scientific reports

OPEN

Aesthetics and neural network image representations

Romuald A. Janik

We analyze the spaces of images encoded by generative neural networks of the BigGAN architecture. We find that generic multiplicative perturbations of neural network parameters away from the photo-realistic point often lead to networks generating images which appear as *"artistic renditions"* of the corresponding objects. This demonstrates an emergence of aesthetic properties directly from the structure of the photo-realistic visual environment as encoded in its neural network parametrization. Moreover, modifying a deep semantic part of the neural network leads to the appearance of symbolic visual representations. None of the considered networks had any access to images of human-made art.

Among the many strands of contemporary Machine Learning, a prominent place is taken by generative neural networks[1–4]. These neural networks aim to generate new, unseen examples based on a given dataset and thus aim to learn the variability structure of the data. Of particular interest for the present investigation are neural networks which generate photo-realistic images of the natural and human environment. In order to do so, they have to incorporate extensive knowledge about the structure of the visual photo-realistic world. This information is encoded in a nontrivial way in the weights of the neural network layers. The goal of this work is to explore global properties of this neural encoding. To the best of our knowledge, such properties have not been investigated so far.

In the present paper, we show two surprising features of these neural encodings. First, moving away from the photo-realistic world in a generic (multiplicative) manner leads in many cases to the emergence of *"artistic rendition"* and aesthetic properties as perceived by humans. Second, upsetting a part of deep semantic information leads to the appearance of imagery which can be interpreted as *symbolic visual representations*. These results may have far reaching interdisciplinary consequences, touching upon our understanding of the neural basis of aesthetics (neuroaesthetics)[5–10], the theory and philosophy of art and, given the similarities between deep convolutional networks and the visual cortex[11–13], these results may inspire novel investigations within cognitive neuroscience.

The two main classes of generative neural networks are Generative Adversarial Networks (GAN)[1] which appear in a multitude of variants[2], as well as Generative Autoencoders e.g. Variational Autoencoders (VAE)[3], Wasserstein Autoencoders (WAE)[4] and many others. At the end of training these constructions provide for us a *generator network* $\mathcal{G}_\theta$ with a given set of "optimal" weights (i.e. neural network parameters) $\theta = \theta_*$. The resulting network generates an image X

$$X = \mathcal{G}_{\theta=\theta_*}(\{z_i\}, c) \tag{1}$$

given as input a set of *latent variables* $\{z_i\}$, and optionally (depending on the particular construction) the class c to which the generated image should belong. The latent variables $\{z_i\}$ are usually drawn from some random distribution and encode the variability of the overall space of images.

Most of the focus in this domain of Machine Learning research is concentrated on finding the optimal architecture and training procedure so that the generated images represent best the space of images given as training data. In the present paper, we would like to pursue, however, an orthogonal line of investigation and study how the generated space of images changes as we move in the space of neural network parameters. In this sense we investigate how the whole space of images is globally encoded in the neural network parameters.

Indeed, a particular generator neural network $\mathcal{G}_{\theta=\theta_*}$ with the specific set of optimal weights θ_*, obtained through training on the *ImageNet* dataset[14], may be understood as providing a *neural network representation* of the space of photo-realistic natural images (including also human-made structures, objects, vehicles, food etc. but no art).

Similarly, we can view each particular choice of parameters θ of the generator neural network $\mathcal{G}_\theta$ as encoding some specific *space of images*. We thus have a mapping

Institute of Theoretical Physics and Mark Kac Center for Complex Systems Research, Jagiellonian University, ul. Łojasiewicza 11, 30-348 Kraków, Poland. email: romuald.janik@gmail.com

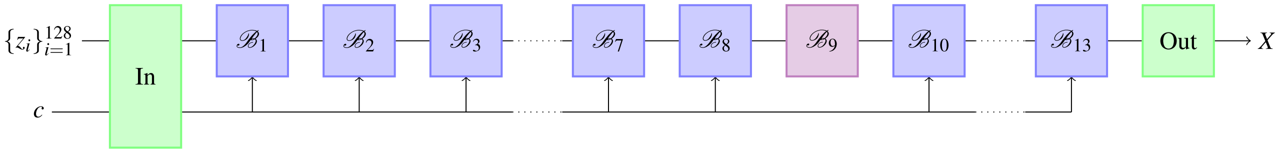

Figure 1. Schematic structure of the `BigGAN-deep-256` generator network $\mathcal{G}_\theta$ taking as input a 128-dimensional vector of latent variables $\{z_i\}$ and one of 1000 ImageNet classes c. The blue blocks are residual blocks with two 1×1 and two 3×3 convolutions as well as four conditional batch normalization layers which receive shortcut connections from the entry stage. The purple block is a "self-attention" block. The blocks $\mathcal{B}_2$, $\mathcal{B}_4$, $\mathcal{B}_6$, $\mathcal{B}_8$, $\mathcal{B}_{11}$ and $\mathcal{B}_{13}$ increase image dimensionality by factors of 2. See[15] for details.

$$\theta \quad \longrightarrow \quad \textit{space of images generated by } \mathcal{G}_\theta \tag{2}$$

The above mentioned optimal weights θ_* get mapped to the space of photo-realistic natural images. The goal of this paper is to investigate in what way the *space of images* changes as we move away from the photo-realistic point $\theta = \theta_*$ in the space of neural network parameters.

Results

Aesthetics and artistic rendition. For the experiments performed in the present paper we utilized the generator part of the `BigGAN-deep-256` network[15] trained on the *ImageNet* dataset[14]. The general structure of the network is shown in Fig. 1. It consists of an entry stage, followed by 13 groups of layers, which also receive shortcut connections directly from the entry stage, followed by the output stage (see the original paper[15] for details). The generator network has around 55 million trainable parameters θ. These are naturally organized into a set of various tensors parametrizing the individual layers of the network. In order not to introduce cluttered notation, we collectively denote all parameters by θ, and we apply all formulas similarly to all the tensors comprising θ. As the photo-realistic point, we take the weights θ_* of the pretrained model[16].

In order to move away from θ_*, we employ a *multiplicative* random perturbation of the neural network parameters

$$\theta = \theta_* \cdot (1 + \alpha \cdot \text{ random}) \tag{3}$$

In the above schematic formula, for each layer of the network, θ denotes a tensor of parameters of appropriate shape and dimensionality, e.g. four-dimensional for the weights of a convolutional layer. θ_* are the corresponding tensors of parameters of the pretrained model. The tensor *random* has the same shape and dimensionality as θ_* and its elements are drawn independently from a normal distribution with zero mean and unit standard deviation. The symbol $\cdot$ denotes element-wise multiplication of the corresponding tensor elements. α is a numerical constant which characterizes the magnitude of the deformation.

In other words, the perturbation (3) amounts to multiplying each parameter of the network by an independent random number drawn from a gaussian distribution with unit mean and standard deviation equal to α, thus increasing or decreasing the value of each parameter proportionally by a random factor, hence the name *multiplicative*. A possible alternative would be an *additive* perturbation, but it would require specific tuning of the magnitude of the additive random perturbation to the typical magnitude of the particular weight tensor. Another drawback would be that very small weights could get disproportionally large shifts. Hence we do not study it here.

The particular perturbation (3), leading to a specific deformed network can be succinctly characterized in a reproducible manner by specifying the random seed of the random number generator used for generating the random numbers in (3). In the following we use small integer numbers as random seeds for constructing particular deformed networks. This allows to assess the genericity of the obtained results, as the considered neural networks are in this way *a-priori* prespecified.

We take the constant $\alpha = 0.35$ so that we move noticeably away from the photo-realistic point θ_* but still retain some link with the original objects or scenes. In Supplementary Materials Sect. A2, we discuss in some detail the dependence on the value of α.

It is important to contrast here moving in latent space $\{z_i\}$ for a fixed generative neural network, which is often studied in the Machine Learning literature, with moving in the space of weights θ, which we do in (3). In the former case, each point $\{z_i\}$ corresponds to a single image generated by the fixed generator network $\mathcal{G}_{\theta_*}$. Thus, when varying $\{z_i\}$, one moves in the given fixed photo-realistic space of images associated to $\mathcal{G}_{\theta_*}$. In contrast, in the case studied here and given by (3), each point θ is a *different* generator network $\mathcal{G}_\theta$, and thus corresponds to a different *visual universe* of images which can be potentially generated by $\mathcal{G}_\theta$.

Before presenting the results, let us make some comments related to the space of neural network parameters θ. Since the dimensionality of this space is very high, of the order of 55 million, this space has very unintuitive properties. Any two randomly chosen points are essentially orthogonal and hence at a fixed cosine distance equal to 1 (the cosine distance is not really a metric, but it is a very convenient and standard measure of distance, especially in high dimensional space). The deformations given by (3) with $\alpha = 0.35$ are roughly at a cosine distance of 0.056 from θ_* (see Supplementary Fig. S4), hence we are exploring a local neighbourhood of θ_*. Yet this local neighbourhood is also of similarly high dimensionality, hence any two randomly chosen deformation directions are orthogonal, leading to potential visual diversity. Indeed the cosine distance between two such perturbed networks is generically around 0.109, so even higher than their distance from the original network.

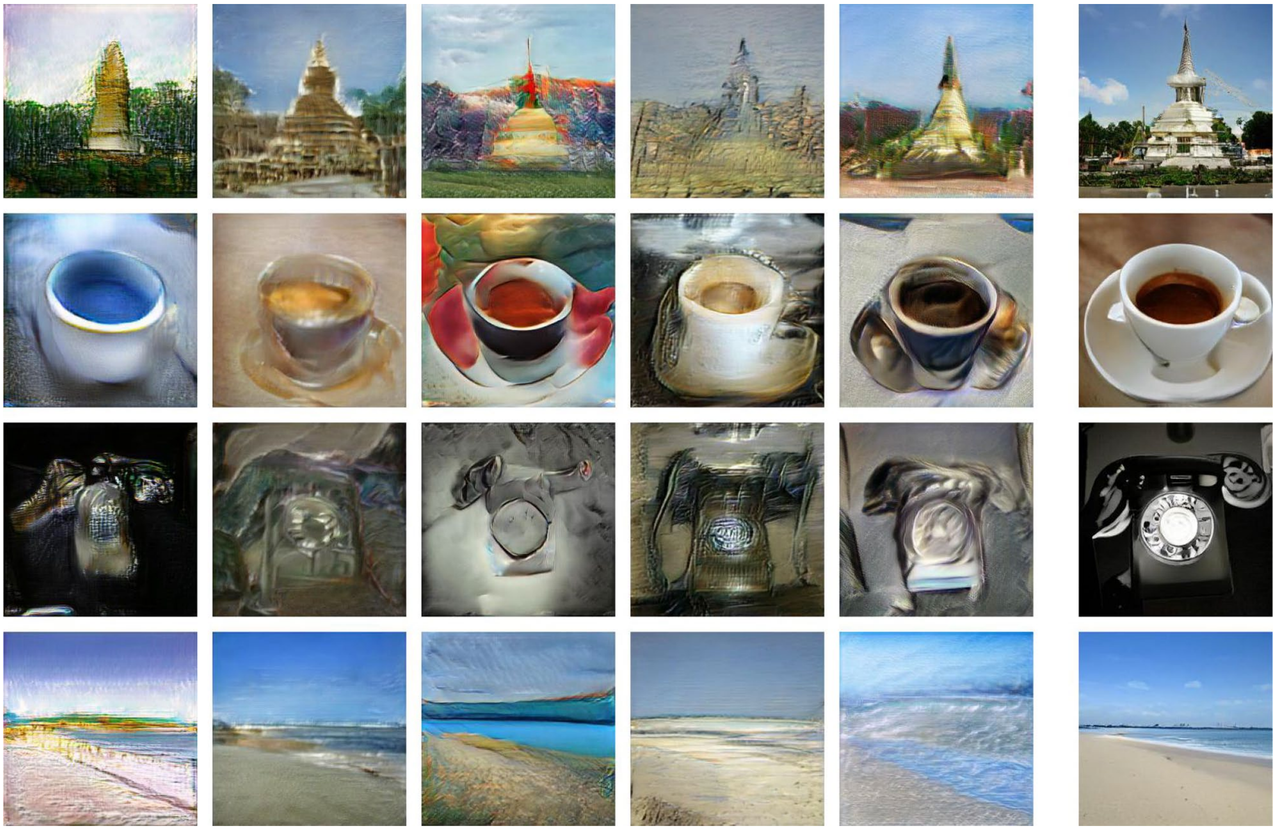

Figure 2. Images generated by neural networks with weights given by (3), realizing various deviations from the photo-realistic point $\theta = \theta_*$. Each column corresponds to a distinct neural network. None of the networks had access to any human-made art. Far right: corresponding photo-realistic images generated by the original `BigGAN-deep-256` network. The inputs to the different networks were identical.

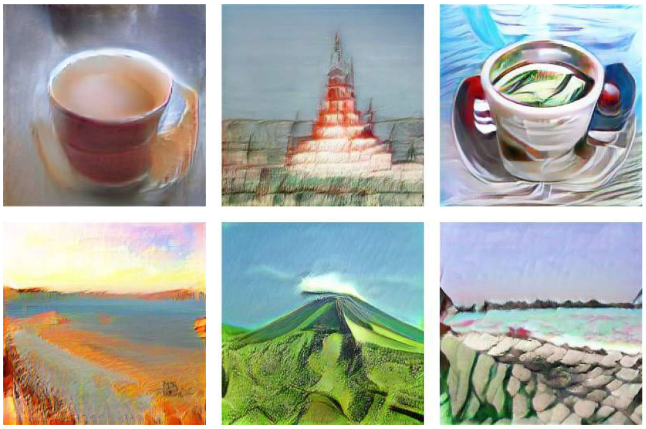

Figure 3. Selected images generated by neural networks obtained through various ways of randomized modifications from a BigGAN network generating photo-realistic images (for further examples see https://neuromorphic.art).

In Fig. 2 we show images generated by five networks $\mathcal{G}_\theta$ with weights θ given by random perturbations of the form (3) with random seeds chosen from the range $0 - 10$ and, for comparison, images produced by the original photo-realistic network $\mathcal{G}_{\theta=\theta_*}$. The images represent *stupa*, *espresso*, *dial telephone* and *seashore*. The respective latent noise $\{z_i\}$ inputs in each row of Fig. 2 for all networks were identical. The two leftmost images in the top row of Fig. 3 were also obtained using (3) (the remaining four images in Fig. 3 were obtained by substituting a specific *subset* of weights by randomly drawn values).

A striking feature of the obtained images is that they seem to give an *"artistic rendition"* of the original photo-realistic objects. The perturbation of the space of parameters θ away from the point θ_* clearly breaks the fine-tuning necessary for the photo-realistic rendition of the images by the original network, which is of course not surprising. What is quite unexpected, however, is that this manner of breaking leads to aesthetically pleasing and interesting images, at least for a range of object classes. Moreover, in the majority of cases the utilized colour palette and colour transitions appear balanced and aesthetic — they do not strike us as artificial, which would be a natural expectation given the random character of the perturbation (3) of the photo-realistic network parameters θ_*. Most probably, the multiplicative character of the perturbation (3) helps in this respect as small weights do not get disproportionally large modifications, which could happen for *additive* perturbations of fixed magnitude.

The deformations of photo-realism are reminiscent of the kind of simplifications that a human artist would employ when painting or making a rough sketch. Indeed, many of the obtained images could arguably be mistaken at first glance for paintings or sketches made by a human. In order to quantify this observation, for the generated images from Fig. 2 we picked the most perceptually similar paintings from the ArtBench dataset[17] comprising 60000 human-made artworks. Perceptual similarity was measured using a technique from Neural Style Transfer[18] comparing the so-called Gram matrix of the internal activations of a pretrained ResNet-152 deep neural network (see *Methods* for details). In Fig. 4 we present four particularly interesting cases. We note a striking similarity in the colour palettes as well as some sketchiness in the depictions, a bit more pronounced in the case of the generated images, but somehow remaining in character.

Some aspects of the apparent sketchiness can be identified by examining the statistics of the Fourier coefficients of the generated images. It is a classical result that natural images have a power-law spectrum[19] of the magnitudes of Fourier coefficients as a function of the wavenumber k. In Supplementary Material Sect. A4 (and Fig. S6(left) therein), we compared the spectra of the images generated by the perturbed networks to the ones obtained from the original one taking the spectra of natural images as a benchmark. We find that the images generated through (3) have distinctly lower low frequency components signifying less pronounced foreground than for both natural and undeformed network images. Sharp edges are also suppressed as can be seen by analyzing a Canny edge detector (see Supplementary Material Sect. A4 and Fig. S6(right)).

We note that the images generated by the deformed networks depict more or less the original objects. The modification of the network parameters leads primarily to a change of rendition and not a change in semantics. In the later part of the paper we will consider a different deformation which will strongly upset the semantic content. In Supplementary Materials Sect. A5 and Fig. S7 therein, we examine how enhancing the semantic deformation interacts with the multiplicative deformation (3). We observe some softening, although the balanced colour palettes of Fig. 2 seem to be missing. This is most probably due to the fact that the source for downstream processing by the deformed network is no longer an encoding of natural images but rather some garbled encoding leaving the strongly deformed block $\mathcal{B}_2$. This behaviour supports the role of natural imagery as a source of some aesthetic qualities.

Another intriguing feature is that quite often one can discern a particular style characteristic of a specific perturbed neural network, which differentiates it from neural networks obtained through other perturbations. This can be seen as a certain visual consistency in the columns of Fig. 2.

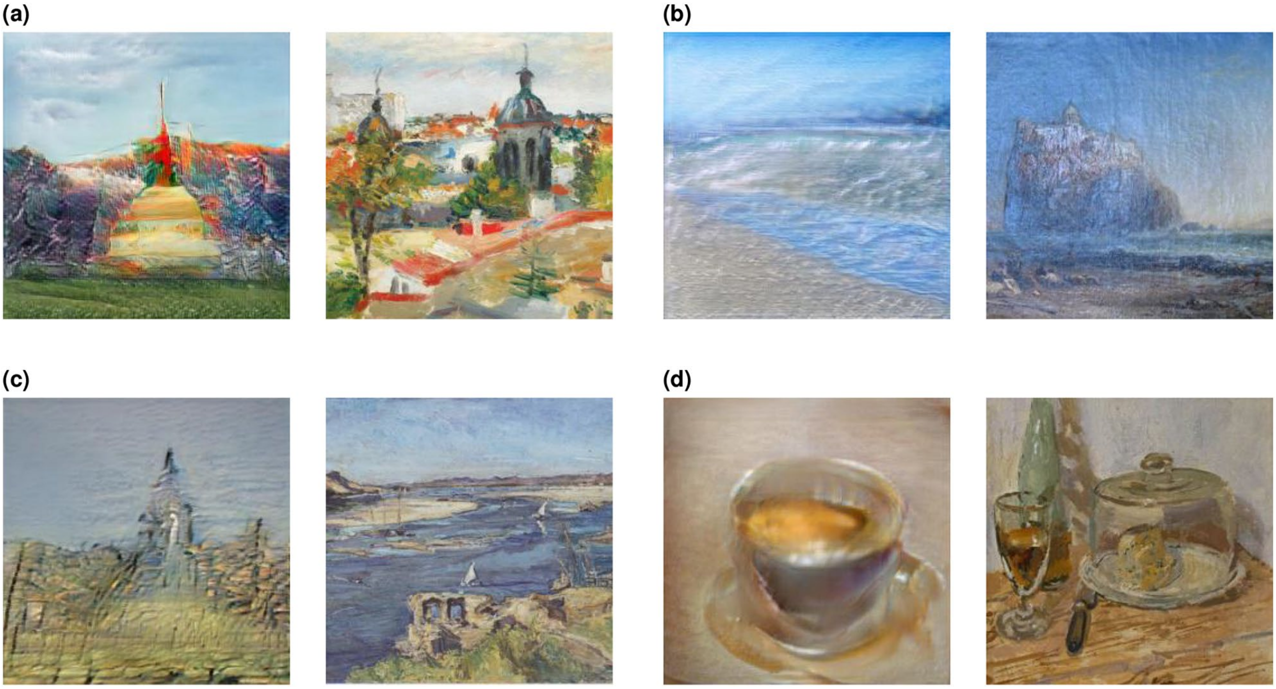

Figure 4. Selected generated images from Fig. 2 juxtaposed with perceptually closest artworks from the ArtBench dataset[17] (shown on the right of each pair): (**a**) Lucian Grigorescu, *Atelier View (Bucharest rooftops)*, (**b**) James Webb, *The Old Castle overlooking the Bay of Naples, Italy* (1875), (**c**) Max Slevogt, *The Nile at Aswan* (1914), (**d**) Walter Sickert, *Roquefort* (1918–1920).

Let us also mention some limitations of these results, which have to be kept in mind. Only a subset of *ImageNet* classes (architectural, some objects, landscapes) behaves equally well under these deformations. Most probably the other ones require more fine-tuned weights. Indeed, generated images for certain classes exhibit some pathologies even at the photo-realistic point θ_*, i.e. for the original pretrained network. Imposing further weight perturbations in these cases may easily "break" the images. Furthermore, we do not claim that *every* randomized perturbation leads to an aesthetic result. However, quite a lot do (in particular, the examples shown in Fig. 2 were chosen just out of consecutive random seeds 0–10). Let us note that occasionally one encounters visually stunning examples such as those shown in Fig. 3 (the two first images in the top row of Fig. 3 were obtained using (3)). One might make here an analogy with the wide spectrum of artistic talent in the human population.

As a test of genericity of the obtained results, in Supplementary Fig. S1, we show examples of a *stupa* and *espresso* generated by deformed networks for the range 0–23 of *consecutive* random seeds, thus without any human selection or cherry-picking. In addition, we implemented the same construction (3) for a different generative neural network architecture `StyleGAN-XL`[20] (see Supplementary Material Sect. A3 and Fig. S5). The results go in the same direction of an *"artistic rendition"*, albeit of a more pictorial character and perhaps with less variety.

Finally, we would like to emphasize that the appearance of aesthetic properties should be interpreted in the context of the immense dimensionality of the space of parameters (~ 55 *million*). In such a high dimensional space any two randomly chosen directions of deformation are essentially mutually orthogonal, and as described above, the perturbed networks are at a larger cosine distance between themselves than from the original undeformed network. Therefore, if any qualitative property *repeats itself under random sampling* even in a subset of cases, it is, in our opinion, a significant observation and the aforementioned property can be considered as being to a large degree generic.

What do the above experiments tell us? Firstly, we may infer that the property of being perceived as aesthetic by a human may be related to the very nature of the photo-realistic world. Indeed, the original network was exposed only to photo-realistic images and did not have any contact with human art. The perturbations of the neural parametrization of the space of images leading to the ones shown in Fig. 2 were generic (multiplicative) random perturbations which were not biased by any further image input or optimization procedure.

Secondly, this property is firmly tied to the *neural network representation* $\mathcal{G}_\theta$ of spaces of images of the human visual environment, which apart from the particular values of the parameters θ, incorporates as a kind of structural prior the specific generator architecture of the `BigGAN-deep-256` model. The imprint of the neural network architecture may be assessed by comparing with the results for the `StyleGAN-XL` network presented in the Supplementary Material.

Thirdly, the observed interplay of aesthetics and neural parametrization ties in with the hypothesis of[5–7] that the perception of aesthetics is linked with features of the human visual system in the brain. This may go beyond being just an analogy as there are already indications that the fMRI activations in the higher stages of human visual processing are quite well correlated with deeper levels of convolutional neural networks[11–13]. We will return to this point in more detail in the *Discussion* section.

Finally, we believe that the above findings could be of potential interest for humanities, in particular for the theory and philosophy of art and aesthetics. In this respect, the results of the experiments performed in the present paper could be treated as providing an unexpected piece of evidence for the possibility of a biological (non-cultural) origin of some kinds of "artistic renditions".

Differences with other approaches. It is important to contrast the results obtained in the present paper with some other approaches linking artistic renditions and neural networks as superficially they may seem similar.

A very well known construction is the so-called *Neural Style Transfer*[18], where a given input image is transformed into the style of a second image (the style image), typically an image of a painting or work of art, with the similarity in style measured by a deep neural network pretrained on an image classification task. Alternatively, GANs have also been trained on art (see e.g.[21]) to generate new images based on the given artistic styles. These techniques use explicit input of human-made art to produce new images similar in style, which was of course their key goal.

The aim of our investigation was, however, quite different and the "artistic character" of the images described in the present paper appeared spontaneously as an *a-priori* unexpected byproduct. Our results were obtained using only photo-realistic images of the natural world and the human environment without any contact with human or machine-made art. They thus provide a realization of the *emergence* of aesthetic properties directly from a neural network parametrization of a photo-realistic world.

An approach perhaps closest in spirit to ours is the hand picking of *exceptional* latent variables $\{z_i\}$ for the photo-realistic model θ_* in order to generate surreal images (see[22] for a discussion). That procedure really exploits the *deficiencies* of the generative photo-realistic model in order to produce artistically interesting images. The more modern text-to-image generators DALL-E and DALL-E2 from OpenAI[23] can generate stunning images especially through paradoxical input text (which can be thought of as an analog of the exceptional latent variables $\{z_i\}$ mentioned above), hence receive essential input from a human. Moreover, their training data is much richer and includes, in particular, human art. Our result is conceptually quite different, as we show the essentially *generic* appearance of aesthetic/"artistic" images under finite deviations of the neural network parameters away from the "photo-realistic" values θ_*. This occurs without any human intervention and without any contact with human-made art.

Visual symbolic representations. Another surprising feature of the generative neural representation of the space of images provided by the `BigGAN-deep-256` model is that it allows to exhibit a certain kind of visual symbolic representations. In order to see that, we first heuristically identify the location of some high level semantic information in the neural encoding of $\mathcal{G}_\theta$.

In contrast to the usual neural networks used for classification, the flow of information in a generative network generally runs from the most semantic/global features (incorporating here the class c given as input) to the low-level pixel-based visual output. We may therefore expect, that closer to the input we have more high-level semantic information.

In a subsequent experiment, we substitute the parameters of the second block of layers, which we denote by $\mathcal{B}_2$ (see Fig. 1), with values drawn randomly from normal distributions:

$$\begin{aligned} \theta_L &= random_L & L \in \mathcal{B}_2 \\ \theta_L &= \theta_{*L} & L \notin \mathcal{B}_2 \end{aligned} \tag{4}$$

The parameters of the normal distributions are taken from the statistics of θ_* for the corresponding weights of the given layer $L \in \mathcal{B}_2$ (for convolutional weights we computed the statistics for each pixel of the 3×3 filter individually (see *Methods*), this last choice is, however, really inessential for the phenomenon discussed here). The modification (4) can be viewed as upsetting only a deep semantic part of the neural representation of the space of images. We will comment on the specific choice of the second block $\mathcal{B}_2$ further below.

Images generated by neural networks $\mathcal{G}_\theta$ constructed through (4) for five random seeds in the range $0 - 10$ are shown in Fig. 5 (in the Supplementary Fig. S2, we show more examples of a *stupa* and an *espresso* for a wider range of *consecutive* random seeds in order to assess the genericity of the discussed phenomenon). At first glance, individually the images may seem haphazard and quite disconnected from the original photo-realistic objects, but viewing them side by side we observe surprising similarities. Indeed, an overall distinctive characteristic of the original object seems preserved – like the round shape of the *espresso* or the triangular form of the *volcano*. It is however articulated using quite different and varying graphics primitives and materials. A similar phenomenon, but on a slightly more subtle level, occurs also for the *stupa* in the first row. There, the overall shape morphs either into some quasi-architectural form or into a person-like depiction. The *dial telephone* in the bottom row is most extreme. Here one cannot really identify *by eye* a strong dominant feature, so its visual representations may be difficult to interpret – although one could perhaps put forward some arguments for certain specific networks.

The above deformations can be thus understood as inducing a *visual symbolic representation*, where a dominant strong characteristic of the original object is realized in terms of completely unrelated materials and

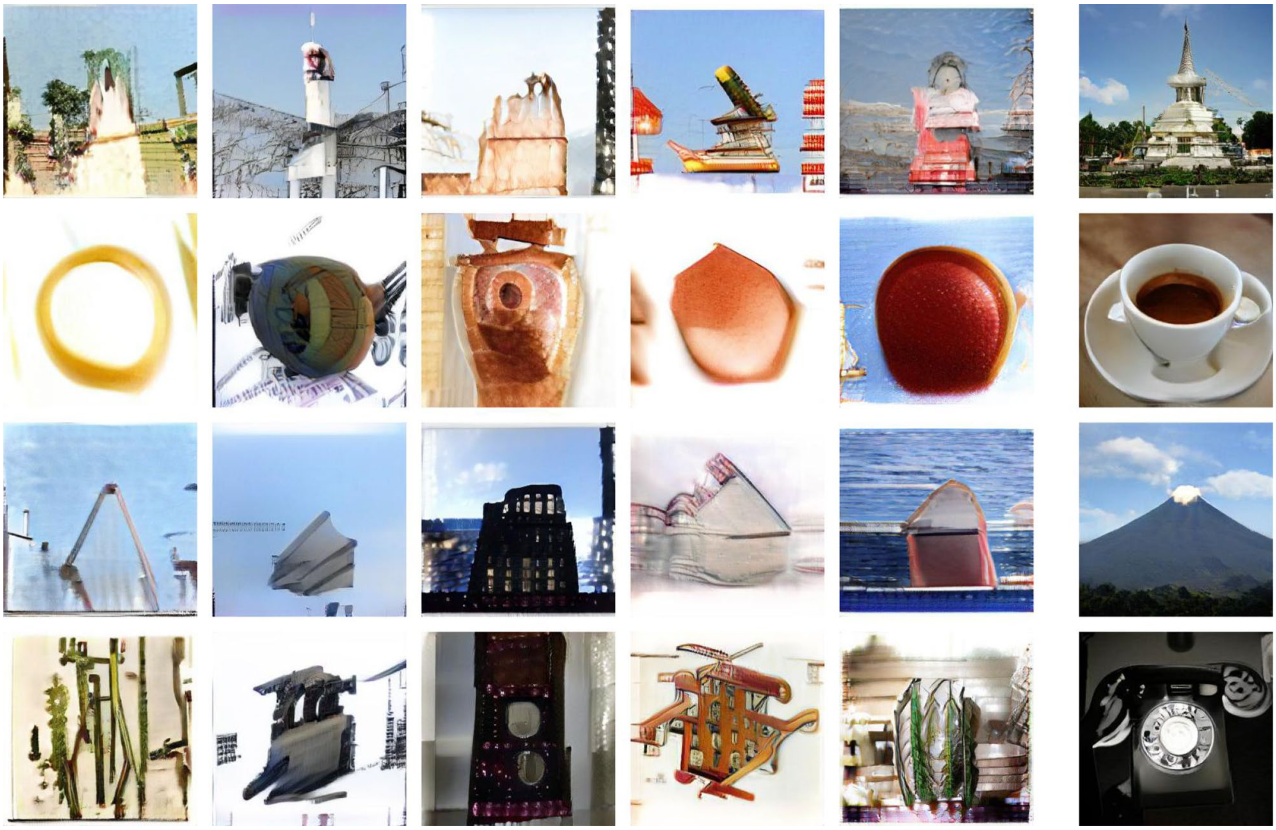

Figure 5. Images generated by neural networks with weights given by (4), upsetting the deep semantic structure of the representation of the space of images. Each column corresponds to a distinct neural network. Most of the images exhibit dominant features of the original object realized in terms of different ingredients (see text). Far right: corresponding photo-realistic images generated by the original `BigGAN-deep-256` network. The inputs to the different networks were identical.

ingredients (for strawberries, not shown here, on the other hand, it is the material – red texture – which seems to be the dominating feature). We expect this interpretation to hold under the condition that such a very strong simple dominant feature exists for the object class in question.

The fact that the dominant visually prominent feature is still present after the modification of the weights in (4), indicates that it must be encoded also in the undeformed parts of the network. From this point of view, the second block of layers $\mathcal{B}_2$ seems to play a privileged role in the neural network representation $\mathcal{G}_\theta$, as it does not destroy that feature but rather swaps in varying local visual ingredients while still preserving some sharpness and locally detailed depiction.

Upsetting similarly only the first block $\mathcal{B}_1$ loses any resemblance to the original objects, while doing the same for further blocks leads first to a loss of the *local* photo-realistic depiction and sharpness still seen in Fig. 5, while for still further blocks the original object becomes more and more directly recognizable. Consequently, from the point of view of the visual symbolic representations, the block $\mathcal{B}_2$ is essentially singled out.

The fact that the phenomenon is mostly restricted to a subset of the neural network architecture should not be understood as a problem. First, we do not expect that all blocks/layers in a deep neural network play an equivalent role. The differentiation of their roles is in fact a very interesting feature (recall that e.g. the visual system in the brain has clear non-interchangeable modular structure). Second, the subset is quite sizeable, as the dimensionality of the $\mathcal{B}_2$ parameter space in (4) is still very large, equal to around 8.5 million. Last but not least, we find it extremely intriguing that examples of such visual symbolic representations can be indeed realized in an artificial neural network context.

Discussion

In this paper we studied the global properties of a generative neural network parametrization of *spaces of images*. We found that essentially generic deviations of the neural network parameters from the photo-realistic point θ_* quite often lead to neural networks which generate images which may appear aesthetic to humans. In many cases these images are difficult to distinguish at first glance from images of paintings or sketches made by a human, even though the neural network did not encounter any human-made art.

The above observation shows that aesthetic properties could arise in an *emergent* way directly from the nature of the photo-realistic visual environment through the character of its neural network representation. What is particularly intriguing about this result arises from tension with the belief that aesthetic perception is intimately linked to the human observer and appears to us as very subjective. Yet, the artificial neural network construction presented in this paper in some sense *objectivizes* this quality. This opens up numerous questions. What is the interplay between subjectivity and objectivity in aesthetic perception? To what extent and at what level can we draw an analogy between aspects of the neural parametrization and the biological roots of aesthetic perception in the human brain[8–10]? In particular, how does this fit with the hypothesis of[5–7] that aesthetic perception is related to the human visual system in the brain?

On the one hand, as already mentioned, there is research showing that activations in the human visual cortex as measured by fMRI are quite well correlated with features in a deep convolutional network[11–13]. This does not mean, however, that there is a direct correspondence between human visual system and convolutional neural networks. Indeed there are numerous crucial differences like the ubiquity of feedback connections in the brain, the whole mechanics of saccades and attention, and of course the training regimen as well as some known pathologies of neural networks like catastrophic forgetting or adversarial examples. The meaning of the results[11–13] is that while the fMRI activations represent coarse grained averages both in space and in time of the visual neuronal activity, the deep convolutional networks can up to a point capture this mesoscopic scale of activity. It is only at this coarse averaged level that we are making an analogy between neural networks and the human visual system.

On the other hand, the cited results[11–13] were obtained for discriminative/classification networks and, as far as we know, there is no similar investigation for generative networks. Indeed in the latter case, the information flow goes in an opposite direction as the generative networks produce images (thus intuitively mimicking visual imagination), while the human brain in the studies[11–13] perceives them. Of course, the human brain visual system with bidirectional information flow is certainly quite different in detail from a standard feedforward convolutional neural network. Paradoxically, this may increase the potential relevance of generative neural networks as they may be considered as modelling the top-down pathway in perception (e.g. along the lines of[24]). In addition, one should note that there are marked similarities between perception and visual imagery seen in neuroimaging studies[25] (see also[26] for an extended discussion).

On a higher, more qualitative level, a common feature of the analysis of aesthetic perception motivated by neuroscience in[5–7] is its emphasis on the essences of particular concepts, characteristic of the brain seeking *constancy* in its environment and thus abstracting away transient particularities. From this point of view, a pictorial representation which is closer to the internalized essence is more likely to be perceived as aesthetic. Photo-realistic details, on the other hand, are specific to particular object instances and tend to lower the aesthetic appeal.

In this sense, one can view the randomized perturbations away from the photo-realistic point as dispensing with the fine-tuned particular details, which due to the uncorrelated nature of the perturbations would get averaged out. The outcome could thus be interpreted indeed as generating more *essence-like* depictions. But this is certainly not the whole story, as just performing gaussian blurring on images does not make them aesthetic or *essence-like*. The neural network randomized perturbations must therefore act in a more subtle way and the concrete form of the generator neural network parametrization $\mathcal{G}_\theta$ somehow manages to capture some finer aspects of human aesthetic perception. Indeed, the emergence of visually appealing forms and colour transitions probably depends crucially on properties of convolutions appearing in the neural encoding, the overall colour structure of the natural environment and the specific mixing induced by the randomized modification of weights.

In addition, the specific randomized perturbations leave an imprint on the overall style of images generated by a particular deformed network. One could think of this as an analog of inter-subject "artistic" variability.

In this respect, it is interesting to speculate to what extent natural randomness and stochasticity in the nervous system[27,28] could be relevant in the context of the present observations. One could expect that randomness would lead to more robust (*essence-like?*) concepts. Indeed, in the artificial neural network context it has been shown that adding random noise to neural networks during training and evaluation increases their resilience to adversarial examples[29,30]. This type of randomness, however, would be associated with intra-subject (or here intra-network) variability and is not directly represented in the constructions of the present paper.

The second main result of this paper is that randomly scrambling a specific part of the deep semantic structure of the neural network parameters θ_* can lead to *visual symbolic representations*, where a dominant visual feature of a particular object is realized in terms of atypical and nonstandard visual ingredients.

This result is quite intriguing, as symbolic representations are an important component appearing throughout human culture, ranging from a key element of artistic expression (see e.g.[31,32]) to the way that psychoanalysis interprets dreams[33,34], with some important psychological concepts manifesting themselves encoded in various proxy objects, persons or scenes. In this context, we should nevertheless emphasize that the type of symbolic representation appearing in the present work is very much simplified, restricted just to some visual characteristics and completely blind to any aspect of cultural meaning, as the original generative neural network's world was just the purely visual environment. Even with these caveats, however, we find it very surprising that an analog of a symbolic representation can arise naturally in an artificial neural network context.

As a side remark, let us note that all the constructions in the present paper involve various kinds of randomized "rewirings" of the connection strengths of the artificial neural network. If one would look for brain states where randomness is enhanced, then a natural example would be the psychedelic state, where increased neural signal diversity was measured[35] in accordance with the "entropic brain" picture[36,37]. Perhaps some analogies could be pursued in this direction.

Finally, we would also like to make a methodological comment. The method of analysis of the neural network encoding used in the above case is in fact akin to the classical practice in neuroscience/neurology of analyzing the cognitive characteristics of patients with various brain lesions as a window on the functioning of the corresponding subsystems of the brain. In the present paper, we basically artificially induced a lesion in the generator network by substituting the values of a subset of neural network weights with completely random numbers. Subsequently, we examined the resulting neural network output. We expect that this technique may be quite useful for analyzing the structure of deep neural network knowledge representations for very complex models. Although here our focus was slightly different, as we emphasized more the qualitatively novel "positive" aspects (the visual symbolic representations) rather than the breakdown of photo-realism.

We believe that the obtained results and the consequent questions could foster new research on the borderline of cognitive neuroscience, (neuro)aesthetics and artificial neural networks. Moreover, we hope that both of the two main results of the present paper would be of potential interest for humanities, wherein they can be considered as *proofs of concept* showing the possible roots of some key human phenomena.

Methods

Image generation. In the present paper we use the implementation of the `BigGAN-deep-256` neural network provided by the package https://github.com/huggingface/pytorch-pretrained-BigGAN. After loading the model, we set `PyTorch` and `numpy` random generator seeds and modify the weights as described either by Eq. (3) or (4). In case of (4) and convolutional layers, we collect statistics for each pixel in the 3×3 filters and use multivariate normal sampling with diagonal covariance from the `scipy` library. The inputs to the networks are constructed using the code

```
truncation = 0.4
classes = one_hot_from_int([832, 967, 947, 46, 528, 949, 319, 978, 980], batch_size=9)
noise = truncated_noise_sample(truncation=truncation, batch_size=9, seed=5678)
```

Classes 832, 967, 528 and 978 are used for Fig. 2, and classes 832, 967, 980 and 528 are used for Fig. 5.

Perceptual similarity. The technique of Neural Style Transfer[18] introduced a measure of artistic style abstracting from the precise image content, based on the Gram matrix of the activations of a deep neural network pretrained on ImageNet. One first inputs the image of interest to the pretrained network and measures the activations of a particular internal layer of the network, which form a tensor x_{ij}^C where C denotes channel number and i, j are coordinates (generically at a lower resolution than the input image). The Gram matrix is defined by

$$Gram^{CD}(image) \equiv \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{ij}^C \, x_{ij}^D \tag{5}$$

which can be subsequently flattened to a vector. A measure of perceptual similarity between images may be defined as the cosine similarity between the flattened Gram matrices of the two images. In the present paper we use a pretrained ResNet-152 deep neural network (with weights included in `PyTorch`) and consider the activations at the output of the third block of residual layers. These activations have dimensionality $1024 \times 14 \times 14$, hence the resulting Gram vectors have dimensionality 1048576. The specific choice of the network and which residual block output to take for computing the Gram matrix is motivated by our earlier results[38], that this combination seems to reproduce best the fMRI activations of the human visual system. This choice of residual block output within the ResNet-152 network also obtained the best results in matching the corresponding *stupa* and *espresso* images in Fig. 2.

Data availability

The key constructions in the present study do not use any dataset or images, but only the weights of the pre-trained `BigGAN-deep-256` neural network which are provided by the package https://github.com/huggingface/pytorch-pretrained-BigGAN, and are the `PyTorch` version of the original https://tfhub.dev/deepmind/biggan-deep-256/1 released by the authors of[15] under the Apache 2.0 license. For the perceptual comparison with human artworks we used the ArtBench dataset[17] available at https://github.com/liaopeiyuan/artbench. For an evaluation of some statistical properties of the generated images we used a small part of the validation dataset of ImageNet, available at https://huggingface.co/datasets/imagenet-1k. The `StyleGAN-XL` network used for Supplementary Fig. S5 is available at https://github.com/autonomousvision/stylegan-xl.

Received: 19 December 2022; Accepted: 8 July 2023
Published online: 15 July 2023

References

1. Goodfellow, I. J. *et al. Generative Adversarial Networks* (NIPS, 2014). arXiv:arXiv:1406.2661.
2. Wang, Z., She, Q. & Ward, T. E. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Comput. Surv.* **54**(2), 37 (2019).
3. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* (ICLR, 2014). arXiv:1312.6114.
4. Tolstikhin, I., Bousquet, O., Gelly, S. & Schoelkopf, B. *Wasserstein Auto-Encoders* (ICLR, 2018) arXiv:1711.01558.
5. Zeki, S. Art and the brain. *Daedalus* **127**(2), 71–103, (1998) reprinted in Journal of Consciousness Studies **6**(6-7), 76 (1999).
6. Zeki, S. *Inner Vision: An Exploration of Art and the Brain* (Oxford University Press, 1999).
7. Ramachandram, V. S. & Hirstein, W. The Science of Art. *J. Conscious. Stud.* **6**(6–7), 15 (1999).
8. Nadal, M. The experience of art: Insights from neuroimaging. *Prog. Brain Res.* **204**(7), 135–158 (2013).
9. Chatterjee, A. & Vartanian, O. Neuroaesthetics. *Trends Cogn. Sci.* **18**, 370–375 (2014).
10. Li, R. & Zhang, J. Review of computational neuroaesthetics: Bridging the gap between neuroaesthetics and computer science. *Brain Inf.* **7**, 16 (2020).
11. Yamins, L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* **111**(23), 8619–8624 (2014).
12. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**(11), e1003915 (2014).
13. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).
14. Deng, J. *et al. Imagenet: A Large-Scale Hierarchical Image Database* (CVPR, 2009).
15. Brock, A., Donahue, J., & Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis (ICLR, 2019). arXiv:1809.11096.
16. PyTorch version of the model https://github.com/huggingface/pytorch-pretrained-BigGAN transformed from the original https://tfhub.dev/deepmind/biggan-deep-256/1 released by the authors under the Apache 2.0 license.
17. Liao, P., Li, X., Liu, X. & Keutzer, K. The ArtBench Dataset: Benchmarking Generative Models with Artworks. arXiv:2206.11404.
18. Gatys, L. A., Ecker, A. S., & Bethge, M. Image style transfer using convolutional neural networks (CVPR, 2016) arXiv:1508.06576.
19. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.* **A4**, 2379–2394 (1987).
20. Sauer, A., Schwarz, K., & Geiger, A. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets SIGGRAPH'22. arXiv:2201.00273].
21. Elgammal, A., Liu, B., Elhoseiny, M. & Mazzone, M. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. ICCC 2017 and arXiv:1706.07068.
22. Hertzmann, A. Aesthetics of neural network art. arXiv:1903.05696.
23. Ramesh, A. *et al.* Zero-shot text-to-image generation. arXiv:2102.12092.
24. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *JOSA A* **20**, 1434–1448 (2003).
25. Dijkstra, N., van Bosch, S. E. & Gerven, M. A. J. Shared neural mechanisms of visual perception and imagery. *Trends Cogn. Sci.* **23**(5), 423 (2019).
26. Albright, T. D. On the perception of probable things: Neural substrates of associative memory, imagery, and perception. *Neuron* **74**, 227–245 (2012).
27. Stein, R. B., Gossen, E. R. & Jones, K. E. Neuronal variability: Noise or part of the signal?. *Nat. Rev. Neurosci.* **6**, 389 (2005).
28. Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. Noise in the nervous system. *Nat. Rev. Neurosci.* **9**, 292 (2008).
29. Liu, X., Cheng, M., Zhang, H. & Hsieh, C.-J. Towards Robust Neural Networks via Random Self-ensemble (ECCV 2018) arXiv:1712.00673.
30. Dapello, J. *et al.* Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Adv. Neural Inf. Process. Syst.* **33**, 13073–13087 (2020).
31. Naumburg, M. Art as symbolic speech. *J. Aesthet. Art Crit.* **13**, 435–450 (1955).
32. Balter, M. On the origin of art and symbolism. *Science* **323**, 709–711 (2009).
33. Freud, S. The Interpretation of Dreams. Various editions.
34. Jung, C. G. *Dreams* (Princeton University Press, 2011).
35. Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K. & Muthukumaraswamy, S. D. Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Sci. Rep.* **7**, 46421 (2017).
36. Carhart-Harris, R. L. *et al.* The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front. Hum. Neurosci.* **8**, 20 (2014).
37. Carhart-Harris, R. L. The entropic brain: Revisited. *Neuropharmacology* **142**, 167–178 (2018).
38. Janik, R. A. & Olesik, M. The visual brain seen through the lens of a deep neural network, poster at CSHL Meeting From Neuroscience to Artificially Intelligent Systems NAISys 2022 and in preparation.

Acknowledgements

I would like to thank Anka Janik, Maciej A. Nowak, Igor Podolak for interesting discussions and comments and Daniel Wójcik for explanations and references on noise in the nervous system. This work was supported by the research project *Bio-inspired artificial neural networks* (Grant No. POIR.04.04.00-00-14DE/18-00) within the Team-Net program of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund and by a grant from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University.

Author contributions

R.J. conceptualized and performed the presented research and wrote the paper.

Competing interests

The author declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-38443-9.

Correspondence and requests for materials should be addressed to R.A.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.