AI & SOCIETY (2021) 36:1133–1144
https://doi.org/10.1007/s00146-020-01058-z

ORIGINAL ARTICLE

Perceptual bias and technical metapictures: critical machine vision as a humanities challenge

Fabian Offert[1,2] · **Peter Bell**[2]

Received: 30 July 2019 / Accepted: 18 August 2020 / Published online: 12 October 2020

Abstract
In many critical investigations of machine vision, the focus lies almost exclusively on dataset bias and on fixing datasets by introducing more and more diverse sets of images. We propose that machine vision systems are inherently biased not only because they rely on biased datasets but also because their *perceptual topology*, their specific way of representing the visual world, gives rise to a new class of bias that we call *perceptual bias*. Concretely, we define perceptual topology as the set of those inductive biases in machine vision systems that determine its capability to represent the visual world. Perceptual bias, then, describes the difference between the assumed "ways of seeing" of a machine vision system, our reasonable expectations regarding its way of representing the visual world, and its actual perceptual topology. We show how perceptual bias affects the interpretability of machine vision systems in particular, by means of a close reading of a visualization technique called "feature visualization". We conclude that dataset bias and perceptual bias both need to be considered in the critical analysis of machine vision systems and propose to understand critical machine vision as an important transdisciplinary challenge, situated at the interface of computer science and visual studies/*Bildwissenschaft*.

Keywords Machine learning · Computer vision · Bias · Interpretability · Perception

1 Introduction

The susceptibility of machine learning systems to bias has recently become a prominent field of study in many disciplines, most visibly at the intersection of computer science (Friedler et al. 2019; Barocas et al. 2019) and science and technology studies (Selbst et al. 2019), and also in disciplines such as African-American studies (Benjamin 2019), media studies (Pasquinelli and Joler 2020) and law (Mittelstadt et al. 2016). As part of this development, machine vision has moved into the spotlight of critique as well,[1] particularly where it is used for socially charged applications like facial recognition (Buolamwini and Gebru 2018; Garvie et al. 2016).

In many critical investigations of machine vision, however, the focus lies almost exclusively on dataset bias (Crawford and Paglen 2019), and on fixing datasets by introducing more, or more diverse sets of images (Merler et al. 2019). In the following, we argue that this focus on dataset bias in critical investigations of machine vision paints an incomplete picture, metaphorically and literally. In the worst case, it increases trust in quick technological fixes that fix (almost) nothing, while systemic failures continue to reproduce.[2]

We propose that machine vision systems are inherently biased not only because they rely on biased datasets (which they do) but also because their *perceptual topology*, their specific way of representing the visual world, gives rise to a new class of bias that we call *perceptual bias*.

Concretely, we define perceptual topology as the set of those inductive biases in machine vision systems that determine its capability to represent the visual world. Perceptual bias, then, describes the difference between the assumed

✉ Fabian Offert
offert@ucsb.edu

Peter Bell
peter.bell@fau.de

[1] University of California, Santa Barbara, CA, USA

[2] Friedrich Alexander University Erlangen-Nuremberg, Erlangen, Germany

[1] As shown, for instance, by the increasing prominence of the FATE-CV workshop, organized by Timnit Gebru at CVPR, one of the major international computer vision conferences.

[2] The most recent (at the time of writing) example of this process at work can be seen in the controversy around the PULSE paper, see Kurenkov (2020) and Offert (2020).

"ways of seeing" of a machine vision system, our reasonable expectations regarding its way of representing the visual world, and its actual perceptual topology. Research in computer science has shown that the perceptual topologies of many commonly used machine vision systems are surprisingly non-intuitive, and that their perceptual bias is thus surprisingly large.

We show how perceptual bias affects the interpretability of machine vision systems in particular, by means of a close reading of a visualization technique called "feature visualization" (Erhan et al. 2009). Feature visualization can be used to visualize the image objects that specific parts of a machine vision system are "looking for". While, on the surface, such visualizations do make machine vision systems more interpretable, we show that the more legible a feature visualization image is, the less it actually represents the perceptual topology of a specific machine vision system. While feature visualizations thus indeed mitigate the opacity of machine vision systems, they also conceal, and thus potentially perpetuate, their inherent perceptual bias.

Feature visualizations, we argue, should thus not be understood so much as direct "traces" or "reproductions" of the perceptual topology of machine vision systems (analog to the technical images of photography) but more as indirect "illustrations", as "visualizations" in the literal sense of forcibly making-visual (and thus making visible and subsequently making interpretable) the non-visual. They should be understood as *technical metapictures* in the sense of W. J. T. Mitchell (Mitchell 1995), as images about (machine) seeing.

We also show how feature visualization can still become a useful tool in the fight against dataset bias, if perceptual bias is taken into account. We describe a case study where we employ feature visualization to discover dataset bias in several ImageNet classes, tracing its effects all the way to Google Image Search.

We conclude that dataset bias and perceptual bias both need to be considered in the critical analysis of machine vision systems and propose to understand *critical machine vision* as an important transdisciplinary challenge, situated at the interface of computer science and visual studies/*Bildwissenschaft*.

2 Deep convolutional neural networks

Our investigation looks at machine vision systems based on deep convolutional neural networks (CNNs), one of the most successful machine learning techniques within the larger artificial intelligence revolution we are witnessing today (Krizhevsky et al. 2012). CNNs have significantly changed the state of the art for many computer vision applications: object recognition, object detection, human pose estimation, and many other computer vision tasks are powered by CNNs today, superseding "traditional" feature engineering processes.

For the purpose of this investigation, we will describe CNNs from a topological perspective rather than a mathematical perspective. In other words, we propose to understand CNNs as spatial structures. While the topological perspective certainly requires the bracketing of some technical details, it also encapsulates the historical development of AI from "computational geometry", as Pasquinelli (2019b) reminds us.

From the topological perspective, we can describe CNNs as layered systems. In fact, the "deep" in "deep convolutional neural network" is literal rather than metaphorical (Arnold and Tilton 2019): it simply describes the fact that CNNs usually have more than two layers, with networks consisting of two layers sometimes being called "shallow" neural networks. In the simplest version of a (non-convolutional) neural network, these layers consist of neurons, atomic units that take in values from neurons in the previous layer and return some weighted sum of these values. So-called "fully connected layers" are thus not at all different from traditional perceptrons (Rosenblatt 1957; Minsky and Papert 1988), with the one exception that they only encode differentiable activation functions, i.e. the computation of the weighted sum of input values is achieved in a differentiable way, most commonly in the form of a so-called rectified linear unit.

Deep convolutional neural networks, then, introduce new classes of neurons, which perform more complex functions. Convolution operations have been used in signal processing way before neural networks regained popularity. Mathematically speaking, a convolution operation is an operation on two functions that produces a third function which measures the influence of the second function on the shape of the first. A more intuitive geometric definition is that of a kernel, a matrix, "scanning over" a second matrix to produce a third (Dumoulin and Visin 2016). A common example is a Gaussian kernel which can be used to blur an image, or a Sobel kernel that detects edges in images. The weights of such a kernel can become learnable parameters of a convolutional neural network. This means that the network, during training, will learn which kind of kernels, and thus essentially which kind of *image filters* are useful for a classification task.

Generally, as with all neural networks, learning in CNNs takes place in three different stages. A labeled input, an image, for instance, is passed through the interconnected layers of the network, until it reaches an output layer where a prediction regarding the input image is made, depending on the task set for the system. Such a task could be to classify an image according to certain categories, find the boundaries of an object in an image, or other problems from computer vision. An evaluation function (called "loss function" in machine learning) then measures how far off the prediction

of the system is. This information "flows back"[3] through the network, and all its internal connections are adjusted accordingly.

All of these steps take place for all images in the training set, and periodically, the system is also tested on previously unseen samples. This validation process is particularly important as it avoids a failure mode called "overfitting", where a network learns indeed to perfectly predict the training set labels, but does not generalize to unseen data, which is the whole purpose of training.

It is because of this incremental process that often spans thousands of iterations, that CNNs are notoriously opaque. Common CNN architectures can have millions of neurons and even more interconnections between these neurons. It is thus close to impossible to infer from looking at the source code, data, weights, or any other aspect of a CNN, either alone or in conjunction, what it does, or what it has learned. Selbst and Barocas (2018) have suggested calling this opacity *inscrutability.*

Complexity, however, is not the only reason for the notorious opacity of CNNs. As Selbst and Barocas argue, CNNs are also *non-intuitive*. The internal "reasoning" of neural networks does not necessarily correspond to intuitive methods of inference, as *hidden correlations* often play an essential role. In other words, it is not only hard to infer the rationale behind a network's decision from "looking at it" because it is inscrutable, this rationale might also be significantly non-intuitive. Selbst and Barocas have argued that non-intuitiveness could be described as an "inability to weave a sensible story to account for the statistical relationships in the model. Although the statistical relationship that serves as the basis for decision-making might be readily identifiable, that relationship may defy intuitive expectations about the relevance of certain criteria to the decision" (Selbst and Barocas 2018, 1097).

3 Interpretable machine learning

This problem has been widely recognized in the technical disciplines as the problem of building *interpretable machine learning* systems, also referred to as *explainable artificial intelligence* systems.[4] Such systems, either by design or with the help of external tools, provide human-understandable explanations for their decisions, self-mitigating both their inscrutability and non-intuitiveness.

In the past 3 to 5 years, research in interpretable machine learning has matured into a proper subfield of computer science (Lipton 2016; Doshi-Velez and Kim 2017; Gilpin et al. 2018; Mittelstadt et al. 2019) and a plethora of statistical tricks (Lundberg and Lee 2017; Ribeiro et al. 2016) has been developed to ensure the interpretability of simpler models like linear regression, particularly in safety-critical or socially charged areas of machine learning like credit rating or recidivism prediction.

Beyond these technical results, however, a larger conceptual discussion has emerged in the technical disciplines as well that "infringes" on the terrain of the humanities. It is centered around attempts to find quantitative definitions for concepts that naturally emerge from the problem at hand, such as "interpretation" and "representation", with the help of methods and concepts from disciplines as diverse as psychology, philosophy, and sociology, building a "rigorous science of interpretable machine learning", as Doshi-Velez and Kim (2017) write.

A concrete example would be that of a "cat" and "dog" classifier. Hypothetically, we can train a standard CNN architecture on a large dataset of cat images and dog images. These images will be processed by the network as described above: the input layer of the network will transform the pixel values of an image into a multidimensional vector, which then flows "forward" through the network, until a prediction is made in the very last layer. This prediction is evaluated by the loss function—for instance by the mean squared error of all neurons—and back-propagated through the network. Weights are changed incrementally, until the performance of the network on a validation set, for instance, 10% of the dataset of cat and dog images that have been held back, stops increasing. Given a large enough dataset, we can assume that we will reach almost 99% accuracy for this task. But, how are the concepts "dog" and "cat" encoded in the system—a system that clearly somehow "knows" what these concepts are?

Research in interpretable machine learning thus requires the consideration of both technical and philosophical notions of interpretation and representation. We propose that, for machine vision systems, this inherent transdisciplinarity

[3] Technically, this "flowing back" of information, also called back-propagation, is enabled by the fact that convolutional neural networks, beyond the topological perspective, are just very large differentiable equations. Hence, for each neuron we can derive the slope of the loss function with respect to this specific neuron. This allows us to then adjust the weights of this neuron into the direction of the slope, thus decreasing the error of the whole function by a tiny amount. This adjustment process, can be implemented in different ways and is most commonly realized through stochastic gradient descent.

[4] The difference between the two terms originally was a matter of geography: "interpretable machine learning" was used in the North American context, while "explainable artificial intelligence" was used in the European context. Today, however, both terms are used interchangeably. We will stick with interpretable machine learning in the context of this paper, to emphasize its focus on machine learning (i.e. technical systems) vs. artificial intelligence (i.e. speculative technologies).

implies linking technical concepts and concepts from visual studies/*Bildwissenschaft*. In particular, it suggests understanding the interpretation of machine vision systems as an act of image-making, both literally and metaphorically. This is why, in the following, we will look at feature visualization.

4 What is feature visualization?

Feature visualization belongs to a range of techniques for the visual analysis of machine learning systems called *visual analytics* (Hohman et al. 2018). The idea of visual analytics is to *show how* a machine vision system perceives the world, and thus how it makes its decisions. Explanations, in visual analytics, thus, take the form of images, not of numbers or sentences. In other words, visual analytics is the visualization of machine learning systems for the sake of interpretability. Within visual analytics, feature visualization (together with attribution techniques) has become one of the most widely used methods. Originally developed by Erhan et al. (2009) and continuously improved since, it has been shown to produce remarkable results (Olah et al. 2017, 2018, 2020).

Technically, feature visualization is a straightforward optimization process. To visualize what a neuron in a deep convolutional neural network has learned, a random noise image is passed through the layers of the network up until the hidden layer that contains the neuron of interest. Normally, during the training or prediction stages, the image would be passed further on to the output layer. For the purpose of visualization, however, we are not interested in a prediction but in the "activation" of a single neuron, its individual response to a specific input image when it reaches the neuron's layer. Hence, instead of utilizing the original loss function of the network, this response is now interpreted as its loss function. In other words, it is now the response of a single neuron that drives the "learning" process.

The important difference is that this new loss flows back through the network beyond the input layer and is used to change the *raw pixel values* of the input image. The input image is thus altered, while the network's internal interconnections remain untouched. The altered image is then being used again as the input image during the next iteration, and so on. After a couple of iterations, the result is an image that highly activates one specific neuron.

5 Perceptual bias as syntactic bias

This process, however, is called "naïve" feature visualization for a reason. In almost all cases, images obtained with it will exclusively contain very high frequencies and will thus be "illegible" in both the syntactic and semantic sense: there will be no visible structure, and no recognizable content (Fig. 1).[5] The images may very well be the *best possible images* with regard to a specific neuron and may very well be the closest possible visualizations of what this neuron has learned. To the human observer, however, they contain no information. They are adversarial examples (Szegedy et al. 2013; Goodfellow et al. 2014a)—images that highly activate specific neurons or classes in a fully trained deep convolutional neural network, despite being utterly uninterpretable.

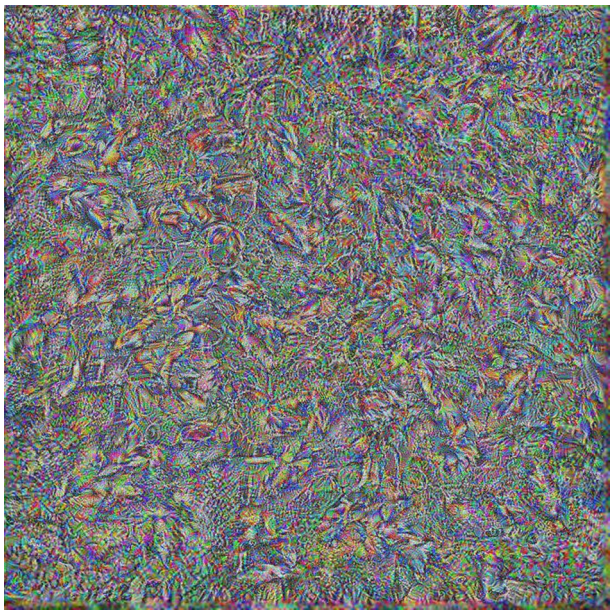

Fig. 1 Unregularized feature visualization of the "banana" class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset. These and the following visualizations have been generated with a custom software written by Fabian Offert in Python/PyTorch, implementing an optimized feature visualization algorithm with regularization but without natural image priors

Naïve feature visualization, then, shows us a first glimpse of the peculiar perceptual topology of CNNs. Perceptual bias, here, takes the form of syntactic bias. This syntactic bias, in turn, manifests as texture bias (Geirhos et al. 2019), an inductive bias in CNNs that "naturally" appears in all common CNN architectures. Inductive biases are "general", prior assumptions that a learning system uses to deal with new, previously unseen data.

[5] Concretely, to generate the feature visualization images in this paper, the following settings have been used: InceptionV3 model pretrained on ImageNet/ILSVRC2012; stochastic gradient descent optimizer (torch.optim.SGD) with a learning rate of 0.4 and an L2 weight decay of 0.0001; optimization target: fully connected prediction layer (layer 17); 3 octaves, with 1.5×resolution increase/octave, leading to a final resolution of 672×672; 2000 iterations per octave; jitter (image is randomly shifted 32 pixels) every octave; total variation filter every 20 iterations.

At this point, it is important to note that we will not consider modifying the inductive biases of the CNN itself as a solution to the problem of perceptual bias, as, for instance, Geirhos et al. (2019) and Zhou et al. (2020) suggest. More precisely, for the purpose of our investigation, we are interested in interpretable machine learning as a narrow set of post hoc methods to produce explanations. Thus, we will also not take the field of representation learning (Bengio et al. 2013) into consideration, which is concerned with the development of mechanisms that enforce the learning of "better" representations. This restriction to the scope of our investigation has three main reasons. The first reason is the post hoc nature of the bias problem. While efforts to build resistance to bias into machine learning models exist, there is, at the moment, no clear incentive for industry practitioners to do so, except for marketing purposes. It can thus be assumed that, in real-world scenarios, the detection and mitigation of bias will be mostly a post-hoc effort. The second reason is a simple historical reason. Thousands of machine learning models based on the exact perceptual topologies under investigation here have already been deployed in the real world. Thus, it is of vital importance to understand, and be able to critique, such models and their perceptual biases. Finally, while impressive progress has been made in other areas of machine learning (Cranmer et al. 2020), in machine vision, controlling and harnessing inductive biases can still be considered an open problem. Recent research suggests that at least one established principle of gestalt theory (the law of closure) does emerge in CNNs (Kim et al. 2019; Ritter et al. 2017; Feinman and Lake 2018) as an inductive bias. Overall, however, the inductive biases of CNNs are still unclear (Cohen and Shashua 2017), and thus un-manageable.

Given these restrictions, the only option to mitigate this specific textural aspect of perceptual bias is to not change the model, but to change our image of it. In the case of feature visualization, it means adding back *representational capacity* to these images. It means introducing constraints—in other words, different biases—that allow the production of images that are images *of something*, instead of "just" images. Importantly, any such constraint, however, automatically moves the image further away from showing the actual perceptual topology of a CNN. It becomes less of a visualization, and more of a reconstruction. This trade-off is the core problem of perceptual bias: it can only be overcome by shifting towards different, "better" biases, i.e. biases that shape *our* perception of the visual world.

One strategy to "add back" the representational capacity to feature visualization is regularization (Fig. 2). Regularization, here, simply means adding additional constraints to the optimization process. This can be achieved either by adapting the loss function—for instance, using a quadratic loss function instead of just taking the mean of some values—or by applying transformations to the input image in regular intervals, for instance, every few iterations in the optimization process.

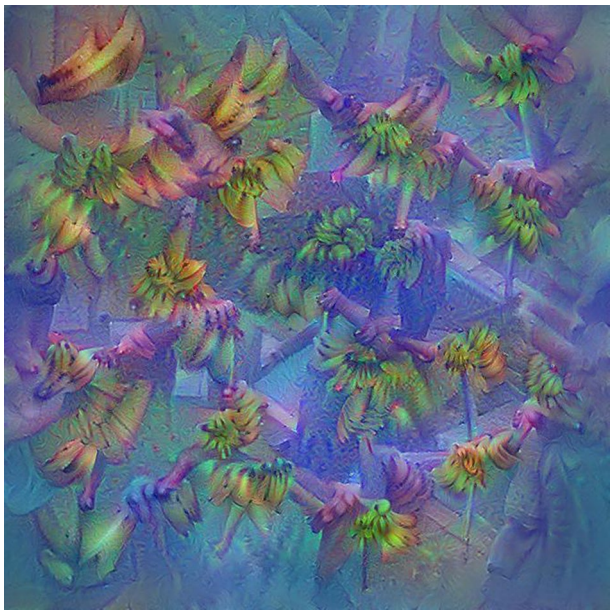

Fig. 2 Regularized feature visualization of the "banana" class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset

Erhan et al. (2009) introduced the concept of activation maximization, the core idea of iteratively optimizing an image to highly activate a selected neuron. From there, more and more elaborate regularization techniques started to appear, each introducing concrete suggestions for signal processing operations on the input image between iterations, on top of more common regularization techniques introduced through the loss function, like L2 regularization. Among these are jitter (Mordvintsev et al. 2015), blur (Yosinski et al. 2015), total variation filters (Mahendran and Vedaldi 2015), bilinear filters (Tyka 2016), stochastic clipping (Lipton and Tripathi 2017) and Laplacian pyramids (Mordvintsev 2016). Mordvintsev et al. (2015) also introduce the idea of octave-based optimization, enabling significantly higher image resolution. What all these techniques have in common is some kind of frequency penalization, i.e. the active avoidance of input images evolving into adversarial examples, either through optimizing for transformation robustness (Mordvintsev et al. 2015) or through direct filtering (all others).

6 Perceptual bias as semantic bias

Despite all regularization efforts, however, feature visualizations often still present "strange mixtures of ideas" (Olah et al. 2018). Visualizing higher level neurons in particular

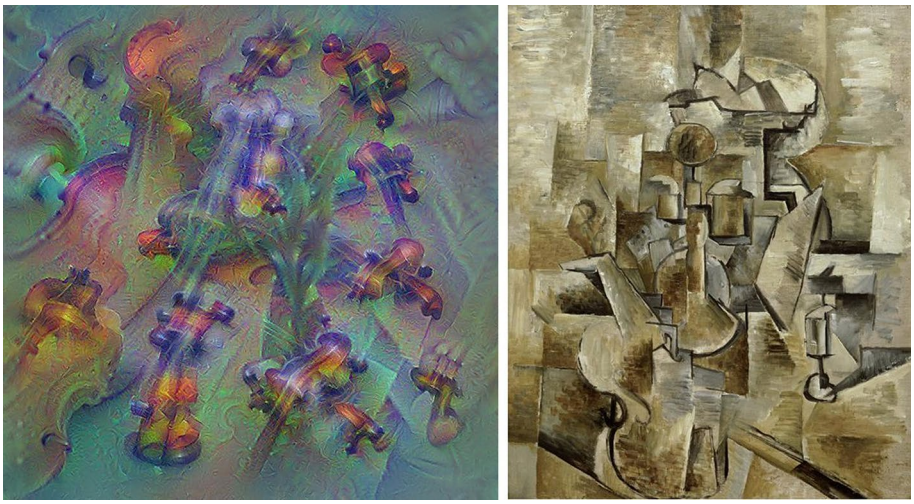

Fig. 3 Left: regularized feature visualization of the "violin" class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset. Right: George Braque, Violin and Candlestick (1910)

produces ambiguous results, images that might, or might not, show proper "objects". To learn more about the logic of representation in CNNs, we thus have to ask: what is the relation between technical and semantic units, between artificial neurons and meaningful concepts, in CNNs?

Trivially, at least for higher level neurons, individual feature visualization images must always have a degree of ambiguity that is directly correlated to the diversity of the training set. After all, the network has to be able to successfully classify a range of instances of an object with very different visual properties. In that sense, reality is "distributed", and it is no surprise that feature visualization images will reflect different manifestations of, and perspectives on, an object, akin to Cubist paintings (Fig. 3).

But, the entanglement of concepts in the internal representations of a CNN goes beyond this "natural" ambiguity. Generally, we can state that, in all predictions of a CNN, all neurons play "a" role. Even if their role is just to stop the information flow, i.e. to pass on zero values to the next layer, these one-way streets are in no way less relevant to the classification accuracy of the whole system than all other neurons. In a way, concepts are thus "dissolved", or "entangled", when they are learned, and represented, by a CNN. Early work by Szegedy et al. (2013) suggests that this entanglement is inevitable and absolute. Later work by Bau et al. (2017, 2018) shows that some neural network architectures (GANs in particular) are less "naturally entangled" than others. Generally, however, significant supervision or, again, artificial inductive biases (Locatello et al. 2019) are required to "disentangle" CNNs, and arrive at a meaningful correspondence of technical and semantic units.[6]

Perceptual bias, here, thus takes the form of semantic bias. Other than in the case of adversarial examples/texture bias, where perceptual bias affects the formal aspects of the visualization, here, it concerns aspects of meaning. Objects, for us, are necessarily spatially cohesive. If they are represented by CNNs, however, they lose this spatial coherence, different aspects of an object are attached to different neurons, which, in turn, get re-used in the detection of other objects. This missing coherence does not interfere with the CNN's ability to detect or classify spatially coherent objects in images but enables it.

For feature visualization, which visualizes CNNs in their "natural", entangled state, reaching semantic interpretability thus implies the introduction of even more constraints. These additional constraints are so called natural image priors. Just as regularization is a syntactic constraint, biasing the visualization towards a more natural frequency distribution, so-called natural image priors are a semantic constraint, biasing the visualization towards separable image objects.

To produce natural image priors, Dosovitskiy and Brox (2016) propose to use a GAN generator. Generative adversarial networks (GANs) have received a lot of attention since 2015/2016 for being able to generate realistic images in an unsupervised way from large datasets. They were originally introduced by Goodfellow et al. (2014) and have since been steadily improved and extended to other tasks beyond image generation. The term "generative adversarial network" refers to an ensemble of two convolutional neural networks that are trained together. The notion of "adversarial" describes the dynamic between the two networks, where a generator attempts to fool a discriminator into accepting its images. Whereas the discriminator is thus a regular, image-classifying convolutional neural network, the generator is a reverse CNN that outputs an image. Its input is a number from a

[6] Unfortunately, we cannot give a full review of the relevant computer science literature regarding representation learning and disentanglement here. Instead, we would like to refer the reader to the review articles by Bengio et al. (2013) and Locatello et al. (2019).

latent vector space, i.e. a high-dimensional space. Concretely, a random sample from this latent space is forwarded through the layers of the generator, and an image is created at its last layer. This generated image, or alternatively an image from a supplied image dataset, is then evaluated by the discriminator, which attempts to learn how to distinguish generator-generated images from images that come from the supplied dataset. Importantly, the discriminator's loss is back-propagated all the way to the generator, which allows the generator to adjust its weights depending on its current ability to fool the discriminator. Eventually, this dynamic results in a generator that has learned how to produce images that look like they come from the supplied dataset, i.e. that have similar features as the images in the supplied dataset but are not part of the dataset. Overfitting, here, is avoided by never giving the generator access to the supplied dataset. Its only measure of success is how well it is able to fool the discriminator.

Nguyen et al. (2016) turn this technique into a dedicated feature visualization method by applying the paradigm of activation maximization to the input of a GAN generator. Instead of optimizing an input image directly, i.e. in pixel space, it is thus optimized in terms of the generator, i.e. in feature space. This technique has three main benefits: (a) Optimization in feature space automatically gets rid of high frequency artifacts. The generator, trained to produce *realistic* images, will never reconstruct an adversarial example from a latent feature representation. (b) Optimization in feature space introduces a strong natural image prior. More precisely, it introduces a natural image prior that corresponds directly to the level of realism that can be attained with the generator. (c) Finally, optimization in feature space requires neither the generator, nor the network that is being analyzed to be trainable, as back propagation just passes through an image: a feature representation is fed into the generator, the generator produces an image, this image is fed into the network that is being analyzed, which in turn produces a loss with regard to the neuron being analyzed, as in regular feature visualization.

This means, however, that the images that can be produced with this feature visualization method are entirely confined to the latent space of the specific GAN generator employed. Where regularization constrains the space of possible images to those with a "natural" frequency distribution, natural image priors constrain the space of possible images to the distribution of a GAN generator. In both cases, interpretable images are the result. These interpretable images, however, do not reflect the perceptual topology of the analyzed CNN. On the contrary, they intentionally get rid of the non-humanness that defines this topology, translating it into a human mode of perception that, in this form, simply does not exist in the CNN. To be images of something, feature visualizations have to be freed from the very mode of perception they are supposed to illustrate.

7 Feature visualizations as technical metapictures

As we have seen, the perceptual topology of machine vision systems, based on CNNs, is not "naturally interpretable". It is biased towards a distributed, entangled, deeply non-human way of representing the world. Mitigating this perceptual bias thus requires a forced "making legible". Feature visualization, as we have seen, is one possibility to achieve this forced legibility. However, feature visualization also exemplifies an essential dilemma: the representational capacity of feature visualization images is *inverse proportional* to their legibility. Feature visualizations that show "something" are further removed from the actual perceptual topology of the machine vision system than feature visualizations that show "nothing" (i.e. illegible noise).

There is thus an irreconcilable difference between the human and machine perspective. As Thomas Nagel reminds us, there is a "subjective character of experience" (Nagel 1974), a surplus generated by each specific perceptual approach to the world that can never be "translated". Even if an external observer would be able to attain all the facts about such an inherently alien experience (analyze it in terms of "functional states"), they would still not be able to reconstruct said experience from these facts.

Feature visualizations, then, should not be understood so much as direct "traces" or "reproductions" of the perceptual topology of machine vision systems (analog to the technical images of photography) but more as indirect "illustrations", as "visualizations" in the literal sense of forcibly making-visual (and thus making visible and subsequently making interpretable) the non-visual.

We thus propose to understand these images as *technical metapictures*, a term we adapt from W. J. T. Mitchell's picture theory (Mitchell 1995). For Mitchell, metapictures are pictures that are "deeper" than "regular" pictures, as they incorporate a form of recursion: they are representations of representation "pictures about pictures" (Mitchell 1995, 36). Mitchell identifies certain abilities of these pictures. "The metapicture […] is the place where pictures reveal and 'know' themselves, where they reflect on the intersections of visuality, language, and similitude, where they engage in speculation and theorizing on their own nature and history" (Mitchell 1995, 82). They are not only self-reflective but reflective on imagery and perception. "The metapicture is a piece of moveable cultural apparatus, one which may serve a marginal role as illustrative device or central role as a kind of summary image, what I have called a 'hypericon'

that encapsulates an entire episteme, a theory of knowledge" (Mitchell 1995, 49).

The technical metapictures that feature visualization produces realize exactly this idea of a "summary image." They promise not a theory of images but a theory of seeing. More precisely, their promise is exactly that of interpretable machine learning: to provide an intuitive visual theory of the non-intuitive perceptual topology of neural networks. In a sense, technical metapictures, and their use in interpretable machine learning, are thus an operationalization of the notion of metapicture itself.

For Mitchell, this epistemological power of metapictures, then, equips them with a sort of agency. Metapictures "don't just illustrate theories of picturing and vision: they show us what vision is, and picture theory" (Mitchell 1995, 57). This agency, however, is actualized only if and when it comes into contact with a viewer. To *make sense*, to actually provide the reflection on images and vision that they promise, metapictures require a viewer. In the case of feature visualization, this interpretation has to happen not only on the level of the viewer but also on the technical level, where a significant effort has to be made to translate the anti-intuitive perceptual topology of a machine vision system into human-interpretable images in the first place. This includes *adding information from the outside*, for instance in the form of natural image priors. In other words, technical metapictures manifest an implicit, technical notion of interpretation, that is inseparable from the explicit interpretation that they also require.

8 Learning from technical metapictures

If we understand feature visualizations as technical metapictures, that is, if we look at them not as representations of machine vision systems but as reflections on the perceptual limits of machine vision systems, we can re-imagine them as a method of critique that can detect deeper forms of bias.

Taking up the idea that "adversarial examples are not bugs, they are features," (Ilyas et al. 2019) the peculiar ways of machine seeing described above would be utilized to detect anomalies in datasets that go beyond problematic or socially charged categories.

One of the main problems of image datasets is the diversity and heterogeneity of the real-life visual world, which is inherently difficult to capture by technical means. How is the diversity and heterogeneity of the real-life visual world to be represented in a set of images of bounded size? Historically, this problem is related to the general epistemological problem of creating taxonomies of what exists. The appearance of solutions to this problem, and their failure, can be traced back to the beginnings of the enlightenment in the seventeenth century, with Descartes, Leibniz, Wilkins and others speculating on scientific approaches to the design of inventories of the world. Importantly, since then, taxonomies have been thought of as symbolic, i.e. as textual representations of the world.

With the introduction of images, however, and the creation of visual taxonomies, conceptual problems already present in symbolic taxonomies have been amplified by an order of magnitude: any claim to completeness is now not only a claim to a completeness of *concepts*, but of *manifestations*. Put differently: while symbolic taxonomies have to deal with an infinity of *abstractions*, visual taxonomies have to deal with an infinity of *individual objects*. Of course, this grounding of the learning problem in "real" data, in "ground truths", to use the terminology of machine learning, is exactly the point. The hypothesis is that useful abstractions will appear automatically by means of the exposure of a system to a manifold of manifestations. This, however, increases the number of potential points of bias injection *to the size of the dataset*. In other words: it is *in every single image* that potential biases come into play, and thus the design and use of visual taxonomies becomes of crucial importance.

ImageNet (Deng et al. 2009) is a large-scale digital image dataset intended to facilitate the automatic classification of images with regard to depicted objects (object recognition). It consists of over fourteen million images in over 21,000 categories. Originally conceived and first presented at CVPR in 2009 by Fei Fei Li and others, ImageNet consolidated an existing taxonomy, WordNet (Miller 1985) and image resources freely available on the Internet. ImageNet stands out as being the first large-scale dataset using distributed labor acquired via Amazon Mechanical Turk to solve the problem of cheap data vs. expensive labels, i.e. the circular problem of first having to hand-annotate images to automate the process of annotating images.[7]

Since 2010, the ImageNet project has run the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a competition to identify the state of the art in image classification (Russakovsky et al. 2015). The challenge usually only requires the participating models to correctly classify images from an ImageNet subset. Importantly, because the images contained in ImageNet are "scraped" (i.e. downloaded) from the Internet and thus might be subject to copyright, the ImageNet project only supplies URLs to the images. As

[7] A proper analysis of the entanglement of machine learning and labor lies outside the scope of this paper. Nevertheless it should be mentioned, as brilliantly analyzed, among others, by Daston (2018) and Pasquinelli (2019a) that machine learning is an essential Taylorist idea. As Charles Babbage writes already in 1832: "[T]he division of labor can be applied with equal success to mental as to mechanical operations, and […] it ensures in both the same economy of time" (Babbage 2010, 295).

the project has been in existence since 2009, many of these URLs are not accessible anymore today. These two factors have facilitated the current real-life use of ImageNet: in most contemporary applications, it is actually just a subset that is being used to train and test computer vision systems, most prominently the subset created for the 2012 ILSVRC.

In the past few years, much has been written about biased datasets. ImageNet has been at the center of this debate (Malevé 2019), with Trevor Paglen and Kate Crawford's "Excavating AI" project (Crawford and Paglen 2019) receiving broad media attention. What Crawford and Paglen rightfully criticized was essentially the historical debt of the dataset, which operates with a taxonomy based on WordNet. WordNet, in turn, includes many categories which are neutral as textual categories (e.g. "terrorist"), but have necessary social and political implications when "illustrated" with images. The failure of the ImageNet team to remove these and similar categories, and even stock some of them with actual images based on the aforementioned distributed micro-labor, rightfully led to a public outcry in reaction to the "Excavating AI" project, and subsequently to the removal of these categories from the dataset.

Nevertheless, there is a need for more elaborate methods of image dataset critique, and feature visualization could provide a potential basis for such a method. One example application is the detection of dataset anomalies in art historical corpora (Offert 2018). A more broad critical approach would be the analysis of highly common datasets like ImageNet, which are not only used "as is" in real-life classification scenarios but even more often used to pre-train classifiers which are then fine-tuned on a separate dataset, potentially introducing ImageNet biases into a completely separate classification problem.

To demonstrate this approach, we visualized and selected the output neurons for several classes of an InceptionV3 model (Szegedy et al. 2016) pre-trained on ImageNet/ILSVRC2012 hand-selecting visualizations that show some non-intuitive properties of the ImageNet dataset. For instance, for the "fence" class output neuron (Fig. 4) we see that the network has not only picked up the general geometric structure of the fence but also the fact that many photos of fences in the original dataset (that was scraped from the Internet) seem to contain people confined behind these fences. This can be verified by analyzing the 1300 images in the dataset class, which indeed show some, but not many scenes of people confined behind fences. Cultural knowledge, more specifically, a concrete representation of cultural knowledge defined by the lense of stock photo databases and hobby photographers, is introduced here into a supposedly objective image classifier. Importantly, this also means that images of people behind fences will appear more fence-like to the classifier. The relevance of this consequence is revealed by a Google reverse image search: for a sample image (Fig. 5)

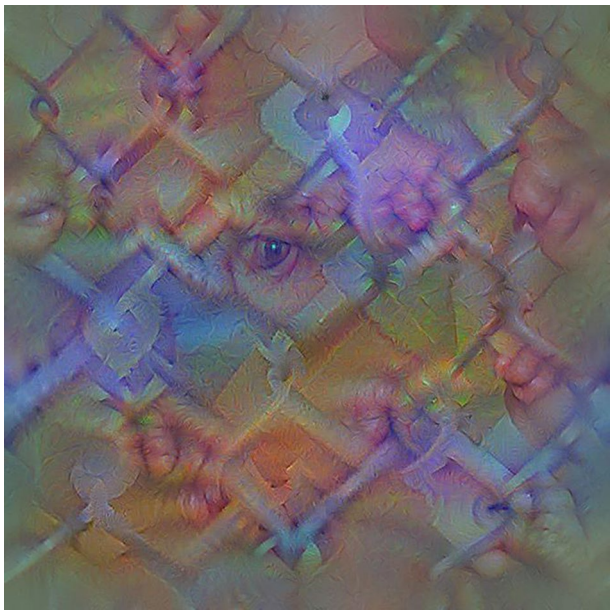

Fig. 4 Regularized feature visualization of the "fence" class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset

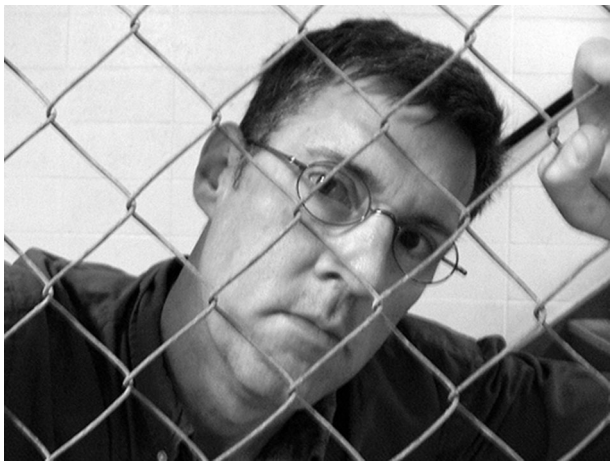

Fig. 5 Sample image from the ILSVRC2012 "chain link fence" class. Note that there are only a few images (between 1 and 5% of the class, depending on what counts as "behind") that show people behind fences

from the "fence class", despite the prominence of the person compared to the actual fence, the search produces the Wikipedia entry for "chain link fencing" (Fig. 6), suggesting an unverifiable but likely connection between the Google image search algorithm and ImageNet/ILSVRC2012.

For the "sunglass" class (Fig. 7), the diversity that the respective output neuron has to deal with becomes obvious in the entanglement of actual sunglasses and body parts. More surprisingly is the inclusion of mirrored landscapes.

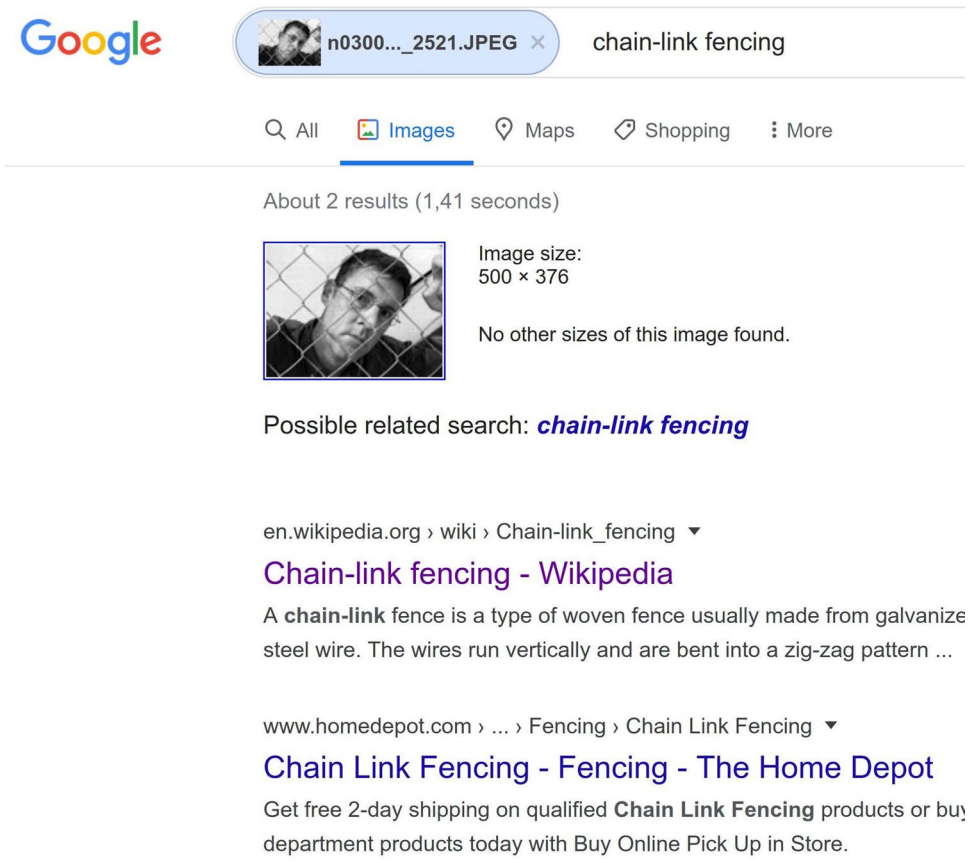

Fig. 6 A Google reverse image search for this specific image, despite the fact that the image does not exist on the Internet anymore, and despite the prominence of the person compared to the actual fence, produces the Wikipedia entry for "chain link fencing", suggesting an unverifiable but likely connection between the Google image search algorithm and ImageNet/ILSVRC2012. A text search for "chain-link fencing" produces no "people behind fences" scenes

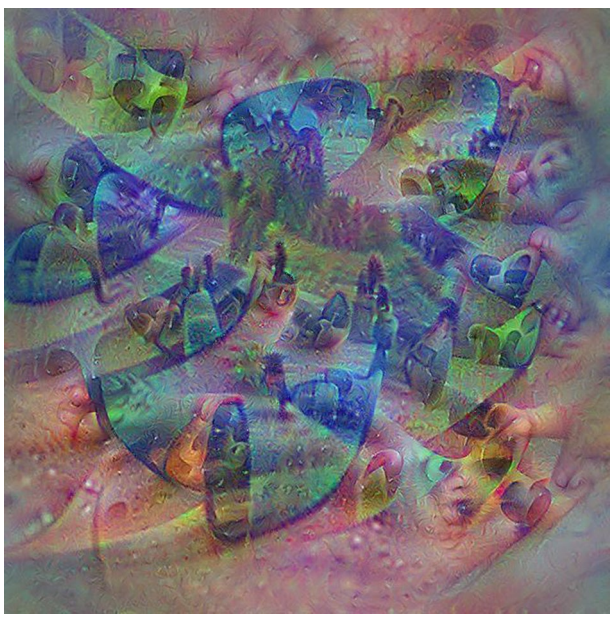

Fig. 7 Regularized feature visualization of the "sunglass" class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset

Fig. 8 All images in the ILSVRC2012 "sunglass" class, sorted by VGG19 fc1 features and plotted with UMAP. Notice the cluster of "landscape mirror"-type images in the center

The original dataset class, as a closer investigation reveals (Fig. 8), is heavily biased towards a specific depiction of sunglasses, popular with stock photo databases and hobby photographers alike: a close-up of a pair of sunglasses that also shows parts of the surrounding landscape (and/or the photographer). The fact that ILSVRC2012 was scraped from the Internet, disregarding aspects like diversity in composition and style, again, leads to an over-specificity of the learned representation. That the mirrored landscapes are desert landscapes (as originally assumed by the authors), however, is a chimera, showing how much artificial contextualization (bringing in additional information from the outside to aid interpretation) matters in both the technical and human interpretations.

9 Conclusion

Analyzing and understanding perceptual bias in machine vision systems requires reframing it as a problem of interpretation and representation, for which we have adapted W. J. T. Mitchells notion of the metapicture. Technical metapictures, we have argued, mirror the act of interpretation in the technical realm: regularization and natural image priors make feature visualization images legible before any interpretation can take place. Paradoxically, however, as the representational capacity of feature visualization images is inverse proportional to their legibility, this pre-interpretation presents itself as a massive technical intervention as well, that disconnects the visualization from the visualized. Nevertheless, feature visualization can also provide a potential new strategy to mitigate bias, if the fact that technical meta-images primarily encapsulate the limitations of machine vision systems is taken into account. All of this suggests that critical machine vision is an essentially humanist endeavor that calls for additional transdisciplinary investigations at the interface of computer science and visual studies/*Bildwissenschaft*.

Funding Open Access funding provided by Projekt DEAL.

References

Arnold T, Tilton L (2019) Depth in deep learning: knowledgeable, layered, and impenetrable. https://statsmaths.github.io. Accessed 8 July 2020

Babbage C (2010) Babbage's calculating engines: being a collection of papers relating to them, their history and construction. Cambridge University Press, Cambridge

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. https://www.fairmlbook.org. Accessed 8 July 2020

Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: quantifying interpretability of deep visual representations. In: IEEE conference on computer vision and pattern recognition (CVPR), pp. 6541–6549

Bau D, Zhu J-Y, Strobelt H, Zhou B, Tenenbaum JB, Freeman WT et al (2018) GAN dissection: visualizing and understanding generative adversarial networks. arXiv preprint arXiv: 181110597

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828

Benjamin R (2019) Race after technology: abolitionist tools for the new Jim Code. Wiley, New York

Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency (FAT*)

Cohen N, Shashua A (2017) Inductive bias of deep convolutional networks through pooling geometry. arXiv preprint arXiv: 16050 6743

Cranmer M, Sanchez-Gonzalez A, Battaglia P, Xu R, Cranmer K, Spergel D et al (2020) Discovering symbolic models from deep learning with inductive biases. arXiv preprint arXiv: 200611287

Crawford K, Paglen T (2019) Excavating AI: the politics of images in machine learning training sets. https://www.excavating.ai/. Accessed 8 July 2020

Daston L (2018) Calculation and the division of labor, 1750–1950. Bull Ger Hist Inst 62(Spring):9–30

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 248–255

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv: 170208608

Dosovitskiy A, Brox T (2016) Generating images with perceptual similarity metrics based on deep networks. In: Advances in neural information processing systems, pp 658–666

Dumoulin V, Visin F (2016) A guide to convolution arithmetic for deep learning. arXiv preprint arXiv: 160307285

Erhan D, Bengio Y, Courville A, Vincent P (2009) Visualizing higher-layer features of a deep network. Université de Montréal, Montreal

Feinman R, Lake BM (2018) Learning inductive biases with simple neural networks. arXiv preprint arXiv: 180202745

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: ACM conference on fairness, accountability, and transparency (FAT*)

Garvie C, Bedoya A, Frankle J (2016) The perpetual line-up: Unregulated police face-recognition in America. Georgetown Law, Center on Privacy and Technology. https://www.perpetuallineup.org. Accessed 8 July 2020

Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv: 181112231

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: An overview of interpretability of

machine learning. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA), pp 80–89
Goodfellow IJ, Shlens J, Szegedy C (2014a) Explaining and harnessing adversarial examples. arXiv preprint arXiv: 14126572
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al (2014b) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
Hohman FM, Kahng M, Pienta R, Chau DH (2018) Visual Analytics in deep learning: an interrogative survey for the next frontiers. IEEE Trans Vis Comput Graph 25(8):2674–2693
Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A (2019) Adversarial examples are not bugs, they are features. arXiv preprint arXiv: 190502175
Kim B, Reif E, Wattenberg M, Bengio S (2019) Do neural networks show gestalt phenomena? An exploration of the law of closure. arXiv preprint arXiv: 190301069
Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
Kurenkov A (2020) Lessons from the PULSE model and discussion. The gradient. https://thegradient.pub/pulse-lessons/. Accessed 8 July 2020
Lipton ZC (2016) The mythos of model interpretability. In: 2016 ICML workshop on human interpretability in machine learning (WHI), New York, NY
Lipton ZC, Tripathi S (2017) Precise recovery of latent vectors from generative adversarial networks. arXiv preprint arXiv: 170204782
Locatello F, Bauer S, Lucic M, Gelly S, Schölkopf B, Bachem O (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. arXiv preprint arXiv: 181112359
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems, pp 4765–4774
Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5188–5196
Malevé N (2019) An introduction to image datasets. Available from: https://unthinking.photography/articles/an-introduction-to-image-datasets
Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. arXiv preprint arXiv: 190110436
Miller GA (1985) Wordnet: a dictionary browser. In: Proceedings of the first international conference on information in data
Minsky ML, Papert S (1988) Perceptrons. MIT Press, Cambridge
Mitchell WJT (1995) Picture theory: essays on verbal and visual representation. University of Chicago Press, Chicago
Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679
Mittelstadt B, Russel C, Wachter S (2019) Explaining explanations in AI. In: ACM conference on fairness, accountability, and transparency (FAT*)
Mordvintsev A (2016) Deep dreaming with TensorFlow. Available from: https://github.com/tensorflow/tensorflow/blob/master/tensorflow/examples/tutorials/deepdream/deepdream.ipynb
Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: going deeper into neural networks. Available from: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Nagel T (1974) What is it like to be a bat? Philos Rev 83(4):435–450
Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in neural information processing systems, pp 3387–3395
Offert F (2018) Images of image machines. Visual interpretability in computer vision for art. In: European conference on computer vision, pp 710–715
Offert F (2020) There is no (real world) use case for face super resolution. https://zentralwerkstatt.org/post_PULSE.html. Accessed 8 July 2020
Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill. Available from: https://distill.pub/2017/feature-visualization
Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, et al (2018) The building blocks of interpretability. Distill. Available from: https://distill.pub/2018/building-blocks
Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S (2020) An overview of early vision in InceptionV1. Distill. Available from: https://distill.pub/2020/circuits/early-vision/
Pasquinelli M (2019a) The origins of Marx's general intellect. Radic Philos 2(6)
Pasquinelli M (2019b) Three thousand years of algorithmic rituals: the emergence of AI from the computation of space. eFlux. Available from: https://www.e-flux.com/journal/101/273221/three-thousand-years-of-algorithmic-rituals-the-emergence-of-ai-from-the-computation-of-space/
Pasquinelli M, Joler V (2020) The Nooscope manifested: artificial intelligence as instrument of knowledge extractivism. KIM HfG Karlsruhe and Share Lab. Available from: https://nooscope.ai
Ribeiro MT, Singh S, Guestrin C (2016) "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, pp 1135–1144
Ritter S, Barrett DG, Santoro A, Botvinick MM (2017) Cognitive psychology for deep neural networks: a shape bias case study. arXiv preprint arXiv: 170608606
Rosenblatt F (1957) The perceptron. A perceiving and recognizing automaton. Cornell Aeronautic Laboratory, Buffalo
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham Law Rev 87:1085
Selbst AD, Friedler S, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: ACM conference on fairness, accountability, and transparency (FAT*)
Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I et al (2013) Intriguing properties of neural networks. arXiv preprint arXiv: 13126199
Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2818–2826
Tyka M (2016) Class visualization with bilateral filters. https://mtyka.github.io/deepdream/2016/02/05/bilateral-class-vis.html. Accessed 8 July 2020
Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv preprint arXiv: 150606579
Zhou A, Knowles T, Finn C (2020) Meta-learning symmetries by reparameterization. arXiv preprint arXiv: 2007.02933