

Low-Dimensional Representations in Generative Self-Learning Models

Serge Dolgikh

National Aviation University, Kyiv 02000 Ukraine,
sdolgikh@nau.edu.ua

Abstract: Informative representations play an important role in learning and intelligence. We analyzed distributions of image classes in low dimensional representations created by a class of deep autoencoder neural network models in unsupervised learning. The representations of real aerial images have been shown to contain higher-level concept structures such as low-dimensional surfaces and higher density clusters that form as a result of unsupervised training with minimization of generative error. Compact and well-defined character of some distributions was demonstrated with a positive correlation between the categorization performance of the model and its classification accuracy. The results provide direct empirical support for the connection between unsupervised learning in models with self-encoding and regeneration and categorization of native concepts in the representations.

1 Introduction: Unsupervised Representations

Study of unsupervised representations with the intent to identify and separate the most informative components in general data has a long history in machine learning. Unsupervised hierarchical representations created with models like Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBN) [1, 2] different types of autoencoder models [3] proved to be efficient and improved the accuracy of subsequent classification [4]. The deep relationship between training of intelligent models and the statistical principles such as minimization of free energy was studied in [5, 6] and other works leading to forming understanding that common methods of training such as gradient descent in deep neural networks and Contrastive Divergence in DBN generally produce configurations compatible with the principles of minimization of free energy and variational Bayesian inference.

On the experimental side, interesting effects of spontaneous high-level concept sensitivity in unsupervised deep neural network models were observed in a number of works. Google Lab team [7] observed an intriguing effect of spontaneous formation of concept sensitive neurons activated by images in certain higher-level category with a massive deep and sparse autoencoder neural network model trained in entirely unsupervised process without any exposure to ground truth with very large arrays of

YouTube images.

In [8] a spontaneous formation of grid-like cells, similar to those observed in mammals was detected in a recurrent neural network with deep reinforcement learning. Higher-level concept-related structures were observed in the representations of deep autoencoder models with strong redundancy reduction with data representing raw Internet traffic in large public telecommunications networks in [9]. The results demonstrated that a density structure in the representations created by such models that emerges as a result of unsupervised training with minimization of generative error it can be used in the iterative approach to training of artificial learning systems that can offer higher flexibility and considerably lower ground truth requirements compared to common methods. Representations of deep variational autoencoder models were studied in [10], demonstrating effective disentangled representations with data of several different types in entirely unsupervised learning under the constraints of redundancy reduction.

These and a number of further results [11, 12] may suggest that certain neural networks whether artificial or biological, in the process of unsupervised learning with an incentive to improve the quality of regeneration of the observable data may naturally structure information by characteristics of similarity in their representations, thus identifying certain natural or native concepts that perhaps can be correlated with higher-level concepts in the observable data. Based on this observation, the hypothesis investigated in this work is that the natural structure in the representations created by certain unsupervised models in self-supervised learning with minimization of the generative error can be correlated with higher-level concepts in the input data, and that relationship can be used in developing approaches to flexible and iterative learning in the environments where prior domain knowledge is scarce or not available.

In this study we are following the line of research outlined in [7, 9] by first creating a compact representation of the observable dataset with a deep self-encoding neural network model (a two-stage stacked autoencoder), then analysing the parameters of distributions of the higher-level concepts in the dataset in the representation created in unsupervised training. But unlike [7] that investigated single-neuron that is, essentially, one-dimensional representations and distributions of concepts (Fig.1), the design of the models in this study, with physical constraints on the dimensionality of the representation layer created low-dimensional representations, that allowed to improve

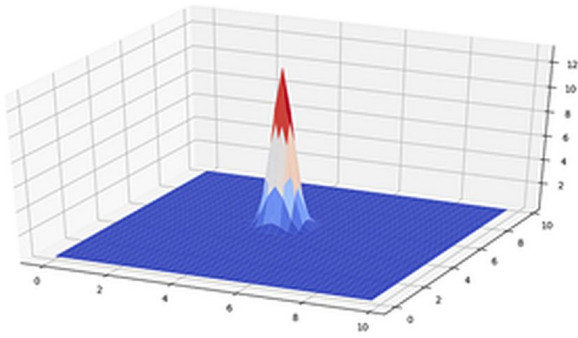


Figure 1: Effective activation of a concept-sensitive neuron (based on [7])

the resolution of the learned concepts from “better than random” across arbitrarily selected pre-known range of higher-level concepts in [7] to “better than random binary” and in a number of cases, “confident binary” classification per concept that was not known previously. It is thought that these results can be of an interest for the research community in unsupervised learning and self-learning systems because, as some recent studies indicate [13, 14] similar low-dimensional representations with only a small number of active neurons can play important role in sensory networks of biologic systems, such as visual and smell processing; as well, the connection between unsupervised learning and concept structures in the representations may suggest approaches to self-learning that would be common for biologic and artificial systems.

2 Methods

The model used in the study is a stacked two-stage autoencoder with strong physical compression in the layer of the final representation. This choice was based on the earlier cited results as well as some strong arguments in favor of neural network models based on generative self-learning being good candidates for producing effective unsupervised representations. Being a universal approximator [15], feedforward neural networks have virtually unlimited versatility and are well suited to model complex data types. And not in the least, deep neural networks are widely present in biologic systems that are also highly successful in self-learning with minimal data [16].

The data was represented by a dataset of raw images obtained in aerial observation of terrain, as described in this section.

2.1 Deep Stacked Autoencoder Model

The diagram of the model is given in Figure 2. The model produced two stages of representations of unprocessed aerial image data. The encoder of the first stage was a convolutional-pooling autoencoder that produced a

numerical representation of dimension 576 from color images with dimensions (32,32) to (128,128). The aim of this stage was to acquire higher scale features in the images via a sequence of convolution-pooling stages.

The resulting numerical representation was used as the input to the second stage autoencoder with a strong reduction of physical dimensionality of the representation layer. The dimension of the representation was chosen based on principal component analysis of the numerical representation of the first stage that revealed three components with combined variation of over 0.95. Hence, the maximum

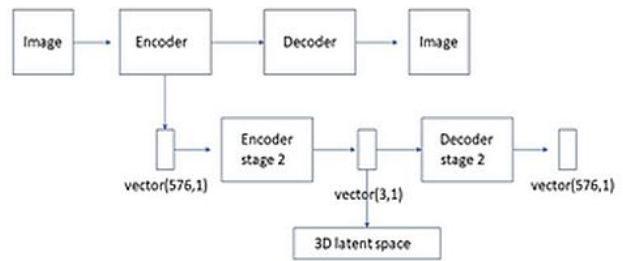


Figure 2: Stacked autoencoder model with physical redundancy reduction

compression of information achieved in the representation layer of the model was approximately 16,000, from input images in the first stage to the final representation. A certain advantage of the studied models is that they allow to measure and visualize the distributions created in unsupervised training directly from the central layer of the latent representation. In feed-forward neural networks, the accuracy of regeneration of input data combined with significant compression in the layer of representation means that the latent representation has retained significant essential information about the original distribution and observing it directly may yield some valuable insights about the character of the concept distributions in the observable data.

The models were implemented with Keras/Tensorflow [17]. For measurement and visualization of distributions we used common libraries and packages such as: sklearn-kit, numpy, matplotlib and others.

Models were trained in an unsupervised autoencoder mode to achieve good reproduction of inputs measured by a cost function such as Mean Squared Error (MSE). 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 calculating a mean deviation of the input sample from the generated output to the mean norm of the input sample, with an average result in the range of 0.1.

In our view these training results and the fact that in feed-forward neural network models the output is generated only from the information that is contained in the repre-

sensation layer indicate that the latent representation has indeed retained significant essential information about the original distribution.

2.2 Data

The dataset consisted of approximately 1,100 color images with resolution (64,64) manually labeled with ten higher-level classes of terrain type such as “trees”, “buildings”, “water”, etc., as described in Table 1. The higher-level classes used in the study represented three different broad categories:

1. Background: the area of the class concept spans the entire image or most of it; an example is “trees” or “field”
2. Structure: the concept area spans a significant part of the image area, such as roads; construction structures, e.g. bridges, power lines; excavations.
3. Object: an object located in compact area relative to the size of the image; vehicles and miscellaneous machinery were in this category. The composition of the dataset

Table 1: Aerial image dataset

Class	Category, Number of samples
Buildings (1)	background/structure, 100
Trees (2)	background, 100
Field (3)	background, 100
Water (4)	background/structure, 100
Roads (5)	structure, 100
Excavations (6)	structure, 100
Vehicles (7)	object, 100
Other (8-10)	varied, 400

with classes of different categories allowed to investigate the character of concept distributions in the latent representations for different types of higher-level concepts.

2.3 Unsupervised Representations

A trained model can perform the encoding transformation from the observable data space to the latent representation obtained with the activations of the central layer of the Phase 2 encoder, and the generative transformation from the latent representation to the observable space as:

$$R(X) = \text{encoder_model.encode}(X) \quad (1)$$

$$X'(Y) = \text{generator_model.decode}(Y) \quad (2)$$

In the latent representation of a trained model, the emergent density structure can be identified by applying a density-based clustering method such as DBScan, Mean-Shift and numerous variations [18]. It allows to identify density clusters of the encoded samples in the representation space without any need for the ground truth data. For

example, the associated density cluster for a sample X in the input data space can be calculated as:

$$K_{nat}(X) = \text{cluster_model.predict}(X) \quad (3)$$

where `cluster_model` is a density-based clustering method trained with a general data sample in the latent representation.

To perform classification, a binary concept classifier can be trained with a subset of labeled concept samples in the latent space. The resulting classifier can be applied to predict the explicit concept class of samples in the input space as:

$$K_{exp}(X) = \text{classifier.predict}(\text{encode}(X)) \quad (4)$$

where K_{exp} is the explicit or external class of the sample X predicted by the trained classifier. Thus K_{exp} and K_{nat} represent respectively, the externally known class of the sample and its native or implicit cluster identified from the density distribution in the latent space needing no external knowledge of the domain, distribution, or any other prior knowledge about the data.

The structure in the latent representation that emerges as a result of unsupervised training, or “unsupervised landscape” can be measured and observed by the following methods:

1. By applying unsupervised clustering in the representation to identify density distribution in general unlabeled data sample as well as concept samples
2. By measuring the parameters of general and concept distributions in the representation space
3. By applying multi-dimensional histogram methods in the representation space to measure density and volume distributions in general and concept samples
4. Via visualization and direct observation of general and concept samples in the representation space.

2.4 Unsupervised Categorizaion

By unsupervised categorization is meant the ability of some models with unsupervised self-encoding and regeneration of the input data to group data samples in the latent representation in compact structures by certain similarity. Such natively similar samples in the representation are then transformed in the generative stage of the model to samples in the observable data space that are related by association to the same, or related native concepts.

To measure categorization ability of models, two types of data samples were used:

- 1) concept samples transformed to the representation space define concept distribution region, that is, the region in the representation space where samples associated with certain higher-level concept can be found.
- 2) a general sample, a set of non-labeled data points that is used to identify and measure the size and shape of the region in the representation space that is populated by all categories, in other words, the image of a representative subset of the input data set in the latent space of the model.

Relative measurements of concept versus general distributions allow to draw conclusions about categorization performance of the model for the given concept, such as the: relative size, density of concept distribution regions, their shape, dimensionality and other parameters that can affect learning of the concept.

Distributions of data in the latent representations, or density landscape created by such models in the process of unsupervised learning can then be analyzed, measured and visualized by transforming marking subsets of labeled concept samples to the latent space with encoding transformation (1), while generative ability of the model can be evaluated by measuring the deviation of the generated output from the input.

The hypothesis that can be drawn from the results discussed earlier is that a structured information “landscape” that emerges in unsupervised training of the models with the incentive to reduce the regeneration error can be correlated with higher-level concepts that have strong representation in the input data.

3 Results

3.1 Visualization Analysis

Concept distributions in the latent space created in unsupervised training with minimization of regeneration error can be visualized and measured directly. To produce visualizations of concept regions, subsets of concept samples were transformed to the latent state of a pre-trained model and visualized with available plotting packages. Continuous approximations of the concept regions in the latent space were obtained with triangular interpolation of the concept samples transformed to the latent space.

Compact Distributions It was observed that the classes in the “background” and some, in the “structure” category, covering significant part of the image area generally produced compact and well-defined concept regions in the form of a two-dimensional surface. These distributions are illustrated in Fig.3. In the diagram, the top plot shows distributions of two concepts of “background” type (classes 3 and 4) in the latent space of a trained unsupervised model. The surface character of the distributions can be clearly observed visually and is confirmed by PCA analysis of the encoded concept samples that yielded over 80% variance for the two highest components. The bottom plot visualizes distributions of several concepts simultaneously in a compact region of the latent space. Interestingly, the distribution regions of the multiple concepts, again in the clear form of two-dimensional surfaces, are layered quite closely together rather than being separated into isolated clusters, as was the case with classes of some other categories. In this pattern, concept regions are stacked closely in the same region of the representation space like an “onion shell”, a strategy that allows to pack

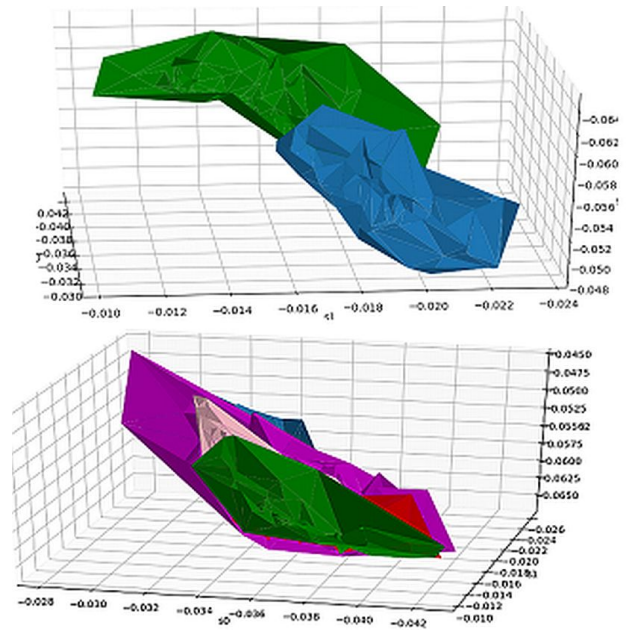


Figure 3: Compact concept distributions in the latent space

data in a very compact and efficient way.

It is worth noting that these results also substantiate the manifold assumption commonly used in unsupervised and semi-supervised learning [19]. For most of studied concepts in this category distribution regions indeed consisted of connected and smooth manifolds or sets of such manifolds. The results of measurements of the distribution parameters for these concepts will be presented in the next section.

Sparse Distributions Distributions of object and structure type concepts showed a different pattern that was noticeably sparser and spread over the latent representation. In Fig.4 concept regions of “structure” and “object” classes are shown with the compact classes that allows to compare relative scales of the variation in the concept regions of classes of different categories: top plot: classes 6 (sparse) and 3 (compact); bottom plot: classes 7 (sparse) and 2,4 (compact). A clear difference in the character of distributions of different categories can be observed the distribution visualizations in Fig.3 and 4. Interestingly, while larger scale background-type concepts appear to occupy compact and well-defined region in the latent space with a small number or single dominant cluster, classes representing local concepts are spread throughout the representation space in multiple clusters. A possible explanation for the latter observation could be that the relationship between the explicit higher-level concepts that label the samples in the dataset and the internal or native concept clusters (3) in unsupervised mode can be more complex than one-to-one. For example, an explicit higher-level concept may encompass a number of different native clusters in which case a distribution of the type seen in Fig.4

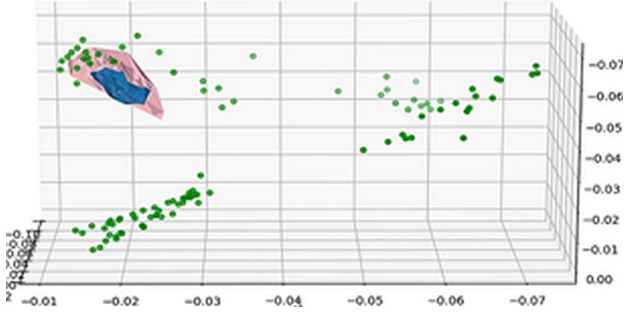


Figure 4: Sparse concept distributions in the representation space

can be observed.

Another logical possibility is that the complexity and depth of the models used in the study, as well as the size of the dataset were not sufficient to identify these more complex patterns with sufficient confidence. This question requires further investigation.

3.2 Categorization and Classification

In this section we attempted to establish the relationship between the categorization properties of concept distributions in the unsupervised representations and the performance of supervised learning with training data in the representation space of a trained model.

As mentioned in the previous sections, the categorizing ability of an unsupervised model can be evaluated with two essentially different approaches: first, in a completely unsupervised mode, where the external concept labels are not provided with samples in the dataset, the parameters of general distribution can be measured, such as the dimensions, shape, the parameters of density distribution. These measurements are important because they provide an a priori evaluation of the categorization ability of the model before any knowledge of external semantics such as known higher-level categories associated with the input data has been applied. For that reason, these methods can be applied to data of any nature in a truly general manner. On the other hand, if external labels for a subset of the data are available (as was the case in this study), it should be possible to train a classifier with labeled data in a supervised mode, but with parameters or “features” being the coordinates in the representation space of a pre-trained model with (4). Comparing the results of the two approaches can indicate how closely the structure that emerges in the representation as a result of unsupervised learning reflects the external concepts used in supervised approach.

The results of measurements of distribution parameters and the accuracy of classification for for selected concepts for each of the scale types are presented in Table 2. The parameters of the concept distribution region in the latent space were defined ([9]) as:

Spread, a characteristic size of the region relative to that of the general distribution

Concentration, the number of concept density clusters relative to the total number of clusters in the general distribution

Density, the density of the structure measured as the population per volume in the latent coordinates, relative to the density of a uniform distribution.

Finally, *Accuracy* for the concept was measured as F1 classification score that accounts for classification errors of both types. The accuracy of a trained classifier was measured with multiples batches of randomly selected in- and out-of-class test samples. Note that the second value in the accuracy column relates to self-learning accuracy that will be discussed in the next section. In the results, a clear

Table 2: Self-learning with unsupervised representations

Class	Categorization	Accuracy
Background		
Trees (2)	0.16, 0.06, 246	0.79, 0.65
Field (3)	0.18, 0.06, 357	0.81, 0.72
Water (4)	0.19, 0.08, 375	0.84, 0.78
<i>Structure</i>		
Roads (5)	0.23, 0.11, 228	0.68, 0.57
Excavations (6)	0.28, 0.14, 292	0.71, 0.54
<i>Object</i>		
Vehicles (7)	0.78, 0.22, 135	0.73, 0.53

correlation can be seen between the parameters of a concept distribution in the representation space and the accuracy of the concept classifier trained with a labeled subset of in- and out-of-concept samples in the latent coordinates. It can be seen as another indication, in addition to already mentioned results that unsupervised training, perhaps under certain conditions and constraints as discussed in [10, 20] can produce configurations of data in the representation space that are correlated with common higher-level concepts in the observable data.

3.3 Self-Learning with Unsupervised Representations

As was shown in [9], the structure emergent in the latent representations as a result of unsupervised training can be used in learning of new concepts with minimal data, down to counted positive samples. The approach has unsupervised and semi-supervised learning phases:

- in the unsupervised phase that requires no labeled data, principal density clusters with significant population are identified as was outlined earlier in Section 2.3, (4); these structures can be seen as principal native concepts in the observable data;
- in the semi-supervised self-learning phase that follows, a small number of positive concept samples is used to tag or

mark the clusters that can be associated with the concept being learned, and creating a small labeled dataset from the genuine concept samples and those obtained from the unsupervised cluster distribution;

- then a binary concept classifier is trained with the dataset and can be used for prediction of the concept being learned for new samples in the input space. Because the genuine labeled samples are used only for tagging of clusters of interest, the method can indeed work with very minimal sets of labeled concept data, down to a single, “signal” sample. In this section the single sample self-learning based on unsupervised density structure was applied to the image dataset, with results for representative classes in each category presented in Table 2, the second value in the accuracy column.

These results show that the concepts with compact representations were learned successfully with a single sample of the concept, while those with more spread and sparse representation achieved only marginally better resolution compared to the random strategy. A possible explanation for this effect can be found in the analysis of the distribution patterns for the concepts in Fig.3, 4. If the representation image of a higher-level concept comprises several native clusters, the datapoints generated in the vicinity of the signal sample wouldn’t sufficiently cover the entire distribution region of the concept in the latent space and the resolution of the classifier would be reduced. This was confirmed by further experiments where it was observed that increasing the size of the learning sample for this type of concepts substantially improves the results of learning. Unlike traditional in machine learning supervised methods, learning with unsupervised density structure or density landscape is more reminiscent of the learning processes in the biologic systems that are often spontaneous, flexible and require minimal data with building accuracy gradually over a sequence of learning iterations. Landscape-based learning can imitate such processes by testing concept distribution regions in an iterative trial and error process as and when learning data becomes available in a close interaction with the environment.

4 Conclusion

The analysis of higher-level concept distributions of image data in the latent space of self-learning models presented in this work is in agreement with the earlier findings that unsupervised training of models with self-encoding and regeneration can lead to emergence of identifiable structure in the latent representation that can be correlated with higher-level concepts in the observable data.

Correlation of classification accuracy with the categorization parameters of the concept distributions in the latent space of such models was shown now with data of different types and nature [7, 10, 9] pointing at the possibility of a general character of this effect.

Low-dimensional representations can be of interest due to

growing evidence that such representations can play an important role in processing of sensory data by biologic systems. Recent results [13, 14] demonstrated that effective representations of sensory data such as images and smells can be produced with a small number of active neurons in biologic neural networks. Linking these results with the findings in this work, where the examples of such low-dimensional representations created artificially were investigated, one can hypothesize that perhaps, the representations in more complex sparse neural networks even of a massive kind [7] can be modeled as a set or a “stack” of low dimensional representation regions indexed by the combination of neurons that collectively participate in creating the latent representation, with surface-like concept regions observed in our results, distributed in them (Fig.5). In such a stacked representation, a concept region for ex-

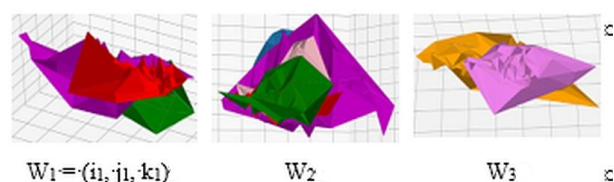


Figure 5: Concept regions in a sparse latent representation

ample, “cats” can be indexed by the indices of activated neurons W_k and the index S_k of the concept surface in the representation subregion of W_k : $I_{cats} = (W_{cats}, S_{cats})$.

Thus, prototypes of native concepts in the observable data can form in an unsupervised observation of the environment via self-learning with minimization of error of regeneration, requiring minimal supervision and prior knowledge of the domain.

Analysing concept distributions in the representations of deep learning models can offer a novel perspective on the program of Explainable AI [21]. Much effort has been invested by the research community in attempts to describe the learning configurations and rules that emerge in complex and deep learning models in training. Understanding the native structure of information in the latent representation created in training can offer a different, and in some cases, very visual interpretation of learning processes in these systems.

All in all, it is believed that the study of native categorization properties of the generative models may lead to better understanding of the underlying principles of self-learning and development of models that could learn in more natural way [16], closer to the spontaneous and iterative learning processes in biologic systems.

Acknowledgements

The author is grateful to Prof. Pilip Prystavka, Chair of the Applied Mathematics, National Aviation University (Kyiv) for valuable discussions of the findings and the opportunity to use the dataset of images used in this work.

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] 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
- [5] Ranzato, M.A., Boureau Y.-L., Chopra, S., LeCun, Y.: A unified energy-based framework for unsupervised learning. *Proc. 11th Intl. Conf. on Artificial Intelligence and Statistics*, **2** (2007) 371–379
- [6] Friston, K.: A free energy principle for biological systems. *Entropy* **14** (2012) 2100–2121
- [7] Le, Q.V., Ransato, M. A., Monga R., et al. Building high-level features using large scale unsupervised learning. *arXiv* **1112.6209** (2012)
- [8] Banino, C., Barry, D., Kumaran D.: Vector-based navigation using grid-like representations in artificial agents. *Nature* **557** (2018) 429–433
- [9] Dolgikh, S.: Categorized representations and general learning. *Proc. 10th Intl. Conf. on Theory and Application of Soft Computing, Computing with Words and Perceptions* **1095** (2019) 93–100
- [10] Higgins, I., Matthey, L., Glorot, X., Pal, A., et al.: Early visual concept learning with unsupervised deep learning. *arXiv* **1606.05579** (2016)
- [11] 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
- [12] Rodriguez, R. C., Alaniz, S., and Akata, Z.: Modeling conceptual understanding in image reference games. In: *Advances in Neural Information Proc. Syst.* (Vancouver, BC) (2019) 13155–13165
- [13] Yoshida, T., Ohki, K.: Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications* **11** (2020) 872
- [14] Bao, X., Gjorgieva, 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
- [15] Hornik, K., Stinchcombe M., White H.: Multilayer feed-forward neural networks are universal approximators. *Neural Networks*, **2(5)**, (1989) 359–366
- [16] Hassabis, D., Kumaran, D., Summerfield C. et al.: Neuroscience inspired Artificial Intelligence. *Neuron* **95(2)** (2017) 245–258
- [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* **21(1)** (1975) 32–40
- [19] Zhou, X., Belkin M.: Semi-supervised learning. In: *Acad. Press Lib. in Signal Proc.* Elsevier (2014) 1239–1269
- [20] Dolgikh, S.: Why good generative models categorize. *Int. Journ. Mod. Edu. Comp. Sci.* (2020) (to appear)
- [21] Gilpin L.H., Bau D., Yuan B.Z. et al.: Explaining explanations: an overview of interpretability of machine learning. *arXiv* **1806.00069** (2018)