Noname manuscript No.
(will be inserted by the editor)

Throwing light on black boxes: emergence of visual categories from deep learning

Ezequiel López-Rubio

the date of receipt and acceptance should be inserted later

Abstract One of the best known arguments against the connectionist approach to artificial intelligence and cognitive science is that neural networks are black boxes, i.e., there is no understandable account of their operation. This difficulty has impeded efforts to explain how categories arise from raw sensory data. Moreover, it has complicated investigation about the role of symbols and language in cognition. This state of things has been radically changed by recent experimental findings in artificial deep learning research. Two kinds of artificial deep learning networks, namely the Convolutional Neural Network (CNN) and the Generative Adversarial Network (GAN) have been found to possess the capability to build internal states that are interpreted by humans as complex visual categories, without any specific hints or any grammatical processing. This emergent ability suggests that those categories do not depend on human knowledge or the syntactic structure of language, while they do rely on their visual context. This supports a mild form of empiricism, while it does not assume that computational functionalism is true. Some consequences are extracted regarding the debate about amodal and grounded representations in the human brain. Furthermore, new avenues for research on cognitive science are open.

Keywords deep learning · visual categories · machine learning · cognitive science

1 Introduction

Artificial neural networks have been regarded for decades as a typical example of a black box among machine learning methods (Benitez et al, 1997, p. 1156; Tu, 1996, p. 1226). It was claimed that it was impossible to understand how the networks

This is a preprint of the paper published in Synthese. The final published version is at https://doi.org/10.1007/s11229-020-02700-5

Ezequiel López-Rubio
Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga (UMA), Bulevar Louis Pasteur 35, 29071 Málaga, Spain
Departamento de Lógica, Historia y Filosofía de la Ciencia, Universidad Nacional de Educación a Distancia (UNED), Paseo de Senda del Rey 7, 28040 Madrid, Spain

transform their inputs into their outputs (Butz and Kutter, 2017, p. 61; Berkeley, 2019, p. 201). It was further emphasized that their functionality was distributed among the units of the network so that it was impossible to identify which parts of the network were responsible for learning specific categories. They have been depicted as the epitome of unintelligibility due to the difficulty of decomposing the operation of the networks into understandable steps (Carabantes, 2019, p. 6). This criticism was typical in the old shallow networks era, where the paradigmatic artificial neural networks had three neural layers at most and was inherited by the new deep networks, which typically contain dozens of layers. Their learning algorithms are understandable, but the specific process which makes a network yield an output for a particular input is exceptionally complex since it involves tracing millions of connections. This additional structural complexity means that deep networks lack functional transparency (Creel, 2020). Building on this drawback, it was possible to argue that the outstanding successes of deep learning did not have any resemblance to the concept based operation of human reasoning (Gomes, 2014). Nevertheless, demonstrations of concept acquisition by shallow networks have been produced for some decades, where concepts are associated with regions in the input spaces of their hidden units (Clark, 1993, pp. 95-98). These attempts have been impeded by the difficulties of assigning meaning to the dimensions of such spaces (Gärdenfors, 2000, section 7.1.4).

This situation has completely changed recently due to the efforts of researchers to analyze the operation of deep networks (Chollet, 2018, p. 160). The networks that are analyzed in this work employ supervised learning. This means that a human supervisor establishes a target (desired) output, and then the parameters (synaptic weights) of the network are adjusted by a learning algorithm to attain that goal. Typically, the human supervisor labels the image of a scene with its class, for example, dining room, kitchen, park, airport, or forest, so that the training set is made of images of scenes. Convolutional neural networks (CNNs) trained to recognize types of natural scenes, and Generative Adversarial Networks (GANs) trained to generate realistic scenes, have been found to contain individual units (neurons) which learn distinct high level (complex) visual categories of objects within the scenes such as floor, chair, airplane, table, or tree (Bau et al, 2019, Qin et al, 2018, p. 173). These visual categories are learned without any hint about them from humans, i.e., the networks are not provided any information about how many categories may exist. For example, the human supervisor marks which images correspond to dining rooms, but she does not provide any indication about chairs, tables, or any other object that may appear in a dining room. The units associated with high level categories appear in specific layers within the deep architecture. It must be noted that the CNNs analyze while the GANs synthesize (Section 4). Deep networks were already known to be able to recognize or generate instances of object classes already defined by a human who serves as the supervisor of the learning system, but these new experiments are entirely different. The emergent visual concepts are learned spontaneously by the deep networks because they are useful as intermediate steps towards the resolution of the final goal established by the human supervisor. Therefore, these intermediate categories are grasped without any human provided information or assistance since they appear spontaneously during the operation of the learning algorithm. In this work, I investigate the consequences of the emergence of internal states in some deep neural networks that are interpreted as visual categories by humans, as detailed in Section 2. I do

not address the issue of neural network transparency for other applications such as decision making for medical diagnosis and judiciary systems, which are most relevant for the general public (Creel, 2020).

The emergent visual categories learned by the deep networks comprise the object themselves and their surroundings. That is, the relevant context of an object type is learned along with the visual aspect of the instances of the object at hand. This is particularly evident for the GANs when interventions are made, after the training is completed, in some units which spontaneously learn categories. It is possible to insert an object artificially somewhere in an image generated by a GAN, by forcing one or more units associated with that kind of object to activate[1]. When such an intervention is made, a freshly generated, new instance of the object class learned by the unit appears in the generated output image, and the vicinity of the new object is modified so that it is visually coherent with the inserted object. It is interesting to note that the network often refuses to insert an object out of place, such as a door in the middle of the sky (Bau et al, 2018, p. 10). Therefore, the GANs learn the shapes associated with an object class, the relative positions of objects of other classes in the scene, and the visual coherence among them. For example, they can even generate reflections on other objects present in a room when a window is inserted. Also, sometimes they change textures while leaving the same kinds of objects. If the unit associated with a category is manually deactivated, a previously existing instance of the associated object class may vanish from the scene, leaving a visually coherent background. As seen, the learned visual categories comprise high level patterns and constraints, i.e., they are not just a miscellaneous collection of low level primitive features. For example, a deep network can learn the pattern that a chair has legs and a seat, and the constraint that the seat is above the legs. Hence there is a true category acquisition rather than a mere adaptation to raw visual similarities (Buckner, 2015, p. 317). This capability resembles the operation of the ventral stream of the human brain (Buckner, 2018, p. 5366).

From a philosophical point of view, these developments are extremely relevant. Since the networks are not provided with any prior knowledge about those high level categories, it follows that the networks learn those categories by themselves. In other words, those categories are naturally present in the image datasets. Here naturalness means that they are not human made categories because humans do not supply the networks with any information about them. Humans interpret the categories as cognitively relevant, but this interpretation is made only after the network has finished its training process. Therefore, these categories represent structures that are intrinsic to visual datasets. That is, high level visual categories like chair, tree or table are not arbitrary, although the signs (words) that are used to name them are.

In this work, I adhere to the deflated notion of a concept advocated in (Buckner, 2018, p. 5341). That is, the focus is on how artificial deep learning neural networks successfully learn representations of categories of objects, where those categories have not been supplied to the networks by humans. Therefore, the question of whether those visual categories fulfill objective criteria of correctness and inter-subjective agreement is not addressed here. Such criteria are associated with more demanding notions of concept.

[1] An online demonstration is available at http://ganpaint.io/

The CNNs and the GANs do not perform any structured linguistic processing at all. Therefore, their acquisition of high level visual categories is made without the participation of any linguistic structures. This suggests that those categories are independent of grammar. In other words, visual categories do not need to be embedded in the syntactic structures which make human language. This position matches the fact that visual information processing is more primitive from the biological evolution point of view and earlier from the child development perspective, than linguistic abilities.

The ability of CNNs and GANs to grasp visual categories from raw image data entails fundamental consequences for the debate between empiricism and rationalism in the philosophy of the mind and sheds some light on the role of language and other prior human knowledge in the formation of visual categories. These aspects are treated in Section 7. Finally, Section 8 concludes this paper with some considerations about possible future developments in cognitive science research.

2 Concepts versus categories

The prevalent notion of concept in cognitive science and psychology states that concepts are bodies of knowledge that are employed to classify entities (Chin-Parker and Ross, 2002, p. 353). Concepts are a primary tool that is used by our cognitive processes (Machery, 2005, p. 446). I will note concepts in **boldface**. In a strict sense, natural concepts are mental representations of classes that arise in nature, i.e., those that are not related to human activity such as **elephants** or **emeralds** (Ross, 2001). In a broader sense, natural concepts are those which are amenable to inductive generalization (Machery, 2005, p. 445), because they comprise a set of properties which are commonly found in the members of the class captured by the concept (Gärdenfors, 2000, section 3.4). In this work, we adhere to the notion of visual category as a set of visual patterns that a human would associate with a visual concept. I will note categories in normal text, as opposed to concepts. The evaluation of the semantic relevance of the classification carried out by a network must be done by a human (Cantwell Smith, 2019, p. 62). In other words, the association of a visual category with a visual concept can only be made by a human. A concept is a body of knowledge that is amenable to constitute an elementary block of human cognitive processing. We acknowledge that many concepts sharply differ in their structure and their suitability for inductive reasoning (Machery, 2005, p. 465). The debate about the nature of concepts is out of the scope of our investigation since we focus on what we will call natural visual categories:

Definition 1 A *natural visual category* is a class of possible visual stimuli whose associated concept is useful for human cognitive processing.

Under the above definition, possible examples of natural visual categories are door, bird, bed, sky, and person, where each of these categories is understood as a class of visual depictions of exemplars of the class. In other words, the door category comprises the infinite set of all possible visual images of doors, both real and imaginary. Of course, the boundaries of such a set are fuzzy and not crisp, i.e., something could look more or less like a door depending on the observer. But the

key point is that natural visual categories are suitable for cognitive processing of the visual information coming from the physical world.

It is worth noting that there is nothing in Definition 1 that requires that a structured linguistic definition is provided for a natural visual category. In order to have a door category, there is no need that a linguistic description of the defining properties of the **door** concept is supplied. A class of visual patterns defines the category, and it is understood that at the boundaries of the class, there are exemplars of visual patterns that have a partial degree of membership to the class. This absence of a linguistic definition and the fuzziness of the boundaries of the class agree with the kind of categories that CNNs and GANs learn. As mentioned in Section 4, neither of these networks perform any symbol processing, since their learning consists in extracting relevant visual patterns from the input images. This does not preclude these networks from learning high level categories since they grasp high level categories by the composition of low level details. Another difference with respect to the categories of the classic approach to cognitive science is that there is a large amount of redundancy since typically several slightly different versions of the same category are acquired by various units of the same network (Section 5). In contrast, the classic account of categories involves unique and sharp linguistic definitions with little room for inconsistencies.

The high level categories learned by CNNs and GANs are natural visual categories in the sense of Definition 1. The interpretable unit detection procedure and the subsequent validation (Section 5) ensure that the classes of visual stimuli that are captured by the networks are meaningful and relevant for humans. Therefore, those categories are amenable to human cognitive processing, as required by Definition 1. The category associated with a given interpretable unit can be defined as follows:

Definition 2 The natural visual category associated with an interpretable unit is the set of visual stimuli for which the unit is active, where activation means that the output of the unit is higher than some predefined threshold. If more than one category surpasses the threshold, then the unit is associated with the category for which the activation is the highest.

This way, a category is understood as a region of the input space of a unit. Deep learning neural networks do not learn prototypes, but regions. In other words, deep learning categories are region based, and not prototype based. As explained in Section 5, typically, a given category is learned by several interpretable units of the network, which may belong to the same or different neural layers. Each unit is activated by a slightly different set of visual stimuli. Hence the representation that the network makes of the category is fuzzy and not crisp. That is, for examples that are at the boundaries of the category, not all the interpretable units will activate. In contrast, for more typical examples, almost all interpretable units associated with the category will agree.

3 Connectionism and artificial neural networks

The classic perspective in cognitive science viewed the human mind as a symbolic processor which manipulates linguistic elements or symbols according to specific rules. These rules include syntactic constraints (Horgan and Tienson, 1996 p. 23).

That is, syntactic structures formed by symbols are passed through some transformation rules to produce other syntactic structures. The dominant symbolic approach to artificial intelligence matched this perspective. Therefore, symbol manipulating Turing machines served both as models of digital computers and human brains (Horgan and Tienson, 1996 p. 24; Butz and Kutter, 2017, pp. 39-40). It was postulated that mental processes are framed in a cognitive architecture that is analogous to a computer programming language so that a language of thought is responsible for the functional characteristics of the brain (Dawson, 2004, p. 110; Piccinini and Scarantino, 2011, p. 14). The characteristics of this postulated mental language were to be determined by cognitive research. Under this framework, computation is equivalent to the manipulation of sequences of symbols that are syntactically valid sentences in the language of thought (Piccinini and Scarantino, 2011, p. 8; Salisbury and Schneider, 2019, p. 311).

However, it has become evident that the individual neurons of the human brain do not process symbols. Consequently, the fundamental assumption of the classic approach has not been properly linked to biological evidence. This is known as the symbol grounding problem (Butz and Kutter, 2017, p. 53), which comes from the difficulty of specifying where mental symbols get their reference (Salisbury and Schneider, 2019, p. 310). Based on this observation, and further supported by the success of artificial neural networks in solving practical problems in science and engineering (Bechtel, 1993, p. 120), the connectionist approach was proposed as an alternative to the classic one. Connectionism draws on artificial neural networks as models of the brain (Horgan and Tienson, 1996 p. 49), which implies that many cognitive tasks are not related to grammar processing. In particular, pattern recognition and categorization might be carried out without any logical reasoning (Bechtel, 1993, p. 126). Therefore, the formation of categories is not necessarily associated with the syntactic manipulation of symbols that represent them. This contrasts with the classical approach, where the manipulation of symbols is typically done according to syntactic rules that preserve the structure of the symbolic representations. From the connectionist point of view, categories are not clear cut and exceptionless linguistic definitions, but fuzzy and robust collections of patterns (Horgan and Tienson, 1996 p. 142). As seen, there is a strong contrast between the classic theories, which are based on crisp categories sequentially manipulated according to rules based on grammars and logic (Newell and Simon, 1976, p. 116), and the connectionist ones, which rely on fuzzy categories which are processed by distributed networks of units (Piccinini and Scarantino, 2011, p. 2; Mayor et al, 2014, p. 1; Buckner and Garson, 2019, p. 80). The latter notion of fuzzy categories agrees with the views of psychologists who think that definitions of everyday categories cannot be completely sharp (Buckner, 2018, p. 5346; Brainerd and Reyna, 2002, p. 164).

Nowadays, distributed, non symbolic representations are directly observed in the human brain. As a consequence of this, a large part of the connectionist research program has already been integrated into the standard approach to cognitive science (Piccinini and Scarantino, 2011, p. 15; Mayor et al, 2014, p. 2) and current psychotherapy methodologies (Neudeck and Wittchen, 2012, p. 4). Nevertheless, the lack of interpretability of artificial neural networks has been singled out as a limit to the chances of connectionism to become an overarching model of the brain (Dawson, 2004, p. 239). It is interesting to note that artificial neural networks do not necessarily learn decomposable representations (Buckner and

Garson, 2019, p. 82). That is, the representation of a high level category does not necessarily contain interpretable representations of low level categories of the parts of the main category.

When connectionism was first proposed, artificial neural networks were typically shallow. This means that they comprised few neural layers, where the neurons of one layer were connected to those of the previous and next layers. In most cases, there were just three layers: the input layer, the hidden layer, and the output layer. Networks with two, three, or four layers were also employed. Still, the performance gain diminished as more hidden layers were added, up to a point where more layers decreased the performance. It was speculated that more layers could furnish better functionality (Buckner, 2018, p. 5351), but it was not known how. In recent years, new activation functions and more advanced learning algorithms have been developed, so that deep networks with tens or hundreds of layers can be successfully trained. These deep architectures share many characteristics of the shallow ones, such as non symbolic, parallel, and distributed processing. Consequently, many of the postulates of connectionism can be applied to deep networks.

For our discussion, a notion of levels of description is considered where low level objects are parts of high level ones (Elber-Dorozko and Shagrir, 2019, p. 211, 213). Hence the objects at different levels differ in their size and complexity: higher level means larger and more complex objects. It has been argued that this compositional hierarchy of levels facilitates the issue of defining abstract categories (Buckner, 2019, p. 9). It seems the compositional hierarchy of levels of categories only emerges when the neural architecture has a certain kind of structure, i.e., other neural architectures learn to accomplish the task (recognize or generate scenes), but they do it without learning high level categories (Qin et al, 2018, p. 173).

Now it is time to clarify the notion of emergence that is employed in this work since its usage varies depending on the discipline. For cognitive science purposes, emergence is conceived as a complex system behavior that comes from the interaction of simple subsystems (Dawson, 2004, p. 63; Stephan, 2006, p. 486). In this sense, the acquisition of high level visual categories from the interaction of the units of a deep neural network can be regarded as emergent (see Section 6). This is a form of self organization since a global pattern arises from the local interaction of a multitude of simple processing units. Alternatively, it can be said that the learning of visual categories is an example of structural unpredictability (Stephan, 2006, p. 496) because the existence of units that represent visual categories of a very complex structure cannot be inferred from the principles of artificial neural computation. It must be noted that the goals which the deep networks are set to accomplish refer to images of scenes with no provided internal structure. That is, the images are supplied to the networks without any indication of which objects might be present on them, i.e., the images are just matrices of pixel values. Therefore, the occurrence of units that comprise structurally complex visual categories is not expected from the optimization of such goals.

4 Two deep learning artificial neural networks

Next, we outline the main characteristics of the two kinds of deep artificial neural networks where the emergence of high level visual categories has been experimen-

tally found, namely CNNs and GANs. Both of them share some critical characteristics. First of all, they are feed forward neural networks. This means that the information always flows from the input to the output, without any feedback loop (Ketkar, 2017, pp. 19-20). In other words, the output of the network depends on the current input only, and not on the past inputs. Consequently, these networks have no internal memory, and they do not process sequences of inputs since all the required inputs are provided simultaneously. They do not do linguistic processing. Instead of this, they are oriented to image processing. They comprise convolutional neural layers, which adapt to specific visual features by learning from a set of training images.

In some sense, CNNs and GANs can be seen as two sides of the same coin. CNNs perform image analysis, while GANs carry out image synthesis. In other words, CNNs are discriminative models, while GANs are generative models (Foster, 2019, ch. 1).

Let us consider CNNs first. They accept an image as input, and they determine whether the image belongs to a specific class of images. The relevant classes are defined by a human before training the network, and they may be classes of objects (dog, cat, person, tree), or classes of scenes (kitchen, living room, church, park, bedroom). The human supervisor must also provide a set of training images and the class labels for them, i.e., for each training image, a human must indicate whether it is a kitchen, a park, or a bedroom. After the CNN is trained with the labeled images, it can classify other test images which differ from those in the training set. Convolutional layers in the CNN are responsible for extracting significant features from the information coming from the previous layer and passing them to the next layer. As the information flows from the input to the output, the computed features are higher level (Ketkar, 2017, p. 78), which matches the hierarchical structure of visual stimuli (Chollet, 2018, p. 123). At the input layer, the raw, low level pixel data are provided, while at the output layer the final decision is produced about the class that the whole image belongs. Often the features computed at the earlier layers are more general, while those extracted at the later layers are more specific to the classification problem to be solved (Chollet, 2018, p. 155). The computed features at each neural layer are learned automatically by the network, i.e., no human intervention is required to define the features that must be extracted from the data. This overcomes the limitations of other machine learning approaches, where the features must be handcrafted by a human expert. Figure 1 depicts a scheme of a CNN.

In contrast to this, the GANs accept a vector of random real numbers as input, and they output a synthetic but realistic image of a specific class. The class of images that a GAN generates depends on a set of training images that must be provided by a human. Typically, a set of training images of the same style is provided, for example, a collection of images of kitchens, if the GAN is aimed to generate synthetic images of kitchens. In this case, no class labels are provided to the network (Langr and Bok, 2019, section 1.4). For each possible input vector, the GAN outputs a different synthetic image, different from all those in the set of training images, but of the same style. The internal structure of a GAN contains two subnetworks. One of them is called the generator, which is responsible for generating a synthetic image from the random input vector, hence the adjective 'generative' of the GAN. The second subnetwork is the discriminator, which determines whether an image is real or synthetic. According to an art forgery

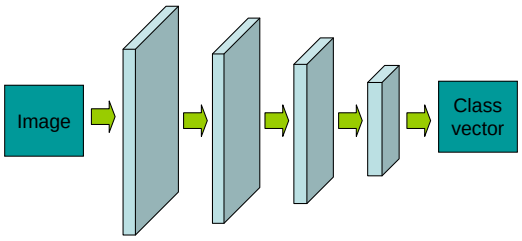

Fig. 1 Scheme of a CNN. An image is given as input to the network, which yields as output a class vector with the likelihoods that the image belongs to each of the predefined object classes. The succesive neural layers (depicted as rectangular boxes) are smaller, i.e. they contain fewer units. Information always flows from the input to the output (shown with arrows).

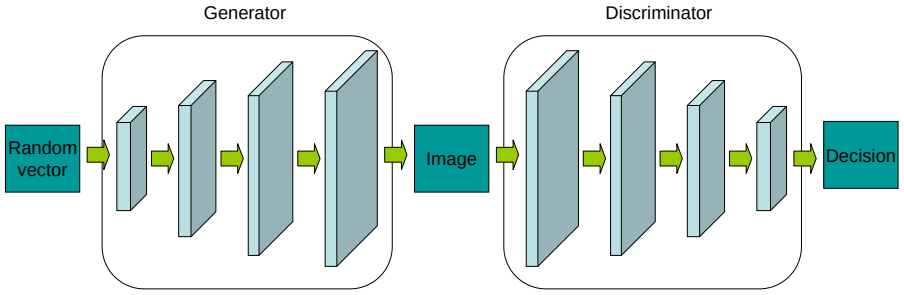

Fig. 2 Scheme of a GAN. A random vector is given as input to the generator, which yields an image as output. Then the image is supplied to the discriminator, which decides whether the image is synthetic or real. The succesive neural layers of the generator (depicted as rectangular boxes) are bigger, i.e. they contain more units. The successive neural layers of the discriminator (also depicted as rectangular boxes) are smaller, i.e. they contain fewer units. Information always flows from the input to the output (shown with arrows).

metaphor, the generator is the forger, and the discriminator is the expert (Gulli and Pal, 2017, ch. 4). This is the reason for the adjective 'adversarial' of the GAN. Both networks are trained alternatively (Foster, 2019, ch. 4). First, the discriminator is adjusted to minimize the classification errors. Then the generator is adapted to maximize the classification errors. A GAN aims to learn the mathematical manifold of the images that have a particular style, within the space of all possible images. Once this manifold is learned, the GAN allows the sampling of realistic images from that manifold (Chollet, 2018, p. 270). A scheme of a GAN is shown in Figure 2.

Both CNNs and GANs stand as two of the best known deep learning neural networks. The CNNs are established as the workhorse of present day computer vision since they attained superhuman performance on image recognition tasks (He et al, 2015, p. 1026; Foster, 2019, ch. 1). The GANs have also received a great deal of attention in recent times due to the ever increasing realism of the images

that they generate[2], which is exploited in creative and artistic endeavors (Foster, 2019, Preface).

Now that the two artificial neural network models which are considered in this work are described, it is time to investigate in the next section the reports of the computational experiments that have been carried out with them.

5 The experiments

During the last years, several experimental reports have been published on deep neural networks about the existence of units that capture human understandable visual patterns. Low level features like colors, edges, and middle level features like textures have been routinely found for some years (Chollet, 2018, p. 172; Qin et al, 2018, p. 163), although this is not particularly interesting because it does not entail the emergence of any high level (complex) category, i.e., only low level categories were found. It could be argued that the results of these early experiments do not imply that the networks are learning any complex categories since the detected visual patterns are relatively simple. Nevertheless, more recent reports indicate that units can arise within a deep network that are associated with highly elaborate kinds of objects.

These units within deep networks have been named interpretable units (Bau et al, 2017, p. 4; Zhang and Zhu, 2018, p. 35), i.e., artificial neurons which have learned to represent a specific human understandable complex visual category. The relevance of these computational experiments is that those interpretable units arise in networks whose tasks do not explicitly require learning those visual patterns. That is, the network learns intermediate visual patterns that are relevant to detect (CNN) or generate (GAN) a more complex kind of object, but humans do not instruct the networks on which patterns are relevant or how they should be detected or generated. For example, let us consider a CNN that is trained to classify images of places into types such as cafeteria, promenade, cemetery, or conference room. It is not interesting that the trained CNN contains units specialized in detecting cafeterias, since the "cafeteria" label has been provided to the CNN by a human supervisor. The critical point is that interpretable units arise in the trained CNN which are associated to intermediate visual categories such as "bed," "car," or "tree" which have not been provided by humans to the CNN in any way, as reported in (Bau et al, 2017, p. 6). In other words, there are no human provided labels about beds, cars, or trees, nor there is any human provided hint regarding such intermediate visual categories that are relevant to classify images into types of places.

The technique which is employed to find out interpretable units in deep networks is called network dissection (Qin et al, 2018, p. 170). First of all, the network is trained to accomplish its goal, detection (CNN) or generation (GAN) of images according to a given training set. The network dissection procedure is carried out on the network in its final trained state. It is necessary to have some test images along with their ground truth segmentation masks, which are annotations made by humans where the objects which appear in each image are marked with their

[2] GANs have been called "the coolest idea in deep learning in the last 20 years" (Metz, 2017) by Yann LeCun, recipient of the Turing award, the computer science equivalent to the Nobel prize.

class labels. For example, for an image of a park, its ground truth segmentation mask indicates which regions of the image correspond to a tree, the grass, a person, a dog, and so on. Each unit of the network has a varying activation strength depending on the regions of each test image. Then for each unit of the network, it is determined whether its high activation strength regions are correlated to the regions associated with a single class of object in the ground truth segmentation masks. If this is the case and the correlation is strong enough, then the unit is likely to represent that class of object, so that a candidate interpretable unit has been found. It must be highlighted that the semantic relevance of the human made region annotations and the correlation threshold required to associate regions affect the identification of interpretable units.

To confirm the results of the candidate interpretable unit detection process, a subsequent validation procedure is conducted (Zhou et al, 2014, p. 6; Bau et al, 2017, p. 4). The validation starts by automatically determining the receptive field of a candidate interpretable unit for the test images, i.e., the region of the image that most strongly activates the unit. Then those receptive fields, which are regions of the test images, are provided to several human workers who do not interact with each other. Furthermore, these workers are distinct from those who created the ground truth segmentation masks for the previous network dissection procedure. Each worker is asked which category is best represented by the set of image regions, without providing a dictionary of possible categories to avoid biases. The workers are also asked to identify which image regions do not associate with the category. Then the precision of the unit is computed as the fraction of image regions that do associate with the category. Finally, the candidate interpretable units whose precision is higher than a given threshold, and such that the category identified by the workers agrees with the category of the ground truth segmentation, are declared as interpretable. It has been reported that for some networks, most of their units are interpretable in this sense (Bau et al, 2017, p. 4). It is worth noting that not all human workers agree on the category that the image regions represent. Therefore, the distinction between interpretable and not interpretable units is not clear cut, since it depends on the visual evaluation of the image regions by the human workers, and the degree of intersubjective agreement that is required to validate the interpretability of a unit. Moreover, it must be highlighted that for a given network and a category, several interpretable units capture that particular category, i.e., there is always redundancy in the learning of the categories. This is typical of artificial neural networks, whose robustness and resilience to errors come from such redundancy.

Reports include interpretable units in classifier networks that detect objects that are parts of the classes to be recognized. For example, wheel detecting units to classify bicycles and face detecting units to classify cats (Gonzalez-Garcia et al, 2018, p. 488). Networks trained to classify scenes contain interpretable units that detect legs or lamps (Zhou et al, 2014, p. 7). Interpretable units specialized in the generation of people, or train tracks appear in networks trained to generate videos (Vondrick et al, 2016, p. 8). The kinds of objects that these units specialize are exceedingly complex to be learned by any simple combination of low level geometric features. In other words, these units contain high level semantic information.

The procedure to determine which units are interpretable must be carefully analyzed since it must be ascertained whether it can lead to some self deception. The researchers might wrongly think that an understandable category emerges in

the network without human help. This may be due to contamination of human supplied information about the purportedly emergent categories, or an erroneous interpretation of what the network does. On the one hand, the network training process does not involve any information transfer from humans to the networks about the objects which might appear in the images or their classes. This implies that the training process does not seem to have any danger of contamination. On the other hand, the subsequent interpretation of the final state of the trained network depends on human categories expressed in natural language by the human evaluators. It is not possible that the degree of interpretability of the network surpasses the degree of agreement among the human evaluators about the linguistic labels assigned to the objects. Given the high proportion of interpretable units found, the strong agreement among the ground truth segmentations and the subsequent validations, and the lack of communication among the human evaluators, it can be said with a high degree of confidence that the interpretations are plausible according to many human evaluators. That is, the networks have learned understandable complex visual categories without any cues from humans.

Nevertheless, another confusion is still possible, namely that no subset of units is responsible for learning a specific category. This might happen if all units work together to learn the same category so that no set of units can be identified as the only one responsible for the acquisition of the category. In other words, there are causal connections among the units which collectively explain the observed acquisition of the category, and these causal connections might span the entire network. If this were the case, the network would still have found a category, but such knowledge would be spread all over the network. This kind of causality analysis requires intervention on the subject of study (Pearl and Mackenzie, 2018, ch. 1), in this case, the trained network. These interventions have already been carried out in GANs, whose results support the claim that identifiable subsets of units learn specific categories (Bau et al, 2018, p. 8). It has been found that typically a small set of units, such as 20 units out of 512 in a layer, are responsible for a given category like "curtain" in a GAN trained to generate conference room images (Bau et al, 2018, p. 9). If the units of this set are ablated (deactivated) one by one, the number of instances of the learned category diminishes progressively. That is, the more curtain representing units are deactivated, the fewer curtains which appear in the generated synthetic images, while the rest of the objects of the scene occur with the same frequencies. Unit ablation can also be employed to enhance the appearance of the generated scenes in case that artifacts arise because such artifacts are often created by defective units (Bau et al, 2018, section 4.2). Conversely, inserting extra units that are associated with a given category increases the proportion of instances of that category which appear in the output images, leaving the ratios of the frequencies of the other categories equal. It must also be highlighted that the network often refuses to remove all instances of objects which are essential for the kind of scene at hand, such as chairs in a conference room. It also usually refuses to insert objects out of place, such as doors on trees (Bau et al, 2018, p. 9). Moreover, the removal and insertion of objects respect the visual coherence of their surroundings (Bau et al, 2019, p. 8). This implies that the semantic knowledge about the relative positions of the instances of the categories is also acquired by the GANs. It must be pointed out that the interventions are not always successful. That is, sometimes, the insertion of units associated with a given category does not result in the appearance of any object. Nevertheless, such

insertions never result in the appearance of an object of a different category, i.e., you never obtain a tree by inserting units associated with domes.

The experiments suggest that the number of layers (depth) and the number of units per layer (width) of the networks influence the number of emergent categories. Counting the number of emergent categories is done by establishing a reference image dataset containing human annotated image regions, where each region is labeled with a normalized category name. Each category name is an English word (Bau et al, 2017, p. 2). It seems that deeper networks are more amenable to finding high level categories (Bau et al, 2017, p. 5). Also, adding units in the layers where high level categories arise often increases the number of learned categories, up to a certain point where no extra categories appear (Bau et al, 2017, p. 8). The overall picture which can be extracted from these results is that the emergence of high level categories is easier in neural architectures whose depth and width are within some ranges.

From the above, it can be seen that the experimental evidence on the ability of CNNs and GANs to learn high level visual categories is overwhelming. In the next section, it is argued that visual categories emerge in these networks, in a sense defined in Section 3.

6 The emergence of natural visual categories

Next, it is argued that the learned natural visual categories are emergent (Section 3). CNNs and GANs are composed of thousands of relatively simple units. The cooperation among the units enables the learning of progressively higher (CNNs) or lower (GANs) level categories in the structure. It is worth noting that units of separate layers have the same structure, while they learn categories of different levels depending on their location within the neural architecture. That is, there is nothing specific in the units which directs them towards learning one category or another. Hence, it is the interaction among units that enables the emergence of the natural visual categories in the neural architecture. Furthermore, structural unpredictability is also found here. The CNNs are trained to recognize several kinds of scenes so that the goal which is set up for their training is maximizing the classification accuracy for the classes of scenes defined by the human supervisor. Therefore, the training goal is not explicitly related to learning any intermediate natural visual categories (door, tree, sky), since the goal only refers to the scene level classification (park, bedroom, kitchen). Similarly, the goal of the GANs is to generate scenes of a specified kind that are maximally realistic. Again this training goal does not imply the generation of intermediate categories whose combination yields the overall scene. Hence the human supervisor does not supply any hints about the visual structure of the scenes.

In other words, for the CNNs and the GANs the training goals do not refer to any objects, but to the overall image of a scene, which implies that the acquisition of natural visual categories is an unexpected effect of the application of the training algorithms to maximize such goals. The training algorithms do not contain any reference to objects, either. Therefore, the emergence of the categories can not be derived from first principles, as required by structural unpredictability. The emergent categories have a level of specificity (Gärdenfors, 2014, section 2.3) that is higher than geometric primitives such as edges, while they are lower level than

the category that comprises the overall kind of scene which is modeled by the network (kitchen, church, bedroom).

7 Discussion

In this section, the implications of the emergence of natural visual categories in artificial deep networks for cognitive science and artificial intelligence are investigated. Four conclusions are extracted from this investigation, which summarize the novelties which derive from the above described experimental findings. As mentioned in Section 1, these visual categories are not understood as full fledged concepts that are consistent according to universal, objective criteria of correctness. Our visual categories are associated with fuzzy sets of images that vary from one subject to another.

The multiplicity of interpretable units that learn the same high level category can be related to perspectivism in conceptual spaces, which emphasizes the importance of alternative views on the same conceptual space (Kaipainen and Hautamäki, 2015, p. 249). For CNNs and GANs this occurs because each neural layer of each neural network defines a different conceptual space to represent the same kind of sensory data, i.e., the images. The dimensionality of the space varies for each neural layer since the dimension is the number of neurons of that layer. Under the perspectivist approach, category construction starts from an onto-space formed by dimensions that are associated with observable features. In our case, the onto-space would be the space of all possible images, so that the dimensions of the onto-space are the raw pixel levels of the image. Then a perspective to the onto-space could be defined as a transformation to a representation space where object recognition is more manageable (Kaipainen and Hautamäki, 2015, p. 251). For CNNs and GANs, this transformation is given by the composition of the transformations carried out by the neural layers of the artificial deep network.

The interpretation that is advocated in this work about deep learning is compatible with but does not assume, a connectionist type of computational functionalism of the kind described in (Buckner and Garson, 2019, p. 79; Piccinini and Bahar, 2013, p. 479; Piccinini and Scarantino, 2011, p. 12), which states that artificial deep neural networks are a good model of the human brain. The two possibilities are:

- If artificial deep networks are good models of the human brain, i.e., connectionist computational functionalism is true, then the emergence of natural visual categories in artificial networks might be caused by the cognitive abilities of the neural structure shared by the human brain and the artificial deep networks. This does not preclude those visual categories from arising in other artificial cognition systems. That is, the possibility that visual categories are not exclusive of brain like structures can not be ruled out.
- If artificial deep networks are not good models of the human brain, i.e., connectionist computational functionalism is false, then the emergence of natural visual categories is not specific to human like cognitive structures. Hence artificial deep networks and the human brain capture the same high level visual categories independently, i.e., they employ different cognitive processes to find such categories. Again, other artificial or biological entities with entirely different cognitive structures might also be capable of learning the same categories.

Empiricism is supported by the ability of CNNs to recognize objects given human provided class labels (Buckner, 2018, p. 5341) since such ability shows that high level knowledge can be abstracted from raw visual data. This ability depends on the supervision by humans, because the CNN is provided a training dataset where each image is accompanied by a class label given by a human supervisor.

The recently observed emergence of natural visual categories without human supervision, which is the topic of our work, goes well beyond in its philosophical consequences. This time the human supervisor does not provide hints about the intermediate categories that should be learned towards the recognition of the human supplied class of scenes. For example, let us consider a CNN to classify rooms. The human supervisor provides the class labels associated with each full image of a room, such as "kitchen," "bedroom," or "sitting room." But the human supervisor does not supply the network with any hints of the objects that appear in the images of the rooms nor their categories, such as "chair," "lamp," "bed," or "table." These intermediate categories are found by the networks as learning progresses. While it must be acknowledged that no machine learning model can work without any prior knowledge, i.e. pure empiricism is out of the question, the debate should focus on how much knowledge is provided to the network and to which extent the network abstracts the knowledge from data (Buckner, 2019, pp. 11-12). Moreover, the absence of linguistic processing in the operation of CNNs and GANs downplays the importance that rationalism gives to linguistic reasoning, which has been associated with the classic approach to cognitive science (Section 3).

We summarize our findings of the operation of CNNs and GANs described in Section 5 in the following claims:

Claim (1) Many natural visual categories are universal and not arbitrary, i.e., they can be learned without any human provided information about them.

The main justification of this claim is that no human supervision is employed to supply the CNNs or the GANs with any information about the categories of objects that appear in the images of the scenes, although the human supervisor does provide the labels for the categories of scenes. Moreover, different units of different networks learn the same category, with small variations, without any information exchange among the networks.

It must be pointed out that the networks are supplied some knowledge in the form of the human provided class labels, the convolution operation (Silver et al, 2017, p. 354; Marcus, 2018, pp. 7-9), the learning algorithm, the activation function, and the architecture of the network which comprises the number of layers and the number of units of each layer. Next, I analyze which prior knowledge is provided to the network by these means.

The convolution operation provides translation invariance (Marcus, 2018, p. 7; Chollet, 2018, p. 321; Gulli and Pal, 2017, ch. 3; Buckner, 2018, p. 5354). For the experiments described in Section 5, this amounts to saying that the network is told that the same object can appear at different positions in the image. This is a general constraint that applies to all kinds of objects, so there is no information concerning any particular visual category here. The learning algorithm tells the network how to adjust its synaptic weights so that the error measure is minimized. The learning algorithms employed in deep learning are variations of gradient descent on the error surface (Langr and Bok, 2019, subsection 2.7; Chollet, 2018, p.

50). These learning algorithms do not provide any insights about specific visual categories. The activation functions employed in deep learning, such as the rectifier linear function, are aimed to facilitate the operation of the gradient descent algorithm by producing stable gradients (Ketkar, 2017, p. 30). Again, this does not supply any information concerning individual visual categories. The architecture of the network has been found to influence the number of emergent categories (Section 5). Still, there is no evidence that changing the number of layers or the number of units in each layer is correlated with finding a particular visual category.

As seen above, the network does not start in a completely blank state. The general notion of an object is inserted into the network architecture by a human. This implies that pure empiricism is not supported by Claim 1. Still, the human knowledge which is injected into the network is not related to the intermediate visual categories which emerge from the network, i.e., those emergent object categories which do not correspond with any human supplied class labels in the training set of images.

Finally, it must be highlighted that the exact same network, comprising the same convolution operators, activation functions, neural architecture, and learning algorithm, finds completely different emergent visual categories depending on the kind of scenes that it is trained on (Bau et al, 2017, p. 6; Bau et al, 2019, p. 5). That is, the same network finds the table category if it is trained with images of kitchens, the bed category if it is trained with bedrooms, and the tree category if it is trained with parks. This strongly suggests that the knowledge that humans inject into the network does not contain any information about specific visual categories.

Claim (2) If connectionist computational functionalism is true, then the innate structure of the human brain spontaneously captures natural visual categories, independently from human knowledge and culture.

Artificial deep networks might be good models of the human brain, i.e., connectionist computational functionalism might be true. If that is the case, so that the processes that underlie visual category formation in CNNs and GANs are similar to the ones in human brains, then the emergence of natural visual categories in these networks would imply, by Claim 1, that such categories are captured by the mind irrespective of human knowledge or culture. It is not completely clear to which extent artificial neural networks are similar to the human brain. The main difficulties in drawing analogies among artificial and biological neural networks are the low biological plausibility of the learning algorithms employed to train artificial neural networks, particularly the backpropagation of the error mechanism (Buckner, 2018, p. 5364; Butz and Kutter, 2017, p. 60). Also, artificial neural networks have much more regular architecture than biological ones. Therefore, it is preferable to leave Claim 2 conditioned on the truth of connectionist computational functionalism.

Claim (3) Some natural visual categories are context dependent, i.e., it is not only the image of the object itself that makes it a part of the category but also the objects that surround it.

This claim is supported by the fact that for the GANs the insertion or deletion of some objects affect their surroundings in a visually coherent way. For example, when a window is inserted or deleted from a kitchen, the wall and the furniture are

often adequately modified to accommodate for the change. That is, the surrounding objects are modified in a way that humans find consistent with their visual constraints on how a kitchen must look. Moreover, the addition of out of context objects, such as a door in the sky, is rejected by the GANs. These experiments suggest that contextual information is learned and integrated into the conceptual spaces.

Claim (4) Natural visual categories are independent of the syntactic structure of language.

This follows from the fact that CNN and GAN deep networks do not perform any grammatical processing. No sequential computation, internal memory, symbols, or tree structures are employed whatsoever (Section 4). That is, artificial deep networks of the CNN and GAN types acquire non syntactic knowledge[3]. This evidence is against the classic approach to cognitive science and artificial intelligence, which states that information processing is done through symbol manipulation according to syntactic rules.

The four claims that have been stated above can be related to one of the most heated debates in cognitive science about the nature of concepts, namely the amodality of the representation of concepts in the human brain (Kiefer and Pulvermüller, 2012, p. 806). This debate deals with the possible concept representation and organization schemes that are employed by the mind (Mahon and Caramazza, 2009, p. 28). The grounded representation approach, also known as neo-empiricism, which adheres to the embodied cognition framework, defends that conceptual content is sensory and motor (Mahon and Caramazza, 2009, p. 41; Kiefer and Pulvermüller, 2012, p. 820). The opposing amodal approach postulates the existence of a concept representation system that is independent of the sensory system so that sensorial information is removed before concepts are stored and managed by linguistic processes (Machery, 2007, p. 21; Kiefer and Pulvermüller, 2012, p. 806). An intermediate stance is that perceptual representations are required for some mental tasks, while amodal ones are more suitable for others (Machery, 2007, p. 36).

The possible contribution of the above described experimental results to this debate depends on the validity of a connectionist form of computational functionalism. If artificial deep networks are good models of the human brain, then the emergence of visual categories in artificial deep networks implies that there is a grounded, sensory representation of their associated concepts in the human brain. It must be noted here that the category representation of CNNs and GANs discussed in this work is perceptual but not motor, since the experiments do not involve any actions to be executed by the agent. But if artificial deep networks are not good models of the brain, then the experiments show that there are ways to learn grounded representations of high level visual concepts, even if those ways do not exist in the brain.

If we assume connectionist computational functionalism, the above claims would only rule out a purely amodal representation of visual concepts as defined in Section 2. A grounded representation would exist, but it could be linked to a

[3] There are artificial deep neural networks that do manage sequential linguistic information, such as Deep Recurrent Neural Networks (DRNNs), including Long Short-Term Memory Units (LSTMs), but they are not considered in this work.

more abstract, amodal representation of the same concepts (Dove, 2009, p. 413). The ability of GANs to generate realistic examples of a learned category can be seen as an artificial networks model of neural reenactment of past experiences by biological networks, also called simulation, which constitutes grounded concepts (Dove, 2009, p. 415; Machery, 2007, pp. 21-22). As required by the grounded representation approach (Machery, 2007, p. 23), the simulated examples of each object class produced by a GAN are not identical to any of the perceived examples in the training set of the GAN. In other words, the acquisition of an embodied concept requires that the representation is flexible enough to accommodate previously unseen instances of the object class, and the GANs possess such flexibility (Kiefer and Pulvermüller, 2012, p. 809). Moreover, the contextual knowledge which is acquired by CNNs and GANs when they learn visual categories provides a plausible mechanism for the context sensitivity characteristic of the human conceptual system (Machery, 2007, p. 30). Visual concepts such as **chair**, **tree**, or **table**, are likely to have a grounded representation. More abstract concepts like **numbers** (**one**, **two**, **seven**) might have none (Dove, 2009, p. 418; Machery, 2007, p. 37), although even this is disputable due to the existence of specific responses to numbers in the human visual system (Shum et al, 2013, p. 6709). In other words, concepts differ in terms of their suitability to be captured by images of exemplars of them (Dove, 2009, p. 426). This points to some kind of hybrid approach (Dove, 2016, p. 1112; Wajnerman Paz, 2018, p. 5249), where CNNs and GANs might be example models of grounded representations.

From a computational functionalist perspective, extreme forms of the amodal theory of concepts, which neglect the existence of grounded concepts, are in conflict with the emergence of visual categories in artificial deep networks. The offloading hypothesis (Machery, 2016, p. 1094) states that concepts are amodal, so that perceptual information is offloaded onto sensorial systems in the brain whenever a task requires it. This contrasts with the emergence of high level categories within purely visual networks such as CNNs and GANs. Other versions of the amodal theory acknowledge that sensorial and motor representations interact with amodal representations (Leshinskaya and Caramazza, 2016, p. 995), which makes them easier to accommodate with the emergence of visual categories in artificial networks. The category representation mechanism of CNNs and GANs can be regarded as a fully detailed model for grounded concepts in the brain, which contrasts with the absence of a precise specification of the representation of amodal concepts (Barsalou, 2016, p. 1127).

8 Conclusion

The success of deep learning in solving many decades-long fundamental struggles in artificial intelligence has sparked a revolution that has affected related disciplines such as cognitive science. It is time to analyze what deep learning has to tell us about the physical world and our minds, ignoring the hype which accompanies these extremely popular victories. It is tempting to say that the prophecies of connectionism have been fulfilled. However, this depends on the validity of artificial neural networks as models of the human brain, i.e., whether some form of computational functionalism is true. There are relevant differences between artificial and biological deep networks, so this line of argumentation is rather risky.

In this work, we have chosen to leave the computational functionalism question aside, and focus on the consequences that can be drawn from the most recent experiments irrespective of its answer.

The emergence of natural visual categories in artificial deep networks is relevant in itself because it teaches us lessons about the nature of visual concepts even if the human brain captures them by other processes. In particular, it tells us that the acquisition of such concepts is not necessarily related to previous human knowledge about those concepts and that they can be acquired without grammatical processing abilities. A mild form of empiricism is supported by these conclusions since some visual categories are found by the networks without human hints, but there are some general visual learning mechanisms which must be provided to the artificial network by a human to enable the network to find high level categories by itself from the raw images. Moreover, the way that the emergent categories interact with each other suggests that visual context is inextricably related to the category itself.

Our claim that natural visual categories are independent of the syntactic structure of language is compatible with the language being the vehicle of abstract mental processes. The experiments which have been discussed here show that object categorization is not necessarily attached to structured linguistic definitions, although it seems obvious that such categorizations are later transformed into structured linguistic knowledge in the human brain. CNNs and GANs might be models of grounded representations of visual categories amenable for sensorial and motor tasks, which could coexist with amodal representations oriented to structured linguistic processing and abstract reasoning.

One of the most important advantages of the artificial deep networks that have been discussed in this work is that they allow experimentation with high level visual categories alone. That is, experimentation on human subjects necessarily implies the concurrent operation of many other mental processes that operate in the brain. In contrast to this, CNNs and GANs are entirely focused on managing visual information, so that other cognitive processes do not complicate the investigation about high level visual categories. This is the first time in the history of cognitive science that such a powerful tool is available to researchers, so it can be expected that surprising revelations will arrive in the near future. New light has been thrown on black boxes that illuminates the path ahead of us.

Acknowledgments

The author is deeply indebted to Dr. Rosa María Ruiz-Domínguez (Universidad de Málaga, Málaga, Spain), for her insightful comments about the connections of this work with psychology and cognitive science. He is also grateful to David Teira (Universidad Nacional de Educación a Distancia, Madrid, Spain) and Emanuele Ratti (University of Notre Dame, Notre Dame, Indiana, USA) for their valuable comments. Finally, he would like to thank the anonymous reviewers for their constructive suggestions, which have greatly improved the original manuscript.

References

Barsalou L (2016) On staying grounded and avoiding Quixotic dead ends. Psychonomic Bulletin and Review 23(4):1122–1142

Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: Quantifying interpretability of deep visual representations. CoRR abs/1704.05796, URL http://arxiv.org/abs/1704.05796, 1704.05796

Bau D, Zhu JY, Strobelt H, Zhou B, Tenenbaum JB, Freeman WT, Torralba A (2018) GAN dissection: Visualizing and understanding generative adversarial networks. CoRR abs/1811.10597, URL https://arxiv.org/abs/1811.10597, 1811.10597

Bau D, Strobelt H, Peebles W, Wulff J, Zhou B, Zhu J, Torralba A (2019) Semantic photo manipulation with a generative image prior. ACM Transactions on Graphics 38(4):59

Bechtel W (1993) The case for connectionism. Philosophical Studies 71(2):119–154

Benitez JM, Castro JL, Requena I (1997) Are artificial neural networks black boxes? IEEE Transactions on Neural Networks 8(5):1156–1164

Berkeley ISN (2019) The curious case of connectionism. Open Philosophy 2(1):190–205

Brainerd CJ, Reyna VF (2002) Fuzzy-trace theory and false memory. Current Directions in Psychological Science 11(5):164–169

Buckner C (2015) A property cluster theory of cognition. Philosophical Psychology 28(3):307–336

Buckner C (2018) Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese 195(12):5339–5372

Buckner C (2019) Deep learning: A philosophical introduction. Philosophy Compass DOI 10.1111/phc3.12625, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12625

Buckner C, Garson J (2019) Connectionism and post-connectionist models. In: Sprevak M, Colombo M (eds) The Routledge Handbook of the Computational Mind, Routledge, New York, pp 76–90

Butz MV, Kutter EF (2017) How the Mind Comes Into Being. Oxford University Press, Oxford, United Kingdom

Cantwell Smith B (2019) The promise of artificial intelligence. The MIT Press, Cambridge, Massachusetts

Carabantes M (2019) Black-box artificial intelligence: an epistemological and critical analysis. AI & Society URL https://doi.org/10.1007/s00146-019-00888-w

Chin-Parker S, Ross BH (2002) The effect of category learning on sensitivity to within-category correlations. Memory & Cognition 30(3):353–362

Chollet F (2018) Deep learning with Python. Manning, Shelter Island, NY

Clark A (1993) Associative Engines: Connectionism, Concepts, and Representational Change. The MIT Press, Cambridge, Massachusetts

Creel KA (2020) Transparency in complex computational systems. Philosophy of Science URL http://philsci-archive.pitt.edu/id/eprint/16669

Dawson MRW (2004) Minds and Machines: Connectionism and Psychological Modeling. Blackwell Publishing, Malden, Massachusetts

Dove G (2009) Beyond perceptual symbols: A call for representational pluralism. Cognition 110(3):412–431

Dove G (2016) Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. Psychonomic Bulletin and Review 23(4):1109–1121

Elber-Dorozko L, Shagrir O (2019) Computation and levels in the cognitive and neural sciences. In: Sprevak M, Colombo M (eds) The Routledge Handbook of the Computational Mind, Routledge, New York, pp 205–222

Foster D (2019) Generative Deep Learning. O'Reilly, Sebastopol, CA

Gärdenfors P (2000) Conceptual spaces: The geometry of thought. The MIT Press, Cambridge, Massachusetts

Gärdenfors P (2014) The geometry of meaning: Semantics based on conceptual spaces. The MIT Press, Cambridge, Massachusetts

Gomes L (2014) Machine-learning maestro Michael Jordan on the delusions of Big Data and other huge engineering efforts. IEEE Spectrum 20 Oct 2014

Gonzalez-Garcia A, Modolo D, Ferrari V (2018) Do semantic parts emerge in convolutional neural networks? International Journal of Computer Vision 126(5):476–494

Gulli A, Pal S (2017) Deep Learning with Keras. Packt Publishing Ltd., Birmingham, UK

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, pp 1026–1034

Horgan T, Tienson J (1996) Connectionism and the Philosophy of Psychology. The MIT Press, Cambridge, Massachusetts

Kaipainen M, Hautamäki A (2015) A perspectivist approach to conceptual spaces. In: Zenker F, Gärdenfors P (eds) Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation, Springer International Publishing, Cham, pp 245–258

Ketkar N (2017) Deep Learning with Python. Apress, New York

Kiefer M, Pulvermüller F (2012) Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. Cortex 48(7):805–825

Langr J, Bok V (2019) GANs in Action. Manning, Shelter Island, NY

Leshinskaya A, Caramazza A (2016) For a cognitive neuroscience of concepts: Moving beyond the grounding issue. Psychonomic Bulletin & Review 23(4):991–1001

Machery E (2005) Concepts are not a natural kind. Philosophy of Science 72(3):444–467

Machery E (2007) Concept empiricism: A methodological critique. Cognition 104(1):19–46

Machery E (2016) The amodal brain and the offloading hypothesis. Psychonomic Bulletin and Review 23(4):1090–1095

Mahon B, Caramazza A (2009) Concepts and categories: A cognitive neuropsychological perspective. Annual Review of Psychology 60:27–51

Marcus G (2018) Innateness, AlphaZero, and artificial intelligence. CoRR abs/1801.05667, URL `http://arxiv.org/abs/1801.05667`, `1801.05667`

Mayor J, Gomez P, Chang F, Lupyan G (2014) Connectionism coming of age: legacy and future challenges. Frontiers in Psychology 5:187

Metz C (2017) Google's dueling neural networks spar to get smarter, no humans required. https://www.wired.com/2017/04/googles-dueling-neural-networks-spar-get-smarter-no-humans-required/

Neudeck P, Wittchen HU (2012) Introduction: Rethinking the model - refining the method. In: Neudeck P, Wittchen HU (eds) Exposure Therapy, Springer, New York, NY, pp 1–8

Newell A, Simon HA (1976) Computer science as empirical inquiry: Symbols and search. Communications of the ACM 19(3):113–126

Pearl J, Mackenzie D (2018) The book of why: the new science of cause and effect. Basic Books, New York

Piccinini G, Bahar S (2013) Neural computation and the computational theory of cognition. Cognitive Science 37(3):453–488

Piccinini G, Scarantino A (2011) Information processing, computation, and cognition. Journal of Biological Physics 37(1):1–38

Qin Z, Yu F, Liu C, Chen X (2018) How convolutional neural networks see the world — a survey of convolutional neural network visualization methods. Mathematical Foundations of Computing 1(2):149–180

Ross B (2001) Natural concepts, psychology of. In: Smelser NJ, Baltes PB (eds) International Encyclopedia of the Social & Behavioral Sciences, Pergamon, Oxford, pp 10,380–10,384, DOI https://doi.org/10.1016/B0-08-043076-7/01489-3

Salisbury J, Schneider S (2019) Concepts, symbols and computation. In: Sprevak M, Colombo M (eds) The Routledge Handbook of the Computational Mind, Routledge, New York, pp 310–322

Shum J, Hermes D, Foster BL, Dastjerdi M, Rangarajan V, Winawer J, Miller KJ, Parvizi J (2013) A brain area for visual numerals. Journal of Neuroscience 33(16):6709–6715

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, et al (2017) Mastering the game of Go without human knowledge. Nature 550:354–359

Stephan A (2006) The dual role of 'emergence' in the philosophy of mind and in cognitive science. Synthese 151(3):485

Tu J (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology 49(11):1225–1231

Vondrick C, Pirsiavash H, Torralba A (2016) Generating videos with scene dynamics. URL `https://arxiv.org/abs/1609.02612`, `1609.02612`

Wajnerman Paz A (2018) An efficient coding approach to the debate on grounded cognition. Synthese 195(12):5245–5269

Zhang QS, Zhu SC (2018) Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering 19(1):27–39

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2014) Object detectors emerge in deep scene CNNs. URL `https://arxiv.org/abs/1412.6856`, `1412.6856`