Synthesizing naturalistic visual textures with multiscale, nonlinear constraints using deep neural samplers

Ludovica de Paolis[1,*], Paolo Muratore[1] and Eugenio Piasini[1]

[1]SISSA - Scuola Internazionale di Studi Superiori Avanzati, Trieste, Italy

Abstract

According to efficient coding, the mammalian visual system conforms to the statistics of its natural input: natural visual scenes. Visual textures are a family of visual patterns that share certain local regularities [1], which have been used to test efficient coding. Still, experimental results are limited to unnatural looking low-order correlations. There are three main texture synthesis algorithms: parametric models [2], maximum entropy sampling methods [3], Gram Matrix-based convolutional neural networks [4]. We propose a new model for texture synthesis: a combination of a Variational Autoencoder (VAE [5]) and a pre-trained convolutional neural network (VGG-16 [6]) to generate realistic textures in the latent space characterized by nonlinear multi-scale representations. The model is trained with unsupervised learning to minimize a Kullback-Leibler divergence between the latent code and a Gaussian distribution, and a Mean Squared Error between pairs of Gram Matrices to define the perceptual difference between original and generated textures. Preliminary results include the replication of [4] with our custom training dataset; the assessment of VAE's image reconstruction ability; the functioning of the Gram Matrices in VGG-16 for generating textures. Preliminary crossvalidation results include finding hyperparameters values consistent with the literature: the latent code dimensionality and the KL loss weight. Future directions include geometric analyses on latent code and VGG-16 representations [7], and the definition of a metric for characterizing the geodesic of the latent code manifold [8]. The metric will be used as a perceptual probe to define the space of textures and to compare it to human perceptual space [3]. We expect our model to build a texture space that aligns with human perception, and the generated textures to show high variance statistical features, which will reveal how efficient coding applies to our behavioral and computational regime.

Keywords

Cognitive neuroscience, computational neuroscience, efficient coding, computer vision

1. Introduction and related works

According to the efficient coding principle, the mammalian visual system conforms to the statistics of its natural input: natural visual scenes. Efficient coding entails that neuronal encoding in the visual systems adapts strategically to the statistics of visual stimuli, retaining relevant information and discarding the rest [9], [10]. In practice, the visual system is expected to attribute less importance to high variance features that are classified as redundant.

However, Hermundstad et al. argue that, in certain coding regimes, the most perceptually salient features are those characterized by high variance. They show that the visual system is most sensitive to high variance statistical features centrally, while it down-weights high variance stimuli peripherally [11]. Crucially, they rely on synthetic texture images to study efficient coding just as in other studies with rodents [12], [10] and humans [11].

Visual textures have been defined as "a family of visual patterns that share certain local regularities" [1], and they represent a tool to understand how efficient coding applies to the mammalian visual system. In fact, textures exemplify the statistical properties of natural images, as they satisfy both local and global statistical constraints. Julesz was the first one to propose not only a formal definition of textures, but also to hypothesize that textures can be statistically determined. His work inspired the development of several texture synthesis models because, as suggested in [13] and in [10], synthetic

Discovery Science - Late Breaking Contributions 2024

*Corresponding author.

ldepaoli@sissa.it (L. de Paolis); pmurator@sissa.it (P. Muratore); epiasini@sissa.it (E. Piasini)

textures are best method for obtaining textures stimuli compared to natural images crops, because contextual information in natural image is a source of bias.

Texture synthesis refers to a group of methods to generate artificial visual textures. The scientific literature reports several kinds of models, such as parametric methods, [2], [3], convolutional neural networks (CNNs) [4], [14], [15], [16], [8], [17], generative adversarial networks (GANs) [18], and attention-based networks [19]. The performance of texture synthesis models is evaluated by humans through several behavioral experiments [2], [3], [20], as well as neuroimaging experiments [21].

Three of these models are particularly relevant for motivating this study: Portilla and Simoncelli developed a parametric model that relies on the parameters of wavelet coefficents. Synthetic textures are generated to statistically match the original ones [2]; Victor and Conte developed a method that samples maximum-entropy textures with fixed correlation statistics among neighboring pixels [3]; Gatys and colleagues used a CNN to compute correlations between feature maps and generated textures with gradient descent on the pixel space in order to reproduce the correlations in the input images [4]. Taken together, these three models are limited in generating textures in linear and low-dimensional spaces, generating textures that appear coarse grained and unnatural, or lacking interpretability. On the one hand Victor and Conte's model provides an elegant parameterization of texture space, which is suitable for highly-controlled experiments, but produces textures that are perceptually distant from those found in natural images [3]. On the other hand, Gatys et al. model generates naturalistic-looking textures but lacks explanations of the model's internal representation and provides fewer affordances towards investigating the preceptual properties of textures [4]. As a consequence, results in testing efficient coding have been limited to non-naturalistic visual textures characterized by low-order correlations among neighboring pixels [2], [3], [11], [12], [10].

2. Research goals

Overall, the literature suggests that texture synthesis constitutes a method to explore the space of features that populate natural visual scenes. For this reason, texture synthesis is a good candidate to test the efficient coding principle in mammalians. However, the most relevant models for texture synthesis still show some limitations not only in generating naturalistic-looking textures, but also for testing efficient coding.

Our goal is to develop a new method to generate textures and explore their relevant statistical features to test efficient coding in humans. We want to overcome the limits of the models found in the literature by proposing a new model: a combination of a Variational Autoencoder (VAE [5]) and a pre-trained CNN (VGG-16 [6]). Its purpose is to generate naturalistic-looking textures characterized by nonlinear multi-scale representations as in [4] in a latent space allowing mathematical exploration of its properties as in [3]. In particular, we think that for each dimension of the latent space there may be a corresponding texture feature. Finally, we want to explore the geometric properties of our model by studying its latent space and to use the generated textures as psychophysical stimuli for human subjects to compare the latent space generated by the model and the human perceptual space.

3. Methods

VAE generates images that show new, synthetic examples of the visual textures found in the input images. It is composed by a Probabilistic Encoder and a Probabilistic Decoder. The Encoder takes as an input images of textures and converts them into the so-called latent code, a vector representing the features of the textures in an ordered fashion. The Decoder generates new texture images from the latent code. In order to generate images, VAE samples a Gaussian distribution and it learns to minimize the difference between the distribution of the latent code learnt by the Encoder and the Gaussian.

At each generative iteration the original and the generated images are given to a pretrained VGG-16 to obtain a representation of the texture characteristics adapting the technique in [4]: computing the Gram Matrices of VGG-16 feature maps for each original-generated texture pairs. The Gram Matrices

are computed as the dot product of the set of feature maps in each convolutional layer, discarding the spatial dimensions and only retaining the statistical features of the textures detected by the filters. We obtain one original-reconstructed Gram Matrix pair for each convolutional layer in VGG-16, for a total of 13.

The model is trained with unsupervised learning to minimize a combination of three losses: a Kullback-Leibler divergence (KL) between the distribution of the latent code and the Gaussian distribution; a Mean Squared Error (MSE) between pairs of Gram Matrices and a MSE between the pixels of image pairs to define the perceptual difference between original and generated textures. In principle, the MSE losses could push the model to reconstruct images as most identical as possible to the original images. In this way, the model would not actually learn the distribution of textures but it would simply reconstruct the images. Therefore the KL loss is weighted to balance out the other two losses and to ensure proper disentanglement of the learnt latent space.

In the training process the model relies on two hyperparameters that we determine with a cross-validation analysis. We fine tune the weight assigned to the KL loss and the dimensionality of the latent code to find out the optimal number of dimensions that best fit the data.

We are using multiple datasets, including MNIST [22], CIFAR-10 [23], NSD [24] and a custom dataset of unpublished naturalistic landscapes taken from different environments.

4. Preliminary results

Preliminary results include the successful replication of the study by [4]. We coded the model from scratch and used our custom dataset to obtain new synthetic textures that display the same perceptual quality as the ones generated by [4].

We assessed VAE's ability to reconstruct images by using MNIST and CIFAR-10. They are datasets constituted by a large number of images that are very small, easy to handle and display semantic content like numbers and objects.

We proved the whole model improvement in generating textures from the custom dataset by tightly controlling hyperparameters search. This confirms the proper implementation and functioning of the Gram Matrix MSE loss in generating textures, not simple images.

These preliminary results suggest that the model is successfully learning to generate good quality textures. Each result provides a glimpse into the functioning of a single component, indicating promising qualitative results in its final version.

Currently we are in the process of guiding the model's training by performing multiple crossvalidations and refining the numbers that the model can explore to find the optimal values of KL weight and dimensionality of the latent code. Preliminary results obtained in the crossvalidation analysis suggest that the KL loss weight values are consistent with those reported the relevant literature; they also suggest that the number of latent dimensions of the latent code are higher than those found previously as in [3].

5. Discussion and future directions

Future directions include studying the properties of the high-dimensional space of the latent code learnt by the model. We plan to carry an Intrinsic Dimensionality (ID) analysis on both the distribution of the latent space and on the distributions of the layers in VGG-16 [7]. We hypothesize that as the number of layers in VGG-16 increases, the dimensionality optimally decreases in such a way to maintain the relevant features of the data and discard the irrelevant ones. We also hypothesize that the latent space of the VAE will show even lower ID properties compared to the last layer of VGG-16, thus revealing that the model can operate in a high-dimensional non-linear space that carries only the relevant features of the texture space in a compact and optimal manner.

Subsequently, we plan to define a local distance metric among close points in the latent space. This metric will be used locally to map the distance of all the neighboring points in restricted portions of

the latent space. It will also be extended to characterize the global space by calculating the distance between arbitrary pairs of points on the texture manifold.

The approach defined above allows to use the metric as a perceptual distance among textures. With our current model we could exploit the Gram Matrix MSE loss, as it measures the perceptual distance among textures. However, the Gram Matrix MSE relies on a Euclidean space, which could limit the description of the latent space manifold. One interesting approach, inspired by [8], includes the implementation of a new loss based on the Wasserstein distance between two distributions. Vacher's method is based on the observation that textures are well described by elliptical distributions, that in turn are a generalization of Gaussian distributions. The Wasserstein distance can be calculated between the target texture vectors found in the CNN filters and a Gaussian distribution. It can be used as a non-Euclidean metric for mapping the geodesic of the texture space. The Wasserstein distance is therefore a good candidate to perform texture interpolation in a curved manifold and to probe human perception of textures, and it is suitable for improving our model.

The final step of this project is to characterize the statistics of the visual scenes in human subjects to test efficient coding. Since the ID and the interpolation analyses described above allow to characterize the space of textures without using an external metric (e.g. Euclidean space) but only relying on the geometrical properties of the manifold, we could in principle use the same metric to test human sensitivity without making further assumptions.

This can be achieved by performing a behavioral task such as in [3], [20], [8]. Victor and Conte implement a four-alternative forced choice in the framework of texture segmentation. They showed participants images composed of textures generated to isolate specific local image statistics. The participants had to identify patches in the images where the texture varied according to the manipulated statistics [3]. Wallis et al. use an oddball paradigm contrasting original and synthesized textures [20]; Vacher et al., as described above, propose the usage of Wasserstein distance to smoothly interpolate two textures. The interpolation makes the transition between different textures less perceptually obvious compared to [3]. The participants had to judge similarity between textures and interpolated textures that are chosen according to the variance of some parameters that define the statistics of the texture image [8].

Given that both our work and that of [8] use CNNs, the textures generated by our models are expected to be similar. Consequently, it is logical for us to upgrade our model by implementing texture generation and space characterization using a Wasserstein loss-based approach. We want to run a behavioral experiment based on Victor and Conte approach, but since the stimuli will be interpolated, the transitions between texture patches will be smooth and will not bias participants in their judgments. We expect our model to be capable of building a perceptual texture space that aligns with human perception: textures that are closer in space according to our metric will also appear more perceptually similar to humans. We also expect our model to generate textures that show high variance features that would allow us to draw conclusions on the efficient coding hypothesis in the framework of our behavioral paradigm and our model.

6. Acknowledgments

I want to thank my supervisor, Eugenio Piasini, for his dedication in advising at any step of this project and for supporting me. I also want to thank my colleague, Paolo Muratore, for contributing to this research.

References

[1] B. Julesz, Textons, the elements of texture perception, and their interactions, Nature 290 (1981) 91–97.

[2] J. Portilla, E. P. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients, International Journal of Computer Vision 40 (2000) 49–70.

[3] J. D. Victor, M. M. Conte, Local image statistics: maximum entropy constructions and perceptual salience, Journal of Optical Society of America 29 (2012) 1313–1345.
[4] L. A. Gatys, A. S. Ecker, M. Bethge, Texture synthesis using convolutional neural networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems, volume 1, 2015, pp. 262–270.
[5] D. P. Kingma, M. Wellig, Auto-encoding variational bayes, in: 2nd International Conference on Learning Representations (ICLR2014), volume 11, 2014.
[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large scale image recognition, in: International Conference on Learning Representations, 2015.
[7] A. Ansuini, A. Laio, J. H. Macke, D. Zoccolan, Intrinsic dimension of data representations in deep neural networks, in: Advances in Neural Information processing Systems, 2019.
[8] J. Vacher, A. Davila, A. Kohn, R. Coen-Cagli, Texture interpolation for probing visual perception, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 22146––22157.
[9] P. Sterling, S. Laughlin, Principles of Neural Design, The MIT Press, 2017.
[10] T. Tesileanu, E. Piasini, V. Balasubramanian, Efficient processing of natural scenes in visual cortex, Frontiers in Cellular Neuroscience 15 (2022).
[11] A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian, G. Tkačik, Variance predicts salience in central sensory processing, eLife (2014).
[12] R. Caramellino, E. Piasini, A. Buccellato, A. Carboncino, V. Balasubramanian, D. Zoccolan, Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes, eLife (2021).
[13] B. Julesz, Visual pattern discrimination, IRE Transactions on Information Theory 8 (1962) 84–92.
[14] O. Sednik, D. Cohen-Or, Deep correlations for texture synthesist, ACM Transactions on Graphics 36 (2017) 161–176.
[15] X. Snelgrove, High-resolution multi-scale neural texture synthesis, in: SA '17: SIGGRAPH Asia 2017 Technical Briefs, 2017, pp. 1–4.
[16] D. Ulyanov, A. Vedaldi, V. Lempitsky, Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, Computer Vision and Pattern Recognition (2017).
[17] E. Heitz, K. Vanhoey, T. Chambon, L. Belcour, A sliced wasserstein loss for neural texture synthesis, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9407–9415.
[18] N. Jetchev, U. M. Bergmann, R. Vollgraf, Texture synthesis with spatial generative adversarial networks, arXiv (2017).
[19] S. Guo, V. Deschaintre, D. Noll, A. Roullier, U-attention to textures: Hierarchical hourglass vision transformer for universal texture synthesis, in: Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production, 2022, pp. 1–10.
[20] T. S. A. Wallis, C. M. Funke, A. S. Ecker, L. A. Gatys, F. A. Wichmann, M. Bethge, A parametric texture model based on deep convolutional features closely matches texture appearance for humans, Journal of Vision 17 (2017) 1–29.
[21] M. M. Henderson, M. J. Tarr, L. Wehbe, A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex, The Journal of Neuroscience 43 (2023) 4144–4161.
[22] Y. LeCunn, C. Cortes, C. J. Burges, The mnist database of handwritten digits, 2009. URL: http://yann.lecun.com/exdb/mnist/.
[23] A. Krizhevsky, V. Nair, G. Hinton, The cifar-10 dataset, 1994. URL: https://www.cs.toronto.edu/~kriz/cifar.html.
[24] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, K. Kay, A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence, Nature Neuroscience 25 (2022) 116–126.