Sparsity Constraint in Unsupervised Concept Learning

Serge Dolgikh

Dept. of Information Technology, National Aviation University, Kyiv 03058 Ukraine,
sdolgikh@nau.edu.ua

Abstract: Informative representations play an important role in learning and intelligent functions of artificial and biological systems. Sparsity constraint in neural network models has been known to be effective in producing informative representations of sensory data. In this work, the structure in low-dimensional representations created by a class of generative neural network models of unsupervised learning was analyzed to establish the relation between the sparsity constraint imposed in the representation layer and the effectiveness of unsupervised concept learning. It was demonstrated that sparsity constraint allows to achieve two essential objectives in successful realistic learning: increasing the effective conceptual capacity of latent representations, while simultaneously limiting the range of latent activations. Sparsity emerges as a simple and effective mechanism of producing functional conceptual representations of complex sensory data in both artificial and biological systems. The results provide empirical support for the connection between unsupervised generative learning and conceptual latent representations correlated with characteristic patterns in the sensory inputs and clarify the role of the sparsity constraint in improving the quality and conceptual capacity in generative representation learning.

1 Introduction

Unsupervised representations obtained with models of generative self-learning were studied for considerable time with the intent to identify and separate informative components and patterns in the data to improve performance of learning models. Hierarchical representations obtained with Restricted Botzmann Machines (RBM), Deep Belief Networks (DBN) [1, 2], and different types of autoencoder models [3, 4] were proven to be effective in improving learning performance in supervised classification [5].

In experimental studies in unsupervised concept learning with artificial intelligent systems interesting results of spontaneous high-level concept sensitivity emerging in the process of unsupervised learning were reported. An intriguing effect of spontaneous formation of concept sensitive neurons activated by images in certain higher-level category was observed [6] with a massive deep sparse autoencoder neural network model trained in entirely unsupervised process with a massive set of real-world images. Representations of deep variational autoencoder models were studied in [7], demonstrating effective disentangled representations with data of several different types in entirely unsupervised learning under the constraints of redundancy reduction. Higher-level concept-related structures were observed in the representations of generative neural network models with strong redundancy reduction with data representing backbone Internet traffic and terrain surveillance images [8].

These and a number of other results [9, 10] demonstrated that latent representations created by models of unsupervised generative learning a result of training under the constraints of minimization of generative error and strong redundancy reduction may acquire non-trivial structure associated and correlated with characteristic patterns, or concepts in the training dataset, assumed to be a representative set of observable sensory inputs.

In a number of results, sparsity constraint [6, 11] introduced based on observations in biological neural networks was suggested to be effective in improving the effectiveness of learning. Essentially, sparsity constraint means limiting the number of active neurons participating in processing inputs, with more details on specific implementation in the studied models provided in the sections that follow. While sparse models became in many cases part of standard architecture of neural network models and growing number of studies supported and highlighted effectiveness of sparse models in producing informative representations of visual data, general principles that explain the effectiveness of sparsity remained less clear. What is the underlying cause of effectiveness of sparsity in processing sensory data? Is it effective in combination with specific architectural choices such as convolutional neural networks widely used with visual data, or has more general nature and can be applied and observed with broader range of models and data?

The importance of these questions is highlighted by very recent results in experimental neuroscience indicating essential and ubiquitous role of low-dimensional sparse representations in interpretation of sensory data by animals and humans [12, 13]. Together with the results in learning of machine systems, they point at a general causes for effectiveness of sparse representations in generative learning, regardless of the nature of the learning system. To address these questions, in this work we follow the direction of research outlined in [14] with the objective to investigate the effect of the sparsity constraint by comparing characteristics and topological structure of latent representations of images of basic geometric shapes obtained with

neural network models of unsupervised generative learning. Comparing characteristics of latent distributions of similar sets of image data obtained with models with and without sparsity constraint in the encoding layer allowed to make essential observations on how sparsity affects the structure of latent representations and effectiveness of unsupervised generative learning and to advance in understanding the causes of effectiveness of sparsity as a mechanism of production of informative low-dimensional representations of sensory inputs.
The paper is organized as follows: in Section 2 the model and data used in the study are described. Section 3 contains the results of experiments with generative models with sparsity constraint (sparse representation layer) and those without it (flat representation layer) are presented. Section 4 contains a brief discussion of results and the connections to the current state of the research in the field.

2 Methods

The models used to produce unsupervised latent representations of images modeling visual sensory inputs had the architecture of a convolutional autoencoder [15] with strong dimensionality reduction in the encoding layer producing latent representation with the coordinates of the activations of encoding neurons. The data was represented by a dataset of images of basic geometric shapes, greyscale and color, as described in this section.

2.1 Convolutional Autoencoder Model

The models were of a common type of a convolutional autoencoder, with the addition of a deep dimensionality reduction stage producing latent representation as shown in Figure 1. The dimension of the latent layer was variable and different in two studied classes of models. The first class (denoted "F") had a flat latent layer with the dimension of 3 to 10 chosen based on discussed results in sensory processing in biological systems. These models did not have sparsity constraint on activations in the latent layer. The second class of models, "S" had the latent dimension of up to 10, with sparsity constraint imposed in the latent layer as L1 regularization [16] limiting the norm of activations of the encoding neurons producing sparse latent representation. The models were implemented with Keras/Tensorflow [17], for measurement and visualization of distributions common machine learning libraries and packages were used.
The compression of information achieved in the representation layer of the model was approximately 1,300 for greyscale and 4,000 for color images of size (64,64). An advantage of the chosen architecture is that it allowed to measure and visualize the distributions of data in the latent representation directly by visualizing activations of neurons in the encoding layer of the model.

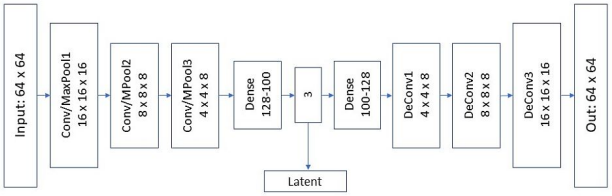

Figure 1: Convolutional autoencoder model with strong dimensionality reduction

2.2 Data

Two datasets of images of basic geometric shapes of the size 64 $\times$ 64 pixels, greyscale and color were used. The first dataset, Shapes-G 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 $\times$ 64 pixels), with variation of contrast of fore- vs. background for each size.
The second dataset (Shapes-C) was a color version of Shapes-G with the total of 1,200 images of the following combinations: red and blue circles of varying size with grey background of varying contrast; red and blue triangles with varying size and background; red and blue horizontal bands (wide red bands, narrow blue bands), with varying background; and empty grey backgrounds with varying contrast.
The composition of the dataset allowed to experiment with data of different complexity: while the conceptual content of the first dataset (Shapes-G, including grey backgrounds) was three concepts, for the Shapes-C data it was 7 (with wide and narrow bands of different color considered as different concepts).

2.3 Training

The models were trained in an unsupervised autoencoder mode to achieve good reproduction of inputs measured by the cost function, such as Mean Squared Error (MSE) and binary cross-entropy (BCE). Several criteria of effectiveness of unsupervised training were used, such as monitoring the cost function and cross-categorical accuracy that both shown significant improvement in unsupervised training with minimization of the generative error. Additionally, generative performance of trained models was measured by comparing a subset of input images to their reproduction by trained models as illustrated in Figure 2.

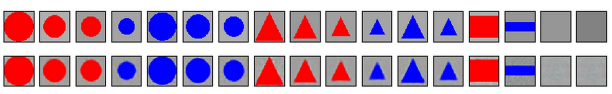

Figure 2: Evaluation of generative quality (top row: input; bottom: interpretation)

2.4 Unsupervised Latent Representations

A trained model performs the encoding transformation from the observable data space $X \in O$ to the latent representation $y \in R$ obtained with the activations of the latent layer of the encoder model $E : O \rightarrow R$, and the generative transformation from the latent representation to the observable space $G : R \rightarrow O$ as:

$$y = E(X);\ X'(y) = G(y) \tag{1}$$

where X' is the generated image or interpretation of input X by the model. The distance between the original input X and its interpretation X' then indicates the ability to interpret inputs in the observable data and training of a generative model is equivalent to minimization of the norm $||X,\ X' = G(E(X))||$ defined by the cost function on the training set of samples in the observable space.
The structure in the latent representation that emerges as a result of unsupervised training can be observed and evaluated with a number of methods:

1. By applying unsupervised clustering such as DBScan, MeanShift and variations [18] in the representation to identify density distribution in general unlabeled data sample as well as concept samples.

2. By evaluating distributions of general and concept data samples in the representation space via direct observation, multi-dimensional histograms and other methods.

3. By producing observable images of latent positions with the methods of generative probing and scanning [14].

2.5 Sparsity Constraint

Sparsity constraint was imposed in the encoding layer of the model as L1 regularization with the value of 10^{-5}-10^{-6}. Level one regularization means a constraint on the sum of absolute values of activations of the neurons in the encoding layer and in most cases resulted in no more than two to three neurons being active at any interpretation task, consistent with the results in visual sensory system in humans that indicated sparse, low-dimensional activation pattern in interpretation of visual data [12].

3 Results

To examine the effect of sparsity constraint on unsupervised latent representations, methods of comparative analysis were applied to representations created by models of similar neural network architecture with sparsity constraint imposed in the representation layer and without it. This approach allowed to clarify the role and demonstrate the significance of sparsity constraint for the ability of the learning model to create effective conceptual representations of the training data.

3.1 Learning Quality

As described in the previous section, the success of generative learning was judged by the improvement in the value of the cost function and by the quality of reproduction of images, similar to those the models were trained with. The outcomes of two evaluations were mostly consistent, that is, models with significant improvement in the cost function had better interpretation quality and vice versa, those with better generative quality also had better training metrics.
In this experiment, training results of models of different architectures, flat and sparse, were compared with the dataset Shapes-C of higher conceptual complexity. Ten models of each type were trained independently with the dataset Shapes-C with evaluation of learning quality factors, such as:

- Learning success, the fraction of models that were able to interpret a representative set of images successfully

- Generative quality indicated by a factor in the range [0,1] measured on the test set, average over the models;

- Average maximum activation recorded on the test set

as shown in Table 1. The number in the type such as "F3" indicates the dimensionality of the latent layer whereas "AL" indicates an activation limit, a constraint on the maximum absolute value of the activations in the encoding layer.

Table 1: Comparative learning metrics, flat and sparse models

Model	Success	Quality	Activation
Sparse (S5)	0.8	0.71	18.6
Sparse (S8)	0.7	0.64	18.3
Sparse (S10)	0.9	0.88	19.6
Flat (F3)	0.4	0.77	90.8
Flat (F3-AL)	0.2	0.8	26
Flat (F5-AL)	0.6	0.8	26
Flat (F10-AL)	0.8	0.9	26

Approximately 80% of flat models without maximum activation constraint (F3) were successful in learning one color concept (i.e., only the red shapes or blue shapes). This is consistent with the results of experiments with a greyscale shapes dataset [14] that indicated that flat models of the type used in the study were generally successful in learning the concepts with the greyscale data of conceptual complexity 3. Note as well that flat models of higher dimensionality were successful in learning more complex data (F5-AL, F10-AL) but it came at the cost of significant increase in the number of active synapses in the case

of fully interconnected layers. For example, F10-AL flat model vs. S10 sparse model would have $(10-3) \times D_h$ more active synapses in the encoding layer, where D_h, the size of the next hidden layer (100 - 300 for the studied models).
These results indicated that sparsity constraint imposed in the representation layer allowed models to maintain the effectiveness of learning with more complex data of higher conceptual content, whereas flat models were either significantly less successful in training, or required more resources such as higher activations and / or the number of active synapses. The causes of this difference will be discussed in the following sections.
Another important conclusion that can be drawn based on results in this section and the earlier ones obtained with the greyscale dataset is that for a given complexity of data and model architecture there exists a certain limit or threshold T_f of conceptual content that can be learned successfully in a given low-dimensional subspace of the representation space defined by the coordinates of activations of active neurons. If and when the threshold is exceeded, in a flat model with realistic activations constraints (F3-AL, Table 1), the learning performance begins to deteriorate. In the next section, the role of activation constraint will be discussed in more detail.

3.2 Activation Constraint

For a realistic learning system, biological or artificial operating autonomously in the real world, the range of activations up to the maximum values can be essential as it directly translates into the operational cost of the system in both energy and physical resources. Particularly, very strong activations expressed in a biological system as an electric voltage may require significant resources to maintain currents under control. Thus keeping activations within a limited range and preventing high activation values can be an essential objective and a constraint for a realistic learning system.
In the experiments with the models in this work it was observed that whereas flat and sparse models were able to attain similar levels of generative quality with the data of limited conceptual content (dataset Shapes-G), an essential difference was found with data of higher conceptual complexity in the dataset of color images, Shapes-C. Successful flat models with good generative quality, without sparsity constraint in the representation layer were significantly more rare (as indicated by the learning success factor, Table 1) and produced significantly higher activations, up to an order of magnitude higher than recorded in the sparse models.
A straightforward explanation of this finding can be given based on the earlier results [14] that demonstrated topological structure of latent representations of greyscale image data as broad polygonal regions, or "columns" associated with characteristic types of images in the training data. If such a structure were to be extended to the data with higher conceptual content, the number of concept regions would need to increase correspondingly and packing of the concept regions in the latent space would require larger volumes, translating, in the latent coordinates, to greater ranges of variation of activations of encoding neurons.
To verify this hypothesis, a constraint on the maximum absolute value was imposed in flat models preventing activations from exceeding set maximum value by absolute magnitude. The maximum was chosen similar to, and slightly greater than the average maximum activations observed in sparse models. The observed result was that flat models were significantly less successful in generative learning (model F3-AL, Table 1), indicated by their ability to interpret images from the dataset.
From these results and the results of section 2.2 it follows that increasing conceptual complexity of the data from 3 to 7 resulted in a strong deterioration of the learning ability of the flat models pointing at the value of the conceptual threshold for low-dimensional flat models, $3 \leq T_f < 7$. It also showed that sparsity offered a simple and effective solution to the high activations problem, allowing to "pack" higher conceptual content within the allowed range of activations.

3.3 Generative Structure of Latent Representations

Methods of topological analysis of latent representations such as latent probing and scanning developed in [14] allow to explicitly describe and visualize the structure of latent representations that emerges as a result of the unsupervised training process under the constraints of generative learning. It was hypothesized [19] that low-dimensional conceptual representations of sparse models are comprised of subspaces or "slices" in the representation space $w = (i, j, k)$ indexed by small subsets of active neurons participating in production of low-dimensional representations, with i, j, k being the indices of active neurons that define the slice w.
Producing latent scans of representations of sparse generative models trained with dataset Shapes-C allowed to confirm this hypothesis and explicitly describe the structure of the resulting sparse representation space. The method involves producing an array (a hypercube) of images generated by a trained model from latent positions in a grid associated with a latent region of interest; by adjusting parameters of the grid it allows to describe the generative structure that emerges in the latent space as a result of unsupervised learning to any level of detail.
To identify the structure, sparse models with dimensionality d of the encoding layer that were trained with Shapes-C were presented with a subset of images representing characteristic patterns in the dataset, T. By recording latent positions $l(T) = E(T)$ as defined by equation (1), it was possible to detect the allocation of concepts in the dataset to slices $w(d, 3)$ in the d-dimensional latent representation space. The examples of distributions of concepts in d-slices for $d = 5, 8, 10$ are shown in Table 2, with slices

indexed by the tuples of indices of participating neurons, $i = 1 \ldots d$.

As can be seen from the results in Table 2, sparse architecture of the representation layer indeed distributed the patterns in the training data quite efficiently between available slices in the stacked representation space. The stacked structure of the latent representations was exhibited by all sparse models though different arrangement of slices could be used by individual models, for example, another S10 model used slices (3,4,10) and (2,9,10) for red and blue circles respectively, indicating that slice allocation is determined by the training process as is not invariant between individual models trained with the same data.

Table 2: Stacked structure in latent representations of sparse models

Concept	S10	S8	S5
Circle, red	(3,4,5)	(3,7)	(2,5)
Circle, blue	(1,2,10)	(1,2,8)	(1,3)
Circle, blue, small	(1,2,10)	(1,3)	(1,3)
Triangle, red	(5,8)	(7)	(2,5)
Triangle, blue	(1,2)	(8)	(1,3,5)
Band, red wide	(3,4,5)	(2,7)	(1,5)
Band, blue narrow	(2,8	(1,3,8)	(3)
Background, grey	(3,5)	(1,6)	(5)

With the slices of interest in the d-dimensional representation space thus identified, in the next step latent scanning was applied in the slices associated with representative samples of visual concepts to examine the generative structure in the slice subspaces identified by neuron indices $w = (i, j, k)$. By producing generative scans of the slices as an array of observable images one can understand how activations of specific neurons are interpreted by a sparse model. The examples of generative scans of a sparse representation, model S10 and a flat three-dimensional representation (model F3, no activation constraint) are shown in Figures 4 and 5 respectively.

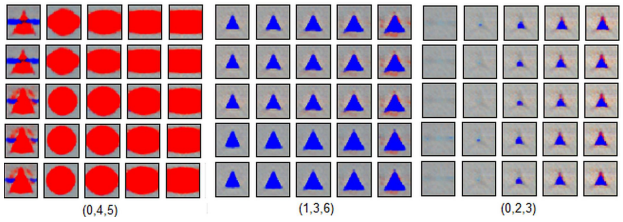

Figure 3: Generative scan, stacked representation

Clearly, the sparse model distributed the concepts in the training data between different low-dimensional slices in its 10-dimensional latent representation space as shown in Fig.4, in direct confirmation of the stacked structure of sparse representations hypothesis discussed earlier. On the other hand, the flat model with a three-dimensional latent representation and no activation limit attempted to pack all concepts into single low-dimensional latent space available to it, resulting in the maximum activation value of 86, compared to under 20.0 for stacked models (Figure 5).

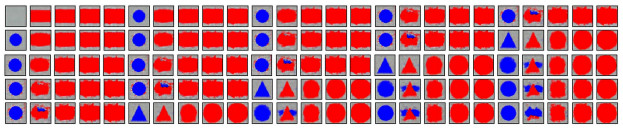

Figure 4: Generative scan, flat representation

From generative scans in Fig.4, 5 it can be observed that the topology of latent regions associated with distinct concepts is continuous and well-defined, shaped as broad slanted "columns" in the flat latent space or slice subspaces, with characteristic parameters of images such as color, shape, size and contrast encoded in the latent position. These observations are consistent with the earlier results on greyscale data [14] as well as the experimental results on the architecture of biological sensory systems [12, 13] pointing to the possibility of a general character of this effect. It is also a direct experimental confirmation of the manifold assumption [20] common in unsupervised and semi-supervised learning.

3.4 Conceptual Capacity of Sparse Representations

The results discussed in the previous sections allow to describe the concept learning ability of sparse models in terms of topological structure of effective latent representations, namely, stacked low-dimensional representations spaces (slices) defined by collections of active neurons participating in producing interpretations of sensory data. In a way, each slice is thus equivalent to a flat low-dimensional representation of constant dimensionality. It was commented earlier based on comparison of interpretation ability of flat models with datasets of different complexity that there appears to be a threshold of maximum conceptual content $T_f(M,D)$ that can be successfully interpreted by a flat model of given architecture M with a given type of observable data D. The results in the earlier sections indicated that for the generative architecture in this work, $3 \leq T_{mx} < 7$.

For a sparse system of dimension d and slice dimension l defined, as commented earlier by the sparsity constraint in the encoding layer, the total conceptual capacity K_m, i.e., the maximum number of independent concepts the model can learn successfully can be described by the formula:

$$K_m = f_{ef} \times N(d,l) \times T_f \tag{2}$$

Here, $N(d,l)$ is the number of slices and $f_{ef} \leq 1$, efficiency factor related to the ability of the model to maximize the use of available slices. In the case of simple single encoding layer models similar to those used in this work, $N(d,l)$

is the binomial coefficient, $B_{d,l}$.
From (2) the bound on the conceptual capacity of a sparse model S5 with the latent dimension of 5 and 3-dimensional slices ($d = 5,\ l = 3$) can be readily estimated as $10 \times T_{mx}$, whereas for a general model with an N-dimensional sparse latent layer it would be:

$$K_m(S_N) \leq \frac{N\ (N-1)\ (N-2)}{6} \times T_{mx} \sim \frac{1}{6}\ N^3\ T_{mx} \qquad (3)$$

with sufficiently large dimensionality of the sparse layer, N. Accordingly, sparsity allows to strongly increase conceptual capacity of generative models with only a small addition of neurons in the latent layer and without significant increase in the cost of processing of sensory inputs as the number of active neurons remains constant.

4 Conclusions

The analysis of distributions of image data in the latent representations of generative self-learning models in this work resulted in several essential findings:

1. For a given type, content and characteristic parameters of data and model architecture D, M there exists, under the constraints of realistic learning, a natural threshold $T_f(D,M)$ for the effective number of characteristic patterns or concepts that can be learned successfully.

2. Models with flat architecture of the encoding layer can maintain learning performance if either: conceptual content of the data does not exceed the threshold T_f; or by expanding the volume of the latent representation space necessary to encode the conceptual content, resulting in higher activations of encoding neurons.

3. Models with sparsity constraint in the encoding layer can bypass the conceptual threshold by creating stacked representation spaces and associating low-dimensional slices with subsets of concepts.

4. An empirical formula for evaluation of maximum conceptual content of sparse models.

These results demonstrated that sparsity allows to achieve two essential objectives in successful realistic learning: increasing the effective conceptual capacity of the latent representations, while limiting the range of latent activations. Thus, sparsity emerges as a simple and effective adaptation that allows to produce functional conceptual representations of complex sensory data in both artificial and biological systems.
Low-dimensional representations, especially those obtained in unsupervised learning can be of interest due to growing evidence that similar systems of sensory neurons can play an important role in processing sensory data by biological systems. Recent results demonstrated that effective representations of sensory data such as images, audio signals and odors can be produced by small sets of active neurons in biological neural networks [12, 13] creating effective low-dimensional representations of the sensory data. Connecting the results in experimental neuroscience with the findings of this work where low-dimensional representations of simple image data were created artificially may offer interesting insights and connections between learning processes in biological and artificial intelligent systems.

References

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

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

[3] Bengio Y.: Learning deep architectures for AI. Found. Trends Machine Learning **2(1)** (2009) 1–127

[4] Welling, M., Kingma D.P.: An introduction to variational autoencoders. Found. Trends Machine Learning **12(4)** (2019) 307–392

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

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

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

[8] Prystavka P., O. Cholyshkina O., S. Dolgikh S., D. Karpenko D.: Automated object recognition system based on aerial photography Proc.10th Intl. Conf. Adv. Comp. Info. Tech. (ACIT) (2020) 830–833

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

[10] Rodriguez, R. C., Alaniz, S., and Akata, Z.: Modeling conceptual understanding in image reference games. In: Advances in Neural Information Proc. Syst. (2019) 13155–13165

[11] Ng, A.: Sparse autoencoder. Lecture notes, Stanford University (2016)

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

[13] 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

[14] Dolgikh, S.: Topology of conceptual representations in unsupervised generative models. Proc. 26th Int. Conf. Info. Soc. Univ. Stud. (IVUS2021) (2021) to appear

[15] Masci, J., Ueli Meier, D.C., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. ICANN (2011).

[16] Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT Press (2016)

[17] Keras: Python deep learning library. https://keras.io/

[18] Fukunaga, K., Hostetler, L.D.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory bf 21(1) (1975) 32–40

[19] Dolgikh, S.: Low dimensional representations in generative self-learning models. Proc. 20th Int. Conf. Info. Tech. – App. Theo. (ITAT-2020) Slovakia **2718** (2020) 239–245

[20] Zhou, X., Belkin M.: Semi-supervised learning. In: Acad. Press Lib. in Signal Proc. Elsevier (2014) 1239–1269