Hierarchical semantic compression predicts texture selectivity in early vision

Mihály Bányai, Dávid G. Nagy, Gergő Orbán
{banyai.mihaly, nagy.g.david, orban.gergo}@wigner.mta.hu
Computational Systems Neuroscience Lab, MTA Wigner RCP, Budapest, Hungary

Abstract

Sensory processing produces hierarchical representations, which according to the semantic compression hypothesis, extract increasingly behaviorally relevant quantities from raw stimuli. Predictions of neural activity in hierarchical systems are most often made in supervised deterministic models, while probabilistic generative models provide a more complete unifying view of sensory perception. Whether unsupervised generative models trained on naturalistic stimuli give rise to representational layers of semantically interpretable quantities is yet unresolved, as is whether such representations can predict properties of neural responses in early vision. We use hierarchical variational autoencoders to learn a representation with graded compression levels from natural images, which exhibits variance according to perceptually relevant texture categories. We predict measures of neural response statistics by assessing the posterior distribution of latent variables in response to texture stimuli. Experimental results show that linearly decodable information about stimulus identity is lost in the secondary visual cortex while information is gained about texture type, which behavior is reproduced by the representational layers of our model. Deep generative models fitted to natural stimuli open up opportunities to investigate perceptual top-down effects, uncertainty representations along the visual hierarchy, and contributions of recognition and generative components to neural responses.

Keywords: visual perception; hierarchical inference; semantic compression

Introduction

Semantic compression in hierarchical systems Animals need to discard the bulk of information acquired through their sensory organs to obtain the tiny portion that will be used for present or future decisions in various tasks. The semantic compression hypothesis (Nagy, Török, & Orbán, 2018) proposes that the efficient way to lose information is to encode the stimulus through the latent variables in a model of the environment. In the ventral stream of the visual cortex, compression is realised in a hierarchical manner where the sequence of compression steps culminates in the recognition of objects and concepts where variance along complex variables such as pose, lighting, and scale are discarded. The Bayesian brain hypothesis suggests that successive layers of representation correspond to latent variables of a hierarchical generative model (Lee & Mumford, 2003). We propose that applying the semantic compression hypothesis to a hierarchical Bayesian model results in a sequence of representational layers that extract increasingly abstract descriptors of the observation from the statistical properties of the stimulus (Fig. 1). Consequently, representations will be invariant to increasingly complex transformations at each layer. For example, as depicted in Fig. 1, when presented by an animal fur pattern, conditioning on the lowest-level inferred latents we can generate the same pattern with different observation conditions (such as lighting), on mid-level latents, different samples from the same type of fur pattern, and on higher-level latents, different types of fur patterns. Measurements of auditory perception have shown the extraction of summary statistics from complex stimuli (McDermott, Schemitsch, & Simoncelli, 2013) in a way compatible with semantic compression. Here we aim to show that unsupervised hierarchical models also extract semantically relevant latent variables in the visual domain.

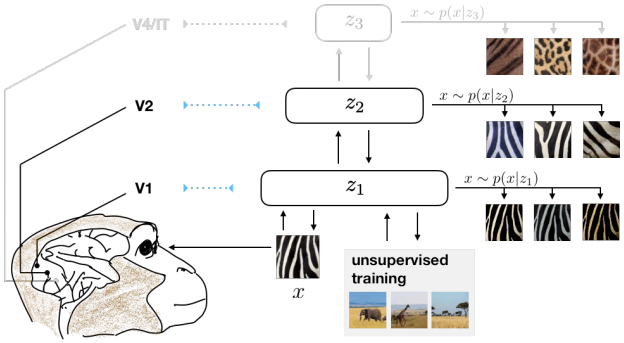

Figure 1: Vision as hierarchical inference. A probabilistic generative model with multiple layers of latent variables ($z_{1..3}$) is trained to compress natural images efficiently (middle). When a specific stimulus (x) is presented to the model, we can infer the corresponding latent representation in each layer. We can condition the generative model to the inferred representation on any specific layer and generate samples from the model that keep the information represented at the given layer and sample lower-level details from the learned distributions (right). Layers of representation in the model can be used to make predictions about measurements from different areas of the ventral stream of the visual cortex (left).

Assessing representations in the visual hierarchy What quantities influence the activity of neurons in various parts of the visual hierarchy beyond the primary visual cortex (V1) is a question far from settled. Recent studies characterise mid-level sensitivities in the secondary visual cortex (V2) (Ziemba, Heeger, Simoncelli, Movshon, & Freeman, 2013). How such

sensitivities constitute a representation can be defined in multiple ways, the simplest of which is linear decodability. Successive processing areas implementing increasing linear decodability of behaviourally relevant quantities is proposed in the hypothesis of representational untangling in higher-level visual areas (DiCarlo, Zoccolan, & Rust, 2012). Furthermore, recent evidence suggests that information related to texture categories is available linearly in V2 but not in V1, while linear information about the stimulus identity available linearly in V1 is lost at V2 (Ziemba, Freeman, Movshon, & Simoncelli, 2016), indicating different degrees of compression being implemented in the early visual hierarchy as well. Here we set out to investigate if semantic compression in a generative model trained on natural images reproduces this signature.

Predicting neural responses in hierarchical systems Recent studies demonstrated impressive performance on predicting neural activity in hierarchical systems (Yamins et al., 2014; Cadena et al., 2019). These models rely on feed-forward deep networks trained to classify images. However, recent evidence suggests that increasing predictive performance will require the consideration of top-down effects (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019), which have also been shown to play a role in cortical computations experimentally (Lee & Nguyen, 2001; Bányai et al., 2019). Probabilistic generative models are well suited to describe such effects (Lee & Mumford, 2003), and have been used to predict response statistics in early vision (Coen-Cagli, Dayan, & Schwartz, 2012; Orbán, Berkes, Fiser, & Lengyel, 2016; Bányai et al., 2019). As opposed to feed-forward networks, hierarchical generative models are trained in an unsupervised way, trying to learn the distribution of inputs as well as possible given capacity constraints instead of trying to perform a specific task well, which is exactly what we expect different layers of representations to do if they are to compress inputs to different degrees. The question of whether semantic compression is a product of task-training or obtainable in an unsupervised way remains open.

There are a number of architectural choices one has to make when building a generative model. The lowest level of visual cortical representations is suggested to be close to linear by Olshausen and Field (1996), formulated as a generative model by Barello, Charles, and Pillow (2018). Beyond V1 we obviously need nonlinear computations, but the constraints on the kind of generative model that would capture this computation are not well characterised, thus warranting the application of a generic machine learning model implementing hierarchical inference. In this study we propose a hierarchical probabilistic generative model fitted to natural stimuli, producing multiple layers of increasingly compressed representations, suitable to make predictions about statistical properties of neural activity in visual cortical areas in response to specific images.

Methods

Texture families to probe layers of representation In order to probe hierarchical representations, we need stimulus sets of compositional nature, such that low-level local features of the image are organized to define an abstract property for the stimulus, which can be treated as a categorical label. Recent results suggest that texture is a relevant abstraction for the secondary visual cortex (Ziemba et al., 2013). Texture images can be synthesised using photographs of natural textures (Portilla & Simoncelli, 2000), enabling us to produce a large number of samples from the same texture family (Fig. 2A). Such texture families are well suited to test semantic compression through linear decodability, since the average image of a family is always zero, all family-specific information being present in higher-order pixel statistics, as opposed to e.g. the digit categories of MNIST which are decodable linearly from the pixel space (Fig. 2B).

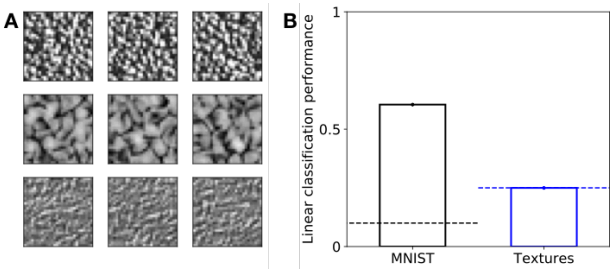

Figure 2: **A.** Samples from three texture families used to evaluate our models. **B.** The digit categories of MNIST are linearly decodable from the data space, while texture families are not characterised by different average pixel patterns, thus are not decodable linearly. Dashed lines indicate chance values, while bars and error bars represent mean and S.E.M. of cross-validation folds respectively.

Hierarchical variational autoencoders Bayesian inference in hierarchical models is computationally intensive, thus the brain is expected to implement efficient approximate solutions. Variational methods use tractable distributions to infer the posterior of latent variables. Recognition-generative models, such as variational autoencoders (VAE) use an explicit feed-forward model to implement variational inference (Dayan, Hinton, Neal, & Zemel, 1995; Kingma & Welling, 2013).

VAEs have been used to describe semantic compression using a capacity parameter to balance the fidelity and the bandwidth of the latent representation (Alemi et al., 2017). They have been demonstrated to capture the abstraction of the digit category in MNIST, but not in more complex categorical stimuli (Fertig, Arbabi, & Alemi, 2018).

A natural extension of the VAE model family is to define hierarchical layers of latent representations in order to capture the stimulus statistics at different levels of abstraction, such as in Fig. 1. Learning such hierarchical representations is a nontrivial problem, for which multiple proposed solutions ex-

ist. One of these is the Ladder Variational Autoencoder (LVAE) (Sønderby, Raiko, Maaløe, Sønderby, & Winther, 2016), which uses a direct feed-forward mapping from the stimulus to all latent layers during inference, allowing for the efficient learning of latent hierarchies while introducing no additional computational steps into the generative model. We used LVAEs as models of the representational hierarchy in the ventral stream, fitting them to naturalistic stimulus statistics and then presenting them texture stimuli to compare properties of the inferred representations to those measured from the visual cortex.

Results

Comparison of model and measurement We fitted a two-layer LVAE to whitened natural image patches of 16x16 pixels obtained from the van Hateren dataset (Hateren & Schaaf, 1998) (shown in Fig. 3A). The architecture consisted of 20 and 5 stochastic units in the two latent layers. The lower level representation was connected to the stimulus through a linear encoder and decoder. The second layer was connected to the first using two densely connected ReLU layers of 32 units each and a batch normalization layer both in the encoder and the decoder. The observation noise was fixed at 0.1. The parameters of the model were fitted to the natural patches using the Adam optimizer for 60 epochs with a learning rate of 0.001, using a burn-in period (Sønderby et al., 2016) of 10 epochs.

As we wanted to compare the properties of the learned representations to those measured in macaques by Ziemba et al. (2016), we inferred the latent representation of texture stimuli shown in Fig. 2A. We constructed linear mixture of Gaussian decoders both to distinguish between the latent representations of specific stimuli and the families they were sampled from, using the representations from both latent layers of the LVAE. The performance of the decoders was calculated as the cross-validated hit rate for either 4 samples from the same family or 4 from different families. The performances were plotted against each other to contrast the properties of the representations learned in the two layers (Fig. 3B). The first layer could be used much better to recover the identity of the stimulus. Most of this information was lost at the second layer, while making the family more linearly decodable. This result is in accordance with the findings from macaque V1 and V2.

Semantic compression can be probed using nonlinear read-out instead of a linear one as well. We used the t-SNE nonlinear embedding method (Maaten & Hinton, 2008) to show that the second-layer representation of texture images is more clustered according to family membership (Fig. 3C), similarly to measurements from macaque V1 and V2.

We explored which architectural choices are essential to produce the results we demonstrate. An indispensable feature of the model is the increasingly compressed representation in the layers. However, the compression levels can be achieved by controlling the information capacity of the layers in multiple ways, such as the dimensionality of the layers, but also the expressive power of the encoders and decoders used in them. The latter property opens up an avenue to train models of much higher latent dimensions with similar semantic compression properties as ours.

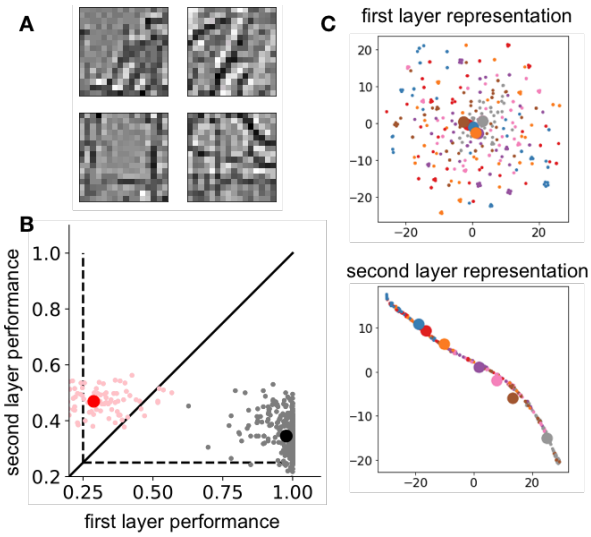

Figure 3: Representations learned by a two-layer LVAE from natural images. **A.** Whitened natural patches used for training the model. **B.** The decodability of the stimulus identity of texture stimuli of the kind shown in Fig. 2 (grey dots) and the family they are sampled from (pink dots). Red and black dots represent the mean of all the decoding comparisons and dashed lines represent chance levels. The representation learned in the first layer of the LVAE contains more information about the identity of the stimulus, while the second layer contains more about the family. Cf. Figure 5B of Ziemba et al. (2016). **C.** T-SNE plot of texture stimuli as represented in the two layers of the LVAE. Colors indicate the family of each sample, large dots indicate the mean of each family. Samples from the same family are more clustered together in the second layer. Cf. Figure 4 of Ziemba et al. (2016).

Visualising semantic compression The learned representations that reproduce experimentally measured untangling effects are expected to compress the stimuli at different semantic levels, as in Fig 1. Since we learn a model of natural images, a high number of latent units would be necessary to learn all the factors of variance that include the ones directly relevant to texture samples, making the levels of variation easily observable visually. Instead of training such a model, we retrain an LVAE using the texture stimuli, directly producing the subset of latents that describe these stimuli in particular. We then infer the latent representation in each layer in response to specific textures, and condition the generative model on the inferred representation in each layer. We indeed observe that conditioning on the lower layer produces samples that reproduce the particular content of the input image and differ only due to the observation noise which is indepen-

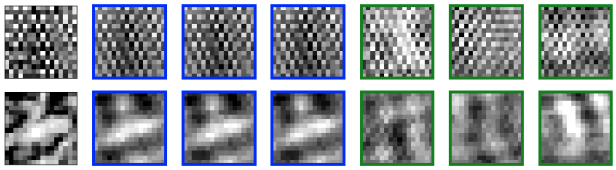

Figure 4: Layer-conditional reconstructions from a two-layer LVAE trained on texture stimuli. For two example stimuli (no border) we take three samples from the posterior distribution of all latent variables, and generate synthetic stimuli conditioning on the latent samples in the first (blue border) and the second (green border) layer. First-layer generated samples differ only in terms of pixel noise, while second-layer samples are different instances of the same texture family.

dent across pixels. Conversely, conditioning the generative model on the higher layer produces samples that come from the same texture family as the input, but vary in terms of the particular realisation of the texture.

Conclusions

Variational autoencoders implement hierarchical Bayesian inference producing a series of increasingly compressed representation of the input. When trained on natural images in an unsupervised way, they reproduce representational untangling of texture stimuli similarly to the visual cortex of macaques, while the learned generative model exhibits variations in the successive representational layers corresponting to perceptual categories.

Acknowledgments

This work has been supported by the National Research, Development and Innovation Fund of Hungary (Grant No. K125343), the Hungarian Brain Research Program (2017-1.2.1-NKP-2017-00002), and a Human Frontier Science Program grant (GO).

References

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*.

Bányai, M., Lazar, A., Klein, L., Klon-Lipok, J., Stippinger, M., Singer, W., & Orbán, G. (2019). Stimulus complexity shapes response correlations in primary visual cortex. *PNAS*, *116*(7), 2723–2732.

Barello, G., Charles, A., & Pillow, J. (2018). Sparse-Coding Variational Auto-Encoders. *bioRxiv*(1996), 1–19.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS Comp Biol*, *15*(4), e1006897.

Coen-Cagli, R., Dayan, P., & Schwartz, O. (2012, mar). Cortical Surround Interactions and Perceptual Salience via Natural Scene Statistics. *PLoS Comp Biol*, *8*(3), e1002405.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. S. (1995, sep). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434.

Fertig, E., Arbabi, A., & Alemi, A. A. (2018). β-vaes can retain label information even at high compression. *arXiv preprint arXiv:1812.02682*.

Hateren, J. H. v., & Schaaf, A. v. d. (1998, Mar). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings: Biological Sciences*, *265*(1394), 359-366.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral streams execution of core object recognition behavior. *Nat Neurosci*, 1.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*, *20*(7), 1434.

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in the early visual cortex. *PNAS*, *98*(4), 1907–1911.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *JMLR*, *9*(Nov), 2579–2605.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nat Neurosci*, *16*(4), 493–U169.

Nagy, D. G., Török, B., & Orbán, G. (2018). Semantic compression of episodic memories. *arXiv preprint arXiv:1806.07990*.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*, *92*(2), 530–543.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.*, *40*(1), 49–70.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). Ladder variational autoencoders. In *Adv neural inf process syst* (pp. 3738–3746).

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016, may). Selectivity and tolerance for visual texture in macaque V2. *PNAS*, 201510847.

Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., Movshon, J. A., & Freeman, J. (2013). A functional and perceptual signature of the second visual area in primates. *Nat Neurosci*, 1–12.