Categorical Perception: A Groundwork for Deep Learning

Laurent Bonnasse-Gahot, Jean-Pierre Nadal

Categorical Perception: A Groundwork for Deep Learning

Laurent Bonnasse-Gahot[1,*] and Jean-Pierre Nadal[1,2]

(1) Centre d'Analyse et de Mathématique Sociales
CAMS, UMR 8557 CNRS-EHESS
École des Hautes Etudes en Sciences Sociales
54 bd. Raspail, 75006 Paris, France
(2) Laboratoire de Physique de l'ENS
LPENS, UMR 8023, CNRS - ENS Paris - PSL University - SU - Université de Paris
École Normale Supérieure
24 rue Lhomond, 75005 Paris, France

(*) Corresponding author (lbg@ehess.fr)

Abstract

Classification is one of the major tasks that deep learning is successfully tackling. Categorization is also a fundamental cognitive ability. A well-known perceptual consequence of categorization in humans and other animals, called categorical perception, is characterized by a within-category compression and a between-category separation: two items, close in input space, are perceived closer if they belong to the same category than if they belong to different categories. Elaborating on experimental and theoretical results in cognitive science, here we study categorical effects in artificial neural networks. Our formal and numerical analysis provides insights into the geometry of the neural representation in deep layers, with expansion of space near category boundaries and contraction far from category boundaries. We investigate categorical representation by using two complementary approaches: one mimics experiments in psychophysics and cognitive neuroscience by means of morphed continua between stimuli of different categories, while the other introduces a categoricality index that quantifies the separability of the classes at the population level (a given layer in the neural network). We show on both shallow and deep neural networks that category learning automatically induces categorical perception. We further show that the deeper a layer, the stronger the categorical effects. An important outcome of our analysis is to provide a coherent and unifying view of the efficacy of different heuristic practices of the dropout regularization technique. Our views, which find echoes in the neuroscience literature, insist on the differential role of noise as a function of the level of representation and in the course of learning: noise injected in the hidden layers gets structured according to the organization of the categories, more variability being allowed within a category than across classes.

Keywords: deep learning; categorical perception; dropout; neuronal noise; categoricality index; Fisher information; mutual information

1 Introduction

One of the main tasks tackled with deep learning methods is the one of categorization: classification of images as representing specific objects, identification of words for speech recognition, etc (see LeCun et al., 2015; Schmidhuber, 2015, for reviews). The identification of categories is as well a central topic in cognitive science. In his book "Categorical Perception: The Groundwork of Cognition", Harnad (1987) shows how central and fundamental to cognition categorization is. A well-studied perceptual consequence of categorization in humans and other animals is characterized by greater cross-category than

within-category discrimination, a phenomenon called categorical perception (CP, see also Repp, 1984, for a review). It originates in the field of speech perception: the seminal work of Liberman et al. (1957) demonstrated that American English-speaking subjects are better at discriminating between a /ba/ and a /da/ than between two different /ba/ tokens, even when the magnitude of the physical difference between the two stimuli is equal. Authors have subsequently shown that such effect is not specific to English, occurring in any language, but with respect to its own structure, as distinct phonemic systems entail different discrimination abilities (Abramson and Lisker, 1970; Goto, 1971). Categorical perception was also found to be not specific to speech (Cross et al., 1965; Burns and Ward, 1978; Bornstein and Korda, 1984; Goldstone, 1994; Beale and Keil, 1995), nor even to human (Nelson and Marler, 1989; Kluender et al., 1998; Caves et al., 2018).

Previous computational work have shown that categorical perception also happens in artificial neural network (Anderson et al., 1977; Padgett and Cottrell, 1998; Tijsseling and Harnad, 1997; Damper and Harnad, 2000). A noisy neural classifier, either biological or artificial, that aims at minimizing the probability of misclassifying an incoming stimulus, has to deal with two sources of uncertainty. One is due to the intrinsic overlap between categories (in the relevant case where classifying these stimuli is not an obvious task). The other one stems from the variability of the response of the neurons to a stimulus. Intuitively, one might want to reduce the neural uncertainty in the regions of the input[1] space where the overlap between categories is already a major source of confusion, *ie* the regions of boundaries between classes. In Bonnasse-Gahot and Nadal (2008), taking an information theoretic approach, in the context of a biologically motivated neural architecture with a large number of coding cells, we quantitatively show how these two quantities interact precisely, and how as a consequence category learning induces an expansion of the stimulus space between categories, *ie* categorical perception. Reducing neuronal noise in the region where the chance of misclassifying a new stimulus is highest, by mistakenly crossing the decision boundary, thus lowers the probability of error. Based on our previous studies of categorical perception (Bonnasse-Gahot and Nadal, 2008, 2012), here we introduce a framework for the analysis of categorization in deep networks. We analyze how categorization builds up in the course of learning and across layers. This analysis reveals the geometry of neural representations in deep layers after learning. We also show that categorical perception is a gradient phenomenon as a function of depth: the deeper, the more pronounced the effect – from the first hidden layer that might show no specific effect of categorical perception to the last, decisional, layer that exhibit full categorical behavior.

So far we mentioned noise as a nuisance that we have to deal with, following the common signal processing tradition. However, many works have conversely shown that neuronal noise, whether small or large depending on the context, can be desirable, helping a system to learn more efficiently, more robustly, with a better generalization ability. In the field of artificial neural network, many studies in the 1990s have shown that input noise helps a network to overcome overfitting (see e.g. Holmstrom and Koistinen, 1992; Matsuoka, 1992; Bishop, 1995; An, 1996). Authors have also shown that noise can have the effect of revealing the structure of the data, as for learning a rule from examples (Seung et al., 1992) or in independent component analysis (Nadal and Parga, 1994), a fact recently put forward by Schwartz-Ziv and Tishby (2017) within the framework of the information bottleneck approach. One important ingredient in the Krizhevsky et al. (2012) paper that ignited the recent revival of connectionism in the past decade is the use of multiplicative Bernoulli noise in the hidden activations, a technique coined dropout (Srivastava et al., 2014). Dropout consists in randomly dropping out a proportion of hidden nodes during training. Bouthillier et al. (2015) proposed that dropout is somewhat equivalent to data augmentation. If we adopt this viewpoint, our analysis of categorical perception in artificial networks leads us to suggest that a key ingredient making dropout beneficial is that this augmentation is not uniform across the input space, but depends on the structure and relation between categories: dropout allows for more variability within category than between classes, which is exactly what one would like to achieve. Crucially, our analysis allows to more generally understand noise in any given layer as being dependent on the current neural representation, hence as not having the same benefit over the course of learning.

The paper is organized as follows. In Section 2, we review empirical and theoretical results on categorical perception, and explain how these motivate the present study of categorization in artificial neural networks. In Section 3, we conduct computational experiments that confirm the relevance of the study of categorical perception in the understanding of artificial neural networks. We consider several

[1] Throughout this paper, we will interchangeably use the terms 'stimulus' and 'input'.

examples gradually going up in task complexity and network sizes. Using illustrative one- and two-dimensional toy examples, we first empirically confirm in Section 3.1 that after learning the variance of estimates of a sample drawn between categories is lower than when drawn within category. Moving to the MNIST dataset (LeCun et al., 1998), a commonly used large database of handwritten digits, we show in Section 3.2 how categorical perception emerges in an artificial neural network that learns to classify handwritten digits. In Section 3.3 we extend our analysis to very deep networks trained on natural images, working with the Kaggle Dogs vs. Cats dataset, and with the ImageNet dataset (Deng et al., 2009; Berg et al., 2011). For our analysis of multi-layer networks trained on large databases, we consider two ways of extracting and visualizing the categorical nature of the encoding resulting from learning. The first one consists in mimicking a standard experimental protocol used for the study of categorical perception in human and other animals. We generate smooth continua interpolating from one category to another, and we look at the structure of the neural activity as one moves along a continuum. We show how, following learning, the stimulus space is enlarged near the boundary between categories, thus revealing categorical perception: distance between the neural representations of two items is greater when these items are close to the boundary between the classes, compared to the case where they are drawn from a same category. We also show that these categorical effects are more marked the deeper the layer of the network. The second one consists in measuring a categoricality index which quantifies, for each layer, how much the neural representation as a whole is specific to the coding of categories. We show that categoricality gets greater following learning, and increases with the depth of the layer, paralleling the results found by means of the morphed continua. We also observe that convolutional layers, while containing categorical information, are less categorical than their dense counterparts. Finally, in Section 4, all these results allow us to discuss many heuristics practices in the use of dropout, notably in terms of the amount of noise used as a function of layer depth, or the difference of amount of such noise in dense vs. convolutional layers. We also suggest that our results might in turn shed light on the psychological and neuroscientific study of categorical perception. We provide technical details in the Material and Methods section and supplementary information in Appendices. In particular, for completeness we give in Appendix A a synthetic view of the results in Bonnasse-Gahot and Nadal (2008, 2012) on the modeling of categorical perception that are relevant for the present paper.

2 Insights from cognitive science

2.1 Categorical perception: empirical and theoretical studies

In the psychological and cognitive science literature, a standard way to look at categorical perception is to present different stimuli along a continuum that evenly interpolate between two stimuli drawn from two different categories. Let us consider a few illustrative examples. In their seminal study on speech perception, Liberman et al. (1957) generated a /ba/–/da/ synthetic speech continuum by evenly varying the second formant transition, which modulates the perception of the place of articulation. Studying categorical perception of music intervals, Burns and Ward (1978) considered a continuum that interpolates from a minor third to a major third to a perfect fourth. Turning to vision, Bornstein and Korda (1984) considered a blue to green continuum, while Goldstone (1994) made use of rectangles that vary either in brightness or in size. In Beale and Keil (1995), the authors generated a continuum of morphed faces interpolating between individual exemplars of familiar faces, from Kennedy to Clinton for instance. As final example, we cite the monkey study by Freedman et al. (2001) that makes use of a dog to cat morphed continuum. In this kind of studies, experimentalists typically measure category membership for all items along the considered continuum, discrimination between neighboring stimuli, as well as reaction times during categorization task. Of interest for the present paper are the behaviours in identification and discrimination tasks observed in these experiments. Although the physical differences in the stimuli change in a continuous way, the identification changes abruptly in a narrow domain near the categories boundary, while discrimination is better near the category boundary than well inside a category, which is the hallmark of categorical perception. If the neural representations have been optimized for the identification of categories, as e.g. in speech perception, these behavioral performances are intuitive: the goal being to decide to which category the stimulus belongs to, far from a category boundary in stimulus space there is no need to have a precise identification of the stimulus itself: two nearby stimuli may not be discriminated. On the contrary, at the vicinity of a boundary, the likelihood of a category is strongly affected by any small shift in stimulus space. Thus, one expects the neural code, if optimized in view of the categorization task, to provide a finer representation of the stimuli near a class boundary than within a category.

These arguments have been formalized and made quantitative through modeling of the neural processing. Most works consider a simplified architecture based on neuroscience data: a coding layer with a distributed representation, and a decision layer. The coding layer represents the last encoding stage before the decision can be taken, and may correspond to a high level in the neural processing. Hence the input to this layer is not the actual stimulus but some projection of it on some relevant dimensions. In the empirical studies, authors find that in the coding layer no single cell carries categorical information, whereas in what seems the decisional layer, single cells are specific to a single category (see e.g. Kreiman et al., 2000; Freedman et al., 2001; Meyers et al., 2008). Taking a Bayesian and information theoretic approach, in previous studies (Bonnasse-Gahot and Nadal, 2008, 2012) we show that efficient coding (targeting optimal performance in classification) leads to categorical perception (in particular better discrimination near the class boundaries). More precisely (see Appendix A for more details), we show that, in order to optimize the neural representation in view of identifying the category, one should maximize the mutual information between the categories and the neural code. This in turn, in the limit of a large number of coding cells, implies that one has to minimize a coding cost which, in qualitative terms, writes:

$$\overline{\mathcal{C}}_{\text{coding}} = \frac{1}{2} \left\langle \frac{\text{categorization uncertainty}}{\text{neural sensitivity}} \right\rangle \tag{1}$$

where the brackets $< . >$ denote the average over the space of relevant dimensions x. In formal terms,

$$\overline{\mathcal{C}}_{\text{coding}} = \frac{1}{2} \int \frac{F_{\text{cat}}(x)}{F_{\text{code}}(x)} \, p(x) \, dx \tag{2}$$

where $F_{\text{code}}(x)$ and $F_{\text{cat}}(x)$ are Fisher information quantities. The quantity $F_{\text{code}}(x)$ represents the sensitivity of the neural code to a small change in stimulus x. Larger Fisher information means greater sensitivity. A crucial remark here is that the inverse of the Fisher information is an optimal lower bound on the variance σ_x^2 of any unbiased estimator of x from the noisy neural activity (Cramér-Rao bound, see e.g. Blahut, 1987):

$$\sigma_x^2 \;\geq\; \frac{1}{F_{\text{code}}(x)} \tag{3}$$

The quantity $F_{\text{cat}}(x)$ represents the categorical sensitivity, that is how much the probability that the stimulus belongs to a given category changes for small variation in x values. Each Fisher information quantity defines a metric over the space x (more exactly, over spaces of probabilities indexed by x). Along a path in stimulus space, F_{cat} quantifies the change in categorical specificity of x, and F_{code} how much the neural activity changes locally. Depending on the constraints specific to the system under consideration, minimization of the cost leads to a neural code such that $F_{\text{code}}(x)$ is some increasing function of $F_{\text{cat}}(x)$. For some constraints one gets $F_{\text{code}}(x) \propto F_{\text{cat}}(x)$ as optimum, but other constraints may lead to other relationships – see Bonnasse-Gahot and Nadal (2008, 2020); Berlemont and Nadal (2020). Efficient coding with respect to optimal classification is thus obtained by essentially matching the two metrics. Since F_{cat} is larger near a class boundary, this should also be the case for $F_{\text{code}}(x)$. A larger $F_{\text{code}}(x)$ around a certain value of x means that the neural representation is stretched at that location. The neural representation paves the space x more finely near than far from the class boundaries. Thus, category learning implies better cross-category than within-category discrimination, hence the so-called *categorical perception*. In a companion paper (Bonnasse-Gahot and Nadal, 2020), we show how the above analysis can actually be extended to multi-layer networks, making explicit that x corresponds to the space of relevant dimensions on which the network projects the stimulus.

2.2 A framework for the study of categorization in artificial neural networks

In biologically motivated models, neural noise is ubiquitous. In the simplest setting, neural spiking activity is described by a Poisson process. Hence, at each instant of time, in a feedforward architecture the input pattern from a layer to the next one has 'missing data'. If we consider rates instead of spikes, one has an activity with a multiplicative noise, such as Poisson noise, very much in the same way as during a run of the dropout heuristic. If we consider that the noise in the neural activity has the effective effect to generate new inputs, we are led to the data augmentation view taken by Bouthillier et al. (2015). However, we see here that this augmentation is not uniform across stimulus space, contrary to what *input* noise would have yielded. Indeed, Cramér-Rao bound, as given by Eq. 3, shows that regions that are within-category allow for more variability, whereas within cross-category regions, the acceptable

variability is much more constrained in order to avoid generating a new input with a wrong label. One key idea motivating the present study is that category learning may precisely have the effect of modulating the noise (e.g. dropout) level efficiently, and this is confirmed by our results presented in this paper.

In the following, we study the building of categorical information in ANN making numerical experiments with a protocol inspired by the experimental approaches mentioned above. We assess the discrimination ability of each layer of an artificial neural network along a continuum of stimuli by considering the distance in neural space between contiguous elements (see Materials and Methods, paragraph "Neural distance", for the precise definition). More distant stimuli are easier to discriminate than closer ones. We make use of this neural distance as a proxy for the Fisher information, as it similarly quantifies how much the neural activity change in average with respect to small variations in the input x. However, contrary to Fisher information, it is straightforward to compute. Note that this quantity reflects sensitivity at the population level (a given layer in this study), and not at a single neuron level. Beside the use of continua, as an alternative and complementary approach to investigate categorical effects, we also consider the measure of a *categoricality* index. Authors have proposed different measures of categoricality, either at the single neuron level (Kreiman et al., 2000; Freedman et al., 2001), or at the population level (Kriegeskorte et al., 2008; Kreiman et al., 2000). The higher this index, the greater the intra class similarity and inter class dissimilarity. Here, we quantify categoricality at the population level. To do so, our choice of categoricality index consists in comparing the distributions of the distances in neural space, as given by the activations in each hidden layer, between items drawn from a same category vs. items drawn from two different categories (see Materials and Methods, paragraph "Categoricality index", for details).

3 Results

In this section we present numerical experiments on classification tasks of increasing difficulty, working with a variety of datasets, from a simple one-dimensional example with two categories to a case that involves natural images with one thousand categories. We consider neural network architectures of complexity congruent with the ones of the tasks, allowing to explore various conditions: multi-layer perceptrons and convolutional networks, a wide range of depths (from one hidden layers to ten or more hidden layers), and learning with different types of multiplicative noise (Bernoulli noise as in dropout, or Gaussian noise as in Gaussian dropout).

3.1 Experiments with toy examples

3.1.1 One dimensional example

We consider a one dimensional input space with two Gaussian categories. The neural network is a multi-layer perceptron with one hidden layer of 128 cells, with sigmoid activation, subject to Gaussian dropout with rate 0.5 – i.e. multiplicative Gaussian noise with standard deviation 1.0. As done in Bouthillier et al. (2015), for a given input x, and a given noisy neural activity $\mathbf{r}$, we compute the estimate $\widehat{x}$ that would have produced an activity as close as possible of $\mathbf{r}$ without noise (details in the Materials and Methods section). In other words, presenting many times the same input to the noisy network is somewhat equivalent to presenting new inputs to a noiseless version of this network.

We present in Figure 1, left panel, the variance of these generated data-points at each point along the x continuum after learning. As expected, the variance is lower at the boundary between categories. In Figure 1, right panel, we present the neural distance (as introduced above) between contiguous inputs that are evenly distributed in stimulus space. We see that the behaviour of this quantity parallels the one of the variance in the left panel.

3.1.2 Two dimensional example

Similarly, we present in Figure 2 a two dimensional example with two Gaussian categories. The neural network is a multi-layer perceptron with one hidden layer of 128 cells, with ReLU activation, subject to dropout with rate 0.3 (proportion of the input units to drop, ie multiplicative noise drawn from Bernoulli distribution with $p = 0.3$). We see once again that the variance of the generated virtual inputs is lower at the boundary between the two categories. Here, we also see how these individual estimates $\widehat{\mathbf{x}}(\mathbf{r})$

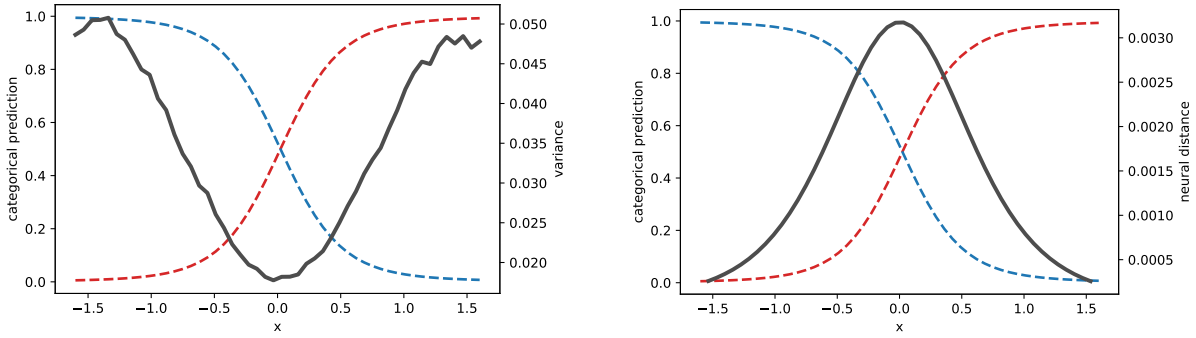

Figure 1: **One dimensional example with two Gaussian categories**, respectively centered in $x_{\mu_1} = -0.5$ and $x_{\mu_2} = +0.5$, with variance equal to 0.25. For both panels, the dotted colored lines indicate the posterior probabilities $P(\mu|x)$. (Left) The dark solid line corresponds to the variance of the distribution of the estimated input $\widehat{x}$ for $n = 10000$ probabilistic realizations of the neural activity given an input x. (Right) The dark solid line corresponds to the distance in the neural space between contiguous stimuli.

are distributed around each input $\mathbf{x}$. As expected, we observe that the variance is not isotropic: it is larger in the direction that is safe from crossing the boundary, and much smaller in the direction that is orthogonal to the decision boundary.

3.2 Experiments with handwritten digits

In this section we move beyond the two simple previous examples to look at the MNIST dataset (LeCun et al., 1998). It is a database of handwritten digits that is commonly used in machine learning, with a training set of 60,000 images and a test set of 10,000 images. Here the goal is to look at categorical perception effects in neural networks that learn to classify these digits.

3.2.1 Creation of an image continuum

We want to build sequences of images that smoothly interpolate between two items of different categories. Directly mixing two stimuli in input space cannot work, as it would just superimpose the two original images. In cognitive experiments, one builds such artificial stimuli keeping each new stimulus as a plausible stimulus for our perception, as in the cases mentioned Section 2.1. One may think of various methods to generate sequences of images. In the present work, we want to obtain sequences that somehow remain in the space generated by the database itself. Taking inspiration from the work of Bengio et al. (2013) to create an image continuum, we used an autoencoder trained to reproduce single digits from the MNIST training set (see Materials and Methods for algorithmic details). The autoencoder consists in an encoder and a decoder. The first part learns to build a compressed representation of the input. The second part learns to reconstruct the original input from this compressed representation. This representation projects the data onto a lower dimensional space that meaningfully represents the structure of the data. By interpolating between two stimuli in this space, then reconstructing the resulting image thanks to the decoder, we obtain a continuum that nicely morphs between two stimuli, along the nonlinear manifold represented by the data.

We provide in Figure 3 an example of such a generated continuum, using two digits from the MNIST test set, one '4' and one '9'. We choose here this example because the 4/9 classes are among the most confused classes, and the present work focuses on cases where the classification is not obvious.

3.2.2 Peak in discrimination at the boundary between categories

We consider here a multi-layer perceptron with two hidden layers of 256 cells, and look at the changes in representation before and after learning on the MNIST database. In this example, we investigate the behavior of the network with respect to the continuum shown in Fig. 3.

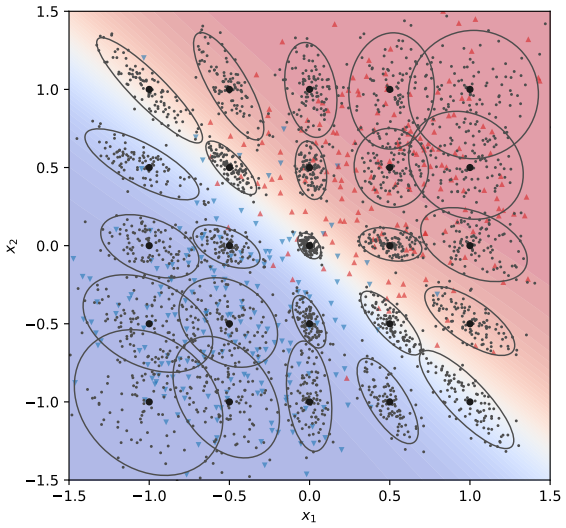

Figure 2: **Two dimensional example with two Gaussian categories**, respectively centered in $\mathbf{x}_{\mu_1} = \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}$ and $\mathbf{x}_{\mu_2} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, with covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}$. Individual training examples drawn from these multivariate Gaussian distributions are indicated as downward blue triangles (category 1) and upward red triangles (category 2). The background color indicates the posterior probabilities $P(\mu|\mathbf{x})$, from blue (category 1) to red (category 2) through white (region between categories), as found by the network. The largest dark dots correspond to a 5×5 grid of stimuli paving the input space between -1.0 and 1.0 in both dimensions. For each one of these inputs, we computed the estimates $\widehat{\mathbf{x}}(\mathbf{r})$ for $n = 100$ probabilistic realizations of the neural activity. We represent these estimates as smaller gray dots, circled for each input by an ellipse that pictures the 2σ confidence ellipse.

Figure 3: **Example of an image continuum**, smoothly interpolating between digits '4' and '9', using the MNIST dataset.

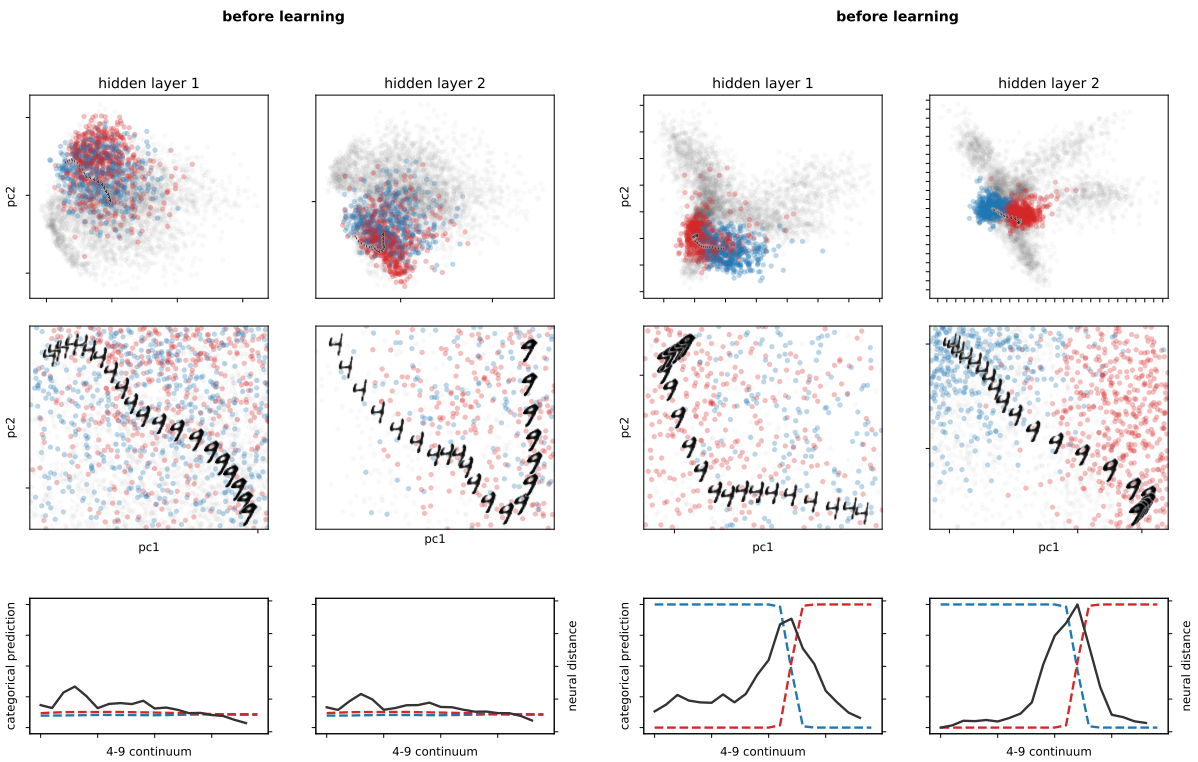

Figure 4: **Changes in the neural representation following learning of categories**: example on a '4' to '9' continuum, using the MNIST dataset. The neural network is a multi-layer perceptron with two hidden layers of 256 cells. (Left) Representation before learning. (Right) Representation after learning. (Top row) Two-dimensional PCA projections based on the activations of the hidden layers on the test set. Items from category '4' are colored in blue, while items from category '9' are colored in red. The rest of the test set is represented in gray. For a lighter representation, only one every two data points is shown. One specific '4' to '9' continuum, connecting two items from the test set is represented in black. (Middle row) Same, zoomed in on the '4' to '9' continuum. (Bottom row) The dotted colored lines indicate the posterior probabilities of predicting category '4' (blue) or '9' (red) along the continuum. The dark solid line indicates the neural distance between adjacent items along the continuum. The scale along the y-axis is shared across conditions.

We summarize the results in Figure 4, for each of the two hidden layers, before and after learning. Before training (left panel), the representation of all 4 and 9 digits are not well separated. If one looks at the neural distance between items along the 4–9 continuum (bottom row), we see that it is rather flat. Conversely, following training (right panel), the two categories are now well separated in neural space. The neural distance between items along the specific 4–9 continuum presents a clear peak at the decision boundary, thus exhibiting categorical perception. We can also already notice that the deeper hidden layer exhibits a more categorical representation.

3.2.3 Gradient categorical perception as a function of depth

Beyond the particular example of this 4-9 continuum, we now look at the pattern of discriminability along many similar continua that interpolates between stimuli from different categories (see Materials and Methods). For each pair of stimuli drawn from two different categories, we computed the neural distance between neighboring stimuli along the interpolating continuum, for each hidden layer in a multi-layer perceptron. As the category might not cross at the same point along the continuum, we aligned all these curves by centering them around the point where the two posterior probabilities cross. As we are interested in looking at the effect of depth, the neural network considered here is a multi-layer perceptron with three hidden layers. We present the results in Figure 5. We first observe that all layers exhibit categorical perception: space is dilated at the boundary between categories and warped within a category. Moreover, we see that the deeper the layer the more pronounced the effect.

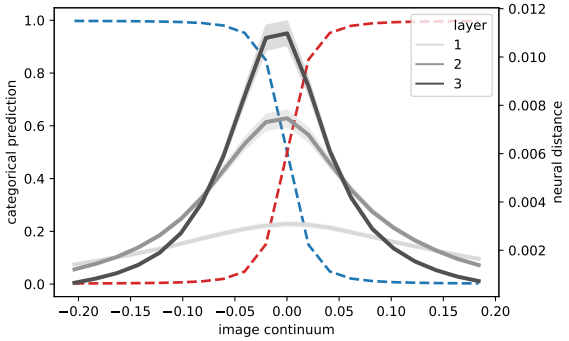

Figure 5: **Gradual categorical perception across layers**: the deeper the layer, the more pronounced the categorical perception effect. The neural network is a multi-layer perceptron with three hidden layers of 256 cells, trained on the MNIST dataset. The dotted colored lines indicate the mean posterior probabilities. For each hidden layer, the solid line corresponds to the mean neural distance between adjacent items, averaged over several continua, and aligned with the boundary between the two classes (error bars indicate 95% confidence intervals, estimated by bootstrap).

3.2.4 Categoricality as a function of depth and layer type

We now turn to a second method for characterizing how much the neural code is specific to the categorization task, making use of the categoricality index mentioned Section 2.2. We recall that this index quantifies the degree of relative intra-class compression vs. inter-class expansion of the neural representation provided by a given layer (see Materials and Methods). A certain level of categoricality implies that some information is lost about individual items far from category boundaries. According to the information processing theorem (Blahut, 1987), feedforward processing cannot increase information, hence the information loss will not be recovered. Thus we expect categoricality to increase with depth, and the issue is to characterize this increase.

In Figure 6 we compare the categoricality index before (broken line) and after (solid line) learning on the MNIST training set for two types of neural network: on the left, (a), a multi-layer perceptron with three hidden layers, and on the right (b), a convolutional neural network with seven hidden layers (see Materials and Methods for full details). Let us first consider the multi-layer perceptron. The results in Fig. 6a echo the one presented in Fig. 5: all layers present categorical effects, and, in line with recent findings (Alain and Bengio, 2016; Mehrer et al., 2020), categoricality increases with depth. Let us now turn to the convolutional neural network. The first convolutional layer does not have a larger categoricality index than an untrained network, meaning that the features it has learned are quite general and not yet specialized for the classification task at hand. The categoricality then increases with depth, the last hidden layer, a dense layer, presenting the largest value. Finally, comparing the two figures, we can see that the dense layers have a larger categoricality index than the convolutional layers, even for the first layer.

3.3 Experiments with natural images

We now go one step further in task and network complexity by considering natural image databases and deeper networks with ten or more of hidden layers.

3.3.1 Categorical perception of a cat/dog continuum

In this section, we consider a deep convolutional neural network trained to classify natural images of cats and dogs. We investigate its behavior with respect to a continuum that interpolates between different cat/dog categories.
Let us first introduce the neural network, the database used for training and finally the continua that are considered (see Materials and Methods for details). The neural network is a convolutional network with three blocks of two convolutional layers and a max pooling layer, followed by a global pooling average layer (introduced in Lin et al., 2013, and that replaces the dense layers by reducing drastically the number of parameters, thus avoiding overfitting when dealing with not so large database such as the one considered here). We used Gaussian dropout during learning. We trained each network instance

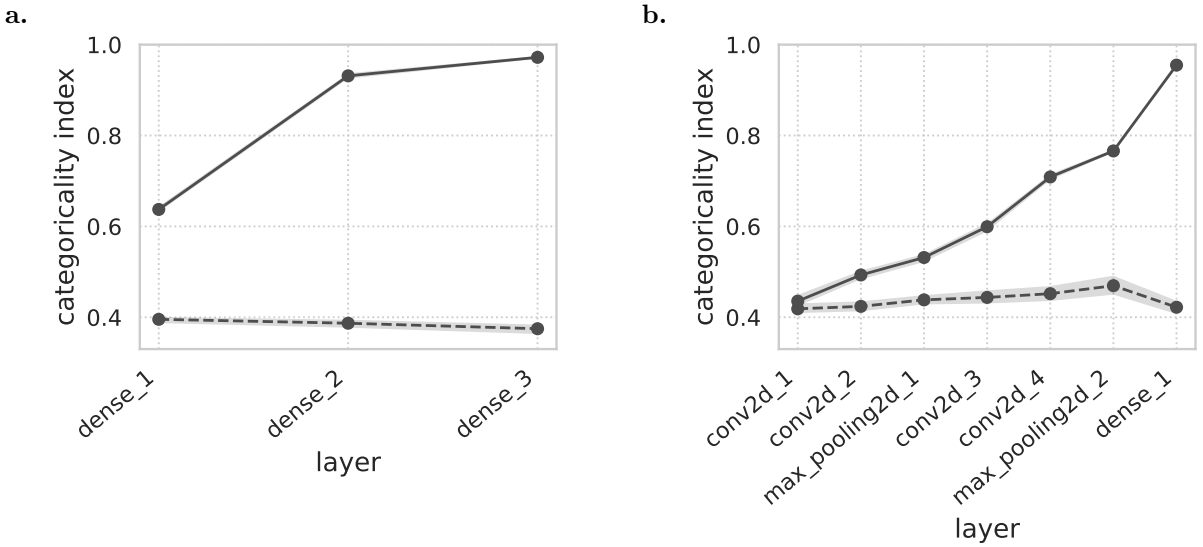

Figure 6: **Categoricality as a function of layer depth, using the MNIST dataset**. Categoricality is a measure that quantifies the distance between items drawn from the same category vs. different categories, thanks to the Kolmogorov-Smirnov statistic between the intra- and inter-category distributions of neural distances. Solid lines correspond to trained networks, broken lines to untrained ones (error bars indicate 95% confidence intervals, estimated by bootstrap). (a) The neural network is a multi-layer perceptron with three hidden layers. (b) The neural network is a convolutional neural network whose layer structure is described in the x-axis of the figure.

on the Kaggle Dogs vs. Cats database, which contains 25,000 images of dogs and cats, with a final classification performance on the test set of about 95%. In order to assess the changes in representation induced by learning for the different layers in the neural network, we considered different continua that either interpolate between items from the two categories, in which case we expect categorical perception to emerge, and between items from the same categories, as a control. The continua that we consider cycle from one dog to the same dog, going from one dog to a cat to another cat to another dog and finally back to the first dog. Each sub-continuum is made of 8 images, thus totaling 28 images for a given full continuum. We considered two such continua, using the same cat/dog categories but considering different viewpoints (close up vs full body – see the x-axes of Fig. 7a and b for the morphed images that were used as input). The different cat/dog, cat/cat or dog/dog morphed continua were generated thanks to the code and pretrained model provided by Miyato et al. (2018) that uses a method based on Generative Adversarial Networks (Goodfellow et al., 2014). Note that these continua on which the neural networks are tested have been generated by a network trained on a completely different database.

We present the results of this experiment in Figure 7. First, one can see that the network well categorizes the different cats and dogs images (see the dotted colored lines) into the correct classes. The last hidden layer, right before the final decision, exhibits a strong categorical perception effect: only items that straddle the categories can be discriminated. At the opposite, the first convolutional layers do not exhibit categorical perception: there is no clear peak of discrimination between categories. Instead, differences between contiguous images appear to mainly reflect differences in input space (as a comparison, see Appendix C, Fig. C.2 for a picture of the distances in input space). For instance, the peak difference in input space and for these first convolutional layers is driven by the tongue sticking out of the mouth, which does not affect the more categorical upper layers. Finally, the last convolutional layers exhibit in-between behavior, with an ability to discriminate between within-category stimuli, but with a clear peak at the cat/dog boundary, thus displaying categorical perception.

3.3.2 Categoricality in deep networks

In this section, we work with the ImageNet dataset (Deng et al., 2009), and more precisely with the subset of images used in the ILSVRC-2010 challenge (Berg et al., 2011). The network that we consider is the VGG16 model described in Simonyan and Zisserman (2014), which has won the ImageNet Challenge 2014. This model is characterized by 16 weight layers, an architecture considered very deep (at the time

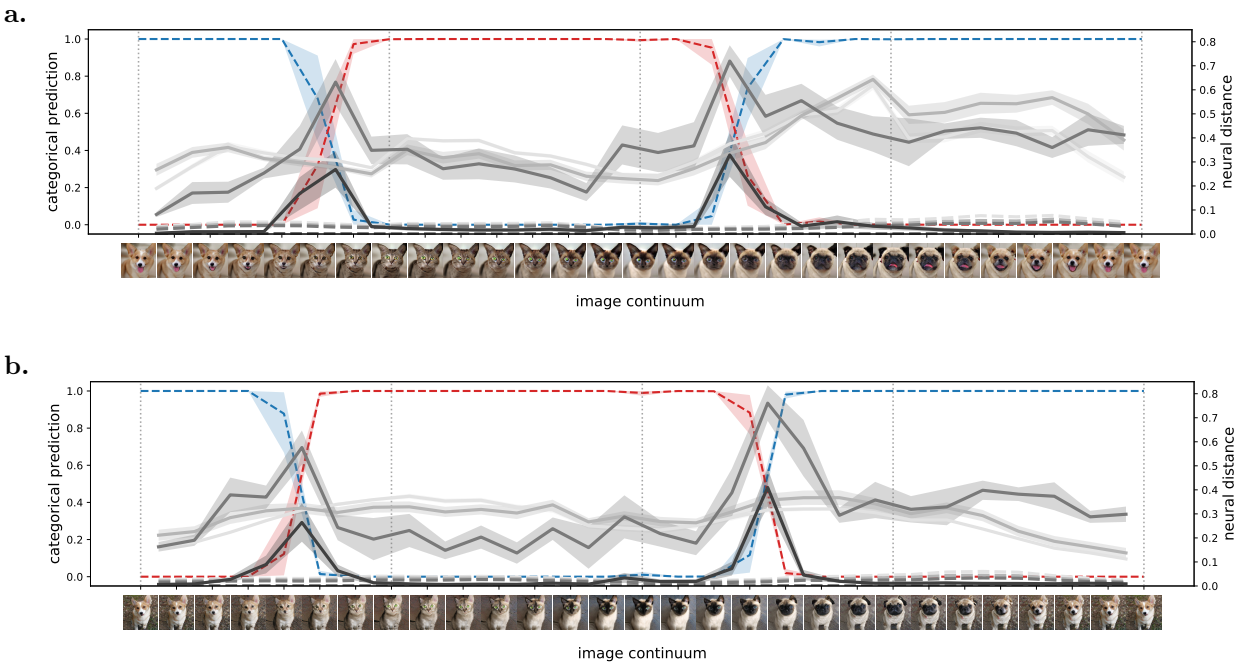

Figure 7: **Categorical perception of a cat/dog circular continuum**. Experiment with continua interpolating between cats and dogs, with two different viewpoints: (a) close up on the face, and (b) full body. The interpolations involve two types of dogs and two types of cats. Each continuum corresponds to a circular interpolation with four sub-continua: from the first dog to a cat, then to the other cat, then to the second dog, and finally back to the first dog. The neural network is a convolutional neural network with three blocks of two convolutional layers and a max pooling layer, finally followed by a global pooling average layer. The dotted colored lines indicate the posterior probabilities (blue is dog, red is cat). The solid lines correspond to the neural distance between adjacent items along the continuum, the darker the line the deeper the layer. Only the last convolution layer of each block and the global pooling average layer are shown. The dotted gray lines are the counterparts for the same networks but before learning. Error bars indicate 95% confidence intervals, estimated by bootstrap. The thin dotted vertical lines indicate the start and end points of each sub-continua.

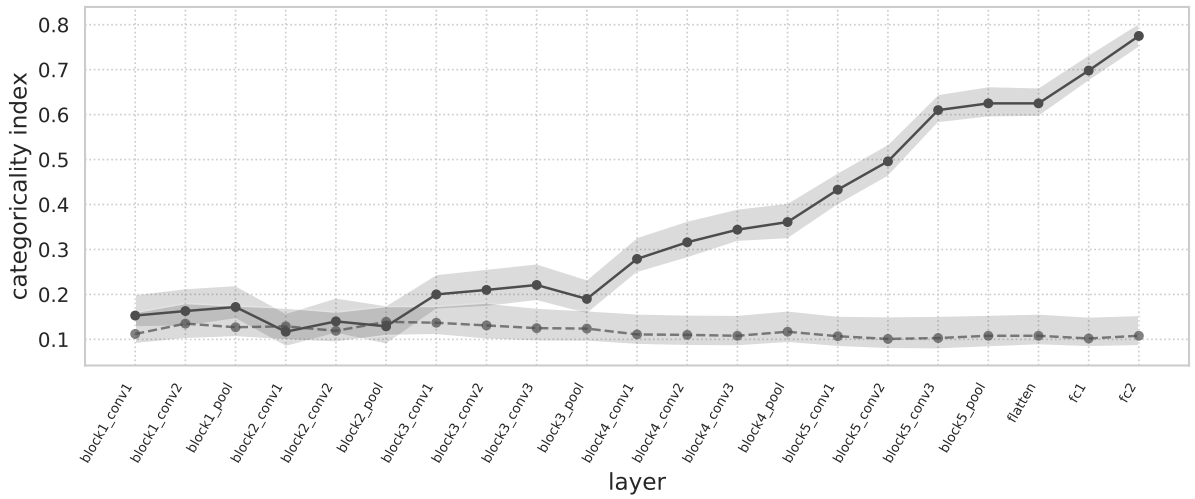

Figure 8: **Categoricality as a function of layer depth, using the ImageNet dataset**. The neural network is the VGG16 model (Simonyan and Zisserman, 2014). The dash line corresponds to a network with random initialization of the weights, whereas the solid line corresponds to a network pretrained on ImageNet (error bars indicate 95% confidence intervals, estimated by bootstrap).

this VGG16 model was published). Here, we compare the categoricality index on randomly initialized networks with the exact same architecture, and on a network that has been pretrained on the ImageNet database (as provided by the keras package[2]) (see Materials and Methods for details).

We show in Figure 8 the results of this comparison. As expected, one can see that for an untrained network, the categoricality index is flat across layers: the neuronal layers do not show any preferential knowledge of the categories. For a trained network, as seen in the MNIST section above, the categoricality increases as a function of depth. We can observe that, for the first convolutional layers, this index is essentially not much different from the case of an untrained network. Intermediate convolutional layers do exhibit some effects of category learning, while for the last convolutional layers the categoricality is much more marked. Finally, the last two dense layers exhibit the greatest categoricality.

4 Discussion

4.1 A new view on dropout

Dropout is one of the most widely used regularization techniques at the time of writing this paper (for instance, it ranks number 1 among the regularization techniques in the paperswithcode website, with 4727 papers referenced, as of December 9, 2020 [3]). This heuristics aims at preventing co-adaptation between neurons by randomly dropping out a proportion of units in the neural network during learning (Srivastava et al., 2014). It was first proposed as a way to train many different networks with shared weights that are averaged in the end. Dropout consists in multiplicative Bernoulli noise, but other types of noise work just as well (see the case of multiplicative Gaussian noise, also discussed in Srivastava et al., 2014).

Bouthillier et al. (2015) propose an interesting interpretation of dropout: it serves as a kind of data augmentation (see also Zhao et al., 2019). In this work, the authors transform dropout noise into new samples, leading to an augmented training dataset, containing much more samples than the original dataset. They show that training a deterministic network on this augmented dataset leads to results on par with the dropout results.

Our framework allows to understand why dropout is particularly beneficial, and explain various empirical observations and heuristics practices in the use of the dropout technique.

[2] `https://keras.io/api/applications/vgg/`

[3] `https://paperswithcode.com/methods/category/regularization`

4.1.1 An adaptive variability

First, we argue here that dropout is particularly advantageous because it adapts to the structure of the data. For instance, as seen in Fig. 2, the generated stimuli, following the method proposed by Bouthillier et al. (2015), do not distribute uniformly over the input space, but adjust their variability to the structure of the categories, with more variability within a category and less variability between categories, where the risk of generating a new stimuli with a wrong label is highest. This is one of the reason why injecting noise in the hidden layers with a technique such as dropout happens to yield better results than techniques that only consider input noise, as this latter type of noise might be too weak within a category and too strong between categories.

Note that we do not claim that there might exist a strict equivalence between dropout and data augmentation. The main point is that for a given layer, after learning, Fisher information is greater at the boundary between categories. Experimentally, we showed that neural distance between neighboring stimuli is indeed greater in these regions, compared to region well within a category. This has the consequence of controlling the impact of neuronal noise as a function of the probability of misclassification (see Section 2.1 and Appendix A).

4.1.2 Amount of dropout as a function of depth

A widespread heuristic practice is to increase the dropout level of a layer as a function of its depth. Typically, shallow layers receive a very small amount of dropout, whereas deeper layers use a much higher amount (see for instance all the examples by Srivastava et al., 2014).

More noise at a given layer requires to better separate the categories so as to be more robust, as implied by Eq. 2 (see also Section A.5). Mehrer et al. (2020) have indeed found that an increase of the dropout probability yields greater categoricality (see their Fig. 8c). This may suggest that categoricality is mainly the effect of using dropout. We checked that the increase in categoricality, as observed in our experiments, is not primarily driven by the use of dropout. To do so, as a control experiment we reproduced the results of Fig. 6 Section 3.2.4, but without the use of dropout (see Appendix B, Fig. B.1). We find that the categoricality index does increase as a function of depth, although the slope of this increase is slightly lower than with the use of noise (dropout or Gaussian dropout), as expected from our analysis.

Thus, although more noise implies greater separation between categories, it also works the other way around: a greater categoricality makes it possible for a given layer to accept a larger amount of noise, for this noise is structured according the categories, with lower variance between classes. Given that the categorical effects are more pronounced the deeper a layer, a good strategy in the tuning of the dropout noise is thus indeed to inject more noise into deeper layers.

4.1.3 Dense vs. convolutional layers

We have seen that dense layers exhibit more categoricality than their convolutional counterparts. This might explain why dropout level is usually stronger for dense layers. Moreover, convolutional layers, especially the one closest to the input layer, are not very categorical. This explains why they receive a small amount of dropout, or even no dropout at all. Still, we observe that the deepest convolutional layers do exhibit categoricality, which suggests that dropout should be beneficial even for deep convolutional layers, a prediction that is indeed backed up by the literature (see the gain in performance described in the original dropout article by Srivastava et al., 2014; see also Park and Kwak, 2016; Spilsbury and Camps, 2019).

4.1.4 Evolution of dropout rate during learning

Importantly, our work shows that the noise in a given layer depends on the representation that is being learned, *ie* the structure of the categories. This means that its effect is not the same before, during, and after learning. At first, the noise is more uniform, whereas after training it is less strong in the cross-category regions, relative to the intra-class regions. Thus, our framework insists on the dynamical aspect of the impact of noise as a function of training.

Several works have considered adapting the dropout rate (that is the fraction of neurons that are made silent) during the course of learning. Relying on the simulated annealing metaphor, Rennie et al.

(2014) suggest to decrease the dropout rate over the course of training, starting from a high initial value to zero. The intuition is that at first high noise makes it possible to explore the space of solutions, avoiding local minima, with a last phase of fine-tuning without noise. On the opposite side, taking inspiration from curriculum learning (Bengio et al., 2009), Morerio et al. (2017) proposes the exact contrary: to start easy, with low noise, and then increase the difficulty by raising the amount of noise.

Our work suggests that a network needs to have already partly learn the structure of the categories in order to allow for the use of a high level of noise – otherwise, as we discussed, it might just be detrimental. We thus expect the curriculum dropout to have better performance than the original dropout (that is with a fixed value of the dropout rate over the course of training), and better than the annealed dropout: this is exactly what is found in the experiments by Morerio et al. (2017) on a variety of image classification dataset (see also Spilsbury and Camps, 2019).

4.1.5 Interaction with the size of the dataset

The interaction that we describe between learning and dropout also predicts that there should be an interaction with the size of the dataset. Obviously, with a very large number of training samples available, there is no need for a regularization technique such as dropout, as there is no worry about overfitting. Less intuitive is the prediction here that if the number of training sample is too low, the structure of the categories cannot be learned, and the use of dropout will be detrimental to the performance. In between, we expect the introduction of noise to be beneficial to the generalization ability of the neural network. All in all, this is exactly what is found by Srivastava et al. (2014, see Fig. 10).

4.2 Back to the fields of psychology and neuroscience

4.2.1 An historical debate on categorical perception

Historically, categorical perception was first presented as an all-or-nothing phenomenon according to which subjects can discriminate between stimuli only through their phonetic labels – or more precisely through their class identification probability (following the view of the classical Haskins model; see Liberman et al., 1957). As a result, discrimination was thought to be almost zero within category. In our setting, it would be as if discrimination was only made possible through the use of the very last layer of the neural network, the one corresponding to the decision process with the softmax function. However, from the very beginning, this idealized form of categorical perception has never been met experimentally (see, e.g., Liberman et al., 1957 or Liberman et al., 1961; see Lane, 1965; Repp, 1984 for reviews): observed discrimination is always above the one predicted from the identification function. Some authors rather talk about *phoneme boundary effect* in order to distinguish it from the original categorical perception proposal (Wood, 1976; Iverson and Kuhl, 2000). As this phenomenon is not limited to speech, we keep here the term *categorical perception*, simply defined as a better ability to perceive differences between stimuli in the cross-category regions than within a category.

Our framework makes it possible to better understand this phenomenon. As we have seen, category learning does not only affect the last decisional layer, but also the upstream coding layers in order to better represent categories and robustly cope with noise. For each hidden layer, we observe a warping of the neural space within category and an expansion between categories. Thus, without the need to even actually compute labels, assuming the coding layers are reused during discrimination experiments, we expect to see categorical perception effects during discrimination tasks, but possibly with an above chance discrimination within category, as indeed found experimentally. This effect is not due to labeling, but a consequence of the optimization of the coding layers upstream of the categorization final process.

As we have seen, a deep network exhibits a gradient of categorical perception, from the first layers having almost no traces of categoricality to the last layers that show an important distortion of the space according to the structure of the categories. Where exactly is the basis for our conscious perception remains an important research question.

4.2.2 Results in neuroscience and imagery

In our work, we notably study categorical representations by looking at the neural distance between stimuli that are equally spaced in stimulus space. Previous experimental works have used the same

technique. Using high-density intracranial recordings in the human posterior superior temporal gyrus, Chang et al. (2010) have found that this region responds categorically to a /ba/–/da/–/ga/ continuum (as used in the original Liberman et al., 1957 study): stimuli from the same category yield indeed more similar neural patterns that stimuli that cross categories. Still in the domain of speech perception, using event-related brain potentials, Bidelman et al. (2013) followed a similar approach in comparing neural activities in response to a vowel continuum. They found that the brainstem encodes stimuli in a continuous way, with changes in activity mirroring changes in the sounds, contrary to late cortical activity that are shaped according to the categories. This is in line with the view of gradient effects of categorical perception as function of depth in the processing stream: input layers show almost no effect of categories, whereas last layers are strongly affected by them.

The monkey study by Freedman et al. (2003) have shown a distinct behavior of the prefontal and inferior temporal cortices during a visual categorization task. The visual processing stream goes from sensory to the prefrontal cortex (PFC) through the inferior temporal cortex (ITC). In order to assess the categorical nature of the activity of each region, Freedman et al. (2003) introduced an index similar to what we use in our study, by comparing the responses to within- vs. between-categories pairs of stimuli, but at the level of an individual neuron. In agreement with the picture proposed here, the authors have found that both regions show significant effects of category learning, with larger differences in the neural responses for pairs of stimuli that are drawn from different categories, and that the PFC neurons show stronger category effects than the ITC neurons. Note though that the categoricality of the ITC is actually underevaluated due to the use of a measure at the single neuron level. A subsequent study by the same team has indeed shown that the ITC contains actually more categorical information when analyzed at the population level (see Meyers et al., 2008). Similarly, using both electrode recordings in monkeys and fMRI data in humans, Kriegeskorte et al. (2008) have shown that the ITC exhibits a categorical representation, contrary to the early visual cortex. Interestingly enough, the categorical information in ITC can only be seen at the population level. All in all, it is once again found that the visual stream is organized along a path that goes from being not categorical (early visual area) to being categorical at the population level while retaining information of within-category individual examples (inferior temporal cortex) to a more fully categorical representation (prefrontal cortex), in agreement with our analysis. From the work presented here, we expect an even wider range of such gradient effects of categorical perception to be found experimentally, either with neurophysiology or imagery.

5 Conclusion

Studying categorical perception in biological and artificial neural networks can provide a fruitful discussion between cognitive science and machine learning. Here, we have shown that artificial neural networks that learn a classification task exhibit an enlarged representation near the boundary between categories, with a peak in discrimination, *ie* categorical perception. Our work further suggests a strong gradient of categorical effects along the processing stream, which will be interesting to investigate experimentally. Our analysis is based on the mathematical understanding of the geometry of neural representations optimized in view of a classification task, with contraction of space far from category boundaries, and expansion near boundaries. Our results find counterparts in the literature of neurophysiology and imagery, with low-level generic regions feeding high-level task-specific regions. Our framework allows to understand various properties and practical uses of dropout. A given layer will be more robustly able to benefit from noise if its representation well separates the different classes. Layers closer to the inputs should thus indeed receive a lower amount of dropout than the deeper layers. We saw that dense layers are more categorical than convolutional layers, which explains why they typically receive more dropout, although as we have seen the convolutional layers also benefit from a small amount of noise. Finally, an important aspect of our work is that noise in the hidden layers has a differential impact as a function of the representation that is learned, and thus changes during learning.

Further work is needed so as to quantify more finely how the amount of noise should depend on the level of representation, in particular for instance on the performance. An interesting perspective is in the domain of transfer learning, where a neural network trained for a specific task is reused for another task. The categoricality of each layer, which is specific to the classification task at hand, gives thus a measure of the degree (the lack) of genericity of the layer. We expect that this quantity, along with a measure of the overlap between the old and the new tasks, can be used to decide where to cut a neural network for reuse, with the lower part left untouched and the deeper part fine-tuned or retrained from

scratch on the new task.

To conclude, our work insists on the the geometry of internal representations shaped by learning categories, and on the resulting positive impact of noise as a way to learn more robustly. Noise is actually structured by learning, which implies an interaction between these two aspects.

Materials and Methods

Neural distance. For a given stimulus $\mathbf{x}$, let us notate $f(\mathbf{x})$ the N-dimensional deterministic function computed by the network in the absence of noise (for a given layer with N neurons). The neural distance $d_{\text{neural}}(\mathbf{x}_1, \mathbf{x}_2)$ between two stimuli $\mathbf{x}_1$ and $\mathbf{x}_2$ is then defined, at the population level, as the cosine distance between $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$. The cosine distance is equal to 1 minus the cosine similarity, which is equal to the dot product between the two vectors, normalized by the product of the norms of each vector:

$$d_{\text{neural}}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{f(\mathbf{x}_1) \cdot f(\mathbf{x}_2)}{||f(\mathbf{x}_1)|| * ||f(\mathbf{x}_2)||} \tag{4}$$

Note that this measure is not mathematically a distance metric as it does not satisfy the triangular inequality. We nevertheless improperly call this dissimilarity a distance, following a common abuse of language.

Categoricality index. The categoricality index quantifies the degree of relative intra-class compression vs inter-class expansion of the representation provided by a given layer. It measures the distance between (i) the distribution of the neural distance of items that belong to the same category, and (ii) the distribution of the neural distance of items that are drawn from different categories. Technically, we compute these distributions thanks to random samples taken either from the same categories or from different categories, and we take as distance between the two distributions the two sample Kolmogorov-Smirnov statistic (using the Python implementation provided by the SciPy package, Virtanen et al., 2020). Other distance measures could be considered as well, as those previously proposed in the literature. We expect that they would yield overall qualitatively similar results. For instance, Mehrer et al. (2020) recently considered a clustering index defined as the "normalized difference in average distances for stimulus pairs from different categories (across) and stimulus pairs from the same category (within): CCI = (across - withing)/(across + within)". Yet, while quantifying similarly the degree of intra class compression vs inter class separation, we believe our measure gives a better account of the categorical nature of the neural representation by considering not just the average values of the 'across' and 'within' distances between pairs but the full distributions. Imagine two cases where the average quantities 'across' and 'within' are equal, but whose distributions exhibit different variances: in the first case, the two distributions of distance are very well separated, with a small variance for each 'across' or 'within' distribution; in the second case, the two distributions overlap substantially, with larger variance for these distributions. By definition, both cases will receive the same clustering index as defined in Mehrer et al. (2020), but our measure assigns a greater categoricality to the first case, as expected by construction of this example.

Section "Experiments with toy examples": Estimate of the deterministic counterpart of a noisy neural activity. For a given layer, we consider the neural activity $\mathbf{r} = \{r_1, \ldots, r_N\}$ from a population of N neurons evoked by a stimulus $\mathbf{x}$ as a noisy version $\widetilde{f(\mathbf{x})}$ of the N-dimensional deterministic function $f(\mathbf{x})$ computed by the network at that level. As an example in the spirit of the dropout heuristic, for a Gaussian multiplicative noise, $\mathbf{r} = \widetilde{f(\mathbf{x})} = f(\mathbf{x}) * \xi$, where $\xi \sim \mathcal{N}(1, \sigma^2)$ (σ^2 being the noise variance). For a given $\mathbf{x}$ and a given $\mathbf{r} = \widetilde{f(\mathbf{x})}$, we make use of gradient descent to compute the estimate $\widehat{\mathbf{x}}$ that minimizes the square error between $\widetilde{f(\mathbf{x})}$ and $f(\widehat{\mathbf{x}})$:

$$\widehat{\mathbf{x}} \equiv \underset{\mathbf{x}^*}{\text{argmin}} \left(\widetilde{f(\mathbf{x})} - f(\mathbf{x}^*) \right)^2 . \tag{5}$$

Section "Creation of an image continuum": Autoencoder architecture and learning. The autoencoder is a made of an encoder chained with a decoder, both convolutional neural networks. The encoder is made of three convolutional layers, each followed by a max-pooling layer. Similarly, the decoder uses three convolutional layers, each followed by an upsampling layer. All cells have ReLU activation function.

The autoencoder is trained using a binary cross entropy loss on the full MNIST training set for 50 epochs, through gradient descent using Adam optimizer with default parameters (Kingma and Ba, 2015).

Section "Gradient categorical perception as a function of depth": Selection of a set of image continua. We explain here how the continua considered in Section 3.2.3 are selected. We first draw 100 samples per class from the test set. Each sample from one category is paired once with a sample from another category, leading to 4500 pairs of stimuli drawn from two different digit categories. For each pair we generate a continuum as explained above. Not all pairs are valid pairs to look at in the context of our study: we are indeed interested in pairs that can be smoothly interpolated from one category to another. Categories that are close one to another are mainly concerned here – for instance, if the generated continuum straddles another third category then it should be dismissed from this analysis. In order to only consider the relevant pairs, we keep a pair only if the sum of the two posterior probabilities is above a certain threshold (0.95 here) all along the continuum. In the end, 1566 pairs fulfill this criterion and are included in the study.

Section "Categoricality as a function of depth and layer type": Details on the numerical protocol. The multi-layer perceptron has three hidden layers of 1024 cells, with ReLU activations. Gaussian dropout is used after each dense layer, with respective rate 0.1, 0.2, 0.4. The convolutional neural network has two blocks of two convolutional layers of 32 cells with ReLU activations, followed by a max pooling layer, these two blocks finally followed by a dense hidden layers of 128 cells. Each block is followed by a dropout layer with rate 0.2, and the dense layer is followed by a dropout layer with rate 0.5. Simulations for the multi-layer perceptrons and the convolutional neural networks share the same framework. It considers 10 trials. Each trial uses a different random initialization. For each trial, learning is done over 20 epochs through gradient descent using Adam optimizer with default parameters (Kingma and Ba, 2015). In order to compute the categoricality index, the distributions of both the within- and between-category distances are evaluated thanks to 1000 pairs of samples drawn from the same category and 1000 pairs of samples drawn from different categories. Samples come from the test set.

Section "Categorical perception of a cat/dog continuum": Details on the numerical protocol. Image size is 180x180. The neural network is a convolutional network with three blocks of two convolutional layers of 64 cells with ReLU activations followed by a max pooling layer, these three blocks finally followed by a global pooling average layer (Lin et al., 2013). Each block is followed by a Gaussian dropout layer with rate 0.1, and the global pooling average layer is finally followed by a dropout layer with rate 0.4. This simulation is repeated 10 times, each time with a different random initialization. Learning is done over 60 epochs through gradient descent using Adam optimizer with default parameters (Kingma and Ba, 2015). The learning database being quite small, we used data augmentation during learning, performing horizontal random flip and random rotations (with an angle in $[-0.1 * 2\pi, 0.1 * 2\pi]$). Each network is trained on the Kaggle Dogs vs. Cats database[4], which contains 25,000 images of dogs and cats, with a classification performance on the test set of about 95%.
Finally, the different cat/dog, cat/cat or dog/dog morphed continua were generated thanks to the code provided by Miyato et al. (2018)[5]. We made use of the 256x256 model pretrained on ImageNet that is provided by the authors.

Section "Categoricality in deep networks": Evaluation of the categoricality index for the ImageNet experiment. The full ImageNet database consists in more than a million images categorized into 1000 different classes. Categoricality is evaluated through the use of 1000 pairs of samples for the within-category distribution of neural distances, and 1000 pairs of samples for the between-categories one. Samples come from the test set.

Computer code. The custom Python 3 code written for the present project makes use of the following libraries: `tensorflow v2.3.1` (Abadi et al., 2015) (using `tf.keras` API, Chollet et al., 2015), `matplotlib v3.1.3` (Hunter, 2007), `numpy v1.18.1` (Harris et al., 2020), `scipy v1.4.1` (Virtanen et al., 2020), `pandas v1.0.1` (McKinney et al., 2010), `seaborn v0.10.0`. The code will be made available on GitHub.

[4] `https://www.microsoft.com/en-us/download/details.aspx?id=54765`, `https://www.kaggle.com/c/dogs-vs-cats`
[5] `https://github.com/pfnet-research/sngan_projection`

Acknowledgments

We thank the Information Systems Division (DSI) of the EHESS for their helpfulness in providing us with an access to computing resources during the covid-19 lockdown.

Appendix A Modeling categorical perception

For completeness, in this Appendix we review and synthesize the results in Bonnasse-Gahot and Nadal (2008, 2012) that are relevant for the present paper. In a companion paper (Bonnasse-Gahot and Nadal, 2020), we show that the analysis presented here can be extended to multi-layer networks, making explicit that x, instead of being the stimulus, corresponds to the projection (achieved by the network) of the stimulus on a space relevant for the discrimination task.

A.1 Model Description

We consider a finite set of M categories, denoted by $\mu = 1, \ldots, M$, with probabilities of occurrence (relative frequency) $q_\mu > 0$, so that $\sum_\mu q_\mu = 1$. Each category is defined as a probability density distribution $P(\mathbf{x}|\mu)$ over the continuous space of stimulus $\mathbf{x}$. The stimulus space is assumed to be of small dimension K, corresponding to the selection of the features or directions relevant to the task at hand. For the sake of simplicity, we only consider here the one-dimensional case, $K = 1$. See Bonnasse-Gahot and Nadal (2008, 2012) for equations in the general case $K > 1$.

A stimulus $x \in \mathbb{R}$ elicits a response $\mathbf{r} = \{r_1, \ldots, r_N\}$ from a population of N neurons. This neural activity $\mathbf{r}$ is some noisy representation of x and aims at encoding properties about the given categories. The read-out is realized by M output cells with activities $g_\mu, \mu = 1, ..., M$. Each output activity is a deterministic function of the neural activity $\mathbf{r}$, $g_\mu = g(\mu|\mathbf{r})$. With the goal of computing an estimate $\widehat{\mu}$ of μ, we consider these outputs as estimators of the posterior probability $P(\mu|x)$, where x is the (true) stimulus that elicited the neural activity $\mathbf{r}$. The processing chain can be summarized with the following Markov chain:

$$\mu \to x \xrightarrow[\text{coding}]{} \mathbf{r} \xrightarrow[\text{decoding}]{} \widehat{\mu} \tag{A.1}$$

A.2 Estimation of the posterior probabilities

The read-out assumes that, given a neural activity $\mathbf{r}$ in the coding layer, the goal is to construct as neural output an estimator $g(\mu|\mathbf{r})$ of the posterior probability $P(\mu|x)$, where x indicates the (true) stimulus that elicited the neural activity $\mathbf{r}$. The relevant Bayesian quality criterion is given by the Kullback-Leibler divergence (or relative entropy) $\mathcal{C}(x, \mathbf{r})$ between the true probabilities $\{P(\mu|x), \mu = 1, ..., M\}$ and the estimator $\{g(\mu|\mathbf{r}), \mu = 1, ..., M\}$, defined as (Cover and Thomas, 2006):

$$\mathcal{C}(x, \mathbf{r}) \equiv D_{KL}(P_{\mu|x}||g_{\mu|\mathbf{r}}) = \sum_{\mu=1}^{M} P(\mu|x) \ln \frac{P(\mu|x)}{g(\mu|\mathbf{r})} \tag{A.2}$$

Averaging over $\mathbf{r}$ given x, and then over x, the mean cost induced by the estimation can be written:

$$\overline{\mathcal{C}} = -\mathcal{H}(\mu|x) - \int \left(\int \sum_{\mu} P(\mu|x) \ln g(\mu|\mathbf{r}) \; P(\mathbf{r}|x) \, d\mathbf{r} \right) \; p(x) \, dx \tag{A.3}$$

where $\mathcal{H}(\mu|x) = -\int dx \, p(x) \sum_{\mu=1}^{M} P(\mu|x) \ln P(\mu|x)$ is the conditional entropy of μ given x.

We can rewrite (A.3) as the sum of two terms :

$$\overline{\mathcal{C}} = \overline{\mathcal{C}}_{\text{coding}} + \overline{\mathcal{C}}_{\text{decoding}} \tag{A.4}$$

respectively defined as :

$$\overline{\mathcal{C}}_{\text{coding}} = I(\mu, x) - I(\mu, \mathbf{r}) \tag{A.5}$$

and

$$\overline{\mathcal{C}}_{\text{decoding}} = \int D_{KL}(P_{\mu|\mathbf{r}}||g_{\mu|\mathbf{r}}) \; P(\mathbf{r}) \, d\mathbf{r} \tag{A.6}$$

$I(\mu, x)$ and $I(\mu, \mathbf{r})$ are respectively the mutual information between the categories μ and the stimulus x, and between the categories μ and the neural activity $\mathbf{r}$, defined by (Blahut, 1987):

$$I(\mu, x) = \sum_{\mu=1}^{M} q_\mu \int \ln \frac{P(x|\mu)}{P(x)} \, P(x|\mu) \, dx \,, \quad I(\mu, \mathbf{r}) = \sum_{\mu=1}^{M} q_\mu \int \ln \frac{P(\mathbf{r}|\mu)}{P(\mathbf{r})} \, P(\mathbf{r}|\mu) \, d\mathbf{r} \tag{A.7}$$

$D_{KL}(P_{\mu|\mathbf{r}}||g_{\mu|\mathbf{r}})$ is the relative entropy between the true probability of the category given the neural activity and the output function g:

$$D_{KL}(P_{\mu|\mathbf{r}}||g_{\mu|\mathbf{r}}) = \sum_{\mu=1}^{M} P(\mu|\mathbf{r}) \ln \frac{P(\mu|\mathbf{r})}{g(\mu|\mathbf{r})} \tag{A.8}$$

Since processing cannot increase information (see e.g. Blahut, 1987, pp. 158-159), the information $I(\mu, \mathbf{r})$ conveyed by $\mathbf{r}$ about μ is at most equal to the one conveyed by the sensory input x, hence we have that $\overline{\mathcal{C}}_{\text{coding}} \geq 0$. This coding cost tends to zero as noise vanishes. The decoding cost $\overline{\mathcal{C}}_{\text{decoding}}$ is the only term that depends on g, hence the function minimizing the cost function (A.4) is (if it can be realized by the network):

$$g(\mu|\mathbf{r}) = P(\mu|\mathbf{r}) \tag{A.9}$$

From a machine learning viewpoint, minimization of the decoding cost (A.8) can be achieved through supervised learning taking as cost function the cross-entropy loss, as shown in Bonnasse-Gahot and Nadal (2012), SI Section 1.

A.3 Coding efficiency

In Bonnasse-Gahot and Nadal (2008), we show that, in a high signal-to-noise ratio limit, that is when the number N of coding cells grows to infinity, the mutual information $I(\mu, \mathbf{r})$ between the activity of the neural population and the set of discrete categories reaches its upper bound, which is the mutual information $I(\mu, x)$ between the stimuli and the categories. For large but finite N, the leading correction simply writes as the average (over the stimulus space) of the ratio between two Fisher-information values: in the denominator, the Fisher information $F_{\text{code}}(x)$, specific to the neural encoding stage $x \to \mathbf{r}$, and in the numerator, the Fisher information $F_{\text{cat}}(x)$, that characterizes the category realizations $\mu \to x$ and does not depend on the neural code. Fisher information is an important concept that comes from the field of parameter estimation in statistics. $F_{\text{code}}(x)$ characterizes the sensitivity of the neural activity $\mathbf{r}$ with respect to small variations of x. The higher the Fisher information $F_{\text{code}}(x)$, the better an estimate of x can be obtained. $F_{\text{cat}}(x)$ quantifies the categorization uncertainty. As a consequence, $F_{\text{cat}}(x)$ is larger in the transition regions between categories, where the identification function $P(\mu|x)$ changes quickly, than within category, where the identification function $P(\mu|x)$ is almost flat.

Explicitly, the coding cost (A.5) writes:

$$\overline{\mathcal{C}}_{\text{coding}} = \frac{1}{2} \int \frac{F_{\text{cat}}(x)}{F_{\text{code}}(x)} \, p(x) \, dx \tag{A.10}$$

where $F_{\text{code}}(x)$ and $F_{\text{cat}}(x)$ are respectively defined as

$$F_{\text{code}}(x) = -\int \frac{\partial^2 \ln P(\mathbf{r}|x)}{\partial x^2} \, P(\mathbf{r}|x) \, d\mathbf{r} \tag{A.11}$$

$$F_{\text{cat}}(x) = -\sum_{\mu=1}^{M} \frac{\partial^2 \ln P(\mu|x)}{\partial x^2} \, P(\mu|x). \tag{A.12}$$

Crucially for the present work, the inverse of the Fisher information is an optimal lower bound on the variance σ_x^2 of any unbiased estimator $\widehat{x}(\mathbf{r})$ of x (Cramér-Rao bound, see e.g. Blahut, 1987):

$$\sigma_x^2 \equiv \int \left(\widehat{x}(\mathbf{r}) - x\right)^2 P(\mathbf{r}|x) \, d\mathbf{r} \;\geq\; \frac{1}{F_{\text{code}}(x)} \tag{A.13}$$

A.4 Optimal decoding

In Bonnasse-Gahot and Nadal (2012), we show that the function $g(\mu|\mathbf{r})$ that minimizes the decoding cost function given by Eq. A.6 is equal to $P(\mu|\mathbf{r})$, which is an (asymptotically) unbiased and (asymptotically) efficient estimator of $P(\mu|x)$. For a given x, its mean is thus equal to

$$\int g(\mu|\mathbf{r})\, P(\mathbf{r}|x)\, d\mathbf{r} = P(\mu|x) \tag{A.14}$$

and its variance is given by the Cramér-Rao bound, that is, in this 1d case,

$$\int \big(g(\mu|\mathbf{r}) - P(\mu|x)\big)^2\, P(\mathbf{r}|x)\, d\mathbf{r} \;=\; \frac{P'(\mu|x)^2}{F_{\text{code}}(x)} \tag{A.15}$$

A.5 Category learning implies categorical perception

The Fisher information $F_{\text{cat}}(x)$ is the largest at the boundary between categories. If the neural code is to be optimized, we therefore expect, as the number N of neurons is limited, Fisher information $F_{\text{code}}(x)$ to be greater between categories than within, so as to compensate for the higher value of $F_{\text{cat}}(x)$ in this region (see Eq. A.10). Another way to look at it is through Eq. A.15: a greater Fisher information $F_{\text{code}}(x)$ in the transition region between categories, where $P'(\mu|x)^2$ is the highest, makes it possible to lower the variance of the estimate $g(\mu|\mathbf{r})$ of the $P(\mu|x)$, hence the probability of misclassifying x given the neural activity $\mathbf{r}$.

The Fisher information $F_{\text{code}}(x)$ represents the sensitivity of the neural code to a small change in stimulus x. Larger information means greater sensitivity. It gives the metric of the representation: a larger $F_{\text{code}}(x)$ around a certain value of x means that the representation is dilated at that location. Thus, in other words, category learning implies better cross-category than within-category discrimination, hence the so-called *categorical perception*.

Appendix B Comparing categoricality on the MNIST dataset with and without dropout

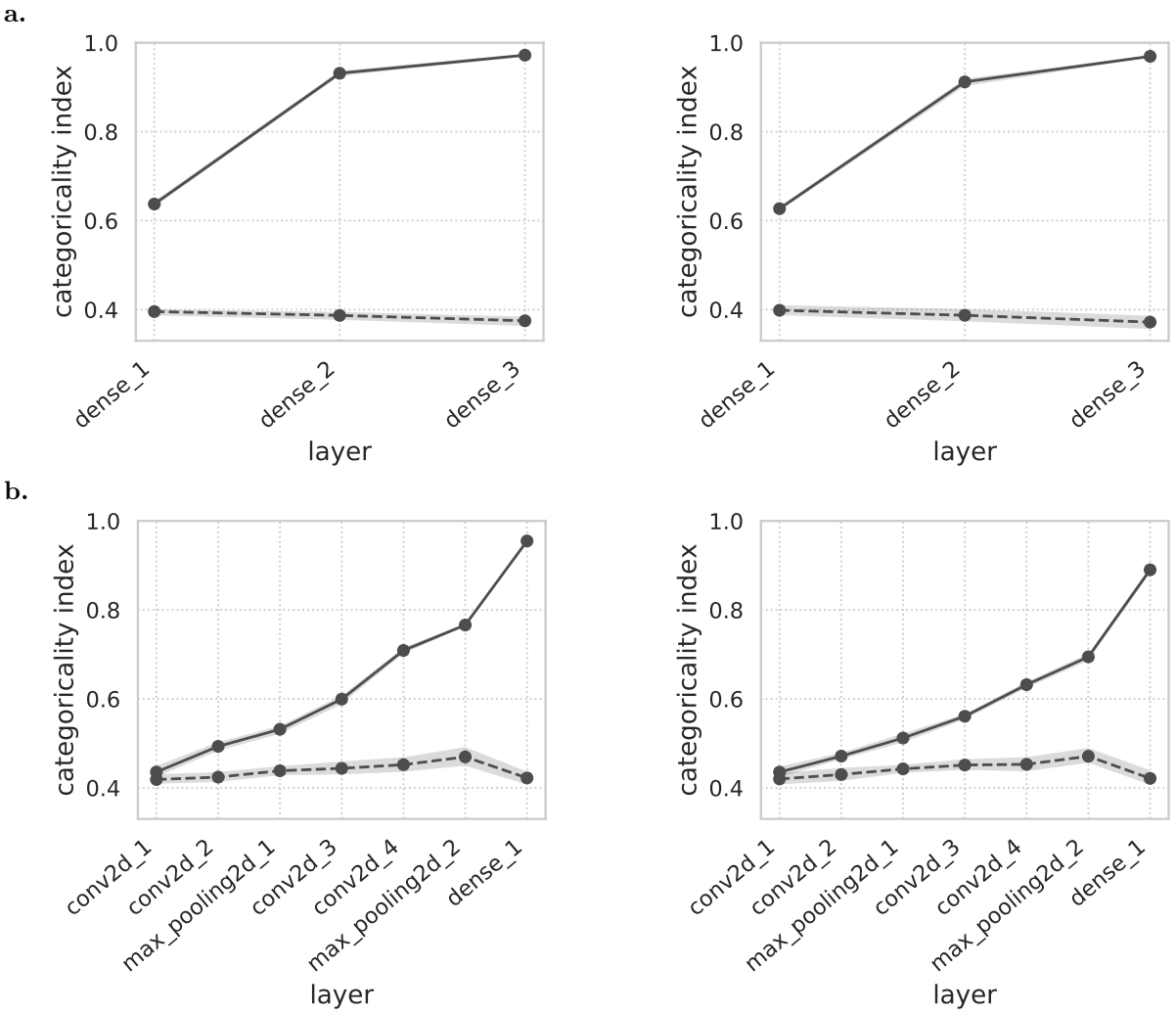

Figure B.1: **Categoricality as a function of layer depth, using the MNIST dataset, with and without the use of dropout**. (Left) Reproduction of Fig. 6. (Right) Same, but without the use of dropout.

Appendix C Supplementary figure for the cat/dog example: Distance in input space

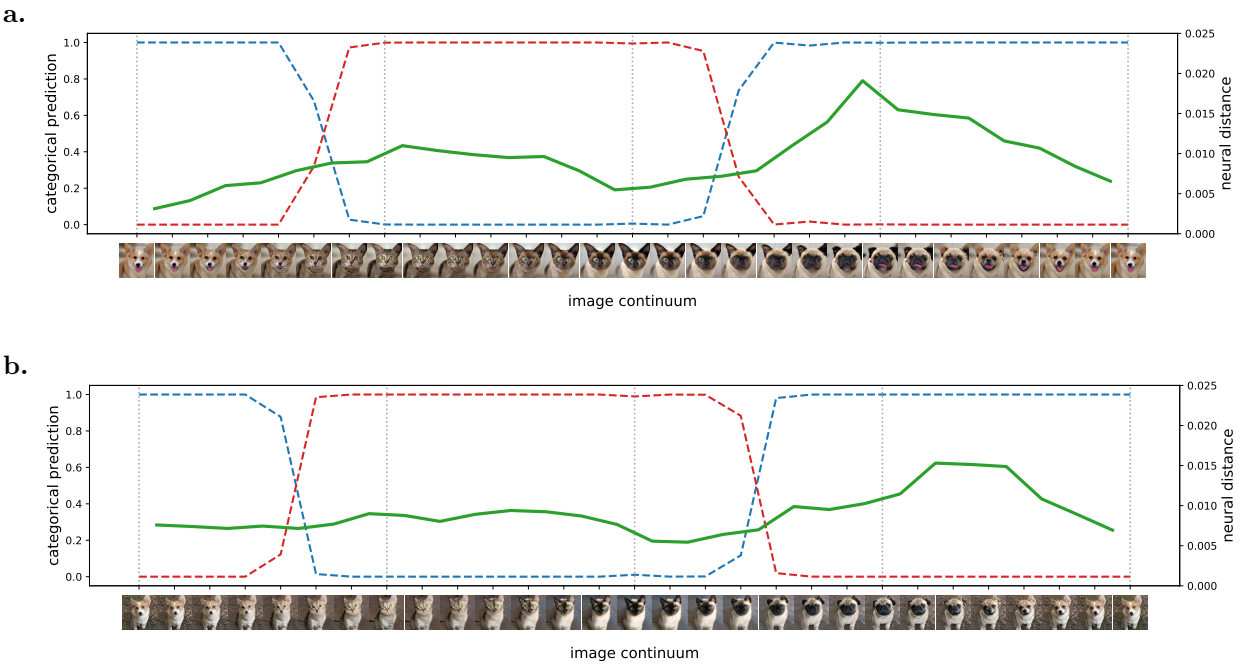

Figure C.2: **Categorical perception of a cat/dog continuum: Distance in input space**. Same legend as in Fig. 7. The green solid line corresponds to the distance in input (pixel) space between adjacent items along the continuum.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abramson, A. and Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In *Proc. of the VIth ICPhS Prague*. Academia.

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

An, G. (1996). The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., and Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological review*, 84(5):413.

Beale, J. and Keil, F. (1995). Categorical effects in the perception of faces. *Cognition*, 57:217–239.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. (2013). Better mixing via deep representations. In *International conference on machine learning*, pages 552–560.

Berg, A., Deng, J., and Fei-Fei, L. (2011). Large scale visual recognition challenge 2010, 2010. *URL http://www. image-net. org/challenges/LSVRC/2010/index*.

Berlemont, K. and Nadal, J.-P. (2020). Confidence-controlled hebbian learning efficiently extracts category membership from stimuli encoded in view of a categorization task. *bioRxiv*.

Bidelman, G. M., Moreno, S., and Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage*, 79:201–212.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116.

Blahut, R. E. (1987). *Principles and practice of information theory*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Bonnasse-Gahot, L. and Nadal, J.-P. (2008). Neural coding of categories: Information efficiency and optimal population codes. *Journal of Computational Neuroscience*, 25(1):169–87.

Bonnasse-Gahot, L. and Nadal, J.-P. (2012). Perception of categories: from coding efficiency to reaction times. *Brain Research*, 1434:47–61.

Bonnasse-Gahot, L. and Nadal, J.-P. (2020). Category learning in deep neural networks: Information content and geometry of internal representations. *In preparation*.

Bornstein, M. and Korda, N. (1984). Discrimination and matching within and between hues measured by reaction times: some implications for categorical perception and levels of information processing. *Psychological Research*, 46:207–222.

Bouthillier, X., Konda, K., Vincent, P., and Memisevic, R. (2015). Dropout as data augmentation. *arXiv preprint arXiv:1506.08700*.

Burns, E. M. and Ward, W. D. (1978). Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *The Journal of the Acoustical Society of America*, 63(2):456–468.

Caves, E. M., Green, P. A., Zipple, M. N., Peters, S., Johnsen, S., and Nowicki, S. (2018). Categorical perception of colour signals in a songbird. *Nature*, 560(7718):365–367.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley & Sons, NY, USA. Second Edition.

Cross, D., Lane, H., and Sheppard, W. (1965). Identification and discrimination functions for a visual continuum and their relation to the motor theory of speech perception. *Journal of Experimental Psychology*, 70(1):63.

Damper, R. and Harnad, S. (2000). Neural network models of categorical perception. *Percept. Psychophys.*, 62(4):843–867.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, 23(12):5235–5246.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178–200.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goto, H. (1971). Auditory perception by normal japanese adults of the sounds "" and "r.". *Neuropsychologia*.

Harnad, S., editor (1987). *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, 585(7825):357–362.

Holmstrom, L. and Koistinen, P. (1992). Using additive noise in back-propagation training. *IEEE transactions on neural networks*, 3(1):24–38.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Iverson, P. and Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & psychophysics*, 62(4):874–886.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kluender, K., Lotto, A., Holt, L., and Bloedel, S. (1998). Role of experience for language-specific functional mappings of vowel sounds. *Journal of the Acoustical Society of America*, 104(6):3568–3582.

Kreiman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature neuroscience*, 3(9):946–953.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, 72(4):275.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Liberman, A., Harris, K., Hoffman, H., and Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–369.

Liberman, A., Harris, K. S., Eimas, P., Lisker, L., and Bastian, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 4(4):175–195.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Matsuoka, K. (1992). Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):436–440.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, 11(1):5725.

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology*, 100(3):1407–1419.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*.

Morerio, P., Cavazza, J., Volpi, R., Vidal, R., and Murino, V. (2017). Curriculum dropout. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3544–3552.

Nadal, J.-P. and Parga, N. (1994). Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 5(4):565–581.

Nelson, D. A. and Marler, P. (1989). Categorical perception of a natural stimulus continuum: birdsong. *Science*, 244(4907):976–978.

Padgett, C. and Cottrell, G. W. (1998). A simple neural network models categorical perception of facial expressions. In *Proceedings of the twentieth annual cognitive science conference*, pages 806–807. Citeseer.

Park, S. and Kwak, N. (2016). Analysis on the dropout effect in convolutional neural networks. In *Asian conference on computer vision*, pages 189–204. Springer.

Rennie, S. J., Goel, V., and Thomas, S. (2014). Annealed dropout training of deep networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 159–164. IEEE.

Repp, B. (1984). Categorical perception: issues, methods, findings. In *Speech and Language: Advances in Basic Research and Practice*.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Schwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv:1703.00810*.

Seung, H. S., Sompolinsky, H., and Tishby, N. (1992). Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spilsbury, T. and Camps, P. (2019). Don't ignore dropout in fully convolutional networks. *arXiv preprint arXiv:1908.09162*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Tijsseling, A. and Harnad, S. (1997). Warping similarity space in category learning by backprop nets. In *Proceedings of SimCat 1997: Interdisciplinary workshop on similarity and categorization*, pages 263–269. Department of Artificial Intelligence, Edinburgh University.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *The Journal of the Acoustical Society of America*, 60(6):1381–1389.

Zhao, D., Yu, G., Xu, P., and Luo, M. (2019). Equivalence between dropout and data augmentation: A mathematical check. *Neural Networks*, 115:82–89.