Sparse generative representations of handwritten digits

Serge Dolgikh

National Aviation University, 1 Lubomyra Huzara Ave, 03058 Kyiv, Ukraine

Abstract

We investigated the process of unsupervised generative learning and the structure of informative generative representations of images of handwritten digits (MNIST dataset). Learning models with the architecture of sparse convolutional autoencoder with constraints to produce low-dimensional representations achieved successful generative learning demonstrated by high accuracy of generation of images. A well-defined, continuous and connected structure of generative representations was observed and described. Structured informative representations of unsupervised generative models can be an effective platform for investigation of origins of intelligent behaviors in artificial and biological learning systems.

Keywords

Artificial neural networks, generative machine learning, representation learning, clustering

1. Introduction

Representation learning with the objective to identify informative elements in the observable data has a well-established record in machine learning. Informative representations were obtained with Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBN) [1, 2], different flavors of autoencoders [3] and other models allowed to improve accuracy of supervised learning [4]. The relations between learning and statistical thermodynamics were studied in [5] and other works leading to understanding of a deep connection between learning processes and principles of information theory and statistics.

In the experimental studies, a range of results was reported, such as the "cat experiment" that demonstrated spontaneous emergence of concept sensitivity on a single neuron level in unsupervised deep learning with image data [6]. Disentangled representations were produced and discussed [7] with a deep variational autoencoder and different types of data pointing at the possibility of a general nature of the effect. Concept-associated structure was observed in latent representations of Internet network traffic [8], images [6,7,9], as well as a number of other results with different types of data and applications [10,11].

These results demonstrated that structure that emerges in the latent representations created by models of generative learning in the process of unsupervised self-learning with minimization of generative error can have intrinsic associations with characteristics patterns in the observable data and perhaps, can be used as a foundation for learning methods and processes that use these associations for improved efficiency.

Interestingly, these observations in unsupervised machine learning were paralleled in the recent works with a number of results in biologic sensory networks [12,13] that demonstrated commonality of low-dimensional representations in processing of sensory information by mammals, including humans.

These previous findings prompted and stimulated an investigation into the process of production and essential characteristics of low-dimensional informative latent representations obtained with neural network models of unsupervised generative self-learning, including formation of a conceptual structure in the latent

IVUS 2022: 27th International Conference on Information Technology, May 12, 2022, Kaunas, Lithuania
EMAIL: sdolgikh@nau.edu.ua (S. Dolgikh)
ORCID: 0000-0001-5929-8954 (S. Dolgikh)

representations of the sensory environment of the learner.

The questions investigated in this work were the following: what are the characteristics of the latent representations of successful generative models? Is there an association between the characteristic patterns (or higher-level concepts) in the input data and latent distributions produced by learning models?

What structure can be identified in the latent representations with entirely unsupervised methods, without prior knowledge of conceptual content of the input data?

These questions were approached with generative models of deep neural network architecture and a dataset of images of real, unprocessed image data of handwritten images (MNIST dataset) used widely in the studies of machine intelligence systems. The intent of the study is to understand how successful common generative models of unsupervised self-learning even of limited complexity, can produce informative and structured representations of input data modeling sensory environments.

The novelty of the presented approach is associated with using "generic" generative architecture with clearly defined directions of possible incremental variation and evolution. Using this type of architecture can provide answers to essential questions of how complex architectures that were reported in the cited results could have developed in realistic learning systems.

Throughout the work, externally known types or patterns in the input data that models observable sensory environment of a learning system will be referred to as "higher-level concepts" or "external concepts", that signify a class of a sample in the input space that is defined by an external process, outside of the model. An example of an external concept for an image with a geometric shape can be word "triangle" or a specific symbol. In contrast, structures in the latent representations of the observable space that can be identified entirely by unsupervised means without any external or prior information, will be referred to as "internal" "natural" or "native" concepts [14].

A priori, there is no reason to assume that external and native concepts are related or correlated, so the relation between the external and native concepts is an interesting and intriguing question in its own right.

2. Methods and data

2.1. Model architecture

A convolutional autoencoder model [15] used in this work had the encoding stage with convolution-pooling layers followed by several layers of dimensionality reduction with a sparse encoding layer of size 20–25 producing an effective low-dimensional latent representation described by activations of neurons in the encoding layer.

Sparse training penalty was applied to latent activations as L1 regularization, resulting in 2 to 4 neuron activations for most images in the dataset. The decoding / generative stage that was fully symmetrical to the encoder. The diagram of the architecture used in this work is shown in **Figure 1**.

Figure 1: Sparse convolutional autoencoder model

Overall, the models had 21 layers and ~ 9×10^4 trainable parameters. The models were implemented in Keras / Tensorflow programming package [16] and trained for minimization of generative error i.e., an average norm of the difference between input images and the output produced by the model on the unsupervised training set defined by the categorical cross-entropy (CCE) cost function.

2.2. Data

The dataset of images used in the study, MNIST [17] consisted of three sets of images (training, validation and test) of handwritten digits, from 0 to 9 produced by different real individuals. The models were trained on a subset of 10,000 images, with approximately equal representation of all digits.

To ensure entirely unsupervised character of latent representations created by trained models, labeled samples were not used in the phase of generative training of the models, but only in the analysis of distributions of higher-level concepts in the latent representations created by trained models.

2.3. Training

The success of unsupervised learning was measured by the characteristics of training performance and generative ability. Training performance was measured by the reduction in the value of the cost function over the period of training. Generative performance was evaluated visually based on the quality of generation of a subset of images in the training dataset. Approximately 70% of models were successful in generative learning by both measures. A clear correlation was observed between the training and generative characteristics. Models with training loss above certain threshold generally did not succeed in acquiring good generative ability.

Success of generative learning, that is, the ability to generate high quality images of the types present in the training dataset indicated that latent representations produced by the learning models retained significant information about the distribution of observable data represented by the training dataset.

2.4. Encoding and generation

A trained model can perform two essential transformations of data: encoding, $E(x)$ from the observable space, i.e., image x to the latent position l; and generative, $G(l)$ in the opposite direction, producing an observable image, y. The objective of generative learning is to minimize the distance between training images and their generations by the model, defined by a training metric (cost function) in the observable space.

2.5. Sparse representations

As a result of a sparsity constraint imposed in unsupervised generative training, the effective latent representations of observable images were low dimensional, that is, an observable image was described by activations of a small number of latent neurons; the observed effective dimensionality with the images in the dataset was 2 to 4 (i.e., two to four non-zero activations of latent neurons).

A sparse latent representation of this type can be described by a stacked space of low-dimensional "slices" [18], indexed by a tuple of activated neurons, (i_1, i_2, i_3). For example, an image of digit "2" can be described in a 24-dimensional sparse representation space by the index (1, 3, 8) with coordinates (0.011, 0.017, 0.019) that translates to corresponding activations of the neurons 1, 3 and 8 in the latent layer, and nil activations of other latent neurons.

3. Results

The results in this section were obtained with several instances of models trained as outlined earlier, that were successful in generative learning. The results pertain to essential characteristics of low-dimensional latent representations produced by generative modes, such as structure, topology, consistency and others.

3.1. Generative latent structure

Examination of the geometrical and topological structure of sparse representations of the handwritten digit images produced by generative models confirmed highly structured character of representations closely correlated with characteristic types of images.

Following the objective of the study to examine the structure of informative generative representations without known concept samples, an approach was developed that allows to investigate the structure in the latent representations produced by successfully learned generative models by purely unsupervised methods that do not require knowledge of the semantics, concept, class or any other prior information about the input data. The process of producing such unsupervised structure (or "generative landscape" of the representation) is based on identification of a density structure, such as density clusters in a general sample of encoded sensory inputs with methods of unsupervised density clustering such as MeanShift [19].

The approach is based on several essential assumptions. The first one is success of generative learning reflected by sufficient accuracy and quality of generation. The second is sparsity of resulting representations, that provides two essential benefits: a lower dimensionality of the encoded inputs, and higher decoupling in the structure of representations making it easier to detect and harness for learning. And finally, an assumption on the composition of the training set to contain a constant number of characteristic types of inputs (i.e., representativity).

To apply methods of density clustering in the latent representation, first a structure of space

slices needs to be identified (Section 2.5). This was done according to the following process:

- For each three-dimensional slice: $l = (i_1, i_2, i_3)$ a subset of significant activations $S(l)$ identified as $\Sigma\ a_j \geq f \times a_{max}$, where a_{max}: maximum activation in the slice (the sum of activations of slice neurons); f: a factor, $f = 0.25$ in the study.
- $S(l)$ projected on the slice coordinates, resulting in a three-dimensional set $S_p(l)$.
- A density clustering method applied on the set $S_p(l)$ producing a sequence of density clusters ordered by size $D(l) = \{ D_k(l) \}$. The length of the sequence is defined by the clustering method and does not have to be known in advance.
- The process is repeated for slices with significant representation of significant activations (i.e., the size of $S(l)$ above certain threshold, relative to other slices) resulting in a stacked structure of density clusters, generative landscape $D = \{ D(l) \}$, with a natural two-dimensional unique index of (l, n); l: the position of the slice; n: the position of the cluster in the slice

With the generative landscape produced with the described process, the first task was to examine how the resulting latent structure is correlated with characteristic types of data in the training dataset. It can be determined by transforming center positions of the clusters of the landscape $D(l)$ to observable images with the generative transformation $G(l)$ (Section 2.4). **Figure 2** shows the resulting "map" of images associated with the identified density structure of the generative landscape.

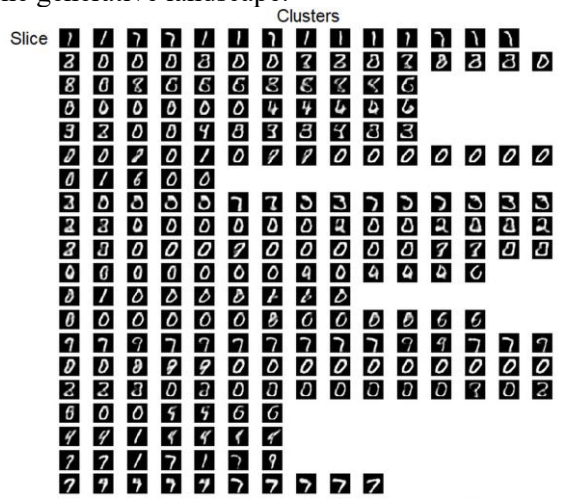

Figure 2: Generative structure of the latent landscape (vertical axis: slice; horizontal axis: cluster (first 15 clusters)

As can be observed in the visualization of **Figure 2**, cluster positions were indeed closely associated with characteristic types of images in the training dataset.

3.2. Latent geometry and topology

The identified landscape of density structure can assist in examination of the geometry and topology of the sparse latent space.

The first objective was to investigate connectedness and continuity of the latent regions associated with characteristic types of observable images. To this end, arrays of random positions were created on the spheres of a given radius from the cluster centers, thus producing a "flow" of latent positions from cluster centers of the landscape outwards. The positions were then transformed to observable space with generative transformation, as in the previous section producing arrays of observable images associated with the latent positions.

Examination of the resulting images allowed to conclude that generative representations produced by models were indeed connected and continuous, with well-defined regions associated with specific types of images (**Figure 3**).

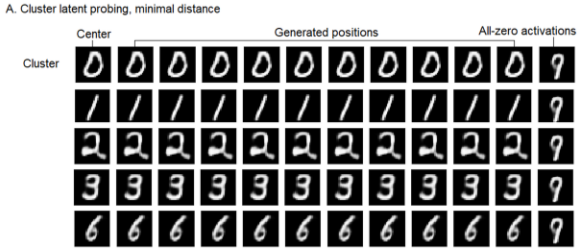

Figure 3: Generative latent landscape, continuity

Examination of different clusters and landscapes produced by different individual models allows to conclude that consistency and connectedness is a general property of generative latent landscapes.

3.3. Structural consistency of latent representations

While latent representations created by generative models can be expected to be specific to individual learning models due to peculiarities of the training process, for example, random selection of training samples. At the same time, some essential characteristics of generative representations appeared to be consistent between the learning models.

To investigate consistency of the latent structure, an analysis of latent landscapes produced with three independently trained generative models was performed.

The models were trained over 40-60 epochs of unsupervised generative learning with a training set of 10,000 samples, achieving a training plateau at validation loss of 0.12-0.14 (with the starting value of ~ 0.7) and good to excellent generative performance on a subset of images and were not selected by any specific criteria. After completion of the training phase several successful independently trained models were selected and characteristics of generative landscapes produced with methods described earlier measured.

The measured characteristics were: the overall size of the landscape as the number of identified density clusters with population above certain margin, relative to the size of the training dataset (~ 2%); recognition, the fraction of the landscape clusters associated with recognizable digits (as discussed in Section 3.1), indicating a correlation of the landscape with the characteristic content of the training set; representativity of the content of the landscape, such as presence of all types of digits (completeness) and distribution of digits between slices and clusters (digits with highest and lowest population of associated clusters in the landscape). The results are presented in Table 1.

Table 1

Consistency of latent structure

Model	Size	Recognition	Completeness	Population: h / l
A	474	0.973	True	0,7 / 4,6
B	396	0.975	True	0,7 / 2,5
C	485	0.971	True	0,4 / 2,8

As can be inferred from these results, latent landscapes of independently trained successful generative models had significant consistency in the size, recognition and representation of characteristic types of images. On the other hand, factors such as distribution of digits in the slices and clusters, highest and lowest representation of digits in the clusters and a number of others tended to be more specific to individual learning models.

Similar results were previously obtained with several different types of image data such as geometrical shapes [9] pointing at the likelihood of a general character of the observed effect of categorization in the latent representations of successful generative models by characteristic types of patterns.

3.4. Unsupervised concept learning

The results of the preceding sections, with strong correlations observed between the emergent latent structure of successful generative models and characteristic types of observable data can be interpreted as distillation of “native” or “natural” concepts in the observable data in the process of unsupervised learning with minimization of generative error. The structure or the latent landscape, as discussed in the preceding sections, can be resolved in an entirely unsupervised process by a number of methods.

It can be concluded from these results that generative learning under certain constraints and the resulting structure in the informative latent representations can be used as a foundation for implicit learning of characteristic patterns in the observable data before and without external contextual information about it. These results can also offer insights into explainability of learning in generative models via association of learned concepts or classes in the observable data and the native information structure that emerges in the latent representations in the process of unsupervised generative learning.

4. Discussion

Highly structured character of low-dimensional generative representations produced by successful models of unsupervised generative self-learning observed in this work provides further support for a growing number of results pointing at importance of informative representations in processing of sensory information by learning systems, of both artificial and biological nature.

In this work the effect was observed with real-world image data of significant complexity, pointing at a general character of the effect. Informative structured representations strongly correlated with characteristic patterns, or concepts in the sensory data can play an essential role in emergence and development of intelligent behaviors including conceptual intelligence, abstraction and communications.

Continuing research in this direction can shed light on common principles of learning for artificial and biological systems and perhaps point a direction to a generation of learning systems capable of more natural and intuitive learning from direct interaction with the sensory environment [20].

5. References

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

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

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

[4] A. Coates, H. Lee, A.Y. Ng, An analysis of single-layer networks in unsupervised feature learning, in: Proceedings of 14th International Conference on Artificial Intelligence and Statistics 15 (2011) 215–223.

[5] M.A Ranzato, Y.-L. Boureau, S. Chopra, Y. LeCun, A unified energy-based framework for unsupervised learning, in: Proceedings of 11th International Conference on Artificial Intelligence and Statistics 2, 2007, 371–379.

[6] Q.V. Le, M.A. Ranzato, R. Monga et al., Building high level features using large scale unsupervised learning, arXiv 1112.6209 (2012).

[7] I. Higgins, L. Matthey, X. Glorot, A. Pal et al., Early visual concept learning with unsupervised deep learning, arXiv 1606.05579 (2016).

[8] N. Seddigh, B. Nandy, D. Bennett, Y. Ren, S. Dolgikh et al., A framework & system for classification of encrypted network traffic using Machine Learning, in: Proceedings of 15th International Conference on Network and Service Management (CNSM), Halifax, Canada, 2019, 1–5.

[9] S. Dolgikh, Topology of conceptual representations in unsupervised generative models, in: Proceedings of 26th International Conference on Information Society and University Studies, Kaunas, Lithuania, 2021, 150–157.

[10] J. Shi, J. Xu, Y. Yao, B. Xu, Concept learning through deep reinforcement learning with memory augmented neural networks, Neural Networks 110 (2019) 47–54.

[11] R.C. Rodriguez, S. Alaniz, Z. Akata, Modeling conceptual understanding in image reference games, in: Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada, 2019 13155–13165.

[12] T. Yoshida, K. Ohki, Natural images are reliably represented by sparse and variable populations of neurons in visual cortex, Nature Communications 11 (2020) 872.

[13] X. Bao, E. Gjorgiea, L.K. Shanahan et al., Grid-like neural representations support olfactory navigation of a two-dimensional odor space, Neuron 102 (5) (2019) 1066–1075.

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

[15] Q.V. Le, A tutorial on deep learning: autoencoders, convolutional neural networks and recurrent neural networks, Stanford University 2015.

[16] Keras: Python deep learning library, last accessed: 2020/11 URL: https://keras.io.

[17] Y. Le Cun, The MNIST database of handwritten digits, Courant Institute, NYU Corinna Cortes, Google Labs, New York Christopher J.C. Burges, Microsoft Research 2007 Redmond, USA.

[18] S. Dolgikh, Low-dimensional represent-tations in unsupervised generative models, in: Proceedings of 20th International Conference Information Technologies – Applications and Theory (ITAT), Slovakia, 2020 239–245.

[19] K. Fukunaga, L.D. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Transactions on Information Theory 21 (1) (1975) 32–40.

[20] D. Hassabis, D. Kumaran, C. Summerfield, M. Botvinick, Neuroscience-inspired Artificial Intelligence, Neuron 95(2) (2017) 245–258.