Understanding Perceptual Bias in Machine Vision Systems

Visual Analytics as a (Digital) Humanities Challenge

Fabian Offert[1] Peter Bell[2]

Abstract: Machine vision systems based on deep convolutional neural networks are increasingly utilized in digital humanities projects, particularly in the context of art-historical and audiovisual data. As research has shown, such systems are highly susceptible to bias. We propose that this is not only due to their reliance 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. Perceptual bias, we argue, affects almost all currently available "off-the-shelf" machine vision systems, and is thus especially relevant for digital humanities applications, which often rely on such systems for hypothesis building. We evaluate the nature and scope of perceptual bias by means of a close reading of a visual analytics technique called "feature visualization" and propose to understand the development of critical visual analytics techniques as an important (digital) humanities challenge, situated at the interface of computer science and visual studies.

Keywords: machine learning; visual analytics; computer vision; bias; interpretability; digital art history

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 [Fr18] and science and technology studies [Se19], but also in disciplines such as African American studies [Be19]. As part of this development, machine vision has moved into the spotlight of critique as well, particularly where it is used for socially charged applications like facial recognition [BG18; Ga16]. In many critical investigations of machine vision, however, the focus lies almost exclusively on dataset bias [CP19], and on fixing datasets by introducing more, or more diverse sets of images [Me19]. 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.

We propose that machine vision systems are often inherently biased not only because they rely on biased datasets (which they do) but also because their perceptual topology, their

[1] University of California, Santa Barbara / Friedrich-Alexander-Universität Erlangen-Nürnberg, offert@ucsb.edu
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, peter.bell@fau.de

doi:10.18420/inf2020_121

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. 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 visual analytics technique called “feature visualization” [Er09]. 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 and other visual analytics techniques, 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 [Mi95], as images about (machine) seeing. The development of critical visual analytics techniques, then, becomes an important (digital) humanities challenge, situated at the interface of computer science and visual studies.

2 Building Blocks of Perceptual Bias

2.1 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 [Kr12]. 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. From the topological perspective, we can describe CNNs as layered systems. In the simplest version of a (non-convolutional) neural network, individual layers consist of neurons, atomic units that take in values from neurons in the previous layer and return some weighted sum of these values. Deep convolutional neural networks, then, introduce new classes of neurons, which perform more complex functions like convolution. 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. [SB18] have suggested calling this opacity "inscrutability".

Inscrutability, however, is not the only reason for the notorious opacity of CNNs. As [SB18] 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. [SB18] have argued that the non-intuitiveness of CNNs 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."

2.2 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. Such systems would, 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 three to five years, research in interpretable machine learning has matured into a proper subfield of computer science [Li16; DK17; Gi18] and a plethora of statistical tricks has been developed to ensure the interpretability of simpler models like linear regression. 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 [Of17]. 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 [DK17] write. We propose that, for machine vision systems, this inherent transdisciplinarity implies linking technical concepts and concepts from visual studies. 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.

2.3 Feature visualization

Feature visualization belongs to a range of techniques for the visual analysis of machine learning systems called visual analytics [Ho18]. Originally developed by [Er09] and continuously improved since, feature visualization has been shown to produce remarkable results [Ol17; Ol18]. Importantly, we choose to investigate feature visualization not for its specific relevance to the digital humanities context – in fact, attribution methods [Se17; Ch19] are more commonly employed in the analysis of cultural data [BO20] – but because it best demonstrates the issue of perceptual bias that affects all visual analytics methods. 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 iterations, the result is an image that highly activates one specific neuron.

3 Manifestations of Perceptual Bias

3.1 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). 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 [Sz13; Go14b] – images that highly activate specific neurons or classes in a fully trained deep convolutional neural network, despite being utterly uninterpretable. 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 [Ge19], 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.

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, [Ge19] 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 [Be13] 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, and the digital humanities in particular rely on these "off-the-shelf" models. 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 [Cr20], 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 [Ki19; Ri17; FL18] as an inductive bias. Overall, however, the inductive biases of CNNs are still unclear [CS17] and thus unmanageable.

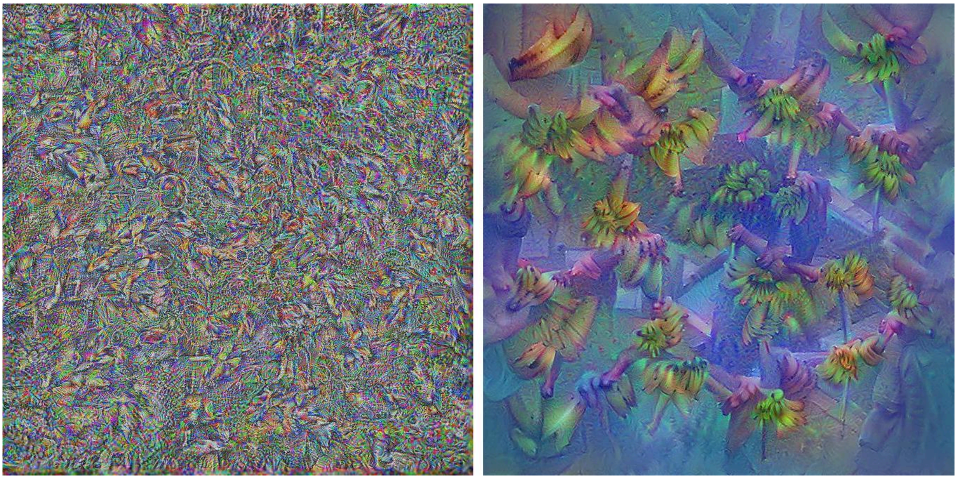

Fig. 1: Left: Unregularized feature visualization of the "banana" class of an InceptionV3 CNN [Sz16] trained on the 1,000 ImageNet classes in the ILSVRC2012 ImageNet subset [Ru15]. Right: regularized visualization of the same class.

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. 1). Regularization, here, simply means adding additional constraints to the optimization process. This can be achieved either by adapting the loss function – for instance, by 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. [Er09] 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, blur total variation filters, bilinear filters, stochastic clipping and Laplacian pyramids. 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 or through direct filtering.

3.2 Semantic bias

Despite all regularization efforts, however, feature visualizations often still present "strange mixtures of ideas" [Ol18]. Visualizing higher-level neurons in particular produces ambiguous results, images that might, or might not, show proper "objects" (fig. 2). 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.

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 [Sz13] suggests that this entanglement is inevitable and absolute. Later work [Ba17] shows that some neural network architectures are less "naturally entangled" then others. Generally, however, significant supervision or, again, artificial inductive biases [Lo19] are required to “disentangle” CNNs and arrive at a meaningful correspondence of technical and semantic units.

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.

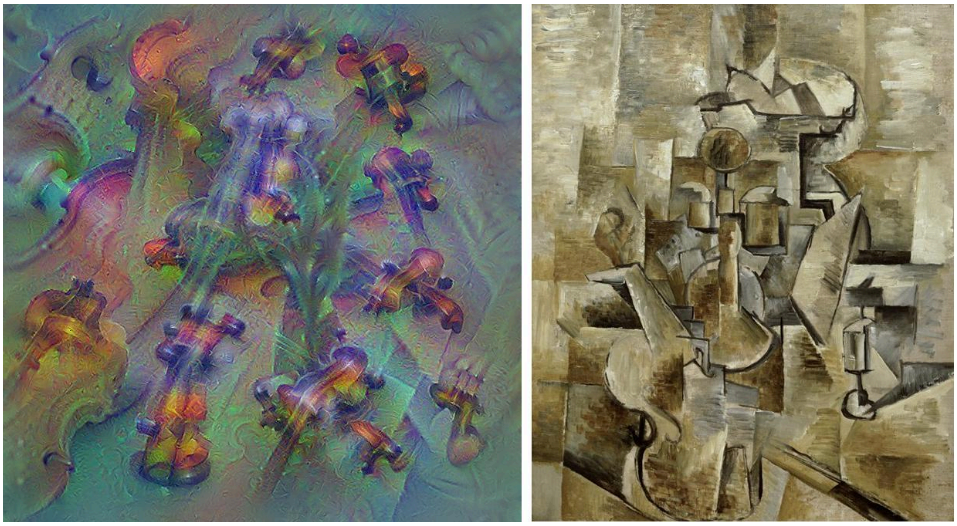

Fig. 2: Left: regularized feature visualization of the “violin” class of an InceptionV3 CNN [Sz16] trained on the 1,000 ImageNet classes in the ILSVRC2012 ImageNet subset [Ru15]. Right: George Braque, Violin and Candlestick (1910).

To produce natural image priors, [DB16; Ng16] propose to use a generative adversarial network (GAN)[3]. This implies, however, that the images that can be produced with this feature visualization method are entirely confined to the latent space of the specific GAN 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

[3] Unfortunately, we cannot explain GANs in detail here, and instead refer the reader to [Go14].

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.

4 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" [Na74], 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 [Mi95]. 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" [Mi95, 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" [Mi95, 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" [Mi95, 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" [Mi95, 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. Interpretations based on feature visualization images thus become (human) interpretations of (technical) interpretations [Of18].

In 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. All of this suggests that visual analytics is an essentially humanist endeavor that calls for additional transdisciplinary investigations at the interface of computer science and visual studies.

Bibliography

[Ba17] Bau, D. et al.: Network dissection: Quantifying interpretability of deep visual representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017).

[BO20] Bell, P., Offert, F.: Reflections on connoisseurship and computer vision. Journal of Art Historiography 23 (2020).

[Be13] Bengio, Y. et al.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence. 35, 8, 1798–1828 (2013).

[Be19] Benjamin, R.: Race After Technology: Abolitionist Tools for the New Jim Code. John Wiley & Sons (2019).

[BG18] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. (2018).

[Ch19] Chen C. et. al.: This looks like that. Deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems. pp 8930–8941 (2019).

[CS17] Cohen, N., Shashua, A.: Inductive bias of deep convolutional networks through pooling geometry. arXiv preprint arXiv:1605.06743. (2017).

[Cr20] Cranmer, M. et al.: Discovering symbolic models from deep learning with inductive biases. arXiv preprint arXiv: 2006.11287. (2020).

[CP19] Crawford, K., Paglen, T.: Excavating AI: The politics of images in machine learning training sets. (2019).

[DK17] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. (2017).

[DB16] Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Advances in Neural Information Processing Systems. pp. 658–666 (2016).

[Er09] Erhan, D. et al.: Visualizing higher-layer features of a deep network. Université de Montréal (2009).

[FL18] Feinman, R., Lake, B.M.: Learning inductive biases with simple neural networks. arXiv preprint arXiv:1802.02745. (2018).

[Fr19] Friedler, S.A. et al.: A comparative study of fairness-enhancing interventions in machine learning. In: ACM Conference on Fairness, Accountability, and Transparency (FAT*). (2019).

[Ga16] Garvie, C. et al.: The perpetual line-up: Unregulated police face-recognition in America. Georgetown Law, Center on Privacy & Technology (2016).

[Ge19] Geirhos, R. et al.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231. (2019).

[Gi18] Gilpin, L.H. et al.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. pp. 80–89 IEEE (2018).

[Go14] Goodfellow, I. et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014).

[Go14b] Goodfellow, I.J. et al.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. (2014).

[Ho18] Hohman, F.M. et al.: Visual Analytics in deep learning: An interrogative survey for the next frontiers. IEEE Transactions on Visualization and Computer Graphics. (2018).

[Ki19] Kim, B. et al.: Do neural networks show gestalt phenomena? An exploration of the law of closure. arXiv preprint arXiv:1903.01069. (2019).

[Kr12] Krizhevsky, A. et al.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012).

[Li16] Lipton, Z.C.: The mythos of model interpretability. In: 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY. (2016).

[Lo19] Locatello, F. et al.: Challenging common assumptions in the unsupervised learning of disentangled representations. arXiv preprint arXiv:1811.12359. (2019).

[Me19] Merler, M. et al.: Diversity in faces. arXiv preprint arXiv:1901.10436. (2019).

[Mi95] Mitchell, W.J.T.: Picture Theory: Essays on Verbal and Visual Representation. University of Chicago Press (1995).

[Mi19] Mittelstadt, B. et al.: Explaining Explanations in AI. In: ACM Conference on Fairness, Accountability, and Transparency (FAT*). (2019).

[Na74] Nagel, T.: What is it like to be a bat? The Philosophical Review. 83, 4, 435–450 (1974).

[Ng16] Nguyen, A. et al.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems. pp. 3387–3395 (2016).

[Of17] Offert, F.: “I know it when I see it”. Visualization and intuitive interpretability. arXiv preprint arXiv:1711.08042. (2017).

[Of18] Offert, F.: Images of image machines. Visual interpretability in computer vision for art. In: European Conference on Computer Vision. pp. 710–715 Springer (2018).

[Ol17] Olah, C. et al.: Feature visualization. Distill. (2017).

[Ol18] Olah, C. et al.: The building blocks of interpretability. Distill. (2018).

[Ri17] Ritter, S. et al.: Cognitive psychology for deep neural networks: A shape bias case study. arXiv preprint arXiv:1706.08606. (2017).

[Ru15] Russakovsky, O. et al.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision. 115, 3, 211–252 (2015).

[Se19] Selbst, A.D. et al.: Fairness and abstraction in sociotechnical systems. In: ACM Conference on Fairness, Accountability, and Transparency (FAT*). (2019).

[SB18] Selbst, A.D., Barocas, S.: The intuitive appeal of explainable machines. Fordham Law Review. 87, (2018).

[Se17] Selvaraju R.R. et. al.: Grad-CAM. Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017).

[Sz13] Szegedy, C. et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. (2013).

[Sz16] Szegedy, C. et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016).