A model of category-learning difficulty

COGS444: HONOURS RESEARCH PROJECT

SOLIM LEGRIS

Supervised by Prof. Stevan Harnad at McGill University

August 31, 2021

Abstract

Categorization is a fundamental cognitive ability both for humans and other animals. Characterizing and quantifying the degree of difficulty of category learning is important to understanding how the brain categorizes. It is known that different categories are learned with varying degrees of difficulty. The present paper seeks to elucidate, using a neural net model of experimental outcomes in categorization tasks, (1) what makes some categories more difficult to learn than others and (2) how does task difficulty relate to categorical perception (CP), the phenomenon in which the internal representation of similarities is modified by category learning such that inputs belonging to different categories come to be perceived as more different after the category is learned and inputs belonging to the same category come to be perceived as more similar. Four parameters are defined in this paper to control and evaluate category learning difficulty. We show how these parameters relate to category-learning difficulty and in turn provide clues as to how category-learning difficulty may be related to CP.

Keywords: categorization, complexity, categorical perception, neural net.

Introduction

A key property of the human brain is its ability to make sense of inputs from an otherwise noisy and chaotic world. One of the fundamental mechanisms underlying this ability is categorization. The categories that an organism recognizes guide its behavior such that it behaves differently towards different kinds of objects and events (e.g., approaching, avoiding, eating, manipulating, and, in the case of humans, naming and describing). Categorization is the process by which living organisms learn to do the right thing with the right kind of thing (Harnad, 2017). In this general sense, categorization underlies much of cognition. It underlies the capacity of living organisms to deal adaptively with the vast variation in their world. Through categorization, continuous analog sensory perception becomes discrete symbolic processing. Identifying the members of a category to which a stimulus belongs is based on detecting the features that distinguish it from members of other categories.

Although there are well-studied examples of innate categories in the scientific literature, (e.g., in color perception; Neitz & Neitz, 2017), most categories are not inborn but learned and acquired throughout a lifetime. An organism's categories depend on their sensorimotor affordances (Gibson, 2014) — the possible interactions that the features of an object "afford" (allow) given the structure of the object and the sensorimotor and somatic structure of the organism. For example, a keyboard affords typing to an organism with hands but not to one with hooves. Identifying relevant features and discarding or ignoring irrelevant ones, a process called *feature detection*, allows an organism to learn categories. Learning to categorize can occur through two types of learning: "unsupervised" and "supervised".

Unsupervised learning is a passive form of learning: the intrinsic structure of a sample of stimuli is learned by detecting statistical regularities such as feature frequencies or correlations. In *supervised* (or reinforcement) *learning* an organism or algorithm responds to inputs and is provided with corrective feedback signaling whether the response was right or wrong. This feedback can come from the environment, a teacher, an error signal, a reinforcement signal, etc. Through supervised learning, an organism learns to assign stimuli to distinct categories as a result of the corrective feedback so it eventually stops making errors (if the categories are learnable).

Learning categories is the process through which the features that distinguish the categories are detected, allowing the learning system to identify which inputs are members and which are not. Of all the features of inputs, only the subset of them that is relevant for distinguishing members from non-members needs to be detected. This subset of features is called the *category-covariant subset* and it can generally be described as a Boolean rule that defines the possible combinations of a category's covariant features. The rule defines one composite feature. The rest of the features of a stimulus can be ignored.

Some categories are inherently more difficult to learn than others: what is it that makes them more difficult? A stimulus space can be *partitioned* in many ways into many different categories. Our working assumption is that properties of the *structure* of a stimulus space together with the complexity of the Boolean rule on the category-covariant features that can partition the space into categories will predict how difficult it will be to learn the partition.

Various complexity measures for category learning have been proposed. Most are restricted to what are known as *explicit rule-based*[1] category-learning tasks. In such tasks, category learning difficulty has been shown to be predicted by the complexity of the rule that describes the categories verbally (Feldman, 2000; Pape et al., 2015; Vigo, 2009). The complexity of these explicit verbal (or symbolic) rules has been evaluated using methods derived from Boolean algebra, computational models and information theory. There are two differences between learning categories based on *explicit* rules that describe their distinguishing features verbally using already known and named features and the learning studied in the present paper:

(1) The features for distinguishing categories in our model do not already have names and would hence be difficult to describe explicitly in words: The rules are nonverbal, hence *implicit* rather than explicit. The feature learning is hence more directly sensory rather than verbal in our model.

[1] An example of an explicit verbal or symbolic category rule: an object is an A if it has features a_1 and a_2 and it is a B if it has features b_1 and b_2.

(2) We are interested not only in category *learning* but also in changes in category *perception* induced by learning. In our model, these take the form of changes in the distances between categories in an internal similarity space generated by *changes in the weights on sensory feature detectors*: As category-distinguishing features are detected and their weights increase, members of different categories become more separated in similarity space and members of the same category become more compressed.

We think most categories that living organisms learn are based on learned perceptual differentiation of this kind. Categorization based on explicit verbal rules, because it requires language, is a special case that is unique to human category-learning. We are accordingly seeking a measure of the difficulty of category learning that is agnostic about whether the rules underlying the features that distinguish categories are verbalized explicitly or are simply implicit in sensory feature detection.

In human experiments it has been found that under certain conditions category learning induces perceptual changes such that after learning, people perceive the members of different categories as looking (or sounding) more different from one another, and members of the same categories as more similar to one another (Harnad, 2003; Notman et al., 2005; Goldstone & Hendrickson, 2010; Pérez-Gay et al., 2017). This phenomenon, called learned categorical perception (CP), seems to reflect subtle perceptual changes that organisms undergo in learning to categorize.

Through CP, organisms as well as machine learning algorithms (Bonnasse-Gahot & Nadal, 2020; Damper & Harnad, 2000) learn to assign inputs to different categories by altering perception to detect and highlight sensory features that differentiate the categories. A question naturally arises about the relation between this modification of similarity space and how difficult it is to learn a category. If what must be learned are the perceptual features that distinguish members of different categories, what makes categories easier or harder to learn? Our hypothesis is that learning difficulty is proportional to how much the internal representations of inputs in similarity space need

to be modified (i.e., separated/compressed, as in learned CP) for error-free categorization.

Figure 2. An example of 6 binary pairs of visual microfeatures (upper and lower are mutually exclusive pairs) used to construct the texture stimuli in the experiments conducted by Pérez-Gay et al. (2019).

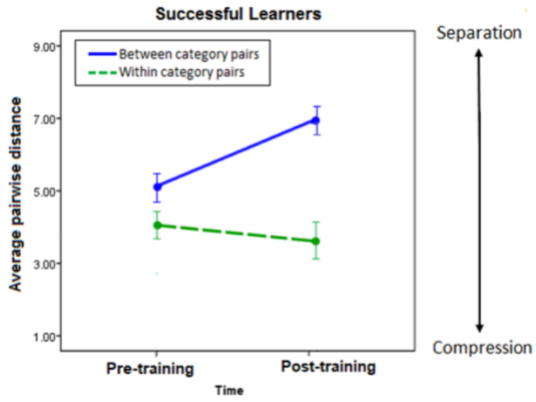

Figure 1. Human experimental data showing how pairwise stimulus dissimilarity judgments changed from before to after successfully learning the category through trial and error training with corrective feedback (adapted from Pérez-Gay et al. 2019).

In the next section we present a framework for quantifying category-learning difficulty that provides correlates and predictors of category-learning difficulty and relates CP and category-learning difficulty using algorithmically generated formal stimuli in a neural net model proposed by Thériault et al. (2018).

Methods

Previous experiments

It has been shown using algorithmically generated texture-like images that (1) human subjects could learn to identify them as members of one or the other of two categories, A and A', based on their distinguishing features with an accuracy of at least 80% and (2) that successful learning resulted in CP effects - between-category separation and within-category compression in pairwise dissimilarity (distance) judgments.

Membership in the two mutually exclusive categories A and A' were based on the presence of binary microfeatures, randomly distributed in each texture. The mutually exclusive binary pairs of microfeatures for each A and A' came from one of three sets: A_k, ${A'}_k$ or I_{N-k} where $A_k$and ${A'}_k$ are the sets of k paired category-relevant microfeatures of category A and A' respectively and I_{N-k} is the set of $N-k$ category-irrelevant microfeatures (noise). Relevant microfeatures are defined as those that covary with category membership meaning that all of the k category-relevant microfeatures from each respective set must be present in each A or A' texture. Irrelevant features are just those that are uncorrelated with category membership and hence provide no information to the learner. The rule that determined category membership of the images was conjunctive: the presence of the k covariant features of A_k and ${A'}_k$ for members A and A' respectively where $A_k \cap {A'}_k = \emptyset$.

In the experiments of Pérez-Gay Juárez et al. (2019) participants were first exposed to the stimuli through a pairwise (dis)similarity rating task. If any learning occurred during this task, it was unsupervised. Next the participants underwent 400 training trials of supervised learning in which at every presentation of a stimulus they were required to indicate what category it belonged to; they were then immediately given feedback on whether their response was correct or incorrect. The a-priori levels of difficulty were defined as $\frac{k}{N}$, the proportion of covariant features relative to the total number of features of each category. The assumption behind this measure of difficulty was that stimuli with a ratio $\frac{k}{N} = 1$ would be the easiest to categorize since all features in such stimuli are relevant to categorization and $I_{N-k} = \emptyset$ meaning that there is no noise. Conversely, as $\frac{k}{N}$ decreases, the number of irrelevant features increases while the number of relevant features decreases, leading to a more difficult categorization task. It was observed that with $N = 6$ and $3 \leq k \leq 6$ as k decreased, the difficulty of the categorization task increased, as indicated by the fact that the number of trials required to learn successfully increased and the number of participants that succeeded in learning with an accuracy of 80% or more decreased. Neural network simulations with unsupervised and supervised learning also showed that the $\frac{k}{N}$ ratio was a determining factor in category learning difficulty for conjunctive categories.

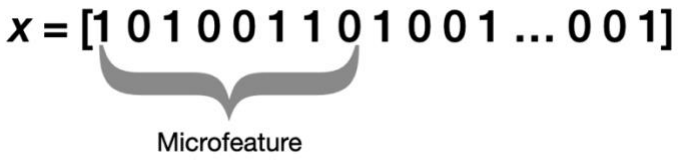

Figure 3. An example of a stimulus fed to the neural nets in our experiments. Each microfeature was a binary vector of length 8. Each stimulus of length 256 contains 32 such microfeatures.

The neural net model

In the present paper, the categorization task was modelled using a general deep neural network (LeCun et al., 2015), inspired by previous simpler models (Harnad et al., 1995; Greco et al., 1997; Damper & Harnad, 2000) and more recent developments in categorization and categorical perception research (Pérez-Gay Juárez et al., 2019; Thériault et al., 2018). Neural nets can be thought of as idealized feature learners that can categorize inputs by detecting and weighting their features. To match human categorization tasks, the neural net architecture used in the present paper consists of a denoising autoencoder (Vincent et al., 2008) that feeds into a supervised classification layer (Thériault et al., 2018). The denoising autoencoder learns through auto-association to generate the inputs it receives during training from compressed and perturbed representations. To do so, the network is forced to learn relevant features that will allow it to reconstruct the denoised representation from the perturbed inputs it is fed. This portion of the model stands in for the pairwise similarity ratings of human participants before category learning. Once the autoencoder has learned the appropriate representation space, the internal representations of the autoencoder are then fed to the rest of the net so that it learns the category labels through supervised learning. During this portion of the training, the net learns the categories through error-corrective feedback in the form of gradient descent backpropagation based on the difference between the output of the net and the correct category label. This results in the learning of the appropriate feature and weightings, leading to successful categorization of the stimuli after enough trials. More details on the technical aspects of the implementation of the neural net model are provided in the Appendix.

The stimuli

The stimuli used in our experiments consisted of binary vectors $\boldsymbol{x} \in \{0,1\}^N$ where $N = 256$. They were constructed using randomly generated sets of binary microfeatures $\boldsymbol{U_M} = \{m_1, m_2, m_3, \dots, m_l\}$ where $\boldsymbol{M} = 8$ is the length of the binary vector representing each microfeature and $\boldsymbol{l} = 32$ is the total number of microfeatures for a given set of mutually exclusive categories $\boldsymbol{A}$ and $\boldsymbol{B}$. In total, 384 sets $\boldsymbol{U_M}$ were randomly generated. For each category a set $\boldsymbol{U_M}$ was chosen at random and from it three mutually exclusive subsets $\boldsymbol{M_A}$, $\boldsymbol{M_B}$ and $\boldsymbol{Z}$ were randomly generated where $\boldsymbol{M_A}$ is the set of relevant microfeatures for category $\boldsymbol{A}$, $\boldsymbol{M_B}$ is the set of relevant features for category $\boldsymbol{M_B}$ and $\boldsymbol{Z}$ is the set of irrelevant microfeatures. For each category, $\boldsymbol{M_A} \cap \boldsymbol{M_B} = \emptyset$ and $|\boldsymbol{M_A}| = |\boldsymbol{M_B}| = \boldsymbol{k}$. Unlike in the case of the textures, the distribution of the microfeatures was parameterized and varied across categories as explained in detail below. Four parameters were used to control a-priori category-learning difficulty: $\boldsymbol{k}$, $\boldsymbol{d}$, $\boldsymbol{p_d}$ and $\boldsymbol{p_{noise}}$. Following is an account of each parameter, its role and hypotheses related to it.

Covariant microfeatures ($\boldsymbol{k}$)

For each category, $\boldsymbol{A}$ and $\boldsymbol{B}$, the number of covariant microfeatures ($\boldsymbol{k}$) varied from one to twelve. At minimum, at least one of the twelve binary microfeatures covaried

with category membership and the rest were category-irrelevant. At maximum, all twelve microfeatures covaried with category membership. Naturalistic categories have potentially an infinite number of features from which a subset is relevant to categorization. Sometimes, very few features are relevant to whether something belongs to a category, other times many features in combination or alone are relevant. It is hypothesized that on average across other parameters, it should be easier to categorize as $\boldsymbol{k}$ increases (as suggested by the results of human experiments). Given higher values of $\boldsymbol{k}$ and a finite number of features, less of the input is noise and as a result it is easier to determine which stimulus belongs to what category. It is important to note that higher values of $\boldsymbol{k}$ in a finite space also imply a reduction in within-category variance because of the lower proportion of possible irrelevant features in the stimuli.

The number of covariant locations ($\boldsymbol{d}$)

Features cannot always occur at every location on an object, as in our textures with spatially distributed microfeatures. Sometimes the location at which a feature occurs is also relevant. For example, consider an object that has something that looks like a nose, two eyes and a mouth but for some reason they are all over the place. This object is unlikely to be categorized as a face, even though it has some of the relevant features of faces. One of the features is missing: that the nose, eyes, and mouth be positioned correctly. As a result, we choose $\boldsymbol{d}$ locations where relevant information is to be found and hypothesize that as $\boldsymbol{d}$ grows, categorization becomes easier because most of the stimulus contains relevant information. Note that this is like the $\boldsymbol{k}$ parameter with respect to the decrease in within-category variance: as $\boldsymbol{d}$ increases, more of the stimulus contains relevant information (e.g., covariant features). For our experiments, $d \in \{2, 4, 6, \ldots, 28\}$.

The invariance distribution parameter ($\mathbf{p_d}$)

At any one of $\boldsymbol{d}$ locations, the parameter $\boldsymbol{p_d}$ determines which covariant feature can occur there. When learning to categorize, the learner is effectively learning the within-category invariance. The rule that defines a category also defines the what invariance the learner must detect for error-free categorization. The rule can define the location of relevant features (as the parameter $\boldsymbol{d}$ does) or what the relevant features are (as the parameter **k** does). The parameter $\boldsymbol{p_d}$ defines the aspect of category rules that determines how the relevant features are distributed at the relevant locations. At one end of the invariance spectrum, there is spatial or local invariance where $\boldsymbol{p_d} = 1$. This condition implies that in all stimuli of a given category, the same $\boldsymbol{k}$ covariant microfeature(s) occur at the same $\boldsymbol{d}$ locations. In these categories, the rule that must be learned can be thought of as a conjunction of the $\boldsymbol{k}$ covariant microfeatures over the $\boldsymbol{d}$ locations. The within-category invariance in this case is just that any member of the category can only have *one* of the $\boldsymbol{k}$ covariant microfeatures at each of the **d** locations (e.g., each relevant feature k_i always occurs at some location d_l for all members of the category). In contrast, distributed invariance, where $\boldsymbol{p_d} = 12$, is the condition in which any of $\boldsymbol{k}$ covariant microfeatures can occur at any of $\boldsymbol{d}$ locations. The rule for these

categories can be thought of as a disjunction of the $\boldsymbol{k}$ covariant microfeatures over the $\boldsymbol{d}$ locations. Here the within-category invariance is that any member of the category can have *any* of the $\boldsymbol{k}$ covariant features at any of its $\boldsymbol{d}$ locations. It is hypothesized that as invariance tends towards maximal distribution, categorization difficulty increases because within-category variance also increases (e.g., the number of possible combinations of relevant microfeatures that fall under the category rule increases). For our experiments, $1 \leq \mathbf{p_d} \leq 12$ *and* $\mathbf{p_d} \leq \mathbf{k}$.

The noise parameter ($\boldsymbol{p_{noise}}$)

The noise parameter allows us to perturb categories by controlling how certain the occurrence of any of the $\boldsymbol{k}$ microfeatures at any of the $\boldsymbol{d}$ locations is. The higher its value, the more uncertainty is introduced in the category. This just means that the parameter allows us to introduce randomness in the stimuli. Consequently, it is hypothesized that as $\boldsymbol{p_{noise}}$ increases, so does category-learning difficulty. For our experiments, $0 \leq \boldsymbol{p_{noise}} \leq 0.4$.

The simulations

In total, 4368 simulations were conducted with different categories with all possible combinations of the four parameters $\boldsymbol{k}, \boldsymbol{d}, \boldsymbol{p_d}, \boldsymbol{p_{noise}}$ within the constraints described above. The hyper-parameters of the neural net (see Appendix) were kept constant. Every neural net was trained for a set number of 30 epochs and with category samples of size 2048. Category-learning difficulty was measured using the minimal loss achieved by the neural net after supervised learning, which amounts to considering how closely the neural net's outputs matched the correct outputs on average after a set number of training epochs. As such, minimal loss achieved is proportional to category-learning difficulty. We consider this measure to be more informative than classification accuracy since it is possible for a neural net to categorize correctly without having fully or sufficiently minimized its loss function.

Results

In our statistical analyses, all four parameters were evaluated independently as predictors of category learning difficulty or minimum loss achieved by the neural net. Moreover, parameters $\boldsymbol{k}, \boldsymbol{d}$ and $\boldsymbol{p_d}$ were evaluated as predictors of global CP scores. Global CP is a measure of CP that accounts for both between-category separation and within-category compression by subtracting the former to the latter (see Appendix). It was found overall that there was a strong correlation between learning difficulty and global CP scores after learning (see Figure 4), $r(4366) = -.81, p < .001, d = 1.24$. The effect size for this analysis ($d = 1.24$) was found to exceed Cohen's (1988) convention for a large effect ($d = .80$).

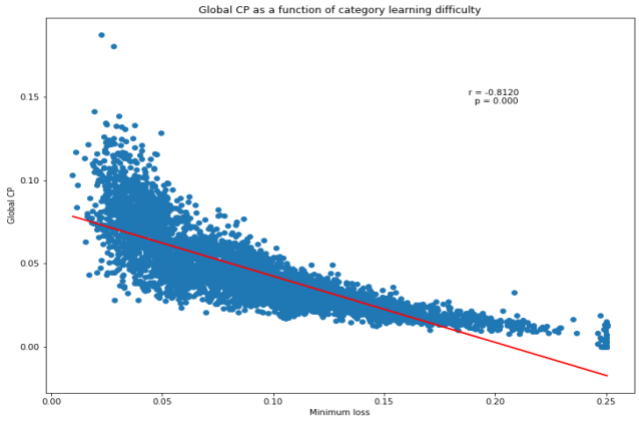

Figure 4. The general trend of the correlation between global CP scores and category-learning difficulty is illustrated here.

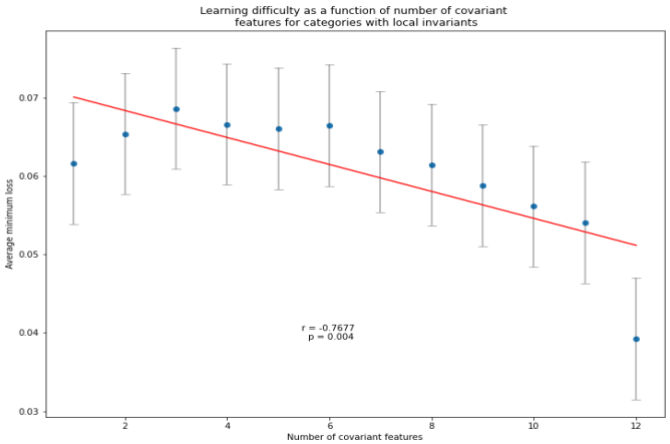

Figure 5. A negative trend is observed for minimum loss as a function of the number of covariant microfeatures for categories with local invariants.

Number of covariant microfeatures ($\boldsymbol{k}$) as a predictor of category learning difficulty and CP

We first analyzed the correlation between the number of covariant microfeatures ($\boldsymbol{k}$) and category learning difficulty for categories with local invariants and categories with distributed invariants. A strong negative correlation was observed for averaged category learning difficulty as a function of the number of covariant microfeatures ($\boldsymbol{k}$) for categories with local invariants (see Figure 5), $r(10) = -.77, p = .004$. In contrast, a strong positive correlation was observed for averaged category learning difficulty as a function of number of covariant microfeatures ($\boldsymbol{k}$) for categories with distributed invariants where $\boldsymbol{k} > 1$ and $\boldsymbol{k} = \boldsymbol{p_d}$ (see Figure 6), $r(10) = .73, p = .01$. We also analyzed the correlation between learning difficulty and the number of covariant microfeatures ($\boldsymbol{k}$) for data points grouped by the number of covariant microfeatures and averaged across all other parameters: this yielded a strong positive correlation (see Figure

7), $r(10) = .72, p = .008$. Last, the correlation between category-learning difficulty and global CP scores averaged across all parameters for data points grouped by the number of covariant features ($\boldsymbol{k}$) was strongly negative (see Figure 8), $r(10) = -.91,\ p < .001$.

Invariance distribution ($\boldsymbol{p_d}$) as a predictor of category learning difficulty and CP

Next, the correlation between category-learning difficulty and invariance distribution ($\boldsymbol{p_d}$) was evaluated. A strong positive correlation was observed for category-learning difficulty averaged across all other parameters for data points grouped by invariance distribution parameter value ($\boldsymbol{p_d}$) (see Figure 9), $r(10) = .85,\ p < .001$. Additionally, a strong negative correlation was observed between average global CP scores across all parameters for data points grouped by invariance distribution ($\boldsymbol{p_d}$) and invariance distribution parameter ($\boldsymbol{p_d}$) value (see Figure 10), $r(10) = -.81,\ p = .001$. In parallel with the results for the number of covariant features ($\boldsymbol{k}$), for data points grouped by invariance distribution, average global CP scores were very strongly negatively correlated with average category-learning difficulty (see Figure 11), $r(12) = -.97,\ p < .001$.

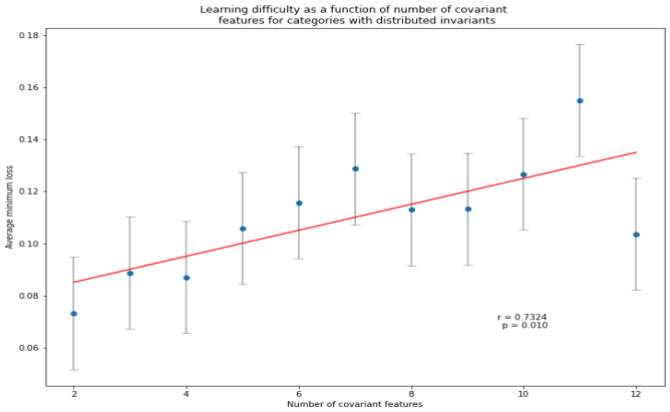

Figure 6. A positive trend is observed for minimal loss achieved as a function of the number of covariant microfeatures for categories with distributed invariants.

Number of relevant locations ($\boldsymbol{d}$) as a predictor of learning difficulty and CP

The correlation between the number of relevant locations ($\boldsymbol{d}$) and learning difficulty was evaluated. The average category learning difficulty was positively correlated with the number of relevant locations ($\boldsymbol{d}$) for data points grouped by number of relevant locations ($\boldsymbol{d}$) and averaged across all other parameters (see Figure 12), $r(12) = .85,\ p < .001$.

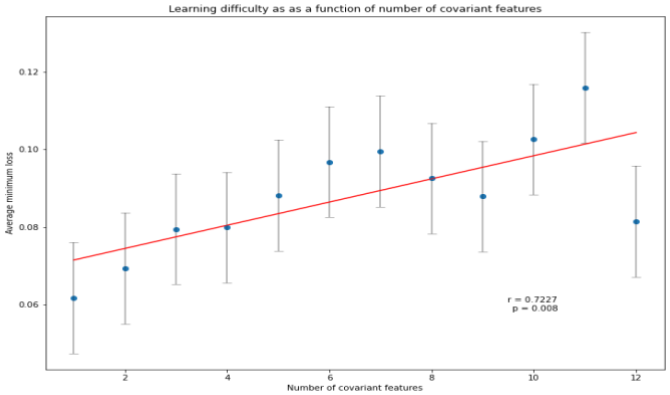

Figure 7. A positive trend is observed for minimum loss reached by the neural net as a function of the number of covariant micro-features for all categories.

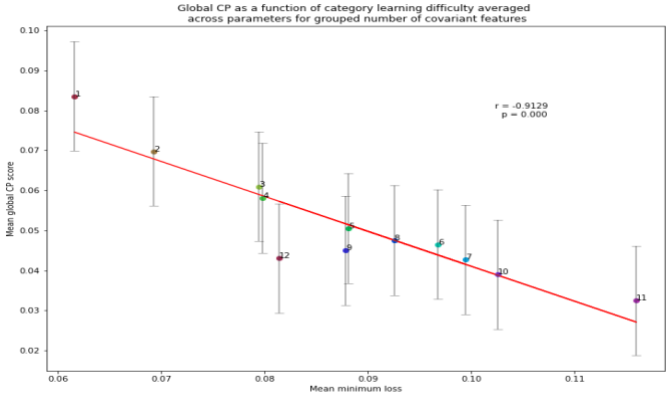

Figure 8. A negative trend is observed for global CP as a function of minimum loss achieved by the net for data points grouped by number of relevant locations.

Further analyses showed that average global CP scores were strongly negatively correlated with number of relevant locations ($\boldsymbol{d}$) for data points grouped by number of

relevant locations ($\boldsymbol{d}$) and averaged across all other parameters (see Figure 13), $r(12) = -.79,\ p = .001$.

Noise ($\boldsymbol{p_{noise}}$) as a predictor of learning difficulty

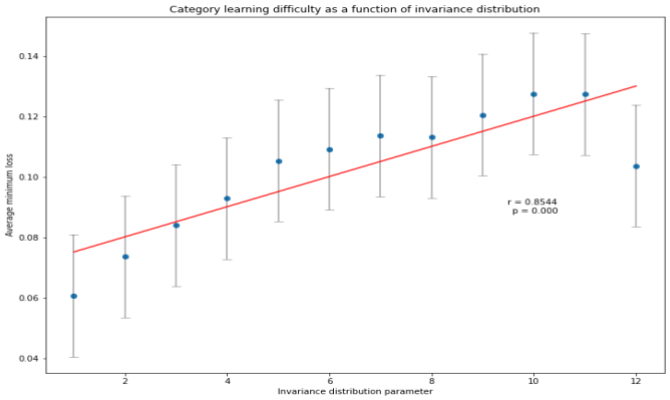

Figure 9. A positive trend is observed for minimum loss achieved as a function of invariance distribution for data points grouped by invariance distribution parameter.

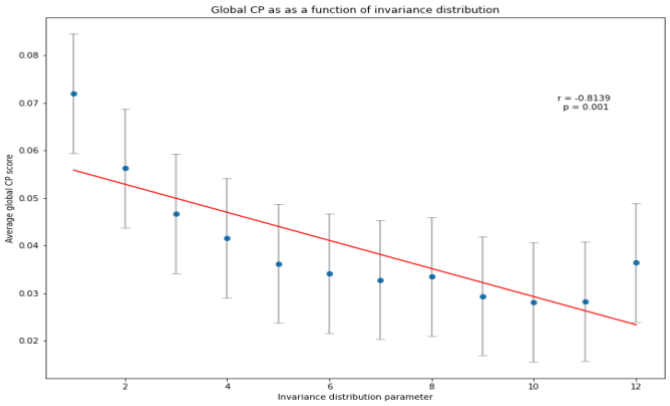

Figure 10. A negative trend is observed for global CP as a function of the invariance distribution parameter for data points grouped by the invariance parameter.

Last, a near perfect positive correlation was found between average learning difficulty and the stimulus noise parameter value ($\boldsymbol{p_{noise}}$) for data points grouped by stimulus noise parameter value ($\boldsymbol{p_{noise}}$) and averaged across all other parameters (see Figure 14), $r(4) = .99, p = .02$.

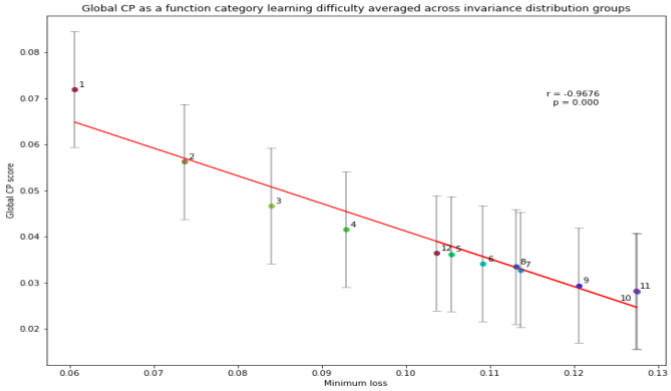

Figure 11. A negative trend is observed for global CP as a function of minimum loss achieved by the net for data points grouped by the invariance distribution parameter.

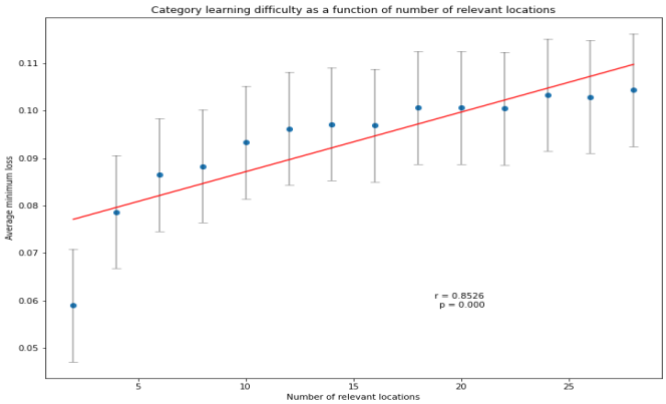

Figure 12. A positive trend is observed for minimum loss achieved by the net as a function of the number of relevant locations for data points grouped by the number of relevant locations.

Discussion

The present study sheds light on the intricacies of determining what aspects of category structure and rule may increase or decrease category-learning difficulty. In previous experiments (Pérez-Gay Juárez et al., 2019; Thériault et al., 2018), the stimuli generated for testing with humans had fixed parameters $\boldsymbol{d}$, $\boldsymbol{p_d}$ and $\boldsymbol{p_{noise}}$ whereas varying parameters $\boldsymbol{d}$ and $\boldsymbol{p_d}$ were used in neural net simulations. For the texture-like stimuli used in human experiments, the number of relevant locations ($\boldsymbol{d}$) was all possible locations in the stimuli, the invariance distribution parameter ($\boldsymbol{p_d}$) was maximal (i.e., features were uniformly distributed) and the noise parameter ($\boldsymbol{p_{noise}}$) was null. As for the stimuli used in the neural net simulations, the number of relevant locations ($\boldsymbol{d}$) was equal to the number of $\boldsymbol{k}$ covariant features, the invariance distribution parameter ($\boldsymbol{p_d}$) varied from 1 to $\boldsymbol{k}$ and the noise parameter ($\boldsymbol{p_{noise}}$) was again null. The a-priori levels of

difficulty were defined as the ratio of k/N as described above (see Methods) and were confirmed both in human and neural net experiments. It is important to note that the simulations in the present study, unlike those in previous studies, involved a fixed number of trials as opposed to a fixed accuracy criterion with unlimited trials since we were attempting to quantify category-learning difficulty and relate it to global CP.

We find that for a constant number of training epochs, as category-learning difficulty increases, measured by the minimum loss value achieved by the neural net, global CP scores decrease (see Figure 4). This trend is also observed for global CP scores

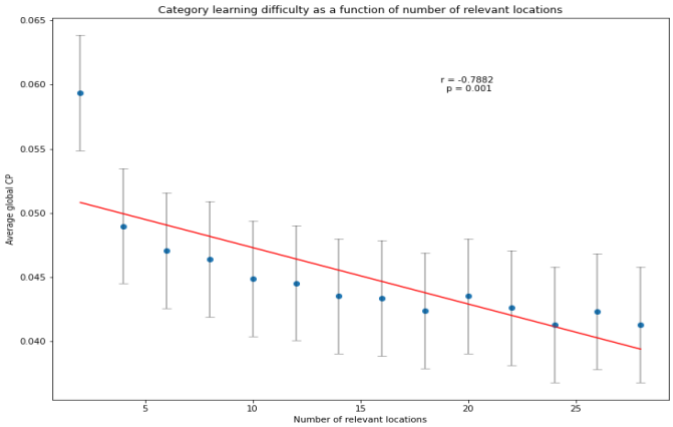

Figure 13. A negative trend is observed for global CP as a function of the number of relevant locations for data points grouped by number of relevant locations.

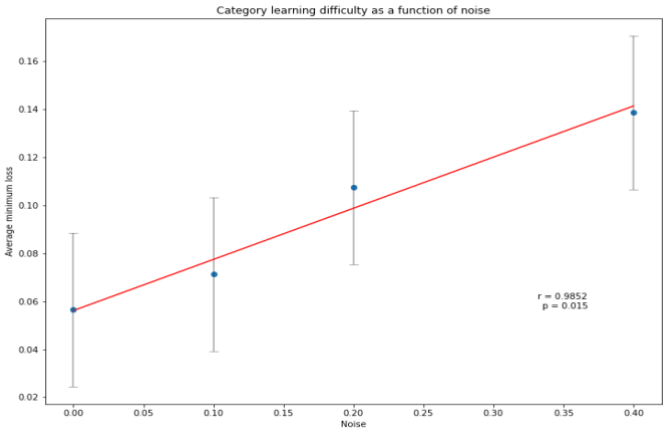

Figure 14. A positive trend is observed for minimum loss achieved by the neural net as a function of noise for categories grouped by noise parameter value.

and category-learning difficulty averaged for data points grouped by parameters $\boldsymbol{k}$ and $\boldsymbol{p_d}$ (see Figure 8 and Figure 11).

This result does not contradict what is expected in experimental outcomes with humans since more learning trials are needed to learn more difficult categories (Pérez-Gay Juárez et al., 2019). Our results suggest that still more reshaping of the stimulus representation space is needed to achieve high levels of categorization accuracy for more complex categories (see Figures 15 and 16). Therefore, we hypothesize that this additional reshaping of the similarity space will result in the higher global CP scores that are expected for more difficult categories when learning trials are unlimited until a fixed accuracy criterion is reached.

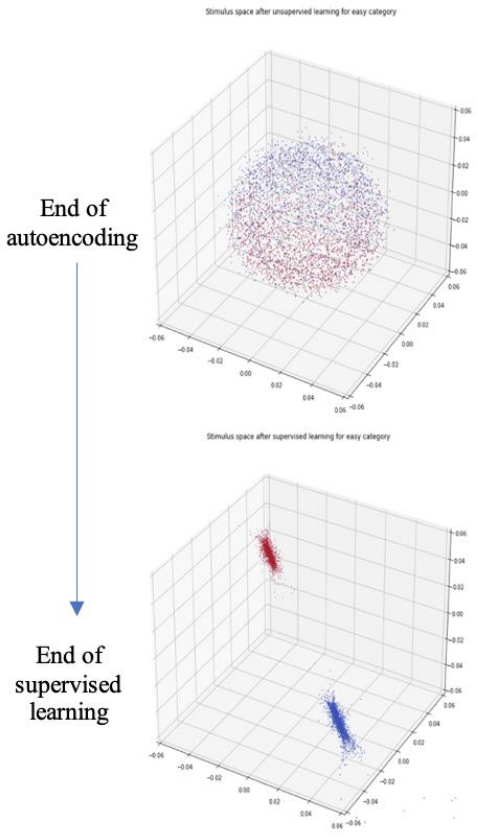

Figure 15. Transformation of stimulus representation space from before supervised learning to after supervised learning for easy category with error-free categorization.

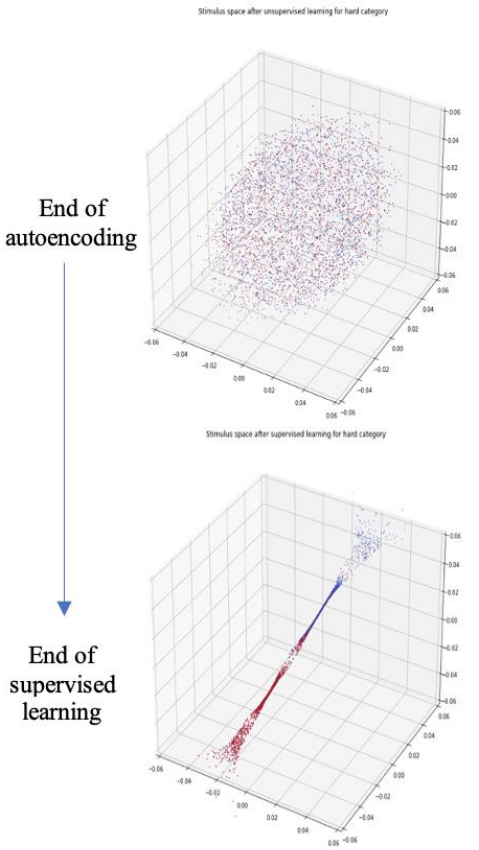

Figure 16. Transformation of stimulus representation space from before supervised learning to after supervised learning for hard category with imperfect categorization accuracy.

Partial confirmation of our hypothesis on the number of covariant features is given by the decrease in category-learning difficulty as $\boldsymbol{k}$ increases but only for *local* invariance where $\boldsymbol{p_d} = 1$ (see Figure 5). As for the extreme case of distributed invariance where $\boldsymbol{p_d} = 12$, we observe the initially counterintuitive result that as $\boldsymbol{k}$ increases, so does category-learning difficulty (see Figure 6). This is explained by the fact that as $\boldsymbol{k}$ increases for categories with distributed invariance, within-category variance increases combinatorially leading to the increased category-learning difficulty. Moreover, we observe that the general trend for increasing values of $\boldsymbol{k}$ is an increase in category-

learning difficulty (see Figure 7). This is explained by the fact that as $\boldsymbol{k}$ increases, on average category-learning difficulty also increases since the possibilities for distributed cases increase. For example, if we consider $\boldsymbol{k} = 1$, there is no possible distributed case since there is only one micro-feature and $\boldsymbol{p_d} = 1$ for all categories. As $\boldsymbol{k}$ increases, so do the possibilities for how distributed the invariance can be. In other words, there is possibility for $\boldsymbol{p_d} \geq 1$ leading to partially and fully distributed invariance far outweighing the difficulty of local cases.

An invariance distribution continuum was initially hypothesized to be correlated with category-learning difficulty, from local invariance as the easiest case to learn to fully distributed invariance as the most difficult case. This hypothesis was confirmed by our results which demonstrated that as the value of the invariance distribution parameter $\boldsymbol{p_d}$ increases, so does category-learning difficulty (see Figure 10). As explained above, an increase in within-category variance necessarily follows from an increase in $\boldsymbol{p_d}$. Moreover, all neural nets were tested with a constant category sample size. This implies that for categories with higher values of $\boldsymbol{p_d}$, the constant sample size is a smaller proportion of the category seen by the neural net, which results in higher categorization uncertainty and thus lower categorization accuracy.

The less conclusive results obtained for the number of relevant locations parameter ($\boldsymbol{d}$) suggest that on average as the number of locations increases, so does category-learning difficulty (see Figure 12). This result was unexpected since we hypothesized that increasing the number of relevant locations would decrease within-category variance. It is unclear whether this correlation is explained by the parameter $\boldsymbol{d}$ alone since further analysis suggests that there is almost no correlation between this parameter and category-learning difficulty when the data are analyzed independently for each possible value of $\boldsymbol{k}$ for categories where $\boldsymbol{p_{noise}} = 0$. It was expected that for small values of $\boldsymbol{k}$, a negative trend would be observed and that as $\boldsymbol{k}$ increased, a gradual reversal in the trend would be observed. For cases where $\boldsymbol{k} = 1$, as $\boldsymbol{d}$ increases, it is expected that category-learning difficulty would increase since in this case one covariant feature is increasingly present in the stimulus and the number of positions containing noise decreases proportionally. Conversely, for higher values of $\boldsymbol{k}$, the opposite trend would be expected since the possible combinations of $\boldsymbol{k}$ covariant features increase combinatorially, especially for highly distributed invariance as is suggested by the analysis above of the $\boldsymbol{p_d}$ parameter. Further testing and experimentation are necessary to elucidate the relation between the number of relevant locations and category-learning difficulty.

Last, we obtained strong evidence for the trivial claim that as perturbation increases within the category sample fed to the neural net, category-learning difficulty also increases (see Figure 15).

Limitations and Further Experiments

An important limitation of the present study is that the neural net architecture used for our simulations was not sufficiently complex to learn the harder categories with accuracy equivalent to that obtained for the easier categories, even with more learning trials. To test for the hypothesis that as category learning difficulty increases more global CP is necessary, a neural net capable of learning error-free categorization for all categories tested is necessary with recordings after some fixed number of learning trials *and* after error-free categorization is achieved. Furthermore, it is unclear what the weight of the invariance distribution parameter is for category learning-difficulty with respect to other parameters as demonstrated by the inconclusive results obtained with respect to the number of relevant locations ($\boldsymbol{d}$) and the negative trend obtained for category-learning difficulty as a function of the number of relevant locations ($\boldsymbol{k}$). As such, further testing and experiments will also need to account for the variance explained by each of the parameters in category-learning difficulty. Lastly, due to time constraints, we were unable to compute categorization task complexity using a priori measures such as those presented in Lorena et al. (2020). We believe that such measures would serve to cross-validate the category-learning difficulty framework presented here as well as increase explanatory power.

Conclusion

In the present study, our results suggest that as category-learning difficulty increases for a constant number of training epochs, global CP scores decrease. This is thought to be a consequence of insufficient reshaping of the stimulus representation space to allow for similar accuracy levels across categories with varying difficulty levels. Therefore, supposing that global CP is a measure of category separation in stimulus representation space, we hypothesize that for a fixed accuracy criterion with unlimited learning trials, global CP scores should be higher for more complex categories. Moreover, our results suggest that as the number of covariant locations increases, there is on average an increase in category-learning difficulty, but it remains unclear why that is the case. The same correlation has been suggested by our data for invariance distribution. That is, the more distributed the invariance is within stimuli of a category, the more difficult it is to learn that invariance for a constant number of epochs and category sample size.

Appendix

Neural net architecture

The neural net architecture used in the present study follows closely the one suggested by Thériault et al. (2018): A denoising autoencoder is fed noisy examples $\widehat{\boldsymbol{x}}$ and

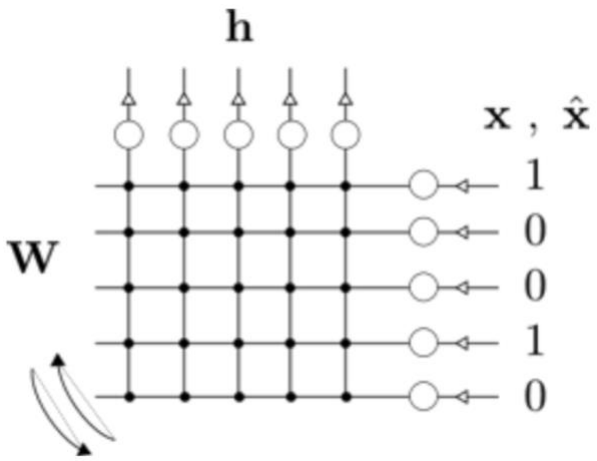

Figure 3. General autoencoder architecture. Adapted from Pérez-Gay et al. 2017.

trained to reconstruct learned examples $\boldsymbol{x}$. This forces some internal feature encoding. In total the model has 3 layers: the input layer, the hidden layer, and the output layer. The forward and feedback activation of layer $\boldsymbol{h}$ and layer **x** are respectively given by

$$\boldsymbol{h} = \boldsymbol{f}(\boldsymbol{W}\hat{\boldsymbol{x}} + \boldsymbol{b_h})$$

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{W^T h} + \boldsymbol{b_x})$$

where $\boldsymbol{f}$ is a non-linear activation function, $\boldsymbol{W}$ is the connection weights between the layers, and $\boldsymbol{b}$ is an activation bias. The weights $\boldsymbol{W}$ are progressively modified through gradient-descent error signal backpropagation to minimize the mean-squared error (MSE) loss function $\boldsymbol{MSE} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \left(\boldsymbol{x} - \boldsymbol{f}\left(\boldsymbol{h}(\hat{\boldsymbol{x}})\right)\right)^{\boldsymbol{2}}$ which is the mean squared difference between the net's predicted output and the correct output. Once the autoencoder has sufficiently minimized this loss, the inner representations (e.g., the encoded inputs) are used by the net to learn the categories through error-corrective feedback given by the MSE loss function and backpropagation. At each trial, the weights $\boldsymbol{W}$ are modified so that category labels are learned.

Global CP computation

To compute global CP scores, we first generate inner representations of all category samples for a given simulation immediately after the autoencoder has been trained. Euclidean distance matrices for within- and between-category distances are then computed using those representations. From these matrices, average within- and between-category distances are computed. Between- and within-category distance differences are then computed by subtracting distances before supervised learning from distances after supervised learning. The global CP scores consist in the difference between the mean between-category distance difference (separation) and the mean

within-category distance difference (compression). Let $\boldsymbol{D_A}$, $\boldsymbol{D_B}$ and $\boldsymbol{D_{A,B}}$ be the average Euclidean distances within A and B and between A and B respectively. We compute as follows between-category separation $\boldsymbol{SP_b}$, within-category compression $\boldsymbol{SP_w}$ and global CP scores $\boldsymbol{CP_G}$:

$$\boldsymbol{SP_w} = \frac{\left((\mathbf{D_A^s} - \boldsymbol{D_A^u}) + (\boldsymbol{D_B^s} - \boldsymbol{D_B^u})\right)}{2}$$

$$\boldsymbol{SP_b} = \mathbf{D_{A,B}^s} - \boldsymbol{D_{A,B}^u}$$

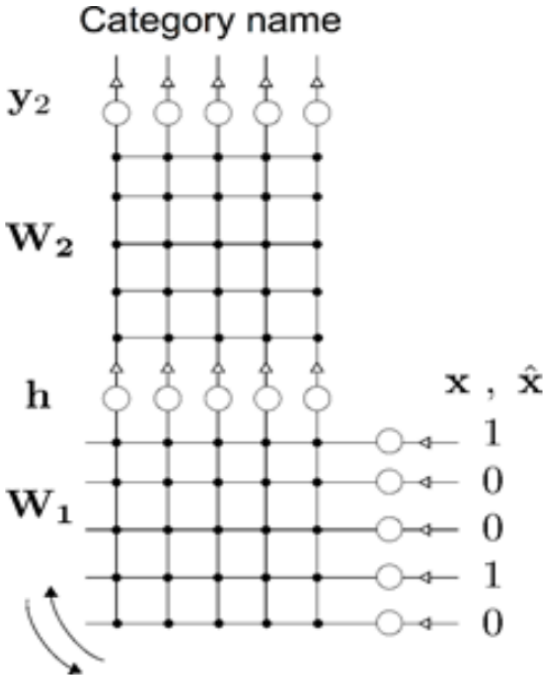

Figure 4. General autoencoder architecture feeding into a categorization layer. Adapted from Pérez-Gay et al. 2017.

$$\boldsymbol{CP_G} = \boldsymbol{SP_b} - \boldsymbol{SP_w}$$

Implementation details

The custom code used for the simulations, analyses and other computations was done in Python 3 using various machine learning libraries and will be made available on GitHub.

Bibliography

Bonnasse-Gahot, L., & Nadal, J.-P. (2020). Categorical Perception: A Groundwork for Deep Learning. *ArXiv:2012.05549 [Cs, Math, q-Bio]*. http://arxiv.org/abs/2012.05549

Damper, R. I., & Harnad, S. R. (2000). Neural network models of categorical perception. *Perception & Psychophysics*, *62*(4), 843–867. https://doi.org/10.3758/BF03206927

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633. https://doi.org/10.1038/35036586

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press. https://doi.org/10.4324/9781315740218

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Greco, A., Cangelosi, A., & Harnad, S. (1997, January). *A connectionist model of categorical perception and symbol grounding.* Proceedings of theAnnual Conference of Proceedings of the 15th Annual Workshop of the European Society for the Study of Cognitive Systems (01/01/97). https://eprints.soton.ac.uk/252865/

Harnad, S. (2003). Categorical Perception. In *Encyclopedia of Cognitive Science: Vol. LXVII* (No. 4; Issue 4). MacMillan: Nature Publishing Group. http://cogprints.org/3017/

Harnad, S. (2017). To Cognize is to Categorize. In *Handbook of Categorization in Cognitive Science* (pp. 21–54). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00002-6

Harnad, S., Hanson, S. J., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In *Symbol processors and connectionist network models in artificial intelligence and cognitive modelling: Steps toward principled integration* (pp. 191–206). Academic Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P., & Ho, T. K. (2020). How Complex is your classification problem? A survey on measuring classification complexity. *ArXiv:1808.03591 [Cs, Stat]*. http://arxiv.org/abs/1808.03591

Neitz, J., & Neitz, M. (2017). Evolution of the circuitry for conscious color vision in primates. *Eye*, *31*(2), 286–300. https://doi.org/10.1038/eye.2016.257

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: A psychophysical approach. *Cognition*, *95*(2), B1–B14. https://doi.org/10.1016/j.cognition.2004.07.002

Pape, A. D., Kurtz, K. J., & Sayama, H. (2015). Complexity measures and concept learning. *Journal of Mathematical Psychology*, *64–65*, 66–75. https://doi.org/10.1016/j.jmp.2015.01.001

Pérez-Gay, F., Thériault, C., Gregory, M., Sabri, H., Rivas, D., & Harnad, S. (2017). How and Why Does Category Learning Cause Categorical Perception? *Scholarly Publishing*, 32.

Pérez-Gay Juárez, F., Sicotte, T., Thériault, C., & Harnad, S. (2019). Category learning can alter perception and its neural correlates. *PLOS ONE*, *14*(12), e0226000. https://doi.org/10.1371/journal.pone.0226000

Thériault, C., Pérez-Gay, F., Rivas, D., & Harnad, S. (2018). Learning-induced categorical perception in a neural network model. *ArXiv:1805.04567 [Cs, Stat]*. http://arxiv.org/abs/1805.04567

Vigo, R. (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology*, *53*(4), 203–221. https://doi.org/10.1016/j.jmp.2009.04.009

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 1096–1103. https://doi.org/10.1145/1390156.1390294