Images of Image Machines. Visual Interpretability in Computer Vision for Art

Fabian Offert[1]

University of California, Santa Barbara, Santa Barbara, CA 93106, USA
offert@ucsb.edu

Abstract. Despite the emergence of interpretable machine learning as a distinct area of research, the role and possible uses of interpretability in digital art history are still unclear. Focusing on feature visualization as the most common technical manifestation of visual interpretability, we argue that in computer vision for art visual interpretability is desirable, if not indispensable. We propose that feature visualization images can be a useful tool if they are used in a non-traditional way that embraces their peculiar representational status. Moreover, we suggest that exactly because of this peculiar representational status, feature visualization images themselves deserve more attention from the computer vision and digital art history communities.

Keywords: interpretability, feature visualization, digital art history, representation

1 Is interpretability necessary?

Contemporary computer vision algorithms – in the context of art and beyond – make extensive use of artificial neural networks to solve object recognition and classification tasks. The most common architecture employed for such tasks is the deep convolutional neural network (CNN) [7, 9, 10]. With the spread of CNNs across domains, however, a problem particular to deep neural networks has resurfaced: while we can train deep neural networks to do very well on specific tasks, it is often impossible to know how a model arrives at a decision, i.e. which features of an input image are relevant for its classification. As a response to this impasse, interpretable machine learning has grown into its own distinct area of research, with visual analytics of CNNs as an emerging field of study [5]. While much of the research in this area is concerned with the development of an empirical approach to interpretability [6,15], one of its open qualitative questions is: which machine learning models need to be interpretable?

While it is obvious that machine learning models deployed in high-stakes scenarios, like credit ratings and recidivism prediction (or predictive policing in general), deserve increased scrutiny and necessitate interpretability [12, 21], it has been put into question [11] if models deployed in less critical contexts require interpretability at all, or if the internal “reasoning” of such models is irrelevant given a good enough error rate on the actual task. The main hypothesis of this paper is that in computer vision for art interpretability is desirable, if not indispensable, despite the lack of a need for normative assessment.

2 Representation and interpretation

One of the most common technical approaches to increase the interpretability of CNNs is feature visualization. Feature visualization has been an important research area within machine learning in general and deep learning in particular at least since 2014 [24,27]. All feature visualization methods rely on the principle of activation maximization: learned features of a particular neuron or layer are visualized by optimizing a random noise input image to maximally activate this neuron or layer.

For instance, an image optimized for an output neuron of a neural network trained on the ImageNet dataset will intuitively show some object from the class associated with this neuron – if it is subjected to proper regularization [19,20,26]. More elaborate methods employ natural image priors to "bias" visualizations even more towards "legible" images [2, 16–18]. In fact, unregularized feature visualization images will often fall into the range of adversarial examples [4] for a given class, i.e. they will not be visually related to natural images from this class but still activate the output neuron for this class with very high confidence. Moreover, as [19] and many others have observed, many feature visualization images are "strange mixtures of ideas" that seem to blend features from many different natural images. This suggests that individual neurons are not necessarily the right semantic units for understanding neural nets. In fact, as [25] show, looking for meaningful features does not necessarily lead to more meaningful visualizations than looking for any combination of features, i.e. producing arbitrary activation maximizations. While some recent results [1] seem to weaken the assumption of a distributed representational structure of CNNs, the assumption has nevertheless given rise to a number of highly visible critical interventions suggesting that it will be necessary to augment deep learning methods with more symbolic approaches [8, 13, 22].

From this indispensability of regularization we can construct a technical argument about the notion of representation as it applies to feature visualization. Johanna Drucker has described the act of interpretation as the collapse of the probability distribution of all possible interpretations [3] for an aesthetic artifact. For feature visualization images, this metaphor applies literally, as feature visualization images are literal samples from the probability distribution that is approximated by the whole model. Somewhat counter-intuitively, feature visualization images, despite being technical images, are thus arbitrary *interpretations* in the exact sense suggested by Drucker. Interpretations based on feature visualization images thus become (human) interpretations of (technical) interpretations. One possible conclusion to draw from this peculiar representational character of feature visualization images would be that visual interpretability as a concept is critically flawed. We propose to draw the opposite conclusion, suggesting that exactly this "subjective" nature of feature visualization images makes visual interpretability useful for computer vision for art.

3 A non-traditional approach to visual interpretability

Our suggestion is to use feature visualization images to "augment" the original dataset under investigation. Concretely this would mean that, in assessing a dataset with the help of machine learning, the digital art historian would not only take the model's results into account but also include a large set of feature visualization images in the analysis. In this "non-traditional" approach, the digital art historian's hermeneutic work would extend back into the very technical system that enables it, operating on both the original dataset and the feature visualization dataset. The technical system, rather than being an opaque tool, would become an integral part of the interpretative process.

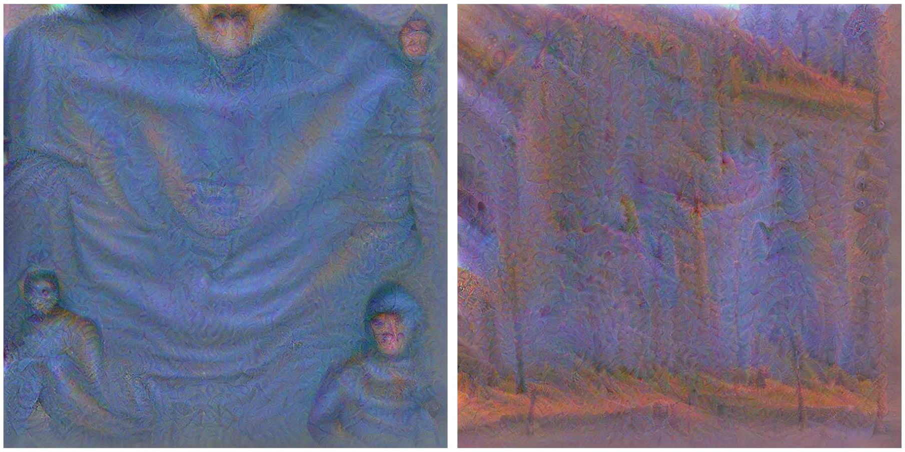

Fig. 1. Feature visualization images for the "portrait" and "landscape" classes of an InceptionV3 neural network. The network was trained on ImageNet and then fine-tuned for ten epochs on an art historical dataset. The dataset, a subset of the Web Gallery of Art dataset, consists of three classes (portrait, landscape, and still life) with 1400 images per class. The resulting classifier reaches 95% validation accuracy. Only minimal regularization was used in the production of the feature visualization images (a 5x5 median filter was applied every four iterations). High resolution was achieved through multi-scale optimization as proposed in the original implementation of the "deep dream" algorithm [14]. The color channels of the final image were normalized independently.

The toy example in figure 1 shows the feasibility of this approach: the model seems to have learned that faces and, surprisingly, drapery are the defining features of a portrait. The highest scoring image from the training dataset, Moretto da Brescia's *Christ with an Angel* (1550) confirms this hypothesis, as it contains two faces and three prominent drapery objects. A defining feature of a landscape painting, according to the model, seems to be an aerial perspective blue shift. Both results point to a subtle (likely historical and/or geographical) bias in the dataset that deserves further analysis. Importantly, however, it is the strangeness, the ambiguity, the "Verfremdungseffekt" of the feature visualization image that is open to the same kind of interpretation as the original image that facilitates this conclusion.

[23] have suggested to understand interpretability as a set of strategies to counteract both the inscrutability and the anti-intuitiveness of machine learning models. Inscrutability is defined as the difficulty to investigate a model with a high number of parameters and a high structural complexity. Anti-intuitiveness, on the other hand, is defined as the fact that the internal "reasoning" of a model does not necessarily correspond to intuitive methods of inference, as hidden correlations often play an essential role. Taking up this distinction, we could say that the specific non-traditional interpretability strategy described above would not try to eliminate the anti-intuitiveness of a machine learning model but put it on its feet by embracing its anti-intuitive nature and exploiting it for the benefit of interpretation.

4 Conclusion

We have shown that the representational status of feature visualization images is not as straightforward as often assumed. Based on this clarification, we have proposed that visual interpretability, understood as a method to render the anti-intuitive properties of machine learning models usable, rather than trying to eliminate them, could benefit computer vision for art by extending the reach of the digital art historian's analysis to include the machines used to facilitate this analysis. Our toy example demonstrates the feasibility of this approach.

Both digital art history and interpretable machine learning are academic fields that only emerged in the past twenty to thirty years, and experienced significant growth only in the past five years. The intimate connection of both fields through their common interest in the analysis and interpretation of images, however, makes a closer collaboration of researchers from both fields reasonable and desirable. The non-traditional interpretability strategy outlined above is only one of many possible non-traditional approaches that could significantly impact both fields, technically, as well as conceptually.

References

1. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3319–3327 (2017)

2. 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), http://papers.nips.cc/paper/6157-generating-images-with-perceptual-similarity-metrics-based-on-deep-networks
3. Drucker, J.: The General Theory of Social Relativity. The Elephants (2018)
4. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014), https://arxiv.org/abs/1412.6572
5. Hohman, F.M., Kahng, M., Pienta, R., Chau, D.H.: Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE Transactions on Visualization and Computer Graphics (2018)
6. Kim, B., Doshi-Velez, F.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017), https://arxiv.org/abs/1702.08608
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
8. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and Brain Sciences **40** (2017)
9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
10. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
11. Lipton, Z.C.: The mythos of model interpretability. In: 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY (2016)
12. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016)
13. Marcus, G.: Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631 (2018), https://arxiv.org/abs/1801.00631v1
14. Mordvintsev, A., Olah, C., Tyka, M.: Inceptionism: Going deeper into neural networks (2015), https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
15. Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., Doshi-Velez, F.: How do humans understand explanations from machine learning systems? arXiv preprint arXiv:1802.00682 (2018), https://arxiv.org/abs/1802.00682v1
16. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems. pp. 3387–3395 (2016), http://papers.nips.cc/paper/6519-synthesizing-the-preferred-inputs-for-neurons-in-neural-networks-via-deep-generator-networks
17. Nguyen, A., Yosinski, J., Bengio, Y., Dosovitskiy, A., Clune, J.: Plug and play generative networks: Conditional iterative generation of images in latent space. arXiv preprint (2017), https://arxiv.org/abs/1612.00005
18. Nguyen, A., Yosinski, J., Clune, J.: Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv preprint arXiv:1602.03616 (2016), https://arxiv.org/abs/1602.03616
19. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill (2017), https://distill.pub/2017/feature-visualization
20. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A.: The building blocks of interpretability. Distill (2018), https://distill.pub/2018/building-blocks
21. Pasquale, F.: The Black Box Society. The Secret Algorithms That Control Money and Information. Harvard University Press (2015)

22. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books, New York, NY (2018)
23. Selbst, A.D., Barocas, S.: The intuitive appeal of explainable machines. Fordham Law Review **87** (2018)
24. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2014), https://arxiv.org/abs/1312.6034
25. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013), https://arxiv.org/abs/1312.6199
26. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. In: Deep Learning Workshop, 31st International Conference on Machine Learning, Lille, France, 2015 (2015)
27. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)