Why Good Generative Models Categorize?

Serge Dolgikh

National Aviation University

Abstract

In this work connections between training processes of unsupervised generative learning with self-encoding and regeneration and information structure in the latent representations created by such models were investigated. Theoretical arguments were proposed leading to the conclusion, confirmed by previously published experimental results, that in generative self-learning under certain constraints latent representations with spontaneous categorization are statistically preferred. The results can provide insights into common principles underlying learning and emergence of intelligence in machine and biologic systems.

Keywords: artificial neural networks; unsupervised learning; general learning, Bayesian inference.

1. Introduction: Categorization in Unsupervised Self-Learning

A regular observation of a backyard naturalist confirms that a common behaviour of animals in nature areas adjacent to human settlements where people visit regularly is to approach a human in the hope of gaining food. It is exhibited even by more primitive species like fish and amphibians and is not specific to a particular human being.

Such a behaviour can have straightforward interpretation in a symbolized concept framework where an observation can be linked to certain general class denoted by a symbol, or concept. On the other hand, attempting to associate each observation directly to a specific behaviour can be complex and expensive in both complexity and performance. For example, it cannot be assured that a particular human or predator would be encountered again and such a direct association could prove useless for the survival, while consuming valuable and limited resources of the memory. Thus, it can be hypothesized [1] that biologic systems can employ conceptual processing of sensory observations for effective and efficient selection of behaviour.

In a number of previously published works, an interesting effect was observed: exposing certain models of unsupervised learning, such as autoencoder neural networks, to real-world data, without any prior knowledge of semantics or content, the law, parameters and characteristics of distribution etc., under certain conditions may lead to emergence of a concept-sensitive structure in the latent representations of the data created by learning models.

In [2] a massive sparse autoencoder model was trained on a large array of unlabelled images (over 10 million raw images) with the observed effect of emergence of concept-sensitive neurons activated by images of a certain higher-level category such as “cat’s face” without any prior knowledge of that concept. The emergence of higher-level concept correlated structure (unsupervised information landscape) was observed in [3,4] in representations created in unsupervised training by deep autoencoder models with Internet and image data. A substantial improvement in the performance of classification with data pre-processed with models of unsupervised learning (unsupervised feature learning) has now become a common practice in machine learning [5,6] among many others.

These results, observed in independent studies with data of different types and nature can lead to a question: is the effect of such native categorization in unsupervised learning coincidental with specific models or scenarios, or can in be caused by some principles of processing information that can be common for a range of learning systems, both artificial and biologic?

In this work we attempted to approach this question from a viewpoint of a theoretical analysis of distributions in the representations of unsupervised models with self-encoding and regeneration, of which autoencoder models represent a subclass. A general structure of such models is illustrated in Figure 1.

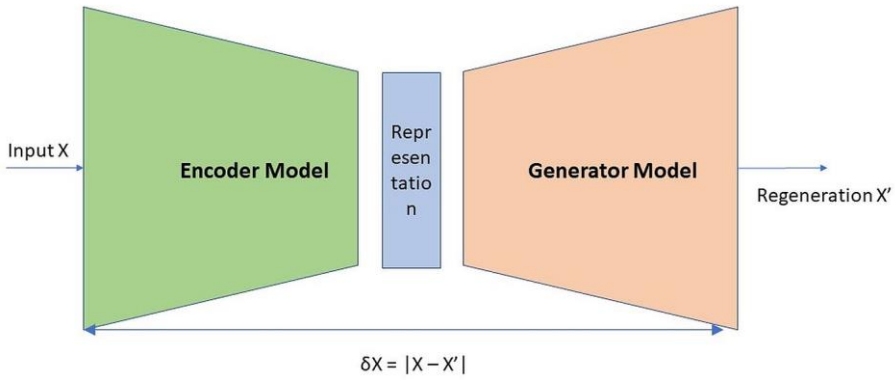

Figure 1. A generative self-learning model

Characteristic for these models is the presence of two essential components: the encoder that performs transformation from the space of input, or observable data to the representation space of the model; and the generator that reproduces, or generates an image in the observable space from a representation sample. Well-known examples of such models are Restricted Boltzmann machines and Deep Belief Networks [7,8] and autoencoder neural networks [9]. In the latter models used in the cited studies these components are combined into a single feed-forward network, with a training process for both encoding and generative components based on minimization of the deviation of regenerated distribution form the original one via one of the loss backpropagation methods, such as stochastic gradient descent (SGD) [10,11].

1.2. Principles of Unsupervised Learning

In energy-based learning [12] configurations with lower energy correspond to lower error of prediction over the training sample, and the aim of learning is to produce the configurations of the model (such as selection of its trainable parameters, weights and biases) with the lowest free energy.

In a number of works over the recent years, the principles of learning of artificial and biologic systems, including unsupervised, were studied. The relations between the principles of Bayesian inherence [13], the bottleneck principle [14] and the principle of minimization of free energy [12] were established, with the conclusion that these principles are essentially equivalent and therefore, compliance of the training process with any of these principles should produce the same result. In [15] methods used commonly for training of artificial neural networks such as Stochastic Gradient Descent were investigated and proven to be an approximation of the Bayesian inference principle. From these results as well as principles of energy learning discussed earlier follows an essential conclusion that will be used in the analysis of categorized representations further in the study, namely that applying training methods compatible with Bayesian inference can lead to configurations of learning models with minimized loss and free energy.

In this work we would like to consider the two sets of results discussed in this section from a common perspective, by asking the question: is there a general, independent of the setup and architecture of the learning model and type of data rule or principle that would connect geometric characteristics of distributions of data in the latent representations created by generative models under certain constraints in the learning process, and their ability to learn from the environment in an unsupervised process without prior knowledge of concepts being learned?

2. Unsupervised Categorization

Based on the empirical observations and results discussed earlier we will formulate the conjecture of unsupervised categorization in generative self-learning as follows:

2.1 Definitions

Data: input or observable data is described by observable parameters $\{x_i\}$ in the observable space *I*. In unsupervised training with self-encoding and regeneration, models create latent representations of the observable data space described by latent parameters $\{r_i\}$ in the representation space *R*.

Hidden (or native) concepts: we will consider the case where data contains significant presence of similar samples (in a certain, unknown to the model explicitly way) described by the set of hidden concepts $\{H_k\}$, possibly with contribution of non-categorized data, or random noise. Neither of: the set of hidden concepts; the relationship of similarity; distribution characteristics of samples in hidden concepts H_k or any other prior knowledge about the composition and distributions in the observable dataset is available to the learning models before or in the process of unsupervised learning.

Categorization: geometric categorization means the type of distributions of hidden concept samples in the latent representation space in which concept regions are compact and well separated from each other, so that both the average size / volume of a concept region and the overall volume of overlap between different concept regions are minimized:

$$Vol(H_i) \rightarrow min \tag{1}$$

$$Vol(O) = \sum (R(H_i) \cap R(H_j)) \rightarrow min$$

Hypothesis of Unsupervised Categorization*:* training of unsupervised models with self-encoding (from observable space to representation) and regeneration (from representation to observable space) with a procedure compliant with the principle of Bayesian inference under the constraints of stable accuracy; generalization; and redundancy reduction; statistically prefers latent representations with stronger geometrical categorization of hidden concepts in the representation space.

2.2 Preliminary Arguments

To start, let us consider two boundary scenarios. In the first scenario hidden concept distributions are well categorized to compact, dense clusters separated from each other.

In this case as can be seen immediately, the dispersion of a hidden concept in the representation space will be minimal $v(h) \sim \sqrt{(x(h) - x(h)_{mean})^2} / G << 1$, G being the characteristic size of the representation space.

The generative mapping from the concept region to its image in the observable space under stated assumptions can be a smooth function with smaller number of parameters that can be modeled with a finite neural network to any accuracy [16]. In the ideal case of a near-perfect categorization, the regenerative mapping for a given concept can be represented by a single point-to-point function – from the center of the concept cluster in the representation space (a representation "template" of the concept) to its image in the observable space, possibly with some local variation that constitutes the manifold of the observable concept region. Such a template mapping would be fully defined by $N_{reg} = D_{rep} + D_{obs} \sim D_{obs}$, (in the case of significant compression, $D_{rep} \ll D_{obs}$) parameters, where D_{rep} and D_{obs} are the dimensionalities of the latent and observable spaces, respectively.

Turning to the second scenario, let's consider the opposite case, whereby representations of hidden concepts are spread over the entire representation space and overlap significantly. Clearly, the dispersion component in this case $d_{spr} \sim G \gg d_{cat}$ in the first example.

Consequently, the mapping from concept regions in the representation space to different continuous manifolds representing concept images in the observable space will need to be more complex and variable, as close points belonging to essentially different concepts would have to be mapped to different concept manifolds in the observable space. This translates into more variation in the mapping function requiring more complex approximation with greater number of independent parameters than in the first example, and consequently, higher variation in the mapping component.

Moreover, as can be seen immediately, this scenario may not be compatible with the constraints of accuracy and generalization imposed simultaneously. Indeed, with each new set of data the density of samples of essentially different concepts in the same region of the representation space would increase continuously meaning that more and more complex approximations would be needed with each installment of data to map them to different regions in the observable space with constant accuracy. Eventually, either the accuracy constraint would need to be foregone, or the generality of the model would suffer, and a new set of data may cause the accuracy to drop.

These examples illustrate two essential points: that representations with better categorization should have lower generation loss and energy configurations of the generative model; and that

only well-categorized representations can support consistent generalization that is, the generative accuracy maintained over the data sets of any size and regardless of the volume or number of new installments, as long as the character of the data remains stable.

2.3 Categorized Configurations Minimize Generative Loss

Let us consider distributions of hidden concepts in the latent representation of a generative model that has been trained to minimize generative error with a representative training sample. We will attempt to prove that models with categorized representations as defined in (1) statistically exhibit lower generative error in empirical trials among the ensemble of learners of similar architecture.

Theorem (minimal loss / energy of categorized representations): *if a configuration with the best categorization of hidden concepts exists in a latent representation created by a generative model trained with minimization of generative error, it is also the configuration with the lowest average loss over a general empirical trial data set in an ensemble of learning models of similar architecture.*

Proof

Suppose that the latent distributions of hidden concepts H_k is controlled by a number of parameters h_k in the latent space, for example, a set of k-dimensional clusters or manifolds. Then,

$$R = U H_k + N$$

N being the random noise component, not associated with any hidden concepts. Let us recall again that hidden concepts are not known to the learning model explicitly, neither before nor in the process of training.

Also, we consider parameters g_i of the generative model performing transformation from the latent space to the observable space, for example, weighs and biases in a neural network model:

$$F_{gen}(g_i, r): R -> I$$

With distributions of hidden concepts H_k and generative parameters g_i, a generative configuration *G* can be defined as a transformation of the latent space to observable space described by the set of parameters $p_k = \{ h_k, g_i \}$ that satisfies the constraint of generative learning on a representative training dataset *T*:

$$| G(T), T | \rightarrow min \qquad (2)$$

Now suppose that in an ensemble of trained models there exists a learner with a generative configuration that maximizes the categorization of hidden concepts $\{H_k\}$ in the latent space as defined in (1), that is, minimizes the volume of concept regions H_k and maximizes their separation from regions of other concepts.

Next, for a fixed concept H_k, consider a small variation of configuration parameters for H_k, $\delta p_k = \{ \delta h_k, \delta g \}$ and evaluate the corresponding differential of the generative loss function $\partial L / \partial p$ in an empirical trial with a set of samples X.

Starting with distribution parameters $\{ h_k \}$ in R, it can be observed that if the area of distribution for a given concept is increased while the model parameters remain constant, some samples outside of the initial concept region can be mapped to a different area in the observable space potentially increasing the error of mapping by placing the observable image in the wrong concept region; as well, samples of different "foreign" concepts can be now present in the extended concept region δH_k, and similarly mapped to a wrong concept region in the observable space, increasing the false positive error.

Hence, as illustrated in Figure 2, both considered scenarios can lead to an increase in the generative error compared to the initial state, with generative model parameters fixed:

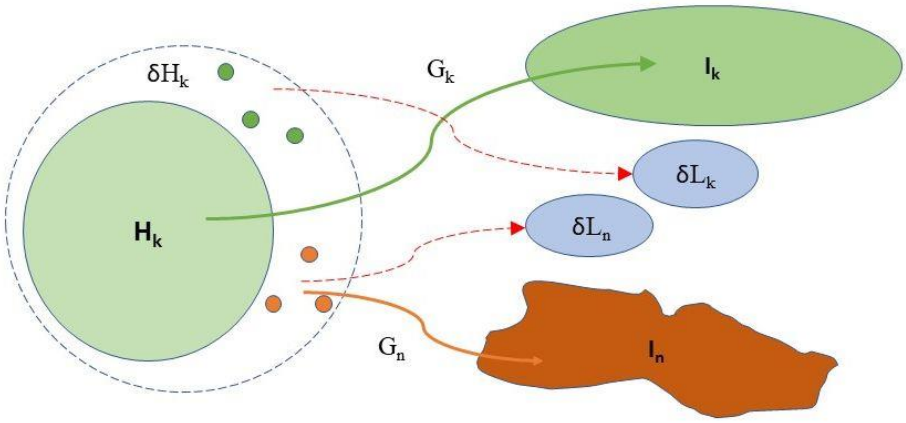

Figure 2 Sources of generative error, distribution parameters

It follows then:

$$\delta L = \delta L_k + \delta L_i = \sum_{i \in K} \delta L_i \geq 0, \ \frac{\partial L(X,\ h)}{\partial h_k} \geq 0 \tag{3}$$

Next we will consider variations of the generative model parameters $\{ g_i \}$, with hidden concept distributions fixed. As a result of generative training with a representative subset T, the resulting model parameters must minimize the generative loss with respect to variation of g_i, so any variation in the model parameter values increases generative error (loss) in the observable space:

$$\frac{\partial L(X,\ g)}{\partial g_i} \geq 0 \tag{4}$$

Then, from (3) and (4) one can conclude that for all generative configuration parameters p_k,

$$\frac{\partial L(X,\ p)}{\partial \mathrm{p}} \geq 0 \tag{5}$$

from which follows that generative configurations with superior categorization minimize generative error in an ensemble of learners in a general empirical trial, completing the proof of the lemma.

In conclusion we will note that in energy-based approach to learning [12], the energy function of the learning model *E(p)* is associated with a mean loss over some representative set of observable data. Then, the result in (5) can be interpreted as minimization of generative energy in configurations with better categorization of latent representations:

$$\frac{\partial E(X,\ p)}{\partial \mathrm{p}} \geq 0 \tag{6}$$

2.4 Assumptions and Limitations

In the proof, a number of implicit assumptions was made which we are going to discuss and attempt to justify.

The first point is the existence of categorized representations. In the proof it was implicitly assumed that a configuration with superior categorization exists in an ensemble of possible generative configurations. It can be noted cited experimental results, as well as empirical experience provide support and justification for this assumption.

The second important assumption is the significance of the observable parameters. Indeed, the parameters that describe data in the observable space must be specific enough to differentiate between samples of essentially different hidden concepts. Consider a simple example: suppose a face image dataset used to train a deep learning model had only one parameter, for example, gender of the person. Clearly, the only possible categorization in this case could be by the value of the gender parameter.

An important consideration is the representativity of the training data that was used implicitly in the proof of (4). Especially artificially constructed datasets can create disbalance between hidden concepts that could affect training of the model and the resulting representations. As well, the populations of hidden concepts must be sufficiently large to establish characteristic structures in the latent representation of the learning model.

Finally, it needs to be noted that the condition of redundancy reduction can be essential to avoid the identity transformation counter-example that is achievable by neural network models if the effective dimension of the representation space is equal or greater to that of the observable space and can describe arbitrary variation in the observable data without categorization of hidden concepts.

3. Categorization and Generality of Learning

The results obtained in this work may provide insights into the relation between the constraints of accuracy and generality in both supervised and unsupervised models. A question can be asked, to what minimal dimensionality limit can a given data be compressed to still satisfy both requirements of accuracy and generality?

Compression of the observable data to a categorized representation can provide significant

reduction in the energy / loss of the generative mapping due to reduction of dimensionality in the representation space that affects the independent degrees of freedom in the effective latent representation of data [17]. This gain can extend all the way to the limit where the dimensionality of the latent representation would reach that of the local variation in the representations of main hidden concepts with strong population in the data, i.e. the number of independent parameters of hidden concept distributions H_k. Beyond this critical dimensionality limit D_{cr} both constraints of accuracy and generality cannot be maintained simultaneously: forcing the model to sustain accuracy would make it overfit that is, memorize or encode the training data literally without categorization up to the limit allowed by its size; while exposing the model to larger empirical trials would lead to deterioration of generative ability or accuracy of classification in supervised learning.

Then, it can be hypothesized that the optimal effective dimension of the representations in both generative and conventional supervised models can be chosen based on the expected or observed variation spectrum in the native concepts of the data, that in some cases can be derived from an analysis of known higher-level concepts.

4. Conclusions and Discussion

In this work we introduced theoretical approaches in the analysis of unsupervised latent representations with the proof of the theorem of unsupervised categorization that links geometrical properties of unsupervised distributions in the latent representations of unsupervised generative models and self-learning with minimization of generative error under constraints of generality and reduction of redundancy.

This connection may have a number of both interesting and important consequences that would need a separate in-depth discussion. But an essential conclusion can be drawn that understanding the emergence of prototypes of native concepts in general data as a result of common principles of information processing may provide insights about the emergence and modeling of the intelligence in both artificial and biologic systems.

Indeed, it is worthy to note that the natural, biologic systems known for their capacity to learn independently, with minimal supervision or prior knowledge as noted by Hassabis et al., “human cognition is distinguished by its capacity to rapidly learn about new concepts from only a handful of examples” [18] demonstrate the same constraints of accuracy of reproduction; generalization; and massive redundancy reduction that were used in the formulation of the principles of unsupervised categorization.

Further, the conclusions about the general character of the effect of spontaneous categorization in unsupervised generative learning can be used as a basis for development of approaches and methods in self-learning and general learning systems that are flexible, interactive, iterative, and require minimal supervision or prior knowledge about the environment [19].

References

1. Rosch, E. H., Natural categories. Cognitive Psychology 4, 328–350 (1973).

2. Le Q.V. Ranzato M.A. Monga R. Devin M. Chen K. et al., Building high-level features using large scale unsupervised learning, arxXiv:1112.6209, (2012).

3. Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., et al. Early visual concept learning with unsupervised deep learning. arXiv:1606.05579 (2016).

4. Dolgikh S., Categorized Representations and General Learning. In: 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions – ICSCCW-2019 1095 93-100 (2019).

5. Zeng N., Zhang H., Song B., Liu W., Li Y et al., Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 643–649 (2018).

6. Ribeiro M., Lazzaretti A. E., Lopes, H. S., A study of deep convolutional auto-encoders for anomaly detection in videos, Pattern Recognition Letters 105 13-22 (2018).

7. Fischer A., Igel C., Training restricted Boltzmann machines: an introduction, Pattern Recognition 47 25 – 39 (2014).

8. Hinton G., Osindero S., Teh Y.W., A fast learning algorithm for deep belief nets, Neural Computation 18(7) 1527 – 1554 (2006).

9. Bengio Y., Learning deep architectures for AI, Foundations and Trends in Machine Learning 2(1) 1–127 (2009).

10. Spall, J., Introduction to Stochastic Search and Optimization, Wiley, 2003.

11. Rumelhart D.E., Hinton G.E., Williams R.J., Learning representations by back-propagating errors, Nature 323 (6088) 533–536 (1986).

12. Ranzato M.A., Boureau Y-L., Chopra S., LeCun Y., A unified energy-based framework for unsupervised learning, In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, 2 371-379 (2007).

13. Bishop C.M., Pattern Recognition and Machine Learning, Springer ISBN 978-0-387-31073-2 (2006).

14. Tishby N. Pereira F.C. Bialek W., The information bottleneck method, arXiv:physics/0004057 (2000).

15. Mandt S., Hoffman M.D., Blei D.M., Stochastic gradient descent as approximate Bayesian inference, Journal of Machine Learning Research 18 1 – 35 (2017).

16. Hornik K., Stinchcombe M., White H., Multilayer feedforward neural networks are universal approximators, Neural Networks 2(5) 359-366 (1989).

17. Greiner W., Neise L., Stöcker H., Thermodynamics and statistical mechanics, Springer (1995).

18. Hassabis D. Kumaran D. Summerfield C. Botvinick M., Neuroscience inspired Artificial Intelligence, Neuron 95 245-258 (2017).

19. Dolgikh S., Spontaneous concept learning with deep autoencoder, International Journal of Computer Intelligence Systems 12(1) 1-12 (2018).