aSeeing with Machines:

Decipherability and Obfuscation in Adversarial Images

Rosemary Lee
IT-University of Copenhagen
Copenhagen, Denmark
rosl@itu.dk

Abstract

Adversarial images, inputs designed to produce errors in machine learning systems, are a common way for researchers to test the ability of algorithms to perform tasks such as image classification. "Fooling images" are a common kind of adversarial image, causing miscategorisation errors which can then be used to diagnose problems within an image classification algorithm. Situations where human and computer categorise an image differently, which arise from adversarial images, reveal discrepancies between human image interpretation and that of computers. In this paper, aspects of state of the art machine learning research and relevant artistic projects touching on adversarial image approaches will be contextualised in reference to current theories. Harun Farocki's concept of the operative image [1] will be used as a model for understanding the coded and procedural nature of automated image interpretation. Through comparison of current adversarial image methodologies, this paper will consider what this kind of image production reveals about the differences between human and computer visual interpretation.

Keywords

Adversarial images, machine learning, deep neural networks, artificial intelligence, invisibility, perception, operative image, algorithms, digital cryptography, decipherability

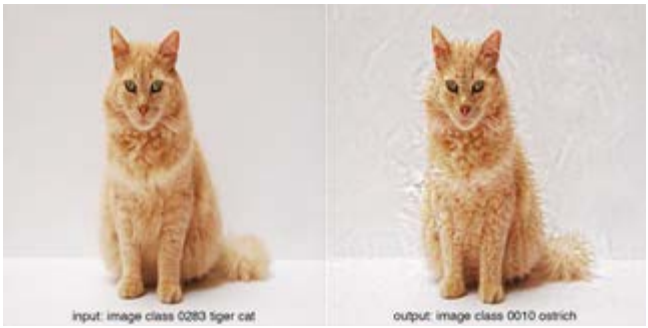
Introduction

Image identification tasks are becoming increasingly automated using machine learning to handle large visual data-sets. Adversarial approaches are developed in parallel to image classification algorithms in order to test algorithms' effectiveness by identifying errors, for example by causing an image to be misclassified. Many adversarial image attacks involve using discrepancies between the limits of human and computer image interpretation to make images human-readable but computer-unreadable. This means that an image may be adversarial in nature while having little to no outward indication that it is so. Adversarial content is thereby obscured from human vision while the image's human-determined meaning is obscured from the computer. Close examination of adversarial images and the errors they trigger help discern the limitations of machine learning systems to decipher digital images. In the following paper, diverse examples of adversarial image approaches will be explored in relation to the concept of the operative image and the relative decipherability and obscurity of digital images.

Adversarial Approaches

While computers can be more efficient than humans at processing large amounts of visual data, it doesn't take much to render a human-interpretable image unreadable to a computer. Because machine learning systems are often black box, it is challenging to develop methodologies to test them and to understand the cause of errors which arise. Many adversarial images are intended to be human-readable at the same time as tricking an algorithm, using human image interpretation as a baseline to check when computer image interpretation diverges from that standard. For this reason, adversarial image attacks often involve image interpretation tasks which are easily performed by most humans but which pose immense technical challenges for current computers [2]. Merely flipping an image upside down, for instance, can render an image incomprehensible to a computer and cause the image to be classified as something other than what it appears to be to humans. On the other end of the spectrum, computers can be tricked into classifying images consisting entirely of visual noise as specific objects, with a high degree of certainty [3].

Transformation, distortion, the addition of visual noise and the use of context cues are a few of the most commonly used techniques to achieve errors in algorithmic image classification. These approaches rely on problematising tasks in which humans excel, which correspond to computer inability. The production of images which are read differently in parallel by humans and computers entails understanding and implementing differences between the two visual processing systems. Internal knowledge about the way that humans see and process images is synthesised with knowledge about the boundaries of computer vision. Much like vision charts used to measure the resolution of biological vision, which test the limit of what size letters a person can read at a distance, adversarial images function as a test to determine the limits of image classification algorithms. If a computer classifies an image of a cat as an ostrich (see Figure 1), for example, a limitation in the image classification algorithm used is thereby established.



Adversarial Image Research

Figure 1: Image processed using the Ostrichinator web demo. The image was initially classified as class 0283: tiger cat, but after processing, the image is classified as class 0010: ostrich.

The past year was an eventful one in machine learning research, with several papers causing shockwaves in the field with adversarial examples which showed that although algorithms for analysing images are human-competitive in many ways, they are not without flaws. Two ground-breaking examples of adversarial image attacks include an approach whereby only one pixel in an image needs to be modified to trigger an error [4] and a 3d-printed turtle, which computers confuse for a rifle [5].

A research project known as the "one pixel attack" [4] demonstrated that significantly changing a single pixel was enough to cause it an image to be miscategorised by otherwise highly successful algorithms. In this project, researchers sought to change images in a human-imperceptible fashion while rendering them unreadable to algorithms. By limiting the transformation of the starting image to a single point, it is possible to maintain an image's semblance for human viewers. Many common adversarial attack approaches apply a slight perturbation to all pixels in an image, but by contrast, this methodology limits the number of pixels altered but allows the perturbation to be unlimited in scale. "Natural images", the data-set of images to be sorted, were evolved using an evolutionary algorithm which tested out different variations on the image. This process was used to determine the placement of the attacking pixel which would be changed and the degree to which it was altered from the original. For each starting image, a set of attacking images was evolved, corresponding to each of the possible target classes used in the experiment. Different versions of an image of a dog, for instance, were thereby made to register as a cat, an automobile, a frog, and so on.

Another innovative project entailed 3d printing a fairly realistic-looking turtle, which, when placed in front of a webcam and analysed by a deep neural network is registered as a rifle [5]. The effect is successful from any angle the turtle is positioned in, whether it is tilted, flipped upside down or turned in any direction. The method used is similar to that of the one-pixel attack, taken a few steps further. In this approach, an adversarial image-texture is developed and validated, then mapped onto a 3d model and made into a physical object. The attacking image is thereby made resistant to variations such as position, lighting and background, making it suitable for real-world application. This project also involved a large amount of practical testing, examining how the objects were "perceived" in various situations. It also moves beyond the still image, toward dealing with visual content "in the wild", by bringing adversarial image approaches into the physical world of objects.

Applications have also been made publicly available for use by non-experts, including the Ostrichinator web demo [6]. The Ostrichinator enables users to automatically transform an ordinary image in such a way that it is classified as an ostrich, using the smallest possible change of pixels. One can also select to change the image so that it registers as another class of images by selecting from a drop-down menu. In the example produced for the purpose of this paper, an image of a cat was transformed so that it would be misclassified as an ostrich. (see Figure 1.)

Artistic Adversarial Image Approaches

Artists have taken on related investigations, considering how computers interpret images and seeking to implement those parameters as a visual language.

Adam Harvey, in his project *CVdazzle* [7], has produced a look-book of suggested styling tips for evading face detection. The makeup and hairstyles presented in the project break the continuity of models' faces with colourful, angular lines, patterns and tufts of hair in unexpected places. By disrupting the symbols which constitute a face for computer vision, these styles render the face undetectable. Another project of Harvey's works toward undermining biometric identification by merging passport photos of multiple people. The images created through this process can thereby be used by one person so they can pass as someone else, without arousing the suspicion of a border control officer.

Sascha Pohflepp's Spacewalk [8] is a generative adversarial network (GAN) [9] which functions like a game played between two neural networks. The one neural network creates images with the goal of tricking the other neural network into classifying the fabricated image as a "real" photograph. The data used to train the networks is a set of images of predatory animals, leading to images of mangled, leopard-printed shapes, suggesting a clash between predator and prey. What is interesting about this approach is that it involves the generative neural network making inferences about what might appear convincing to the judging neural network.

Richard Overill's *Image Steganography* consists of a fairly ordinary-looking snapshot of a girl leaning against a railing [10]. What is significant about this image is that it contains another image embedded within it. Using the technique of digital steganography, Overill concealed a secret message within the least important pixels which define the visible image by offsetting their values nominally. Using a special code one can unlock the hidden message, which was the source code of the same image, from the picture.

Relation to Own Artistic Practice

The author has conducted explorations in adversarial image approaches within her own research-based artistic practice. One such exploration looked at abstraction as a weak spot in image classification algorithms. Due to the absence of image classifiers accounting for abstraction, non-representative images are a great challenge for algorithms to classify. As a critique of the over-determination in machine learning research, the author compiled a data-set of abstract paintings from the Metropolitan Museum of Art's online database and subjected the images to algorithmic analysis using the Wolfram image identifier [11], a successful online image classification program. (see Figure 2) Surprisingly, the abstract images returned a similar success rate compared to expressly-designed adversarial images. 98% of the images were categorised incorrectly as being objects ranging from axes to eggbeaters, even a sea snake and an igneous rock. 2% of the images were correctly categorised as paintings. The abstract paintings analysed in this experiment bore little or no visual resemblance to the classes assigned to them by the computer, but each misclassification added layers of poetic meaning to the respective image. The results of the experiment point toward the conclusion that adversarial images owe their success not to being specially designed to trick algorithms, but rather to being abstractions for which there is no image class in a system which insists all images belong in a category. While abstraction for humans appears to be a continuum along which representations vary, image analysis algorithms lack such notions of conceptual connection across visually diverse images.

Operative Images

Harun Farocki's theory of the operative image is useful in understanding the processes at work in adversarial images. He describes a turn away from representation in favour of implementing visual procedures which may or may not be intelligible to human viewers.

"These are images that do not represent an object, but rather are part of an operation." [1]

In an operative image, what is displayed on the screen is merely a by-product of the operation the image helped to perform. Thus, there is less need to make digital images interpretable to "meat-eyes" [12] and we find ourselves immersed in an image culture where humans are at times a secondary audience. A drone, for example, searches for a flight path using digital video, verifying landmarks as it flies, and adjusting its course accordingly. In the footage Farocki used for his piece "Eye/Machine" and which he uses as an example of an operative image, there's frankly not much for human viewers to see. The video is primarily used as input to guide the flight of a drone. Here, the visual is subjugated to the procedure of navigation and the work fluctuates between visualisation and non-visual processes. While operative images are not necessarily intended to communicate with the human senses, they do so nonetheless. There is a digital residue for us to look at,

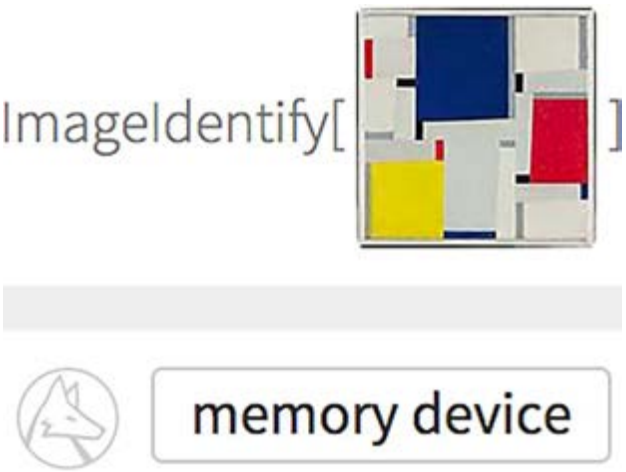


Figure 2: Abstract painting as classified by the Wolfram image identifier.

though we may not understand what it represents. Harun Farocki's work on operative images has been described as an exploration of how to see like a machine [12] and it offers a useful perspective on the human interpretation of images intended for computers.

Decipherability & Obfuscation

The parallel interpretation which occurs in adversarial images entails a two-sided invisibility. There are two levels on which the image functions upon: the human-decipherable image, which is obscured from the computer, and the computer-decipherable image, which is obscured from the human viewer. To render an image indecipherable, in terms of adversarial images, is to ensure it can be read in more than one way. Thus, an image may be unreadable while its intended meaning is hidden in plain sight. Often machine learning systems are opaque, even to their creators, so the errors which arise from adversarial images offer useful insights into their functioning. Adversarial images can help to visualise incongruencies between biological vision and automated image processing by pointing out errors in interpretation. Images which are categorised by algorithms as a different image class than their human-designated category allow us to examine what signifiers are involved in the process of miscategorisation. The kinds of methodologies used in adversarial image attacks are predicated on assumptions as to how human perception and machine learning image analysis relate to one another. For example, in order to create an image which will pass for a representation of a face for either a computer or a human viewer, it is necessary to know what signifiers indicate faciality to the respective receiver and what defines the parameters of the given perceptual system. Situations where image-class signifiers are not aligned with one-another demonstrate the gulf of difference between human visual processing and that of computers.

Adversarial approaches demand a certain level of objectivity in the image interpretation process, for an image

to be deciphered as an intended, “correct” interpretation, upon which a human audience can easily form a consensus. “No, that's not a rifle, it's clearly a turtle.” This requires that images be human-decipherable while being deciphered by computers as a different, pre-selected "target class". In this kind of interaction with images, human and computer visual processing tasks occur on different planes, based on fundamentally different processes and criteria. The overlap between human and computer interpretation of symbolic classes is thereby explored using human-oriented signifiers and parallel, but different, computer-oriented symbolic classes.

Conclusion

Digital images are palimpsests of information, containing far more than what is intelligible to the human eye. The photographic paradigm prevails, yet there is more at work below the surface of the image, their basis in algorithms concealed by a veneer of realism. The turn toward indecipherability and obscurity seen in adversarial images marks a shift in the culture of the screen [13], moving away from traditional pictorial representation toward that of actionable images [14]. Similarly, the progression which can be seen through the examples mentioned in this paper show how the operation of solving a visual task, causing an error for instance, takes priority over an adversarial image's visual content. It is necessary for the image to be transformed while remaining human-readable, but the human-readable representation is merely a cover for the message it sends to the computer or the reverse. In many cases, digital image processes are obscured from the view and understanding of humans, yet the errors visible in adversarial approaches reveal the disjunction between parts of images intended for human eyes and those for computer processing. The black box nature of image processing algorithms is rendered visible by identifying points where neural networks come out of alignment with human vision. The diverse adversarial practices examined here trace the boundaries between visibility, invisibility, human and computer, revealing hidden nuances of each. The image performs a function, which ends up deciphering human image interpretation at the same time as that of computers.

References

1. Farocki, H. (2004), 'Phantom Images', 12-22 in Saara Liinamaa, Janine Marchessault and Christian Shaw (eds) PUBLIC 29: New Localities.

2. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, 321(5895), 1465–1468.

3. Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Computer Vision and Pattern Recognition (CVPR). IEEE.

4. Su, J., Vargas, D. V., & Sakurai, K. (2017). *One pixel attack for fooling deep neural networks*. CoRR, abs/1710.08864. http://arxiv.org/abs/1710.08864

5. Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2017). *Synthesizing Robust Adversarial Examples*. CoRR, abs/1707.07397. http://arxiv.org/abs/1707.07397

6. Tsai, C.-Y., & Cox, D. (2015). *Are Deep Learning Algorithms Easily Hackable?* http://coxlab.github.io/ostrichinator

7. Harvey, A. (2010). CVdazzle [Look book].

8. Pohflepp, S. (2017). Spacewalk [Custom generative adversarial neural network, transfer print on mylar, LED floodlight].

9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Ward-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Nets. *ArXiv:1406.2661v1*, (stat.ML). https://arxiv.org/pdf/1406.2661.pdf

10. Overill, R. (2014). Image Steganography: Digital Images as Covert, Subliminal Channels. *The New Everyday*, (The Operative Image).

11. The Wolfram Language Image Identification Project. (2015). https://www.imageidentify.com

12. Paglen, T. (2014). Operational Images. E-Flux, 59. http://www.e-flux.com/journal/59/61130/operational-images/

13. Manovich, L. (2001). A Screen’s Genealogy. In The Language of New Media (pp. 95–103). Cambridge: MIT Press.

14. Hoelzl, I., & Marie, R. (2015). Softimage: Towards a New Theory of the Digital Image. Bristol/Chicago: Intellect Ltd

Author Biography

Rosemary Lee is an artist and researcher whose work investigates interrelations between technologies and processes of natural science. Their work brings together influences from media geology, hybrid ecology and posthumanism through theory-driven practice-led research. A selection of their notable exhibitions include *machines will watch us die* (Holden Gallery, GB, 2018), *A New We* (Kunsthall Trondheim, NO, 2017), *Hybrid Matters* (Nikolaj Kunsthal, DK, 2016), *TRANSART* (Dome of Visions, DK, 2015) and *Artifacts* (Palais des Beaux Arts, AT, 2015).

Rosemary Lee is currently a PhD fellow at the IT University of Copenhagen. She has also acted as artist-researcher in residence in international contexts including the Ayatana Arist Research Residency (CA, 2017), rural.scapes - Laboratory in Residence (BR, 2016) and the transmediale Vilém Flusser Archive Residency for Artistic Research (DE, 2014).