ARTICLE IN PRESS
SCHRES-07125; No of Pages 4
Schizophrenia Research xxx (2017) xxx–xxx
Contents lists available at ScienceDirect
Schizophrenia Research
journal homepage: www.elsevier.com/locate/schres

Deep dreaming, aberrant salience and psychosis: Connecting the dots by artificial neural networks

Matcheri S. Keshavan [a,b,*], Mukund Sudarshan [c]

[a] BethIsrael Deaconess Medical Center, United States
[b] Harvard Medical School, United States
[c] Cornell University, United States

ARTICLE INFO

Article history:
Received 23 October 2016
Received in revised form 8 January 2017
Accepted 8 January 2017
Available online xxxx

Keywords:
Psychosis
Deep dreaming
Artificial neural networks
Schizophrenia

ABSTRACT

Why some individuals, when presented with unstructured sensory inputs, develop altered perceptions not based in reality, is not well understood. Machine learning approaches can potentially help us understand how the brain normally interprets sensory inputs. Artificial neural networks (ANN) progressively extract higher and higher-level features of sensory input and identify the nature of an object based on a priori information. However, some ANNs which use algorithms such as the "deep-dreaming" developed by Google, allow the network to over-emphasize some objects it "thinks" it recognizes in those areas, and iteratively enhance such outputs leading to representations that appear farther and farther from "reality". We suggest that such "deep dreaming" ANNs may model aberrant salience, a mechanism suggested for pathogenesis of psychosis. Such models can generate testable predictions for psychosis.

1. Introduction

The connections between creativity, madness and dreams is well known (Benson and Park, 2013), but defy explanation. Given the fact that all of these originate from the brain, the logical place to "connect" these dots is by looking at how our neural networks interact with each other to produce our experiences and beliefs. This line of thinking led me and Mukund Sudarshan, a computer science undergraduate in my lab, to glean some interesting, yet speculative insights using principles of machine learning. Mukund applied the Google's "deep dreaming" algorithm, an artificial neural network program (Szegedy et al., 2014), to a landscape painting of mine (Fig. 1, top). Tantalizingly, the realistic picture of lakes and mountains transformed to a Dali-esque dream-like picture with hidden animals, lamps and other obscure objects (Fig. 1, bottom). Given that I did not ask for this particular surrealism effect but the computer generated it on its own, we wondered what might have happened behind the "black-box" of this computer based artificial neural network (ANN) program.

2. How artificial neural networks "learn"

Current thinking in the field of machine learning might help us understand how the brain normally interprets sensory inputs. Artificial neural networks (ANN) are seen as a series of sieves with progressively small holes stacked on one another; these sieves progressively "strain" higher and higher-level (more and more fine-grained) features of sensory input (Fig. 2). In a neural network that has been trained to identify a cat, for example, the lower input layers interpret basic (or coarse) features, like edges or corners (like the outline of a face) and the intermediate layers look for overall shapes. The final layers then combine the details to assemble these into an "answer" so that the image best depicts a "cat". Thus, after training, each layer progressively extracts higher and higher-level features of the image, until the output layer essentially makes a decision on what the image shows. The figure below is a conceptual representation of a neural network. Essentially, the image is processed in overlapping pieces. The first layer analyzes each piece of the image (convolution layer). The subsequent layer (pooling layer) pools together the knowledge gained from pieces that are close to one another and creates a more dense representation (the smaller the square in the image, the more dense the representation). This process repeats until we have a dense representation that captures the absolute essence of the image, rather than minute details. This final representation is used to determine what the image actually contains. In this case, the image contains the realistic representation of a "cat".

3. "Deep dreaming" and aberrant salience

Typically, we feed an input into the network and ask it to identify the image based on whatever it has "seen before". If all is well, the neural network outputs the correct label for the image. However, this "deep-

* Corresponding author at: Room 610, Massachusetts Mental Health Center, 75 Fenwood Road, Boston, MA 02115, United States.
E-mail address: mkeshava@bidmc.harvard.edu (M.S. Keshavan).

http://dx.doi.org/10.1016/j.schres.2017.01.020
0920-9964/© 2017 Elsevier B.V. All rights reserved.

Fig. 1. Application of a deep dreaming algorithm to a painting by the first author. Top: before applying the algorithm; bottom: after applying the algorithm.

dreaming" neural network allows the network to take parts of the image and over-emphasize some objects it thinks it recognizes in those areas. Whatever this neural network thinks it sees is then shown in the image. For example, if the neural network thinks the ear on a cat face is a butterfly wing, it allows it to draw a butterfly wing on top of the eyebrow. This modified image is fed back into the neural network and this process repeats. After numerous iterations, we end up with the painting on the right with unexpected hidden images that were gradually brought out in the image. Thus, the network has wandered off far from "reality" by paying too much attention to less relevant areas. We see what the neural network thinks exists in the inputted image. By observing the output of each step in this iterative process, we can see the transition of the neural network's dream from a normal image into one of fantasy. Such a phenomenon is reminiscent of the progressive changes in the artistic depiction of cats by the famous 19th century English painter Louis Wain (McGennis, 1999).

4. Is psychosis related to "deep-dreaming" neural networks?

The above model can potentially help us understand psychotic phenomena. We suggest that an aberrant enhancement of "salience" of incoming information in selected layers of ANNs may create unexpected, and internally generated informational representations; these are not just based on external inputs, and thus "not matched to reality" akin to what is seen in psychosis.

Indeed, Kapur (2003) has proposed that psychosis may be related to aberrant assignment of salience to the elements of one's experience. At a neural network level, it has been suggested that such aberrant salience may be related to an abnormal neural activity in the ventral striatum, and a failure of feedback regulation by prefrontal structures in individuals with schizophrenia and those at risk for this illness (Roiser et al., 2013). Thus, our response to sensory inputs is regulated by feedback systems, whereby higher brain regions may modulate assignment of