Critical challenges for the visual representation of deep neural networks

Kieran Browne Ben Swift Henry Gardner

2018

Introduction

The internal patterns and processes of artificial neural networks are notoriously difficult to interpret. The advent of deep neural networks has heightened this challenge and rendered many existing interpretive methods obsolete. This has prompted new research into methods for interpreting neural networks. One of the most fruitful areas of this research, and the focus of the present chapter, is visual representation. Central to this research is the concern that neural networks are black boxes. Growing awareness and criticism of machine learning in public discourse has transformed the explanation of these algorithms into a social and political as well as technical concern. Interrogating the black box is a compound problem. Its constituent parts cross disciplinary boundaries to raise questions of engineering, epistemology, aesthetics and semantics. We will argue that it is valuable for researchers aiming to explain neural networks through visual representation to become familiar with the interdisciplinary critical scholarship on this topic. We begin this chapter with a discussion of the black box problem which draws upon this research. We then situate the visual representation of deep neural networks in data visualisation and interface theory, and discuss the specific challenges it poses. In section we outline the diagrammatic representations favoured by researchers prior to 2006 and offer an explanation for their rapid obsolescence following the rise of deep neural networks. In section we present six diverse case studies in contemporary visual representation of neural networks. The case studies come from research, industry and individual makers. They have been selected for their potential to highlight critical challenges rather than their citation metrics. In the final section we summarise the ideas raised in the case studies as a list of takeaways for students or researchers engaging in this area.

The black box problem

The term “black box” describes a system with clearly observable inputs and outputs, but with inscrutable internal processes. Neural networks are considered black boxes not because we cannot see inside as such; the relationship between input and output is observable but unintelligible. An apparently strong relationship in one layer of the network may be cancelled out or inverted in the next or simply diluted by countless other smaller relations. Like many machine learning (ML) techniques, neural networks trade interpretability for predictive power (Breiman 2001).

The black box problem is an ongoing concern for researchers and a growing concern for institutions and individuals who use trained models but are estranged from their development. If it is difficult to understand how neural networks make decisions, then it becomes difficult to trust the decisions they make.

The black box problem has been cited many times as a barrier to the adoption of neural networks (Benitez, Castro, and Requena 1997)(Duch 2003)(Tzeng and Ma 2005). This concern has proved unwarranted as the successes of deep neural networks in countless disparate fields have led to pragmatic adoption despite difficulties explaining their behaviour. Over the past decade deep neural networks have received massive investment from research councils, industry and government and have been applied to problems as broad-ranging as translation (Bahdanau, Cho, and Bengio 2014), gameplay (Mnih et al. 2013), fine art (Elgammal et al. 2017), stock trading (Längkvist, Karlsson, and Loutfi 2014) and object recognition (Krizhevsky, Sutskever, and Hinton 2012).

Despite some early claims to have solved the black box problem (Benitez, Castro, and Requena 1997), concern for explainability remains. Indeed the black box has become a central metaphor for questioning how and whether neural networks can be explained. Notably, researchers have made attempts at “illuminating” (Olden and Jackson 2002), “coloring” (Duch 2003), “opening” (Tzeng and Ma 2005)(Sussillo and Barak 2013) and “greying” (Zahavy, Ben-Zrihem, and Mannor 2016) black boxes.

Since 2006, the use of deep architectures, i.e. neural networks of many layers, has become prevalent (Bengio 2009). The comparatively tiny neural networks used by researchers in the 90s have been replaced by massively deep, massively multivariate networks. AlexNet (Krizhevsky, Sutskever, and Hinton 2012), for example, contains 650,000 neurons and 60 million parameters. This enormous growth has rendered many existing modes of visual representation ineffective.

The invention of new types of networks has created additional challenges for explainability. Much of the recent success of neural networks has been made with alternative architectures such as convolutional neural networks and long short term memory (LSTM) networks (Sainath et al. 2015). These models augment the standard feedforward neural network with structures that enable new kinds of modelling but introduce additional behavioural complexities.

Interdisciplinarity

As a potentially transformative technology, ML has consequences which reach far beyond computer science. As a novel way of representing knowledge ML raises questions for epistemology (Tunç 2015). Because ML is subject to human biases, anti-discrimination law must be reformed to account for it (Barocas and Selbst 2016). As a technology driving socially consequential mechanisms such as news trends and credit scores, the opacity of ML becomes sociological concern (Burrell 2016). There are more examples of interdisciplinary research into ML than can be enumerated here. In each case, the authors describe the critical challenges of machine learning in the nomenclature of their field. This seemingly disparate scholarship provides a useful lens with which to understand ML itself and its effects in the world. In section there are a number of cases where interdisciplinary research is leveraged to make sense of some of the particular critical challenges posed by the visual representation of neural networks.

Historical Precedents

The defining trope of pre-2006 neural network visualisation was the structure of the network itself. This was most commonly represented as a graph of nodes and edges arranged in neat layers left-to-right or bottom-to-top (Fig. 1, 2). In these, the network's topology is central to the representation.

Craven and Shavlik (1992)'s review paper surveyed the contemporary cutting edge of artificial neural network visualisation. Notable amongst these were the Hinton and Bond diagrams, which have a structural focus. Curiously the authors criticise the Hinton diagram for not showing the network's "topology" despite its elements being arranged in layers that mimicked the network's structure. Although the Hinton and Bond diagrams can theoretically include any number of inputs, they are practically uninterpretable for large networks (Tzeng and Ma 2005).

What followed was a general convergence to and then refinement of a particular representation which centred around a structural depiction of the network. We call this type of image a "neural interpretation diagram" (NID) borrowing from the name used by Özesmi and Özesmi (1999). In NIDs, neurons are represented as dots or circles, and synapses are represented as lines with thickness and colour indicating value. Variations on this theme can also be seen in the work of other researchers (Özesmi and Özesmi 1999)(Streeter, Ward, and Alvarez 2001)(Olden and Jackson 2002)(Tzeng and Ma 2005). In their 2005 paper, Tzeng and Ma cite the Hinton and Bond diagrams as precedents to their work and note that these had failed to scale to larger networks. They go on to apply their take on

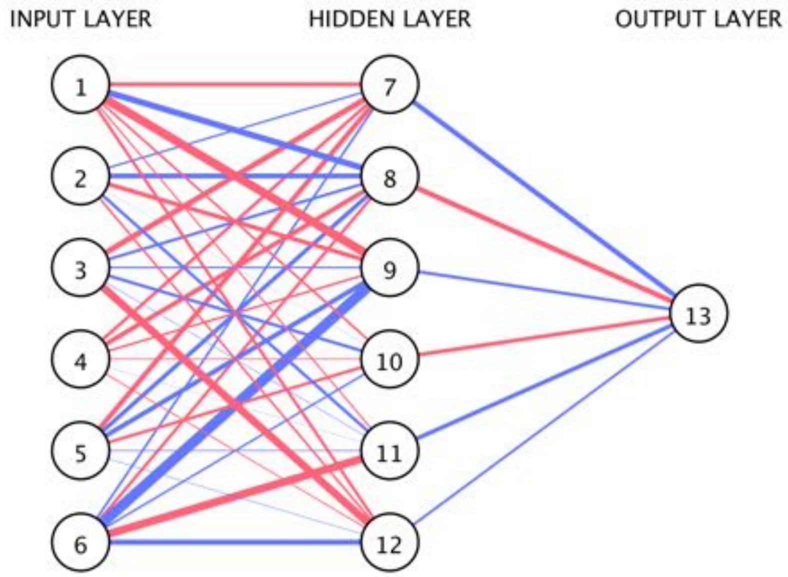


Figure 1: Neural interpretation diagram based on Özesmi and Özesmi (1999) — the visual representation mirrors the structure of the schematic representation introduced by Rosenblatt (1962). The thickness of each edge is relative to the absolute value of the synapse weight. Blue edges are represent positive weights and red edges represent negative weights. Source code for image available at (Browne 2017b).

the NID to a network with 8300 synapses. Although it doesn't appear to be the authors' intent, the resulting image demonstrates that NIDs share a similar scaling problem to Hinton and Bond diagrams. In the multitude of criss-crossing lines it is impossible to make sense of individual synapse values (Fig. 2).

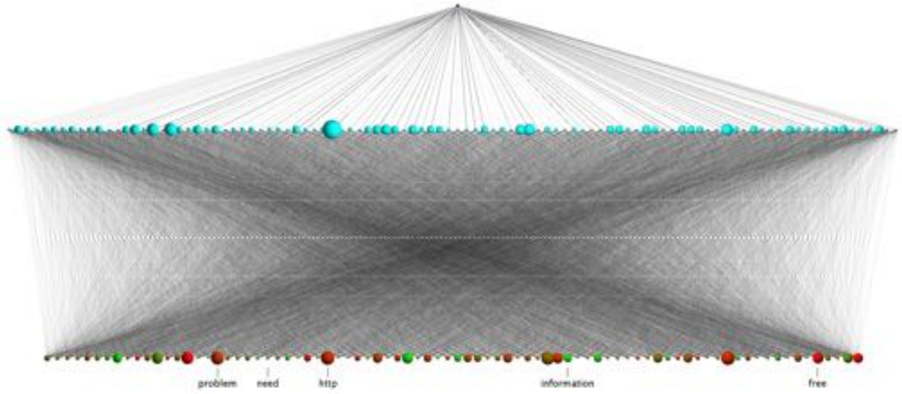


Figure 2: Diagram for a large spam-classifier neural network based on Tzeng and Ma (2005). The size of nodes indicates their relative importance in identifying spam. Node color is used to indicate the mean and standard deviation of a given node. Source code for image available at (Browne 2017c).

The NID and its variants went out of fashion because they failed to scale to the very large, very deep networks that became the focus of research after 2006. More significantly however, these diagrams fail to target the core of the problem. Neural networks are not black boxes because we cannot see inside at all; the value of any weight, bias or activation can be easily accessed. It is clear that seeing the network does not in itself create understanding. The real challenge for explaining neural networks is untangling meaningful relationships from the multitude of connections.

By *meaningful*, we mean relating to representations for which people have concepts. It is not useful to explain how a pixel value relates to steering instructions. For self-driving cars we want our explanations in terms of semitrailers, cliffs and pedestrians. Because deep neural networks produce their own intermediate representations we require means of mapping these back to meaningful concepts and testing how robust these mappings are.

Additionally it is not possible to explain the behaviour of a network by its structure as structure does not dictate behaviour; neural networks are universal approximators (Hornik, Stinchcombe, and White 1989). Two structurally identical networks will approximate different behaviours given different data. A neural network's behaviour is latent in the dataset, not the network.

Case Studies

In this section we explore contemporary developments in visual representation of neural networks, examining six case studies drawn from research, industry and individuals. These cases have been selected for their potential to highlight critical challenges of representing neural networks visually.

AI Brain Scans

The *AI Brain Scans* (2016-ongoing) (Fyles 2017), are a collection of visualisations by Matt Fyles of Graphcore. Initially referred to as "large scale directed graph visualizations" (Fyles 2016), the images were dubbed "brain scans" in a report by *Wired* in early 2017 and the title stuck (Burgess 2017).

The *AI Brain Scans* visualise the edges and nodes of a neural network's computational graph. They are produced with Graphcore's proprietary ML framework, Poplar, and the open source graph visualisation software, Gephi (Bastian et al. 2009), which is used primarily for social and biological network analysis. The "brain scans" are structural like the NID, but rather than representing the network's topology they represent the graph of computations required to train and run a neural network. Unlike other neural network graph visualisations such as those produced by TensorFlow's TensorBoard (Tensorflow 2017), the "brain scans" do not abstract nodes into higher dimensional representations, namely tensors. The result beautiful, but enigmatic; the network is presented in its vast complexity, often containing millions of nodes and edges. These are unidentifiable in the emergent global form, making the image appear more photographic than diagrammatic. In these images, the edges are all but noise and appear as a fine grain, evocative of a micrograph.

The graph layout demonstrates patterns of growth similar to bacteria on a petrie dish. Force-directed layout is used to arrange the nodes, which produces growth-like properties. There are a multitude of ways to lay out a graph visualisation, and no single "correct" way to do so. In this case as in any, the choice to use one layout over another is an aesthetic judgement. The "brain scans" revel in the complexity of deep neural networks. No longer are nodes arranged in even rows and parallel layers with neatly criss-crossing edges. Rather, the sub-structures grow radially but distort due to competition for space.

The "brain scans" appear to be a rejection of structure and clarity of contemporary visualisation which seeks to render phenomena beyond the scale of human perception accessible to our senses. Manovich (2002) calls this the "anti-sublime ideal". Contrary to this, the *AI Brain Scans* are deeply sublime, they are an image of the complexity of contemporary neural networks, whose internal patterns and processes, as we have seen, are at least partially unknowable. By leveraging a biomorphic representation they present a metaphor for artificial neural networks that is complex, esoteric and uncanny.

This is not to say that the "brain scans" do not help us to understand neural networks. It is possible for example to identify convolutional and fully connected layers in the emergent structures of the images (see (Fyles 2017)). Importantly the *AI Brain Scans* help us to think about neural networks because they offer a visual metaphor that actually represents their complexity and obscurity rather than representing an incomplete or inconclusive visual explanation in false clarity.

Optimal Stimulus Images

One of the key strengths of deep neural networks over shallow ones is their capacity to build abstraction over successive layers (Bengio 2009). Accordingly, one of the greatest challenges for explaining the behaviour of neural networks is interpreting these intermediate representations learned by a network and encoded in hidden layers.

Because deep neural networks are often trained on low-level representations (like pixels or characters) which have little semantic value, the use of numerical explanations such as rule extraction are undermined.

Le (2013) demonstrates the use of numerical optimisation to find the optimal stimulus for a given neuron in an unsupervised neural network. The author includes three instances of interpretable high-level features discovered with this method. This suggests that neural networks have the capacity to discover salient features in the pursuit of higher goals, even when these are not encoded as part of a classification. Le uses gradient descent in training, keeping the weights and biases constant in order to to optimise the activation of a given neuron with respect to the input variable. Because the resulting optimal input data is in pixel space, it can be rendered directly as an image. Fig. 3 is an visualisation created in this way that clearly contains a human face.

The power of this and similar methods is that they accumulate the information which encodes this representation and is distributed throughout preceding layers. In doing so, they shift the focus of explanation from network weights with no inherent semantic value to distributed high-level features.

This kind of visualisation is extraordinarily expressive. At later layers in the network it theoretically makes it possible to simplify the tangled mess of relations encoded within. Used in combination with a technique such as rule extraction, it allows us to give coefficients at the level of meaningful concepts. However, this also requires that researchers engage critically with semantics. Tunç (2015) problematised the epistemic status of ML, noting that statistical learning is a form of inductive inference. It is therefore subject to the problem of induction. This idea can help researchers to better understand the nature of the semantics encoded in neural networks.

Fig. 3 shows a blurry image of a human face which is the optimal input for one of the network's neurons. The face is clearly white and clearly male. Its eyes contrast strongly with the background. Its lips are rosy and there is a hint of

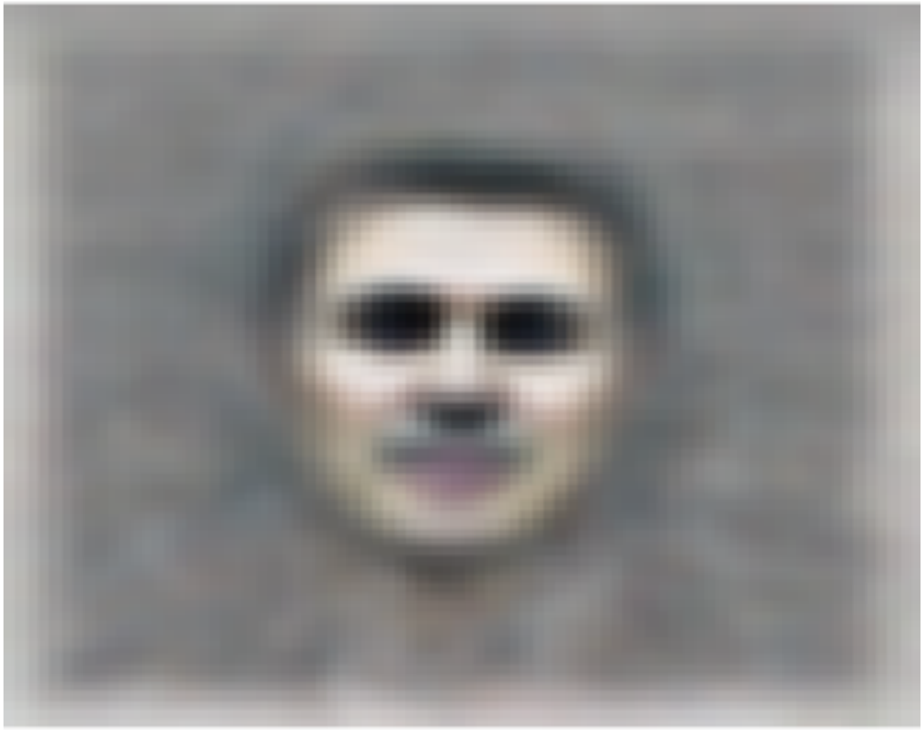


Figure 3: Optimal stimulus of interpretable neuron from Le (2013). (Copyright © 2013 IEEE, reprinted by permission of IEEE)

stubble on the upper lip. The clarity with which the face can be seen in the image led the author to claim that the tested neuron has learned “the concept of faces” (Le 2013). However, with the problem of induction in mind, this claim becomes less certain. The image appears to represent the basic notion of a face because whiteness and maleness are unmarked categories in English. We understand the image through the lens of our preexisting linguistic categories. If the pictured face were feminine, or non-white, or that of a child we would be be less inclined to assume it represents the general concept of faces.

The optimal stimulus is an important datum but it represents an archetype not a category. Categories are defined as much by what is excluded as what is included. It is possible that this neuron activates strongly only for faces that are also white, adult and male, or even only to those that resemble the man pictured. Alternatively this really could be the generic category of “face”, its features representing only an overrepresentation of white adult males in the dataset. It is not possible to know how far this representation of a face will stretch without experimentation. To test the equality of meaning between this neuron and the Anglophonic definition of a face, we need to measure how quickly the activation declines as correlates of femaleness, age and ethnicity change.

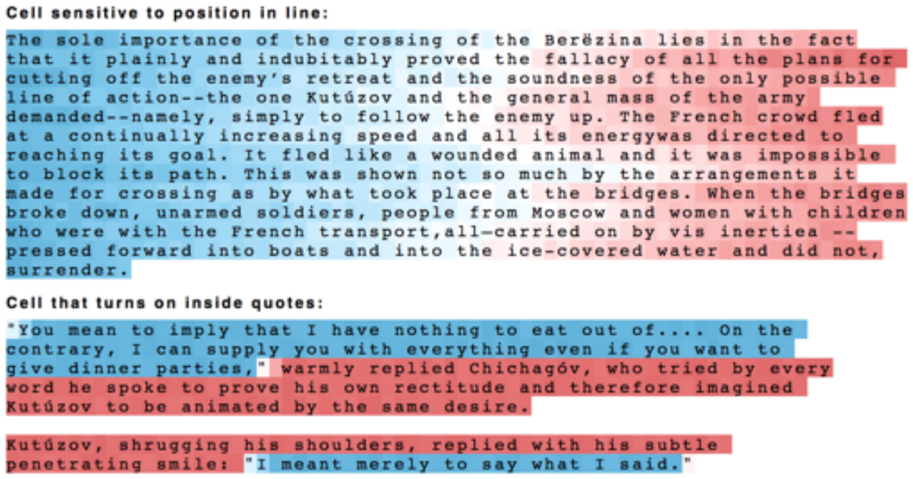
The optimal stimulus images and other methods that visualise the internal semantics of neural networks are crucial to our understanding of these systems. Nonetheless, it is necessary to take a critical approach when dealing with notions as slippery as meaning. It is valuable for researchers aiming to represent semantic encodings in neural networks to become familiar with the critical issues of semantics from philosophy and linguistics.

Interpretable, long-range LSTM cells

Semantic relationships in textual neural networks are explored in Karpathy, Johnson, and Fei-Fei (2015). The authors visualise the activation of a particular neuron across a passage of text generated by an LSTM network to look for interpretable relationships between the neuron’s activation and the composed content. The visualisation highlights each character with a colour mapped to the activation of a given neuron and look for interpretable patterns.

In Fig. 4 this technique was applied to an LSTM trained on Leo Tolstoy’s *War and Peace*. Two interpretable neurons are shown. The first can be interpreted as relating to the carriage return which must be used approximately every 70 characters. The second turns on inside quotes, allowing the network to remember to close them. Interpreting meaningful relationships in the content of the prose itself was not achieved. It is possible that an appropriately comparative string of characters that would reveal the pattern simply did not emerge. It’s also possible that the relationships that produce prose are more complex than the viewer can discern.

Unlike the optimal stimulus images, these visualisations engage with a softer notion of meaning. Here, meaning emerges from the consistent relation between



Cell sensitive to position in line:

The sole importance of the crossing of the Berëzina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutúzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energywas directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport,all—carried on by vis inertiea --
pressed forward into boats and into the ice-covered water and did not,
surrender.

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to
give dinner parties," warmly replied Chichagóv, who tried by every
word he spoke to prove his own rectitude and therefore imagined
Kutúzov to be animated by the same desire.

Kutúzov, shrugging his shoulders, replied with his subtle
penetrating smile: "I meant merely to say what I said."

Figure 4: Visualisation of interpretable activations of neurons from LSTM network trained on *War and Peace* based on Karpathy, Johnson, and Fei-Fei (2015). Text color represents the activation of the interpretable neuron $tanh(c)$, where blue is positive and red is negative. Source code for image available at (Browne 2017a).

the neuron activation and the output. The semantic meaning of a cell is inferred by the viewer in the context of real data. The visualisation exists only to service comparison.

The insight of this visualisation is to integrate neuron activation with the data it consumes or produces. By placing the abstract representation of activations in context, the viewer can discover patterns without the author's curation.

Fooling Images

The "fooling images" of Nguyen, Yosinski, and Clune (2015) are a collection of images produced with genetic algorithms that are unrecognisable to humans but produce high confidence predictions from state-of-the-art deep neural networks. The works expose a significant divide between human and computer vision.

Nguyen, Yosinski, and Clune (2015) introduce two methods for generating "fooling images" based on evolutionary algorithms (EAs). We will focus on the second form which uses Compositional Pattern Producing Networks (CPPNs) (Stanley 2007) to breed images which optimise for a given fitness function, in this case a single classification of a convolutional neural network trained on ImageNet. In order to simultaneously target all 1000 classes of ImageNet the researchers used the *multi-dimensional archive of phenotypic elites* algorithm, but noted that results were unaffected for a simpler EA.

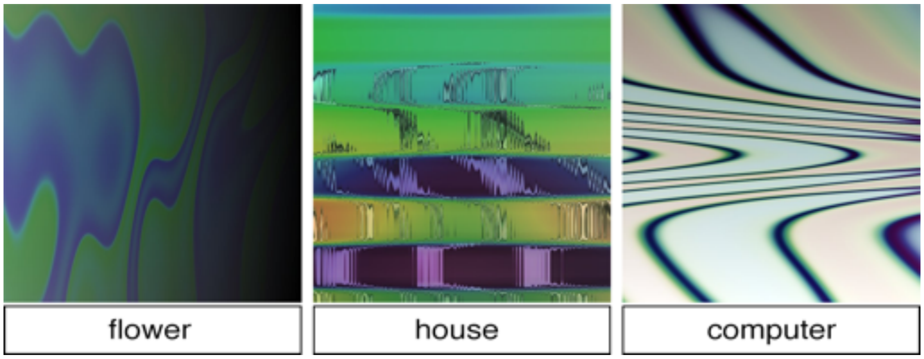


Figure 5: “Fooling images” produce > 99% confidence predictions but are unrecognisable to humans. This process and the notion of a fooling image are introduced in Nguyen, Yosinski, and Clune (2015). The examples above were produced using a script operating on Felix Andrew’s CPPN clojurescript implementation (Andrews 2016). The source code is available at (Browne 2017d).

The “fooling images”, unlike the previous case studies, contain no image of the network’s weights, structure, training set or indeed any data about the network at all. Despite this they do foster understanding. The images probe the network with targeted experiments to seek out unusual and revelatory behaviour.

The “fooling images” are critical cases that force the viewer to reconsider assumptions about the network. The researchers found that test subjects were able to reason about why an image was classified a certain way after its class was revealed to them. It’s clear however, that global coherence does not affect the network’s prediction. Instead, simply having low level features from the class seems to be sufficient to predict a category with high certainty.

The “fooling images” show very clearly that despite high scores on the ImageNet benchmark, neural networks do not “see” in the same way that humans do. It is natural to assume when we see a neural network working correctly that the network perceives and reasons as we do. Critical cases such as the “fooling images” reveal glimpses of the semantic mappings that the network has learned.

The “fooling images” are powerful because they break our natural assumption that performing a complex task with low error means thinking like a human. They destabilise the default anthropocentric notion of seeing. This semantic non-equivalence is likely a property of neural networks in general and gives grounds for skepticism of any neural network that appears to be acting like a person.

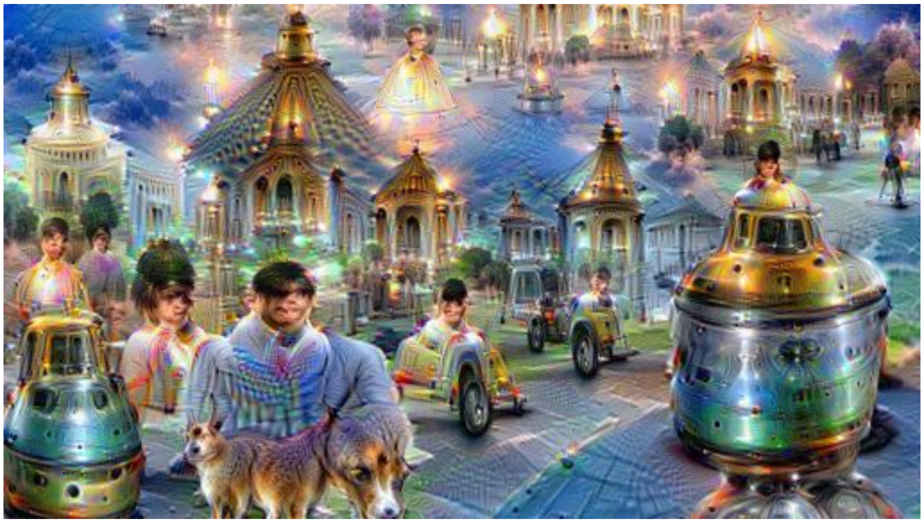


Figure 6: Image produced by DeepDream — by Mordvintsev, Tyka, and Olah (2015). Used under Creative Commons Attribution 4.0 International: https://creativecommons.org/licenses/by/4.0/legalcode

DeepDream

DeepDream is a method for visualising the internal representations of neural networks developed by Mordvintsev, Tyka, and Olah (2015). It is also likely the best known neural network visualisation, having reached viral status.

DeepDream approaches visualisation with the same basic aim as the optimal stimulus images of Le (2013); to visualise the semantic representations encoded by a network. The algorithm was modified from the research of Simonyan, Vedaldi, and Zisserman (2014) and others but provides two notable variations on existing work. First, instead of maximising a single neuron or class representation, DeepDream reinforces whatever is activating the highest to begin with. In doing so, it can hold the representations of hundreds of classes in a single image. These need not be distinct or complete and morph seamlessly from one to another (Fig. 7). Second, DeepDream applies its activations at different "octaves" creating a visual repetition of similar forms at many scales (Fig. 6).

In describing their work the authors make use of the language of conscious human experience. The networks are said to be "dreaming" perhaps in reference to the Phillip K. Dick novel. Later, the process and its emergent images are described as being like children watching clouds (Mordvintsev, Olah, and Tyka 2015).

DeepDream's emergent structures bear similarities to the drawings of MC Escher. Labyrinths grow out of thin air and form strange loops or seemingly infinite layers. Like Escher's tessellation works, representations morph from

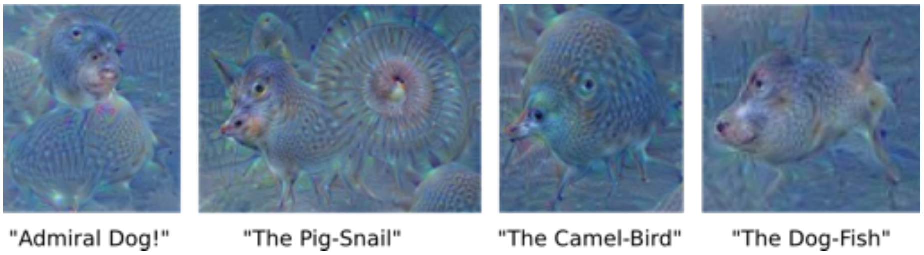


Figure 7: DeepDream's compound animals — by Mordvintsev, Tyka, and Olah (2015). Used under Creative Commons Attribution 4.0 International: https://creativecommons.org/licenses/by/4.0/legalcode

one to another, or change in scale. The eye can trace a path around the image and end up back where it started, or in a vastly different representation or a different scale. Representations morph from one to another but at every stage appear locally coherent.

The key point here is that like Escher's work, the DeepDream images are locally coherent but are globally incoherent. The images support the implication of the "fooling images", that semantic representation in neural networks does not depend on the global form.

The authors are careful to encourage the spread of their work. Alongside the published source code, readers are encouraged to make their own and share them with the hashtag #deepdream. It is also clear that the authors are cognisant of wider cultural implications of the images.

> [we] wonder whether neural networks could become a tool for artists — a new way to remix visual concepts — or perhaps even shed a little light on the roots of the creative process in general. (Mordvintsev, Olah, and Tyka 2015)

Unlike the algorithms it was based on, which visualise the representation of a single class or neuron, DeepDream combines any number of representations in the image. Because of this it is not possible to learn about particular features of a given class, or to understand how features relate to one another. DeepDream is arguably not a technical image but a cultural one. It is a picture of the strangeness and inconsistency of neural networks. Although it uses the language of conscious human experience it presents an uncanny image of neural networks that bears little resemblance to dreams or seeing.

Pix2Pix and FaceApp

Pix2pix (P2P) by Christopher Hesse is an online interface to a neural network that converts images of one kind to another. The work is an implementation of the Pix2Pix neural network designed by Isola et al. (2016). The interface

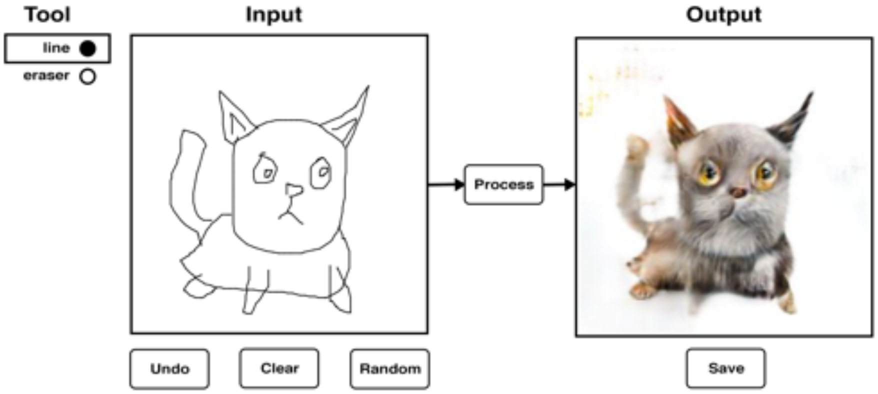


Figure 8: Schematic drawing of Christopher Hesse's *edges2cats* interface https://affinelayer.com/pixsrv/ Users are afforded simple drawing tools to produce a black and white line drawing. This is then processed by a server to which uses a pix2pix network to produce an image with photographic qualities, inferring form, colour and texture from the line drawing.

allows for user driven exploration of the trained network.

P2P is a behavioural visualisation like the "fooling images" in the sense that it does not directly represent information about the network itself but rather facilitates comparison between input and output. Unlike the "fooling images", it does not provide a curated list of inputs. In fact the initial state provided by the demo is rather unremarkable. It is the interaction here that is most central to the work's explanatory power. Users, over successive attempts, can test the limits of the network's semantics.

Fig. 8 shows a P2P demo which converts an outline drawing of a cat to a photographic image based on that outline. With their outline, the user can explore representations. Users can follow their own line of inquiry to learn about the network. Can the cat have more than two eyes? How is a body distinguished from a head? What happens if I don't draw a cat at all?

The interface allows the user to intelligently explore the space of possibility of the network. Though the interface enables individual sense-making, it is on social media that the images have been most successfully interpreted. On Twitter and other social networks curious, bizarre and revelatory images are selected for. Images that create the most interest are transmitted the farthest. In comments users share discoveries and attempt to make sense of the system collectively.

A similar pattern of collective sense-making can be seen in the response to *FaceApp* (FaceApp 2017). The app uses neural networks to transform a face in intelligent ways. It provides filters that make the subject smile, change gender or age, and increase “attractiveness”. Again, the interface allowed users to experiment with the network and seek out patterns, and again the most successful sense-making happened on social networks, which allow revelatory behaviours to spread quickly. Users of social networks quickly discovered that the filter intended to make users more attractive was turning skin white (Cresci 2017).

By allowing the network to infer its own semantics from the training set, it falsely equated beauty with white skin. With this unfortunate pattern in mind, it is possible to post-rationalise the existence of this bias in the training data. Datasets of this kind are labelled by people and thus imbibe the biases of the people who create them. Neural networks are not immune to this kind of bias, in fact it is almost impossible to prevent it. As universal approximators, neural networks make use of any salient patterns in data, including cultural patterns. If the application of labels such as beauty are correlated with whiteness, the network will learn to reproduce that pattern.

How is it possible that a powerful pattern of cultural bias that is completely obvious to users was invisible to those who developed the network? This pattern surprised FaceApp because the design of a network does not produce the behaviour, the data does. Cultural biases are easily learned and repeated by neural networks when we take data uncritically; as an objective representation of what *is*.

In contrast, despite being completely estranged from the neurons and synapses of the network itself, and without requisite knowledge of how neural networks function or learn, the users of social networks were able to discover and make sense of this pattern.

Interfaces that enable exploration and socially mediated interpretation are a powerful explanatory method. There is an opportunity here for researchers to design for collective sense-making, to make it easy for users to share curious behaviours of networks and facilitate collective interpretation.

Takeaways

In this section, we summarise the ideas raised in the case studies as a list of takeaways for researchers engaging in this area.

Structure does not explain behaviour

The structure of a neural network does not explain its behaviour. The shape of a neural network is a design consideration, it has an effect on learning, but not learned behaviour. Instead, behaviour is latent in the training data and approximated by the network.

It is not true that structure is irrelevant. Thinking about structure can help researchers to design better networks, increase capacity for abstraction and restrict overfitting. But these choices do not explain a network’s output.

In comparison, it is demonstrable that users can infer patterns in the network without any knowledge of the network itself or even a technical understanding of how neural networks function. Simply presenting inputs alongside outputs for comparison can allow viewers to spot patterns.

We understand better when things break

We learn more about how neural networks work when they fail. When neural networks do what we expect, it's easy to assume that they are thinking like a person. In the absence of a useful metaphor for how neural networks think we imagine ourselves performing the task. Given the extraordinary results achieved in benchmarks such as ImageNet, where neural networks have equalled or surpassed human accuracy, we tend to assume that the network uses the same features to identify images as we do. Indeed, it is difficult to comprehend how a system could achieve human or superhuman ability for a given task *without* thinking like a human. However, critical cases like the "fooling images" break this assumption.

Examples that break with expectations force the viewer to question their understanding of the system. By comparing input and output, the viewer can reason about which features produced the result and form a new theory for how predictions are made.

Interfaces for exploration

Interfaces such as Pix2Pix and FaceApp allow users to learn about a network by experimenting with it. These interfaces allow users to control input easily and see output immediately. This is a powerful pattern because it allows users to seek out critical cases. Users are able to continually adjust their mental model of how the network behaves by testing their hypotheses immediately.

We understand better together

The visual representations we have discussed, if created for a user at all, have been designed for individuals. Many of these, notably DeepDream, Pix2Pix and FaceApp, have been interpreted significantly on social media. Social networks enable collective sense-making, inspiring users to try similar things and add their results to the conversation. In comments users put into words their questions and theories about the system, where they can be discussed with others. Social networks also select for interesting or surprising content. This allows critical cases to be spread further.

It is possible to design for collective sense-making in neural network interfaces. An interface for collective sense-making might allow users to bookmark and share surprising behaviours and provide a place for users to discuss the content, share explanations and theories. It could also recommend recently viewed and commented bookmarks to encourage users to engage with one another.

Conclusion

The black box problem remains an ongoing challenge for researchers. Visual representation has proved to be a powerful tool with which to manage complexity and an important means of interpreting neural networks. Researchers in this space are making progress in extracting semantic encodings, developing interactive interfaces, discovering critical cases and negotiating the cultural conception of neural networks, however there is still much work to be done. The interdisciplinary interest in ML underscores the consequences of this technology beyond computer science and the importance of finding explanatory methods. The visual representation of neural networks crosses disciplinary boundaries. In this chapter we have outlined some emerging critical challenges for this research and demonstrated that they can be understood in the context of existing scholarship from disciplines considered far removed from computer science. To solve the black box problem will require critical as well as technical engagement with the neural network.

References

Andrews, Felix. 2016. “CPPNX.” https://floybix.github.io/cppnx/. https://floybix.github.io/cppnx/.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. “Neural machine translation by jointly learning to align and translate.” *arXiv Preprint arXiv:1409.0473*.

Barocas, Solon, and Andrew Selbst. 2016. “Big Data’s Disparate Impact.” *California Law Review* 104 (671): 671–732. https://doi.org/10.15779/Z38BG31.

Bastian, Mathieu, Sebastien Heymann, Mathieu Jacomy, and Others. 2009. “Gephi: an open source software for exploring and manipulating networks.” *Icwsm* 8: 361–62.

Bengio, Yoshua. 2009. “Learning Deep Architectures for Ai.” *Foundations and Trends in Machine Learning* 2 (1): 1–127. https://doi.org/10.1561/2200000006.

Benitez, J.M., J.L. Castro, and I. Requena. 1997. “Are artificial neural networks black boxes?” *IEEE Transactions on Neural Networks* 8 (5): 1156–64. https://doi.org/10.1109/72.623216.

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16 (3): 199–231. https://doi.org/10.2307/2676681.

Browne, Kieran. 2017a. “Interpretable Long-Range Lstm Cells Visualisation Redrawn from Karpathy Johnson and Fei-Fei 2015.” https://gist.github.com/kieranbrowne/70d39b2d46a2444cb64e21f38b81c578.

———. 2017b. “Neural Interpretation Diagram Redrawn from Ozesmi and Ozesmi 2005.” https://gist.github.com/kieranbrowne/a8d30f80484aebae796d62b85793dcc.

———. 2017c. “Neural Interpretation Diagram Redrawn from Tzeng and Ma 2005.” https://gist.github.com/kieranbrowne/8ca74d07adce15f39f0c59fe7bf76f17.

———. 2017d. “Script for Fooling Images as in Nguyen Yosinski Clune (2015).” https://gist.github.com/kieranbrowne/4f9fec38396e56cef88227c91283f242.

Burgess, Matt. 2017. “Gallery: ’Brain scans’ map what happens during inside machine learning.” http://www.wired.co.uk/gallery/machine-learning-graphcore-pictures-inside-ai. https://web.archive.org/web/20170223085426/http://www.wired.co.uk/gallery/machine-learning-graphcore-pictures-inside-ai.

Burrell, Jenna. 2016. “How the machine ‘thinks’: Understanding opacity in machine learning algorithms.” *Big Data & Society* 3 (1). http://journals.sagepub.com/doi/full/10.1177/2053951715622512.

Craven, Mark W., and Jude W. Shavlik. 1992. “Visualizing Learning and Computation in Artificial Neural Networks.” *International Journal on Artificial Intelligence Tools* 01 (03): 399–425. https://doi.org/10.1142/S0218213092000260.

Cresci, Elena. 2017. “FaceApp apologises for ’racist’ filter that lightens users’ skintone.” https://www.theguardian.com/technology/2017/apr/25/faceapp-apologises-for-racist-filter-which-lightens-users-skintone.

Duch, W. 2003. “Coloring black boxes: visualization of neural network decisions.” In *Proceedings of the International Joint Conference on Neural Networks*, 3:1735–40. IEEE. https://doi.org/10.1109/IJCNN.2003.1223669.

Elgammal, Ahmed, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. “CAN: Creative Adversarial Networks, Generating ”Art” by Learning About Styles and Deviating from Style Norms.” *arXiv Preprint arXiv:1706.07068*.

FaceApp. 2017. “FaceApp - Free Neural Face Transformation Filters.” https://www.faceapp.com/.

Fyles, Matt. 2016. “Neural network structure, MSR ResNet-50 - large directed graph visualization [OC] : dataisbeautiful.” https://www.reddit.com/r/dataisbeautiful/comments/5eowv6/neur
https://web.archive.org/web/20170813021434/https://www.reddit.com/r/dataisbeautiful/comments/5eowv6/

———. 2017. “Inside an AI ’brain’ - What does machine learning look like?” *GraphCore*. https://www.graphcore.ai/posts/what-does-machine-learning-look-like. https://web.archive.org/web/20170813022326/https://www.graphcore.ai/posts/what-does-machine-learning-look-like.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. “Multilayer feedforward networks are universal approximators.” *Neural Networks* 2 (5): 359–66. https://doi.org/10.1016/0893-6080(89)90020-8.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2016. “Image-to-image translation with conditional adversarial networks.” *arXiv Preprint arXiv:1611.07004*.

Karpathy, Andrej, Justin Johnson, and Li Fei-Fei. 2015. "Visualizing and understanding recurrent networks." *arXiv Preprint arXiv:1506.02078*. https://arxiv.org/pdf/1506.02078.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems 25 (NIPS2012)*, 1–9. https://doi.org/10.1109/5.726791.

Längkvist, Martin, Lars Karlsson, and Amy Loutfi. 2014. "A review of unsupervised feature learning and deep learning for time-series modeling." *Pattern Recognition Letters* 42: 11–24.

Le, Quoc V. 2013. "Building high-level features using large scale unsupervised learning." In *Acoustics, Speech and Signal Processing (Icassp), 2013 Ieee International Conference on*, 8595–8. IEEE.

Manovich, Lev. 2002. "The Anti-Sublime Ideal in Data Art." http://meetopia.net/virus/pdf-ps_db/LManovich_data_art.pdf.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. "Playing atari with deep reinforcement learning." *arXiv Preprint arXiv:1312.5602*.

Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. 2015. "Inceptionism: Going Deeper into Neural Networks." https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Mordvintsev, Alexander, Michael Tyka, and Christopher Olah. 2015. "Deep Dreams (with Caffe)." https://github.com/google/deepdream/blob/master/dream.ipynb. https://github.com/google/deepdream/blob/master/dream.ipynb.

Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2015. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images." In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 427–36. http://www.cv-foundation.org/openaccess/content{_}cvpr{_}2015/html/Nguyen{_}Deep{_}N http://ieeexplore.ieee.org/document/7298640/.

Olden, Julian D., and Donald A. Jackson. 2002. "Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks." *Ecological Modelling* 154 (1-2): 135–50. https://doi.org/10.1016/S0304-3800(02)00064-9.

Özesmi, Stacy L, and Uygar Özesmi. 1999. "An artificial neural network approach to spatial habitat modelling with interspecific interaction." *Ecological Modelling* 116 (1): 15–31. https://doi.org/10.1016/S0304-3800(98)00149-5.

Rosenblatt, Frank. 1962. *Principles of neurodynamics: perceptrons and the theory of brain mechanics*. Spartan Book.

Sainath, Tara N, Oriol Vinyals, Andrew Senior, and Ha\csim Sak. 2015. “Convolutional, long short-term memory, fully connected deep neural networks.” In *Acoustics, Speech and Signal Processing (Icassp), 2015 Ieee International Conference on*, 4580–4. IEEE.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.” *Iclr*. http://arxiv.org/abs/1312.6034.

Stanley, Kenneth O. 2007. “Compositional pattern producing networks: A novel abstraction of development.” *Genetic Programming and Evolvable Machines* 8 (2): 131–62.

Streeter, Matthew, Matthew Ward, and Sergio A Alvarez. 2001. “NVIS: an interactive visualization tool for neural networks.” *Proc of Visual Data Exploration and Analysis Conference*. https://doi.org/10.1117/12.424934.

Sussillo, David, and Omri Barak. 2013. “Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks.” *Neural Computation* 25 (3): 626–49. https://doi.org/10.1162/NECO_a_00409.

Tensorflow. 2017. “TensorBoard: Graph Visualization.” https://www.tensorflow.org/get_started/graph_viz.

Tunç, Birkan. 2015. “Semantics of object representation in machine learning.” *Pattern Recognition Letters* 64: 30–36. https://doi.org/10.1016/j.patrec.2015.03.016.

Tzeng, F Y, and K L Ma. 2005. “Opening the black box-data driven visualization of neural networks.” *Visualization, 2005. VIS 05. IEEE*, 383–90. https://doi.org/10.1109/VISUAL.2005.1532820.

Zahavy, Tom, Nir Ben-Zrihem, and Shie Mannor. 2016. “Graying the Black Box: Understanding Dqns.” In *International Conference on Machine Learning*, 1899–1908.