Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks

Kieran Browne[a,*], Ben Swift[b]

[a]Research School of Humanities & the Arts
Australian National University
[b]Research School of Computer Science
Australian National University

Abstract

Recent papers in explainable AI have made a compelling case for counterfactual modes of explanation. While counterfactual explanations appear to be extremely effective in some instances, they are formally equivalent to adversarial examples. This presents an apparent paradox for explainability researchers: if these two procedures are formally equivalent, what accounts for the explanatory divide apparent between counterfactual explanations and adversarial examples? We resolve this paradox by placing emphasis back on the semantics of counterfactual expressions. Producing satisfactory explanations for deep learning systems will require that we find ways to interpret the semantics of hidden layer representations in deep neural networks.

Keywords: explainable AI, counterfactual explanation, adversarial examples, semantics
2010 MSC: 00-01, 99-00

1. Introduction

Deep neural networks (DNNs) will not be explainable without first addressing the scarcity of semantics. Computational methods already exist to pro-

[*] Corresponding author
Email address: kieran.browne@anu.edu.au (Kieran Browne)

duce model-agnostic explanations that are understandable to laypersons. These methods simply do not function as explanations when applied to ambiguous or low-level representations that are common to DNNs. We will argue that this is not simply a limitation of existing explanatory methods, but rather that *there can be no explanation without semantics*. Because deep learning (DL) typically operates on "raw data", with little semantic content (e.g. pixels and characters), this realisation serves to clarify the explainability challenge; we either find a way to extract the semantics presumed to exist in the hidden layers of the network or concede defeat.

Recent papers in explainable artificial intelligence (XAI) have identified problems with the field's theoretical bases. Tim Miller [1] argues that the field typically operates with only an intuitive notion of what explanation is; and one which is divorced from how humans explain and understand explanation. He proposes that XAI adopt "everyday explanations", based on a set of principles from psychological and social scientific research. Similarly, Sandra Wachter et al. [2] propose "counterfactual explanations" which are consistent with the principles identified by Miller. Wachter et al. additionally specify a method for generating counterfactual explanations. Counterfactual explanations, as Wachter et al. demonstrate, are model-agnostic, automatically computable and comprehensible to laypersons. The authors argue that these counterfactual explanations offer the path to explaining complex algorithmic systems to anyone. However, equivalent computations have been used in DL research since 2014, though not to produce explanations. Instead, in the context of DL research, the counterfactual computation produces "adversarial examples"; imperceptibly modified inputs which cause the network to inexplicably and confidently misclassify.

This should give us pause for thought; how is it possible that the same method can on the one hand represent a promising new means of explaining the decisions of a DNN to anyone, and on the other hand represent a confounding brittleness in that same decision making process? We call this phenomenon, *the explanatory divide*. We will argue that this divide reveals a blind spot in XAI research with regards to semantics.

1.1. The argument

This article proceeds as follows: we begin in Section 2 by outlining the history of XAI research and its revival in the era of deep learning. In Section 2.1 we describe the novel social challenge posed by this technology and how this has affected the problem of explainability. In Section 2.2 we introduce two recent papers that have challenged the prevailing methods in XAI by introducing human-centric modes of explanation to the field. In Section 3 we show the equivalence of counterfactual methods proposed by Wachter et al. with those used to generate adversarial examples and examine the explanatory divide apparent in the two usages. In Section 3.1 we refute Wachter et al.'s account of the explanatory divide. In Section 3.2 we argue that the explanatory divide is instead a consequence of the semantic content of the perturbed vector. In Section 4 we show that semantics is a blind spot in XAI research attributable to researchers' concern for a computational solution. In Section 4.1 we argue that semantic issues are endemic to DL due to operating on "raw data". In Section 4.2 we examine the existing research in extracting the semantics of hidden layers and discuss the ongoing challenges. We conclude by proposing a possible path forward for combining existing explainability methods with the partially known semantics of DNNs.

1.2. A note on "semantics"

Although "semantics" is a common term in DL literature, its use is ambiguous. In the early literature on DL it is commonly claimed that DNNs learn semantic features automatically in order to solve problems (e.g. Bengio [3], LeCun et al. [4, p. 441]). This is based on the assumption that in order to solve complex problems like image classification, the network must generate intelligible intermediate representations (e.g. whiskers and paws used to identify cats). However this kind of semantics has not been reliably shown to exist and it remains a significant challenge to find a mapping between the latent spaces of DNNs and human concepts. Other researchers appear to use "semantics" to refer to any kind of internal representation, that is, any way of carving up the world whether or not it maps to something a human might understand. These duplicate uses of "semantics" cause significant ambiguity and some authors have resorted to tautology to distinguish the two. Biran and Cotton [5] for example use "semantically meaningful representations" to distinguish internal representations which correspond to categories which humans (or perhaps specifically English speakers) find meaningful.

Following the language of semiotics (see [6]), we take semantics to be the relation between sign and signified. This is distinguished from syntactics (relations between signs) and pragmatics (the relation between signs and the interpreter). Of course the representations in DNNs are not really signs, at least not in the standard sense. Their relation to meanings are correlative and continuous rather than discrete as in symbolic systems. Although some accounts of meaning disallow fuzzy concepts, others (e.g. later Wittgetnstein) argue that many of our concepts have "blurred edges" and we are able to use them productively nonetheless [7, sect. 71]. This is the sense in which we suggest "semantics" should be understood in DL. Whether we consider a hidden unit to *mean* "whiskers" then, depends on how reliably it correlates to the English language concept "whiskers".

2. Background

DL has afforded significant advances in a broad range of problems. However, little progress has been made in explaining the behaviour and decision-making processes of these systems. Although the reinterpretation of machine learning as artificial intelligence in the 1990s revived the decades-old field of XAI (eXplainable Artificial Intelligence), DL remains a "black box"—a descriptor that has followed DNNs and their precursory methods since the 1990s [8][9].

This desire for explanations of algorithmic decisions predates DL, beginning in the context of rule-based expert systems as early as the 1970s [5]. Research on XAI has been tied to AI such that it has endured the same periods of disenchantment known colloquially as "AI winters". An era-agnostic survey of explainability is provided in [5]. We will focus our account on the contemporary (deep) neural network paradigm of AI and XAI.

Since the rise of DL in the mid 2000s, artificial neural networks (ANNs) have become more complex by orders of magnitude. Unlike simpler statistical models, ANNs are generally considered to be "black boxes" because the representations they generate are not readily interpretable [10].

In machine learning, the goal of explainability has often been pursued through visualisation [5]. In late 80s and early 90s, a number of diagrammatic visualisations emerged for ANNs; these usually relied on a traditional graph-theoretic "nodes and edges" representation augmented with edge-weight information. As networks increased in size throughout the 90s these images became increasingly difficult to interpret [11]. For deep architectures of contemporary scale they are essentially obsolete e.g. Microsoft's Turing Natural Language Generation T-NLG has 17 billion parameters [12]. Today, DL visualisations tend to represent only single layers or single neurons rather than an entire network [13].

The other common approach in explainability research is to approximate the behaviour of an ANN with a more "interpretable" model. In the late 80s early 90s this was usually referred to as *rule extraction* [14]. This meant distilling the many calculations of a neural network into a series `IF...THEN` rules akin to symbolic AI. Rule extraction is rarely mentioned in the "deep" era of neural networks, but similar methods are still used under alternative names such as *knowledge distillation* [15], which notably omits any reference to explainability. Knowledge distillation, like rule extraction, uses a trained DNN to train a simpler, *interpretable* model such as a decision tree [16]. From an XAI perspective a lingering conceptual problem remains; if the simpler model is similar enough to capture the decisions of the DNN, why use DL at all?

2.1. Why now?

It is still common to see XAI papers use adoption as a motivation for explainability. If users do not understand or trust the model, we are told, they will choose not to use it [17]. While this justification has been made many times, it has little to do with the need for explainability as it exists today. While some opt-in AI-branded DL services do exist, the significant growth area for DL is in systems that people are *subject to* through institutional means. The application of DL in institutions of social and political importance; e.g. banks, courts, media distribution, political campaigning etc. has naturally drawn increasing attention from social scientists and increased scrutiny from law makers [18][19][20]. Explainability matters now more than ever because DL is being used to determine social realities, e.g. by banks to distribute credit or by the justice system to decide parole. Here we are in danger of conflating *prediction* with *prescription.* To borrow the language of speech acts, prediction is *constative*; that is, it makes a claim about the world, e.g. "the house price will be $1,000,000", "this is a picture of a tennis ball" or "this digit is a 6", which may be evaluated by independent observation as more or less accurate. In normal use, the truth is independent of the prediction, and the predictions may be judged as more or less accurate. However, used prescriptively, e.g. job applications, loan decisions, bail decisions, social reality is wholly determined by the prediction. In these cases the need for an explanation of the network's outputs/outcomes is paramount—there is no independent ground truth outside the algorithmic decision.

2.2. A human turn in explainable AI?

XAI has been significantly siloed from other disciplinary understandings of explanation. However two recent papers propose new approaches sensitive to the human factors of explanation. Both draw on bodies of knowledge from outside computer science to propose modes of explanation inspired by human-to-human explanations. The first we will discuss, from XAI researcher Tim Miller [1], draws on research from philosophy, psychology, social science and cognitive science to provide a theoretical framework for XAI sensitive to how humans explain and understand explanation. The second, from an interdisciplinary team led by Sandra Wachter [2], proposes a practical method for providing a legally-compelled explanation for those subject to algorithmic decision making. These two approaches appear to have emerged independently but are complementary. Both argue for a shift in explainability research toward social and context-dependent modes of explanation.

Miller's *everyday explanations* provide a conceptual foundation for this development. Raising a concern about the theoretical underpinnings of XAI, he claims that most research is guided merely by researchers' "intuition" for what constitutes a good explanation and argues that computational solutions are not sufficient for explainability [1]. Miller argues that XAI should take inspiration from the way humans explain to each other. He surveys existing literature on explanation in philosophy, psychology, social science and cognitive science in order to draw four conclusions about explanations:

1. Explanations are *contrastive*; that is, they "explain the cause of an event *relative to some other event* that did not occur".

2. They are *selective*; that is, we rarely if ever give an explanation that describes the "complete" cause of an event.

3. They are *social*; that is, they are presented relative to who the explainee is, and what they can be expected to understand.

4. *Probabilities probably don't matter*; that is, statistical explanations of events are unsatisfying unless accompanied by causal explanations.

Around the same time, Wachter et al.'s *counterfactual explanations* [2] appeared in the *Harvard Journal of Law & Technology*, motivated by the looming challenge of the "right to an explanation" under the European Union's General Data Protection Regulation's (GDPR) and the competing technological, social and legal challenges therein. The paper proposes the counterfactual explanation as a way to offer meaningful explanations of algorithmic decisions to those affected. Counterfactual explanations are a model-agnostic method for generating explanations of algorithmic decisions for a lay audience based on the notion of the *counterfactual* from the philosophy of causation.

A counterfactual explanation is defined by Wachter et al. as a minimal set of changes to the input data found to produce a desired decision in the network.

More formally:

> "Score p was returned because variables V had values $(v_1, v_2, ...)$ associated with them. If V instead had values $(v'_1, v'_2, ...)$, and all other variables had remained constant, score p' would have been returned."

Wachter et al. go on to present a small number of case studies which demonstrate the efficacy of this method in real-world cases. The method is extremely effective in the scenarios described. When applied to a specific example, as in the paper's example of a bank loan, the above formal definition is translated into something which reads as plain English:

> "You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan."

The counterfactual explanation embodies the principles of everyday explanations articulated in Miller's paper. Where existing research treats explanation axiomatically, counterfactuals are conscious of the audience, i.e. they are *social.* The counterfactual explanation is for a lay audience, specifically the GDPR's "data subjects". Counterfactual explanations are also *selective*; as the title of the paper suggests, they allow for explanations "without opening the black box," or in other words, without completely revealing how the algorithm works. Most strikingly, the counterfactual explanation is *contrastive*; it points to the changes in the input which would have resulted in an alternative outcome.

3. Adversarial examples are counterfactual explanations

The counterfactual explanation is only nominally new to DL/XAI. Since 2014 the generative perturbation of input vectors to probe at decision boundaries has been the topic of a significant body of research under the banner of "adversarial examples" [21][22][23][24]. Adversarial examples, like counterfactual explanations, are algorithmically generated perturbations to input data which are optimised to alter the DNN's output in a particular manner. Much of this research has concerned itself with image classification, such as convolutional neural networks, however, successful adversarial examples have been produced in audio [25] and text [26] domains as well.

In adversarial examples, imperceptible (to a human observer) changes to the input cause the network to entirely and confidently misclassify. The originators of this line of research, Szegedy et al. [21] call this phenomenon "intriguing" and "counter-intuitive" while later researchers have commonly described the adversarial example in terms of a vulnerability to attack [27][24].

A more formal definition helps to clarify the equivalence between adversarial examples and counterfactual explanations. Both are defined as a constrained optimisation problem where the objective is to change the network's output to a some other output by minimally altering the input. Consider a DL classifier

$$f_w(x) = y \tag{1}$$

where y is the predicted class of input x. For both adversarial examples and counterfactual explanations we seek an input x' as close as possible to x such that our network f_w classifies x' as a different target class y'. This can be written as an optimisation problem:

$$\begin{aligned} \underset{x'}{arg\,min} \quad & d(x, x') \\ \text{subject to} \quad & f_w(x') = y' \neq y \end{aligned} \tag{2}$$

The distance metric d is measure of the distortion of x' relative to x—the distance between the original input and the altered one. The objective is produce the target output y' while minimising the distance $d(x, x')$. As a consequence the definition of d will influence the resulting of adversarial or counterfactual.

The first paper to propose adversarial examples uses the Euclidian distance (L_2 norm):

$$d(x, x') = |x' - x|_2 \tag{3}$$

Later papers including Wachter et al.'s use other metrics, including the Manhattan distance (L_1 norm), but these metrics do not fundamentally change the nature of the method.

This presents an apparent paradox for explainability researchers: if these two procedures are formally equivalent, what accounts for the explanatory divide apparent between counterfactual explanations and adversarial examples?

3.1. Making sense of the explanatory divide

Wachter et al. acknowledge that an adversarial example is "a counterfactual by a different name," [2] but appear unconcerned by this, and propose two grounds for the explanatory divide (our term). The first is that "none of the standard works on adversarial perturbations make use of appropriate distance functions" and the second is that adversarial examples are invalid because they do not come from the "space of real-images" and therefore do not qualify as "possible worlds". The first amounts to a challenge over the "correct" definition of the distance metric d in Equation 3, the second is a metaphysical claim.

The claim that none of the "standard works" on adversarial examples use an appropriate distance metric needs to be understood in the context of Wachter et al.'s own discussion of the properties of an appropriate distance metric. While they stress that case-specific considerations must be taken into account, they suggest as a first approximation to use the L_1 norm weighted by the inverse median absolute deviation (MAD). The MAD is chosen for its robustness to outliers, while the L_1 norm is chosen for its sparsity-inducing properties; i.e. it restricts differences to as few input dimensions as possible.

From these properties it is possible to understand the concern the authors have with the distance metrics favoured by the adversarial example research community. Wachter et al. are correct that the majority of adversarial example research "favour[s] making small changes to many variables" so that the difference is diluted across the inputs and this contributes to the indistinguishableness of the perturbations. We agree that this would theoretically make these less useful as explanations; in particular that it contravenes Miller's principle that explanations should be *selective.* However, although sparse counterfactuals may be preferable to dense ones, sparsity is not in itself sufficient for explanation. This is clear from the research of Su et al. [24], who do restrict their variations to a single input feature, in this case a single pixel, whose impact is visually identifiable.

Su et al. demonstrate that in the majority of cases, changing a single pixel is enough to cause a network to misclassify to at least one other class. The generated perturbations are sparse and salient and in spite of this, they remain distinctly adversarial.

Wachter et al.'s second account for the explanatory divide is that adversarially-perturbed images do not represent "possible worlds". The intuition here is that standard adversarial perturbations appear as very slight "noise" are distinctly *not* random; instead they encode the signature of a class that is not present in "natural" images. Perhaps there is some truth to this—it does seem unlikely that noise with these very specific properties would occur by chance. However, there is something distinctly unsatisfying about this account. Should we regard crafted or manipulated images as "impossible"? We live amongst a proliferation of unnatural images. Additionally, adversarially-perturbed images have been shown to work in the real world even when printed out and photographed through low-quality cameras [27][29][30]. Whether or not adversarial examples are "possible" without contrivance, researchers must take seriously the possibility of encountering adversarial examples that have been intentionally planted in the world.

3.2. The explanatory divide and semantics

If the explanatory divide cannot be accounted for by poor distance metrics or impossible worlds, how else can we make sense of it? We believe the answer lies in the semantics of the perturbed vector.

An example helps to clarify this assertion. The input data to the AI decision processes in Wachter et al.'s examples are expressed using semantically dense and contextually relevant dimensions: income, grade-point average, body-mass index, etc. Some of these also represent factors that we (humans) might consider appropriate evidence to base a loan decision on, others are certainly not (e.g. age, and race). Regardless, a counterfactual explanation that operates on semantically dense dimensions helps us to understand the decision even if it causes us to question its validity. A counterfactual for a hiring decision that identifies race, gender or age as deciding factors is explanatory even if it only provides a justification for disregarding the results. In the framework of Miller's everyday explanations, these dimensions are *social* as they can be expected to be understood by the explainee.

In contrast, adversarial examples are produced when the same computation is applied to data with little semantic content. Much of DL operates on "raw data", i.e. individual pixels, letters, waveform samples, bits etc. Reductio ad absurdum; explaining an image classified as "building" based on the redness value in a particular pixel is unsurprisingly unhelpful. Instead, the factors that a human would consider to be relevant are dispersed and discontinuous in pixel space, *they are not discoverable using sparse or dense perturbations*. Debiasing the network is also extremely challenging when operating with low-level semantics, because factors we would wish to disallow are equally dispersed and discontinuous.

Mathematically speaking, there is no difference between a vector of pixel values and a vector of semantically rich features. Therefore, the crucial relationship from an XAI perspective is not between the network and the computation producing the explanation, but between the semantics in the network and the human explainee.

4. Semantics is the core challenge of explaining deep learning

As we have discussed, the efforts of XAI researchers have been significantly focused on finding ways to reduce the complexity of a given network. These methods share the same fatal flaw as the counterfactual, no computation can get around the semantic problem. Any explanatory technique will produce a non-sequitur if it, for example, attempts to explain driving instructions from pixel values. For XAI this appears to be a significant blindspot.

4.1. The culture of Deep Learning

Of course, the simplest solution to this explainability problem would be to apply DL only to higher-level, contextually relevant representations. But this would require us to forego what LeCun, Bengio and Hinton [4] call the "key advantage" of DNNs; that they can operate on "raw data" and do not require feature engineering to produce useful results.

Whether counterfactual explanations or any other one of a trove of abandoned methods (rule-extraction, random-forests, etc.) may again be used to explain DNNs rests on whether the network's learned semantics, in the sense we define in Section 1.2, can be discovered.

4.2. In search of semantics

Although significant early papers in the DL literature presume that DNNs discover their own semantics in order to solve problems [3][4] the community as a whole appears to have quietly abandoned this assertion in recent years. We believe this to be a fatal error if we hope to explain these systems.

Let us assume for a moment that semantics can be discovered. Given a clear knowledge of the semantics of hidden layer neurons in a network, it would be possible to generate counterfactual explanations consisting only of semantically dense and contextually relevant dimensions in the network's feature space, perhaps even without needing to synthesise inputs in pixel space. A counterfactual explanation at this level might read:

> The input image was labelled "building" because hidden neuron 41435, which generally activates for hubcaps, had an activation of 0.32. If hidden neuron 41435 had an activation of 0.87 the input image would have been labelled "car".

This is a contrived example, but it illustrates what should be possible if the hidden represenations in DNNs were interpretable. The problem is that we are either yet to develop appropriate tools to discover these DL's internal semantics, or they do not exist.

4.3. Revealing hidden layers

Researchers have made some progress in identifying the semantics of the hidden units ("neurons") (see e.g. [31][13]). In one of the earliest cases, Erhan et al. [31] identifies a neuron that appears to represent "faces", although it remains unclear how general/specific this category actually is (see [11]). A number of visualisation methods have been developed which serve to interpret hidden representations in DNNs. One technique is simply to collate dataset examples that maximise the activation of a given hidden unit [21]. Humans are often able to perceive commonalities between these high ranking examples. However, this method is susceptible to confounding factors. Saliency maps [32][33] serve to avoid some of these pitfalls. They visualise the area(s) of images which contribute significantly to a hidden unit's activation, allowing a viewer to identify contributing visual features. Feature visualisations [31][34][22][35] synthesise images which maximally activate a hidden unit. Olah et al. [36] demonstrate that using many of these methods in tandem can be particularly enlightening.

Using methods described above, Olah et al. [13] discover a neuron that responds to different kinds of sports balls (e.g. golf balls, tennis balls, footballs, baseballs) (Fig. 2). This is particularly compelling because the neuron appears to have captured something approximating the human category that might be called "sports ball" in spite of their differing appearances and contexts. However, this is a particularly favourable example, as units with clear semantics appear to be exception and not the rule. Olah et al. found a number of cases where representations were "poly-semantic" [36] e.g. a unit was discovered that activates for cats and foxes but also cars. This is difficult to make sense of given that there is no apparent visual, contextual or categorical similarity between the cats and cars. In other instances units appeared to have no discernible semantics whatsoever [13]. In summary, while there have been a number of very promising cases, the extraction of semantics from hidden units is far from a solved problem.

Olah et al. note that single neurons (i.e. standard basis vectors) may not necessarily be the vector directions with a clear one-to-one mapping with English semantics, and that other basis vectors seem to be just as meaningful [13]. While this increases the likelihood of some arbitrary direction mapping to a human concept, it also makes the search space essentially infinite (floating point precision notwithstanding).

5. Conclusion

The equivalence of adversarial examples and counterfactual explanations demonstrates the necessity of semantics to the problem of explainability. Semantics appears to be a blindspot for XAI, which has instead focussed on computational innovations. The necessary computational methods to explain DL already exist—provided we use semantically rich and contextually relevant representations as inputs, or we can discover the semantics in hidden layers. With current research these semantics have not been shown consistently to exist and be discoverable.

References

[1] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[2] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law & Technology 31 (2).

[3] Y. Bengio, et al., Learning deep architectures for ai, Foundations and trends® in Machine Learning 2 (1) (2009) 1–127.

[4] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.

[5] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), Vol. 8, 2017.

[6] C. W. Morris, Foundations of the theory of signs, in: International encyclopedia of unified science, Vol. 1, Chicago University Press, 1938, pp. 1–59.

[7] L. Wittgenstein, Philosophical investigations, Basil Blackwell Ltd, 1958.

[8] M. Spining, J. Darsey, B. Sumpter, D. Nold, Opening up the black box of artificial neural networks, Journal of chemical education 71 (5) (1994) 406.

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

[10] L. Breiman, Statistical modeling: The two cultures, Statistical science 16 (3) (2001) 199–231.

[11] K. Browne, B. Swift, H. Gardner, Critical challenges for the visual representation of deep neural networks, in: Human and Machine Learning, Springer, 2018, pp. 119–136.

[12] C. Rosset, Turing-NLG: A 17-billion-parameter language model by Microsoft, library Catalog: www.microsoft.com (Feb. 2020).

[13] C. Olah, A. Mordvintsev, L. Schubert, Feature visualization, Distilldoi:10.23915/distill.00007.

[14] R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-based systems 8 (6) (1995) 373–389.

[15] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: NIPS Deep Learning and Representation Learning Workshop, 2015.
URL http://arxiv.org/abs/1503.02531

[16] G. Hinton, N. Frosst, Distilling a neural network into a soft decision tree, 2017.
URL https://arxiv.org/pdf/1711.09784.pdf

[17] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[18] S. Barocas, A. D. Selbst, Big data's disparate impact, California Law Review 104 (2016) 671–732.

[19] J. Burrell, How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society 3 (1).

[20] J. Stilgoe, Machine learning, social learning and the governance of self-driving cars, Social studies of science 48 (1) (2018) 25–56.

[21] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations, 2014.

[22] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436.

[23] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.

[24] J. Su, D. V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks, IEEE Transactions on Evolutionary Computation.

[25] M. Alzantot, B. Balaji, M. Srivastava, Did you hear that? adversarial examples against automatic speech recognition, arXiv preprint arXiv:1801.00554.

[26] N. Papernot, P. McDaniel, A. Swami, R. Harang, Crafting adversarial input sequences for recurrent neural networks, in: MILCOM 2016-2016 IEEE Military Communications Conference, IEEE, 2016, pp. 49–54.

[27] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, arXiv preprint arXiv:1607.02533.

[28] N. Narodytska, S. Kasiviswanathan, Simple black-box adversarial attacks on deep neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017, pp. 1310–1318.

[29] T. Brown, D. Mane, A. Roy, M. Abadi, J. Gilmer, Adversarial patch.
URL `https://arxiv.org/pdf/1712.09665.pdf`

[30] S. T. Jan, J. Messou, Y.-C. Lin, J.-B. Huang, G. Wang, Connecting the digital and physical world: Improving the robustness of adversarial attacks.

[31] D. Erhan, Y. Bengio, A. Courville, P. Vincent, Visualizing higher-layer features of a deep network, Technical Report 1341 1341 (2009) 1–13.

[32] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: ICLR, 2014.

[33] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.

[34] Q. V. Le, Building high-level features using large scale unsupervised learning, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 8595–8598.

[35] A. Karpathy, J. Johnson, L. Fei-Fei, Visualizing and understanding recurrent networks, ICLR Workshop.

[36] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, A. Mordvintsev, The building blocks of interpretability, DistillHttps://distill.pub/2018/building-blocks. doi:10.23915/distill.00010.