Topology of Conceptual Representations in Unsupervised Generative Models

Serge Dolgikh[a,b]

[a] National Aviation University, 1, Lubomira Huzara, Kyiv, 03058, Ukraine
[b] Solana Networks, 301 Moodie Dr., Ottawa, Canada

Abstract

In this work latent representations of image data were investigated with neural network models of generative self-learning. A convolutional autoencoder with strong redundancy reduction was used to create latent representations of images of basic geometric shapes in the process of unsupervised generative learning and the characteristics of distributions of concept regions in the latent space of models investigated. It was demonstrated that conceptual representations with good separation of latent regions can be produced with generative models of limited complexity and that characteristic types of data, or "concepts" form well-defined, continuous and connected regions in the latent space. Geometric structure of the latent representations was described in detail confirming connected and continuous topology of latent concept regions clearly associated with characteristic types of observable data, such as shape, size and contrast. The results indicate that conceptual representations created in the process of unsupervised generative learning can form a natural basis for the emergence of abstract concepts in intelligent systems.

Keywords 1

Machine learning, unsupervised learning, concept learning, representations

1. Introduction

Representation learning with the objective to identify the informative structure in general real-world data has a well-established record in the field of Machine Learning. Hierarchical representations of different types of data were obtained with Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBN) [1][2], different flavors of autoencoders [3][4] and other models and architectures allowed to improve accuracy of supervised learning. Different types, architectures and flavors of generative models were investigated since including autoencoder neural networks, Generative Adversarial Networks (GAN) and others [5][6][7].

A number of interesting experimental results were obtained in concept learning with artificial systems, such as the "cat experiment", that demonstrated spontaneous emergence of concept sensitivity on a single neuron level in unsupervised deep learning with image data [8]. Disentangled representations were produced and studied with deep variational autoencoder models and different types of image data [9] pointing at the possibility of a general nature of the effect. Concept-associated structure was observed in latent representations of diverse real-world data such Internet in large public networks, aerial surveillance images [10][11] and other results [12][13].

The relations between learning and statistical thermodynamics was studied in the theory of learning systems [14][15] leading to understanding of a deep connection between learning processes in artificial systems and principles of information theory and statistical thermodynamics.

These results demonstrated that training of artificial learning models under certain constraints such as minimization of predictive error can lead to emergence of structure associated and correlated with characteristic patterns in the observable data. This approach, known as representation learning, was developed in a number of studies [16]. Of interest in this work is unsupervised conceptual representation

IVUS2021: Information Society and University Studies 2021, April 23, 2021, Kaunas, Lithuania
EMAIL: sdolgikh@nau.edu.ua (A. 1)
ORCID: 0000-0001-5929-8954 (A. 1)

learning with models of generative learning that does not use explicitly labeled data but rather attempts to minimize the error of reproducing the observed distribution from a representation in the latent space created in the process of generative self-learning. As some of the earlier results have indicated, such entirely unsupervised process can produce non-trivial structure in the latent representations that is correlated with characteristic patterns in the observable data and can be used as a foundation for learning methods and processes based on such structured low-dimensional representations.

Interestingly, these observations in unsupervised learning of artificial systems were paralleled very recently by several results in the studies of biologic sensory networks [17][18] that demonstrated commonality of low-dimensional representations in processing sensory information by mammals, including humans.

Based on and inspired by these results in conceptual representation learning, the questions investigated in this work are the following: what properties characterize latent representations of successful generative models? Is there a stable association between characteristic patterns in the observable data and the structure that emerges in the latent distributions in the process of generative self-learning? What structure can be identified in the latent representations with entirely unsupervised methods, without prior knowledge of conceptual content of the observable data?

These objectives were approached with artificial neural network models of deep convolutional autoencoder that showed effectiveness in producing informative representations [19] and a dataset of images of basic geometric shapes that are described in the following sections.

In conclusion, a brief clarification of terminology used throughout this work. We will refer to the characteristic types or patterns in the dataset that are labeled with known types or classes as “external” concepts, that signifies that the type or label of the input is defined outside of the model based on some external or prior knowledge about the observable data. An example of an external concept for an image with a geometric shape can be its general type, “a triangle”. In contrast, a characteristic structure in the latent representation that can be identified reliably by unsupervised means without any external or prior information, will be referred to as “internal” or “native” concept. Thus, a question of the relation between the external and native concepts can be of interest as well.

2. Methods

In this section the models and data used in the study are described. A dataset of basic shapes such as circles, triangles and greyscale backgrounds, was used to produce low-dimensional latent representations that were studied with several methods as described in the following sections.

2.1. Deep Convolutional Autoencoder

We used artificial neural network models with the architecture of convolutional autoencoder [3] with added strong dimensionality reduction in the representation layer to produce three-dimensional latent representations of image datasets of basic geometric shapes. The advantages of the autoencoder architecture are that it allows to learn essential characteristics of the input data in an unsupervised process without labeled samples, while virtually unlimited depths and complexity allows it to be used with complex real-world data, demonstrating successful learning in many applications [8]-[11],[19].

The architecture diagram of the models used in the study is presented in Figure 1.

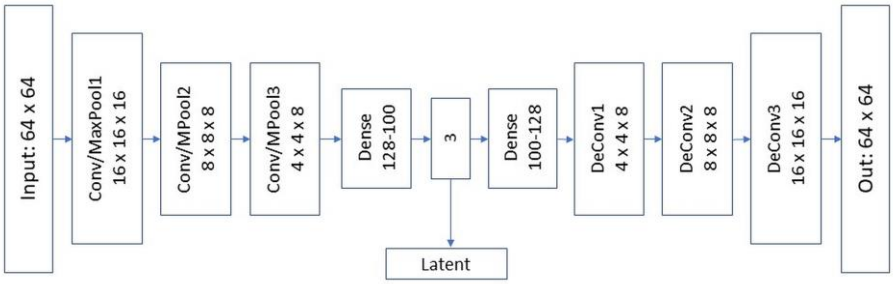

Figure 1: Convolutional autoencoder with deep dimensionality reduction

The models had three stages of convolution-pooling layers followed by several layers of dimensionality reduction in a symmetrical layout. The total number of layers was 21, with approximately 40,000 trainable parameters.

Unsupervised training was performed with standard methods such as Stochastic Gradient Descent [20] to minimize average distance between the input and generated output with binary cross-entropy cost function. The models were implemented with Keras / Tensorflow [21] and several standard machine learning libraries were used in the analysis of latent representations.

The latent representation was produced by activations of the neurons in the central, encoding layer of the model (Figure 1). The activations of neurons were interpreted as coordinates in a three-dimensional latent representation space.

2.2. Data

Three datasets of geometric shape images with different variance and complexity were used to investigate the structure of the latent representations. The images were greyscale, of the size 64 × 64 pixels.

The first, generated dataset, Shapes-1 consisted of 600 greyscale images of circles, triangles and greyscale backgrounds with two representative samples per type with difference in the size and contrast of fore / background.

The second generated dataset, Shapes-2 contained the total of 600 – 1,000 greyscale images of circles, triangles and backgrounds with variation in size in the range 0.3 – 1.0 of the image size (that is, 0.3 × 64 pixels), with variation of contrast of fore- vs. background for each size.

In the generated datasets the images were centered, symmetrical and had no rotation. Only darker foregrounds relative to the background were used.

The third dataset, Shapes-3 contained 1,500 images of randomly generated colored geometric shapes and backgrounds varying in: size; position relative to the center of the image; rotation; and the color of the fore- and background [22]. The images in this dataset were converted to greyscale.

2.3. Latent Representations

Following the process of unsupervised training in which significant reduction in the value of cost function was observed, the models were able to perform two essential transformations of data:

The encoding transformation, from the input data space, that is, images to the three-dimensional latent representation, with coordinates represented by activations of neurons in the central, "encoding" layer of the model (Figure 1):

$$r_x = E(x) = encode(x) \quad (1)$$

The generative transformation operates in the opposite direction, i.e. from the latent representation into the observable (image) space and is performed by the generative part of the autoencoder:

$$g_y = G(y) = generate(y) \quad (2)$$

where y is a latent position, i.e. a point in the latent space represented by latent coordinates (y_1, y_2, y_3).

Transformations of encoding and generation allow to study the structure in the latent representation by measuring and visualization of the distributions of samples encoded to the latent space, as well as generative properties of the models via producing images of latent positions and regions of interest.

3. Results

The subject of this investigation was topological structure in the latent representations created by generative models in the process of unsupervised learning. To verify the hypothesis of the study, methods of analysis of latent distributions and generative ability of models were applied as described in this section. It is essential to note that the process of production of latent representations via

generative learning was entirely unsupervised and no labeled data was used in configuration and / or training of the models.

3.1. Structure of Latent Representations

Latent representations of generative models displayed well structured, connected and continuous distribution of data by characteristic type as illustrated in Figure 2. The visualization data, a subset of Shapes-1 and Shapes-2 datasets described in Section 2.2 represented a selection of images with variation of size and greyscale contrast, color-coded by characteristic type (i.e., circle, triangle or greyscale background).

Visualization of latent distributions by characteristic type showed that the shapes present in the training dataset were assigned specific well-defined regions in the latent space, with latent coordinates encoding essential characteristics of shapes, such as size and contrast.

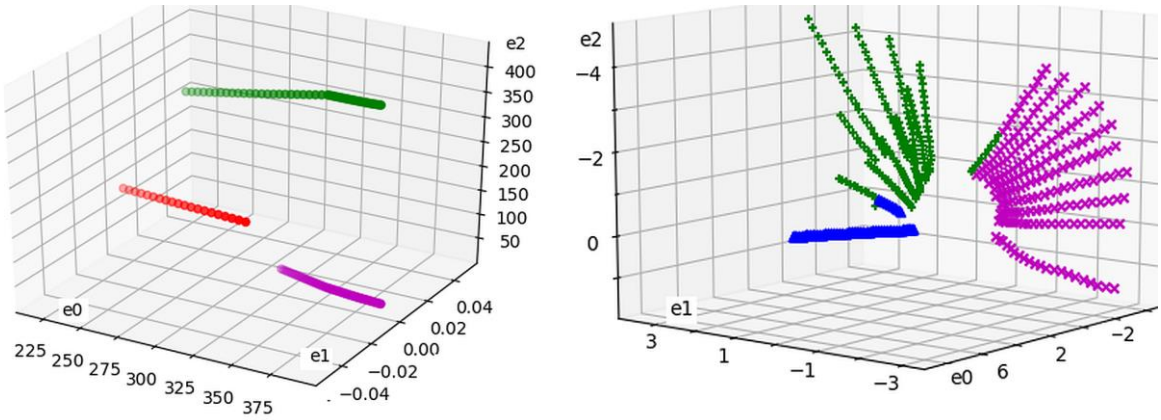

Figure 2: Latent distributions by shape type. Models: Shapes-1 (left), Shapes-2 (right). Legend: magenta: circles; green: triangles; blue/red: greyscale background.

The observed pattern of latent distributions was consistent between different individual models training with the same dataset, that was verified with several instances of trained models. These findings indicate that the emergent concept-associated structure in the latent representations of generative models reflects essential general characteristics of the data in the training data and is to a large extent invariant to individual model.

3.1.1. Connectedness of Latent Representations

As discussed above, latent representations created by generative models in unsupervised training displayed connected and continuous topology allowing to conclude that there exist an association between well-defined latent regions and characteristic types, or general concepts in the observable data. This conclusion was further substantiated with two experiments.

In the first experiment, a set of images S of the same type was selected, with variation in characteristics, such as size and contrast. The images were transformed to the latent space as in (1) and the mean of the resulting set of latent coordinates calculated as:

$$r_{mean} = mean(E(S)) \quad (3)$$

The mean latent representation r_{mean} of the input set was then propagated to the observed space with generative transformation (2) and the resulting image produced. In the experiments, the generated image of the input set was of the same type (Table 1).

Table 1

Generated mean of input set of shape type

Input set	Encoded mean (example)	Result
Circles-2	(-1.979, -1.635, -0.817)	Circle
Circles-3		Circle
Triangles-2		Triangle

Triangles-3	(1.437, 0.073, -1.358)	Triangle*
Background-3	(3.974, 0.011, -0.176)	Background

* An exception was observed in the boundary region (Section 3.1.2), where the mean of a certain input set produced mixed result.

In the second group of experiments, a set of points in a close neighborhood of a selected latent position of a given type was selected, and images corresponding to these positions generated with (2). The resulting images were observed to be of the same type as that of the position of origin (Figure 3).

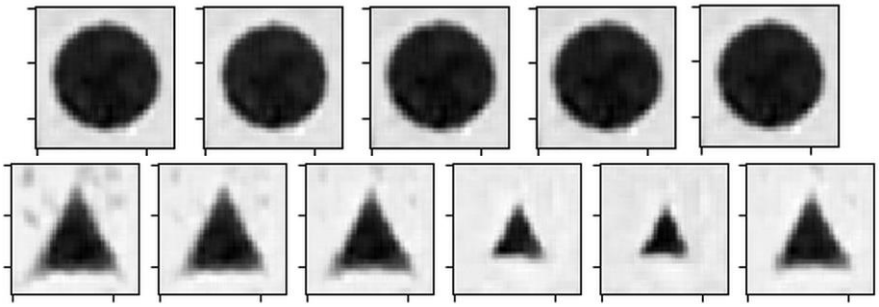

Figure 3: Latent neighborhood experiment. Position of origin: leftmost.

These experiments confirmed the connected and continuous topology of latent representations produced by generative models.

3.1.2. Boundary Regions

Well defined character of latent regions associated with characteristic types of shapes was confirmed by observation of boundary areas between the regions of different types, that produced generated output of mixed form with features from both types. The examples of such boundary generative regions are shown in Figure 4.

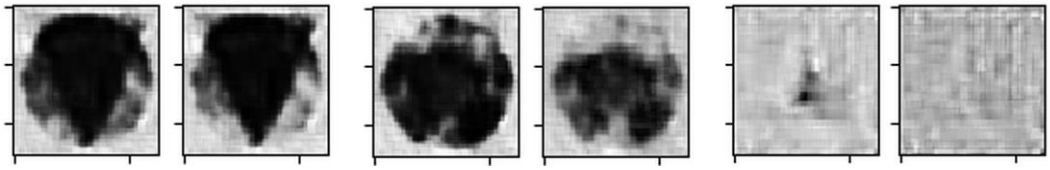

Figure 4: Latent boundary areas. Left to right: circle/triangle; circle/background; triangle/background.

Observation of the boundary areas between the regions associated with characteristic shapes supports the hypothesis of a continuous and connected topology of the latent space.

3.2. Semantics of Latent Coordinates

Generative transformation (2) allows to investigate semantical meaning of the latent coordinates defined as activations of neurons in the encoding layer of the generative model. A method of latent probing was developed that allowed to produce and evaluate generated image of a specified position in the latent space, after conceptual latent representation was produced in the unsupervised learning phase. By applying probing to latent representations of generative models the structure that formed in the process of generative learning under the constraints discussed earlier can be studied in practically any level of detail.

In the input to the method, a set of coordinates of a latent position of interest was provided. The position was propagated via generative part of the autoencoder with generative transformation (2) producing an output as an observable image. Evaluation of generated images associated with a set of latent positions, such as a neighborhood, a line, a surface and so on, in a latent region of interest allowed to obtain essential observations on the semantical meaning of the latent coordinates. The examples of probing along the latent axes are shown in Figure 5.

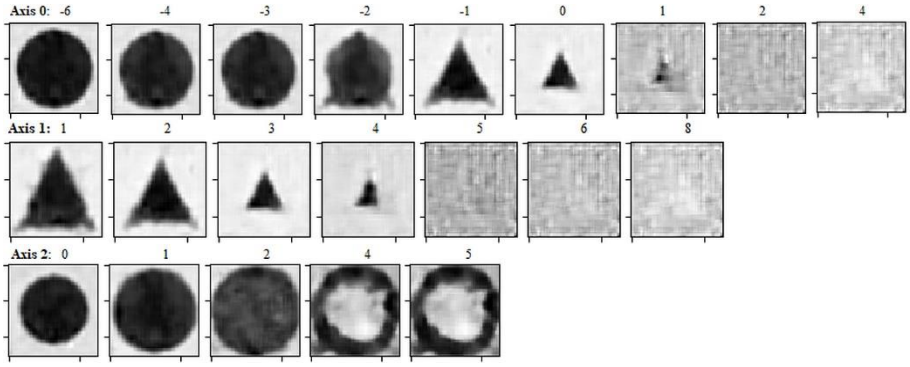

Figure 5: Generated output along latent axes, 0 – 2.

The results of generative probing in the latent space of generative models in this section support the manifold assumption [23] and confirm that latent coordinates in low-dimensional representations can be associated with essential characteristics of input data, such as in this case, the size, contrast and the type of the shape. It can be seen also that the semantics of latent coordinates are local, rather than global: the same axis, at different latent position can signify either of these characteristics. Consequently, investigation of topological structure of latent representations, including with more complex data merits further attention.

4. Conclusion

In this work latent representations of images representing basic geometric shapes were analyzed with methods of unsupervised machine learning. The analysis produced a number or essential results and observations.

It was demonstrated that successful generative learning with stable association of characteristic observable inputs to a structure in the low-dimensional latent representations can be achieved with models of limited complexity, both in size and architecture, well within a range of simple biologic systems. The complexity of models in this work was roughly equivalent to a nervous system of a jellyfish that has a comparable number of neurons and synapses [24].

By applying unsupervised methods that did not use any externally labeled data in the analysis of latent representations it was demonstrated that generative learning can lead to emergence of well-defined latent regions associated with essentially different types of observable data. These results points to possible origin of higher-level abstract concepts in unsupervised generative learning under the constraints of generative accuracy and redundancy reduction, both having clear evolutionary advantage for the learner.

The structure of latent regions associated with characteristic patterns, native concepts in the observable data was investigated and described, confirming connectedness and continuous topology of the latent regions, with distinct boundaries between different concept regions. An analysis of generative capacity of models with the developed method of latent probing allowed to make essential observations on the semantics of the latent coordinates in the representations produced by the models, showing an association with essential characteristics of the data such as type, size and contrast. However, as the analysis of the semantics of latent coordinates has shown (Section 3.2) there may not be a fixed global semantics to the latent coordinates and their significance is determined by the locality, i.e. the position and the region, again confirming manifold-like topology of latent representations emergent in generative self-learning.

Overall, the results of this work demonstrated that conceptual representations can be a natural result of unsupervised observation of the environment under the constraints that are both natural and common for a learning system [25], of artificial or biologic nature.

5. Future Work

Following the direction outlined in this work, further studies can focus on investigation of latent representations of more complex visual data with greater variation of conceptual content in the

observable environment. It is likely that advances in this direction would require models of considerably greater complexity, in size, depth and architecture [8]. For example, investigation of neural network models with sparsity constraints can be a promising direction due to apparent success of this architecture in the sensory input processing neural networks of biologic systems as reported in recently published results [17].

6. References

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

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

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

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

[5] Welling, M. and Kingma, D.P., An introduction to variational autoencoders, Foundations and Trends in Machine Learning 12 (4) (2019) 307–392.

[6] Creswell A., White T., Dumoulin V., Arulkumaran K., Sengupta B. and Bharath A.A., Generative adversarial networks: an overview, IEEE Signal Processing Magazine 35 (1) (2018) 53–65.

[7] Partaourides, H., Chatzis, S.P., Asymmetric deep generative models, Neurocomputing 241 (2017) 90 – 96.

[8] Le, Q.V., Ransato, M. A., Monga, R. et al., Building high level features using large scale unsupervised learning. arXiv 1112.6209 [cs.LG] 2012.

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

[10] Dolgikh, S., Categorized representations and general learning, in: Advances in Intelligent Systems and Computing, volume 1095, Springer, Cham 2019, pp. 93 – 100. doi:10.1007/978-3-030-35249-3_11.

[11] Prystavka P., Cholyshkina O., Dolgikh S., Karpenko D., Automated object recognition system based on aerial photography, in: Proceedings of 10th International Conference on Advanced Computer Information Technologies, ACIT-2020, Deggendorf, Germany, 2020, pp. 830 – 833.

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

[13] Wang, Q., Young, S., Harwood, A., and Ong, C. S., Discriminative concept learning network: reveal high-level differential concepts from shallow architecture, in: Proceedings of 2015 International Joint Conference on Neural Networks, IJCNN, 2015, pp. 1–9.

[14] Hinton, G. E. and Zemel, R.S.: Autoencoders, minimum description length and Helmholtz free energy, Advances in Neural Information Processing Systems, 6 (1994) 3 – 10.

[15] Ranzato, M.A., Boureau Y.-L., Chopra, S., LeCun, Y.: A unified energy-based framework for unsupervised learning, in: Proceedings of 11th International Conference on Artificial Intelligence and Statistics 2, 2017, pp. 371–379.

[16] Bengio, Y., Courville A., Vincent, P., Representation Learning: a review and new perspectives, arXiv:1206.5538 [cs.LG] 2014.

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

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

[19] Holden, D., Saito, J., Komura, T. and Joyce, T., Learning motion manifolds with convolutional autoencoders, in: Proceeding of Asia 2015 Technical Briefs, SA '15 SIGGRAPH, 2015, p.18.

[20] Spall, J. C., Introduction to stochastic search and optimization: estimation, simulation, and control, Wiley Hoboken, NJ, ISBN 0-471-33052-3 2003.
[21] Keras: Python deep learning library, https://keras.io/, last accessed: 2020/11/21.
[22] El Korchi, A., 2D geometric shapes dataset, Mendeley Data V1, 2020, doi:10.17632/wzr2yv7r53.1.
[23] Zhou X., Belkin M., Semi-supervised learning. In: Academic Press Library in Signal Processing, Elsevier 1, 2014, pp. 1239–1269.
[24] Garm, A., Poussart, Y., Parkefelt, L., Ekström, P., Nilsson, D-E., The ring nerve of the box jellyfish Tripedalia cystophora, Cell Tissue Research 329 (1) (2007) 147–157.
[25] Dolgikh, S., Why good generative models categorize? International Journal of Modern Education and Computer Science (2021). To appear.