distinctions (referring to the later as *complexity*), and, as a third dimension, they distinguish between model-agnostic and model-specific interpretability methods. **Inconsistencies** We argue that the distinction between model-specific and model-agnostic methods is suboptimal in that state-of-the-art models are moving targets, and so is what counts as model-specific. This may lead to inconsistencies over time. **Challenge** How do we classify a method that applies to all known methods, but not to all possible methods?

Carvalho et al. (2019) [8] introduce four dimensions in their taxonomy: (a) *scope*, which coincides with the local-global distinction; (b) *intrinsic-posthoc*; (c) *pre-model, in-model, and post-model*, with *in-model* corresponding to intrinsic methods, and *post-model* corresponding to post-hoc methods, whereas *pre-model* comprises various approaches to data analysis. We argue below that (c) is both redundant and inconsistent. Finally, they introduce (d) a *results* dimension, which concerns the *form* of the explanations provided by the methods. **Inconsistencies** In addition to the inconsistency of intrinsic-posthoc, including pre-model explanations leads to further taxonomic inconsistency in that pre-model approaches cannot be classified along the other dimensions in that they do not refer to models at all. For the same reason, one might argue they are not model interpretation methods in the first place. **Redundancies** The redundancy of (c) follows from the observation that the distinction between in-model and post-model explanations is identical to the distinction made in (b), as well as the observation that pre-model explanations do not refer to models at all. **Challenge** What is an intrinsic interpretability method that presents post-model explanations, or a post-hoc interpretability method that presents in-model explanations?

Molnar (2019) [9] distinguishes between local-global and intrinsic-posthoc, between different *results*, and between model-specific and model-agnostic methods, making their taxonomy very similar to [8]. **Inconsistencies** See discussion of [7]. Also, the results dimension is also inconsistent in that explanations can, simultaneously, be intrinsically interpretable models and feature summary statistics. LIME [19], for example, presents local explanations as the linear coefficients of a linear fit, i.e., an intrinsically interpretable model that consists solely of feature summary statistics. **Redundancies** The most important redundancy is that all model-agnostic interpretability methods are also post-hoc, since intrinsic methods require joint training, which in turn requires compatibility with model architectures. Moreover, model-agnostic interpretability methods are all grounded in input features and thus lead to explanations in terms of feature summary statistics or visualizations. Moreover, all explanations in terms of intrinsically interpretable models are, quite obviously, intrinsic. **Challenge** What is a post-hoc interpretability method whose explanations are intrinsically interpretable models?

Zhang et al. (2020) [10] rely on these dimensions: (a) global-local; (b) intrinsic-posthoc (which they call *active-passive*; and (c) a distinction between four *explanation types*, namely *examples*, *attribution*, *hidden semantics*, and *rules*. **Inconsistencies** The *explanation type* dimension in [10] conflates (a) the model components we are trying to explain, and (b) what the explanations look like. Hidden semantics, e.g., is a model component, whereas

[3] Consider the difference between the two global interpretability methods, concept activation vectors and probing classifiers: CAV are trained jointly, probing classifiers sequentially. These are extremes of a (curriculum) continuum, which is hard to binarize: If a probing classifier is trained jointly with the last epoch of the model training, is the method then intrinsic or posthoc? For a real example, consider TracIn [43], in which influence functions are estimated across various training check points. Again, is TracIn intrinsic or posthoc? That the binary distinction covers a continuum, makes the distinction hard to apply in practice.

examples and rules refer to the (syntactic) *form* of the explanations. The distinction between hidden semantics and attribution is also apparent. Hidden semantics can be used to *derive* attribution (a results type in [8] and [9]), e.g., in LSTMVis [45]; this is because hidden semantics is not a type of explanation, but a model component. Attribution, examples, and rules *are* types of explanations, but this list is not exhaustive, since explanations can also be in terms of concepts, free texts, or visualizations, for example. **Challenge** What is a passive interpretability method that does not provide local explanations?

Danilevsky et al. (2020) [11] only distinguish between global-local and intrinsic-posthoc (which they call *self-explaining* and *ad-hoc*) methods. **Inconsistencies** [11] say most attribution methods are global and ad-hoc. We argue attribution methods are necessarily local, and while aggregate statistics can of course be computed across real or synthetic corpora, little is gained by blurring taxonomies to reflect that. All local methods can be used to compute summary statistics. **Incompleteness** [11] admit their survey is biased toward local methods, and many global interpretability methods are left uncovered. **Challenge** What is a local interpretability method that cannot be used to compute summary statistics?

Das et al. (2020) [12] distinguish between local and global methods, gradient-based and perturbation-based methods on the other (methodology), and intrinsic and post-hoc methods (usage). Their taxonomy is both *incomplete* and *redundant*: **Incompleteness** Several approaches are neither gradient-based or perturbation-based. **Redundancies** All gradient-based approaches are classified as post-hoc approaches in [12]; similarly, all intrinsic methods are classified as global methods. Of course these cells may be filled with methods that were not covered, but in particular, it seems that gradient-based approaches are, almost always, post-hoc? **Challenge** What is an intrinsic, gradient-based approach?

Atanasova et al. (2020) [14] distinguish between three classes of interpretability methods: gradient-based, perturbation-based, and simplification-based methods. **Inconsistencies** The distinction between gradient-based and perturbation-based methods is similar to [12], but the two classifications are inconsistent, with [14] citing LIME [19] as a simplification-based method. It seems that the distinction between perturbation-based and simplification-based methods is in itself inconsistent in that both perturbations and gradients can be used to simplify models; similarly, perturbations can be used to baseline gradient-based approaches. **Incompleteness** Clearly, not all interpretability methods are gradient-based, perturbation-based or simplification-based: Other methods are based on weight magnitudes, carefully designed example templates, visualizing and quantifying attention weights or gating mechanisms. **Challenge** How would you classify attention roll-out [30], for example?

	[6]	[7]	[8]	[9]	[10]	[11]	[12]
GradCAM	L				L-H		L-H
DeepLift	H				L-H		L-H
LRP		L/G-H			S-H	L-I/H	L/G-H
LIME	L	L-H	L-H	L	L-H	L-H	L/G-H
TCAV				G-I	G-H		G-H
IF			L/G	L-H			

	Forward	**Backward**
Local	Attention, Attention roll-out, Attention flow, Time step dynamics, Local uptraining, Influence functions	Gradients, Layer-wise relevance propagation, Deep Taylor decomposition, Integrated gradients, DeepLift
Global	Weight pruning, Correlation of representations, Clustering, Probing classifiers, Uptraining	Dynamic sparse training, Binary networks, Sparse coding, Concept activation, Gradient-based weight pruning

Table 2
Left: 4/6 methods (bottom half) are classified incoherently across taxonomies. *Explanation:* local (L), global (G), intrinsic (I), and posthoc (H). **Right:** Our novel taxonomy.

Kotonya and Toni (2020) [15] distinguish between attention-based explanations, explanations as rule discovery, and explanations as summarization. **Incompleteness** Using gating mechanisms to interpret models, e.g., does not fit any of the three categories. **Inconsistencies** One class of methods is defined in terms of the model components being interpreted (attention-based), and another class in terms of the *form* of explanations they provide (rule discovery and summarization). Mixing orthogonal dimensions is inconsistent, i.e., methods can belong to several categories, e.g., attention head pruning [46] (attention-based and summarization), or when rules are induced from attention weights [47].

Chen et al. (2021) [13] introduce the global-local distinction, but not the intrinsic-posthoc distinction. In addition they distinguish between interpretability methods that present explanations in terms of training instances, approximations, feature attribution and counterfactuals. **Inconsistencies** The second dimension again makes orthogonal distinctions. Approximations, for example, can be used to attribute importance to features (LIME). **Incompleteness** Concepts, attention weights, gate activations, rules, etc., are not covered by the second dimension. **Redundancies** All methods that present explanations in terms of training instances are necessarily local. **Challenge** What's a global interpretability method providing explanations in terms of training instances?[4]

Inconsistent Classifications Table 2 shows that tax-

[4]Several of the above taxonomies include dimensions that pertain to the *form* of the output of interpretation methods. We argue such distinctions are orthogonal to the methods and should therefore not be included in taxonomies. To see this, note that most interpretability methods, e.g., LIME, can provide explanations of different form: aggregate statistics, coefficients, rules, visualizations, etc.

onomies are not only internally inconsistent, but also inconsistent in how they classify methods. Six methods were mentioned by more than one survey, 4/6 of which were classified differently.

3. A Novel Taxonomy and Observations

Our taxonomy is two-dimensional: One is local-global, the other a distinction between explanations based on *forward* passes, and explanations based on *backward* passes. The forward explanations correlate intermediate representations or continuous or discrete output representations to obtain explanations, whereas backward explanations concern training dynamics. We define forward-backward:

Definition 3.1 (Forward-backward). *A method $\mathscr{M}$ is said to be* backward *if it contains backward functions; otherwise, $\mathscr{M}$ is said to be* forward.

Local backward methods include gradients [21, 22, 23], integrated gradients [26, 27], layerwise relevance propagation [48], DeepLIFT [28], and deep Taylor decomposition [25], which all derive explanations for individual instances from what is normally used as training signals, typically based on derivatives of the loss function (gradients) evaluating h on training data, e.g., $d(\ell(h(\mathbf{x}_i), y_i))$. Global backward methods rely on such training signals to modify or extend the model parameters $\mathbf{w}$ associated with h, typically extracting approximations, rules or visualizations.[5]

Observation 3.1. *Local backward methods are always attribution methods (presenting feature summary statistics).*

Since local methods have to provide explanations in terms of input/output (as they do not modify weights), and since backward passes do not generate output distributions, they have to present explanations in terms of attribution of relevance or gradients to input features or input segments. §3.1 is empirical. It follows naturally, but not necessarily, from the fact that backward methods reverse the direction of connections, thus returning quantities that hold for the input nodes. Pre-input quantities are not interpretable.

Observation 3.2. *Only global methods can be unfaithful.*

§3.2 follows from the definition of faithfulness: $\mathscr{M}$ is faithful if the inductive functions of $\mathscr{M}$ have $\ell(\mathbf{v}, P) = 0$ and $S\tilde{P}$. $\mathscr{M}$ can only be unfaithful with respective to inductive component functions; local methods can therefore not be unfaithful.[6]

Observation 3.3. *Global methods can at best be epsilon-faithful and only on i.i.d. instances.*

§3.3 follows from the fact that standard learning theory applies to the inductive component functions of global interpretability methods. Since their faithfulness is the inverse of the empirical risk of these inductions, it follows that global methods can at best be ϵ-faithful, with ϵ the expected loss of these inductions. Note that when the explanation is a model approximation θ', $\epsilon = \mathbb{E}[\ell(\theta(x), \theta'(x))]$.

Observation 3.4. *Only forward methods are used for local layer-wise analysis.*

Since local backward methods are attribution methods (§3.1), and layer-wise analysis concerns differences between layers, local backward methods cannot be used here, simply because they only output attributions at the input level. §3.4 thus follows from §3.1, making it, too, an empirical observation, not a formal derivation.

Observation 3.5. *No equivalence relations can hold across the four categories of methods.*

§3.5 follows from the disjointness of the three sets of component functions, and how the four classes are defined, i.e., that global functions cannot be local, and forward functions cannot be backward. Equivalences between methods have already been found [28, 26, 27, 60, 61], but consistent taxonomic classification effectively prunes the search space of possible equivalences.

Observation 3.6. *Local methods can always characterize models globally on i.i.d. samples.*

§3.6 states that any local method that derives quantifies for an example can be used to aggregate corpus-level statistics for appropriate-level samples. See [19] for how to do this with LIME. It should be easy to see how this result generalizes to all other local methods.

[5] Local forward methods either consider intermediate representations, e.g., gates [49], attention [29], attention flow [50], etc.; continuous output representations, e.g., using word association norms [51] or word analogies [52, 53]; or discrete output, such as when evaluating on challenge datasets [33, 34, 35, 36], or when approximating the model's output distribution [19, 54, 37]. In the same way, global forward methods can rely on intermediate representations in forward passes, e.g., in attention head pruning [46], attention factor analysis [55], syntactic decoding of attention heads [50], attention head manipulation [56], etc.; continuous output in forward passes, including work using clustering in the vector space to manually analyze model representations [57, 58], probing classifiers [16], and concept activation strategies [18]; or on discrete output, e.g., in uptraining [44] and knowledge distillation [59].

[6] Local methods compute quantities based on forward or backward passes, but these quantities are not induced to simulate anything. Global methods induce parameters to simulate a distribution and can be more or less faithful to this distribution, but since local methods simply 'read off' their quantities, they cannot be unfaithful. Only, the quantities can be misinterpreted.

4. Conclusion

We examined 10 taxonomies of interpretability methods and found *all* to be inconsistent. We introduces a two-dimensional taxonomy and showed how it can be helpful in deriving general observations and results.

References

[1] S. Bacher, Still not enough taxonomists: reply to Joppa et al., Trends in Ecology Evolution 27 2 (2012) 65–6; author reply 66.

[2] S. Thomson, R. Pyle, S. Ahyong, M. Alonso-Zarazaga, J. Ammirati, J.-F. Araya, J. Ascher, T. Audisio, V. Azevedo-Santos, N. Bailly, W. Baker, M. Balke, M. Barclay, R. Barrett, R. Benine, J. Bickerstaff, P. Bouchard, R. Bour, T. Bourgoin, H.-Z. Zhou, Taxonomy based on science is necessary for global conservation, PLoS Biology 16 (2018). doi:10.1371/journal.pbio.2005075.

[3] M. S. Brewer, P. Sierwald, J. Bond, Millipede taxonomy after 250 years: Classification and taxonomic practices in a mega-diverse yet understudied arthropod group, PLoS ONE 7 (2012).

[4] H. Fraser, G. Garrard, L. Rumpff, C. Hauser, M. McCarthy, Consequences of inconsistently classifying woodland birds, Frontiers in Ecology and Evolution 3 (2015) 83.

[5] B. Jones, A few bad scientists are threatening to topple taxonomy, Smithsonian Magazine (2017).

[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models (2018).

[7] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.

[8] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.

[9] C. Molnar, Interpretable Machine Learning, 2019. https://christophm.github.io/interpretable-ml-book/.

[10] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, 2020. arXiv:2012.14261.

[11] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: AACL-IJNLP, Suzhou, China, 2020.

[12] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020. arXiv:2006.11371.

[13] V. Chen, J. Li, J. S. Kim, G. Plumb, A. Talwalkar, Towards connecting use cases and methods in interpretable machine learning, CoRR (2021). arXiv:2103.06254.

[14] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, A diagnostic study of explainability techniques for text classification, in: EMNLP, Online, 2020.

[15] N. Kotonya, F. Toni, Explainable automated fact-checking: A survey, in: COLING, Barcelona, Spain (Online), 2020.

[16] Y. Belinkov, Probing Classifiers: Promises, Shortcomings, and Advances, Computational Linguistics (2021) 1–13.

[17] J. Ba, R. Caruana, Do deep nets really need to be deep?, in: NeurIPS, 2014.

[18] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)., in: ICML, volume 80, 2018.

[19] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: KDD, New York, NY, USA, 2016.

[20] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, 2018. arXiv:1802.07623.

[21] P. Leray, P. Gallinari, P. Gallinari, P. Gallinari, Feature selection with neural networks, Behaviormetrika 26 (1998) 16–6.

[22] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. arXiv:1312.6034.

[23] M. Denil, A. Demiraj, N. Kalchbrenner, P. Blunsom, N. de Freitas, Modelling, visualising and summarising documents with a single convolutional neural network., CoRR abs/1406.3830 (2014).

[24] L. Arras, F. Horn, G. Montavon, K.-R. Müller, W. Samek, Explaining predictions of non-linear classifiers in NLP, in: RepLNLP, Berlin, Germany, 2016.

[25] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep Taylor decomposition, Pattern Recognition 65 (2017) 211–222.

[26] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, 2017. arXiv:1703.01365.

[27] P. K. Mudrakarta, A. Taly, M. Sundararajan, K. Dhamdhere, Did the model understand the question?, in: ACL, Melbourne, Australia, 2018.

[28] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: ICML, 2017.

[29] M. Rei, A. Søgaard, Zero-shot sequence labeling: Transferring knowledge from sentences to tokens,

in: NAACL, New Orleans, Louisiana, 2018.
[30] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, in: ACL, Online, 2020.
[31] T. Mikolov, Q. V. Le, I. Sutskever, Exploiting similarities among languages for machine translation, CoRR abs/1309.4168 (2013). arXiv:1309.4168.
[32] H. Strobelt, S. Gehrmann, H. Pfister, A. M. Rush, Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks, 2017. arXiv:1606.07461.
[33] M. Richardson, C. J. Burges, E. Renshaw, MCTest: A challenge dataset for the open-domain machine comprehension of text, in: EMNLP, 2013.
[34] J. Mullenbach, J. Gordon, N. Peng, J. May, Do nuclear submarines have nuclear captains? a challenge dataset for commonsense reasoning over adjectives and objects, in: EMNLP, Hong Kong, China, 2019.
[35] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, C. Cardie, DREAM: A challenge data set and models for dialogue-based reading comprehension, Transactions of the Association for Computational Linguistics 7 (2019) 217–231.
[36] N. F. Liu, R. Schwartz, N. A. Smith, Inoculation by fine-tuning: A method for analyzing challenge datasets, in: NAACL, Minneapolis, Minnesota, 2019.
[37] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: ICML, 2017.
[38] N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis – connecting the branches of systems neuroscience, Frontiers in Systems Neuroscience 3 (2008).
[39] T. A. Trost, D. Klakow, Parameter free hierarchical graph-based clustering for analyzing continuous word embeddings, in: TextGraphs, Vancouver, Canada, 2017.
[40] D. Yenicelik, F. Schmidt, Y. Kilcher, How does BERT capture semantics? a closer look at polysemous words, in: BlackboxNLP, Online, 2020.
[41] R. Aharoni, Y. Goldberg, Unsupervised domain clusters in pretrained language models, in: ACL, Online, 2020.
[42] C.-K. Yeh, J. S. Kim, I. E. H. Yen, P. Ravikumar, Representer point selection for explaining deep neural networks, 2018. arXiv:1811.09720.
[43] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, Z. C. Lipton, Learning to deceive with attention-based explanations, in: ACL, Online, 2020.
[44] S. Petrov, P.-C. Chang, M. Ringgaard, H. Alshawi, Uptraining for accurate deterministic question parsing, in: EMNLP, 2010.
[45] H. Strobelt, S. Gehrmann, H. Pfister, A. M. Rush, Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks, IEEE Transactions on Visualization and Computer Graphics 24 (2018) 667–676. doi:10.1109/TVCG.2017.2744158.
[46] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, I. Titov, Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, in: ACL, Florence, Italy, 2019.
[47] T. Ruzsics, O. Sozinova, X. Gutierrez-Vasques, T. Samardzic, Interpretability for morphological inflection: from character-level predictions to subword-level rules, in: EACL, Online, 2021.
[48] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, PLoS ONE (????).
[49] Y. Lakretz, G. Kruszewski, T. Desbordes, D. Hupkes, S. Dehaene, M. Baroni, The emergence of number and syntax units in LSTM language models, in: NAACL, Minneapolis, Minnesota, 2019.
[50] V. Ravishankar, A. Kulmizev, M. Abdou, A. Søgaard, J. Nivre, Attention can reflect syntactic structure (if you let it), in: EACL, Online, 2021.
[51] K. W. Church, P. Hanks, Word association norms, mutual information, and lexicography, in: 2ACL, Vancouver, British Columbia, Canada, 1989.
[52] T. Mikolov, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NeurIPS, 2013.
[53] N. Garneau, M. Hartmann, A. Sandholm, S. Ruder, I. Vulic, A. Søgaard, Analogy training multilingual encoders, in: AAAI, 2021.
[54] D. Alvarez-Melis, T. Jaakkola, A causal framework for explaining the predictions of black-box sequence-to-sequence models, in: EMNLP, Copenhagen, Denmark, 2017.
[55] G. Kobayashi, T. Kuribayashi, S. Yokoi, K. Inui, Attention is not only a weight: Analyzing transformers with vector norms, in: EMNLP, Online, 2020.
[56] S. Vashishth, S. Upadhyay, G. S. Tomar, M. Faruqui, Attention interpretability across NLP tasks, 2019. arXiv:1909.11218.
[57] K. Heylen, D. Speelman, D. Geeraerts, Looking at word meaning. an interactive visualization of semantic vector spaces for Dutch synsets, in: LINGVIS & UNCLH, Avignon, France, 2012.
[58] E. Reif, A. Yuan, M. Wattenberg, F. B. Viegas, A. Coenen, A. Pearce, B. Kim, Visualizing and measuring the geometry of BERT, in: NeurIPS, volume 32, 2019.
[59] Y. Kim, A. M. Rush, Sequence-level knowledge distillation, in: EMNLP, Austin, Texas, 2016.
[60] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, in: ICLR, 2018.
[61] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applica-

tions, Proceedings of the IEEE 109 (2021) 247–278.
doi:10.1109/JPROC.2021.3060483.