International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 14 (2022) pp. 239-248

Received: 21 December2021; Accepted: 15 March, 2022; Published: 28 April, 2022

Generative Conceptual Representations and Semantic Communications

Serge Dolgikh[1]

[1] Department of Information Technology, National Aviation University,
1 Lubomyra Huzara Ave, Kyiv 03058, Ukraine
sdolgikh@nau.edu.ua

***Abstract*: Representations are essential in learning of natural and artificial systems due to their ability to identify characteristic patterns in the sensory inputs. In this work we examined latent representations of images of basic geometric shapes and handwritten digits as a basis for sharing semantic information about observations in a collective of unsupervised generative learners. Individual models trained in an unsupervised process with minimization of generative error were exposed to a process of synchronization of symbolic tokens associated with characteristic regions in the latent representations identified with two different strategies. It was demonstrated that conceptual representations with good decoupling of characteristic patterns can be produced reliably and consistently with models of unsupervised generative self-learning; and that a simple process of conceptual synchronization can enable effective sharing of information between individuals in a collective by associating shared symbols with latent regions correlated with characteristic patterns in the sensory inputs. The results demonstrate the potential of conceptual latent representations as a natural platform for development of abstract concept intelligence and communications.**

***Keywords*:** machine learning, unsupervised learning, representation learning, concept learning, clustering.

I. Introduction

Representation learning with the objective to identify patterns in the observable data has a well-established record in the discipline of machine learning. Informative representations obtained with Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN) [1, 2], different flavors of autoencoders [3] and other models in unsupervised learning (unsupervised feature extraction) allowed to improve accuracy of subsequent supervised learning with conventional methods [4].

Informative representations produced with models of unsupervised generative self-learning were used in a growing number of applications to identify characteristic patterns, or concepts, classes of interest in the observable data. Generative models based on artificial neural networks have a strong potential in such problems due to their capability of universal approximation [5], making them suitable for processing data of virtually any type and complexity including live images, video streams and other types of complex sensory data.

Studying underlying structure of latent representations of generative models can be instrumental due to the observed effect that they can capture essential characteristics of distributions in the sensory environment represented by training datasets. Understanding this structure can offer essential insights into how to improve learning ability of models especially in problems and environments where large amounts of confident prior knowledge is not available.

A. Related Work

Concept learning from unsupervised observation in artificial learning systems was attempted in a number of studies beginning from works in the late 1990s – early 2000 that demonstrated noticeable improvements in the performance of supervised training after unsupervised processing (unsupervised feature selection, etc.) with self-learning models such as RBM, DBN, different types of autoencoder models and other types and architectures. This standard practice, commonly used as a mean to achieve higher accuracy of subsequent supervised training and classification of known concepts, can be seen, in fact, as an intriguing effect as generative models have no access to the externally defined classes in the process of unsupervised training, and an improved correlation between features obtained in unsupervised processing of training data and the external classes could therefore point at a possibility of a link between unsupervised generative learning, and “pre-known” concepts de-fined externally. The existence of such a link is not entirely obvious, and its nature and origins merit, in the authors view, an in-depth investigation.

Earlier results obtained with generative neural network models include applications of deep autoencoder models of different architectures such as sparse, variational, convolutional [6,7] and others to produce informative representations of complex data such as different types of images [8-10], network and Internet [11] and other types of data [12,13]. These results demonstrated that structured latent representations correlated with external higher-level concepts can be produced under certain conditions imposed in training, such as generative accuracy and redundancy reduction by models of unsupervised generative learning.

An unsupervised structure of this kind that does not require massive amounts of labeled data to produce and it can be hypothesized that it could be harnessed to develop effective methods of learning in the environments with scarcity of prior knowledge about the content or characteristics of the

distributions in sensory data. Given the constraints of the problem, informative representations obtained with methods and models of unsupervised generative learning can be used as a platform, framework or "landscape" in which learning can progress effectively even with scarce learning data by using the informative structure in informative latent representations.

The effect of categorization by higher-level concept in unsupervised learning, that is, emergence of latent structures correlated with external concepts as a result of unsupervised generative learning was reported in a number of studies. Le et al., [8] observed spontaneous formation of concept-sensitive neurons activated by images in certain higher-level classes with a deep sparse autoencoder architecture trained with massive arrays of images in an entirely unsupervised process without exposure to known concept semantics. Higgins et al [9] demonstrated decoupled structure of latent representations obtained with variational autoencoder models with different sets of images and underlined the importance of constraints such as redundancy reduction and generative ability in successful unsupervised learning by demonstrating unsupervised models capable to learn without massive supervision with images of different types. In [14] a spontaneous formation of grid-like navigation cells, similar to those observed in mammals was detected in a recurrent neural network with deep reinforcement learning. In [15] structure and topology of generative representations was described with a dataset of images of geometric shapes.

Though undoubtedly significant, a common "chicken and egg" observation related to these results can be made: we have to know what to look for, before being able to verify how well a model has learned it. In other words, these models can provide some insights into the relationship between internal or native information structures that emerge in unsupervised learning and the external, explicit concepts but could not explain how it originates without certain domain knowledge, such as pre-labeled concept data.

While important in proving the ability of generative models to learn successfully from unsupervised processing of sensory data of different types, origin and complexity, the question of "conceptual bootstrap" remained less explored: what are the internal, native information structures that can be associated with known external concepts? Besides, complex and specific architectural features of these models give rise to questions about generality of the observed effect and its applicability to other models and learning scenarios.

In this work we attempted to approach challenges of both conceptual and practical nature in the investigation of conceptual structure in generative representations from several perspectives: first, by choosing a generative neural network architecture of limited complexity we intended to verify the general character of the effect of categorization in unsupervised generative learning; the second objective was to examine the question of origin of higher-level concepts; to approach it methods of evaluation and measurement of distributions of data in unsupervised latent representations were developed and verified. Finally, we investigated possible mechanisms of sharing information about sensory observations in a collective of generative learners, based on the structure emergent in the process of unsupervised generative learning.

The rest of the paper is organized as follows: Section II contains a description of generative models and data used in the study. In Section III we describe the process of production of structured latent representations and synchronization of latent structure between individual learners. In Section IV the results of synchronization experiments are presented with different types of image data and methods of identification of latent structure. Finally, Sections V and VI contain a discussion of the results, their significance and relation to other results in the field of unsupervised generative learning and a synopsis of the work.

II. Materials and Methods

Models based on the architecture of convolutional autoencoder neural network [16] with strong dimensionality reduction to a low-dimensional latent representation were used to produce structured latent representations of a dataset of images of basic geometric shapes as described in this section. Neural networks are good candidates as generative learners of complex data types such as images, due to their capacity of universal approximation [5].

Once the ability of the models to learn characteristic patterns ("concepts") in the data has been established, the objective of the study was to demonstrate and verify their ability to synchronize individual concept structures identified in the process of unsupervised generative learning via a process of information exchange in a collective of learning models trained individually and independently).

A. Convolutional Autoencoder Model

Generative models of the architecture of a convolutional autoencoder had the encoding stage with convolution-pooling layers followed by several layers of dimensionality reduction with a single latent layer of size *M*. The resulting latent representation of the same dimension was defined by activations of the neurons in the latent layer.

The decoding / generative stage was fully symmetrical to the encoder. Overall, the architecture had 21 layers and approximately 40,000 trainable parameters. The models were implemented in Keras / Tensorflow [17] and trained for minimization of the deviation between the training batches of images and their generations by the model (generative error) with categorical cross-entropy cost function (CCE). An architectural diagram of generative models used in the study is presented in Figure 1.

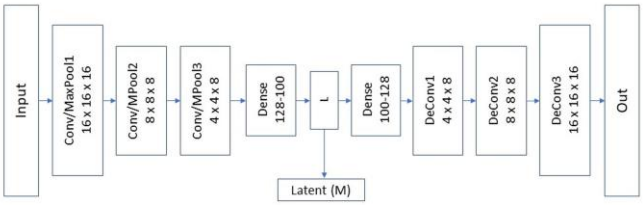

Figure 1. Convolutional autoencoder architecture with dimensionality reduction.

Depending on the type of data, two flavors of generative architecture were used: flat and sparse. Flat models had a low-dimensional latent layer of dimension $M = 3 .. 5$, i.e., 3 to 5 latent neurons. These models were used with data of lower complexity representing images of geometric shapes (Section 2.2). Sparse models had higher dimensionality of the latent layer, $M = 20 .. 25$, with L1 sparsity activation penalty imposed in training. These models were trained with the images of handwritten digits of higher conceptual complexity. Architectural parameters of generative models are provided in Table 1.

Model	Flat	Sparse
Adaptation	Convolution, 2-3	Convolution, 3
Latent size	3–5	20–25
Total layers	11	15
Trainable parameters	4×10^4	9×10^5
Sparsity [(2)]	No	Yes
Cost function	MSE, CCE	CCE

[(1)] MSE: mean squared error; CCE: categorical cross-entropy
[(2)] L1 regularization sparsity activation penalty

Table 1. Architectural parameters.

Both types of models produced low-dimensional representations of the training data. With flat models, representations were defined by activations of all latent neurons producing a latent vector space of dimension *M*. With sparse models, as a result of a sparsity penalty imposed in training, activations for most inputs had 2 to 4 active neurons, describing effective latent subspaces or "slices" of effective dimension $F = 2–4$ in the *M*-dimensional latent space (Fig. 2), indexed by neurons activated by inputs.

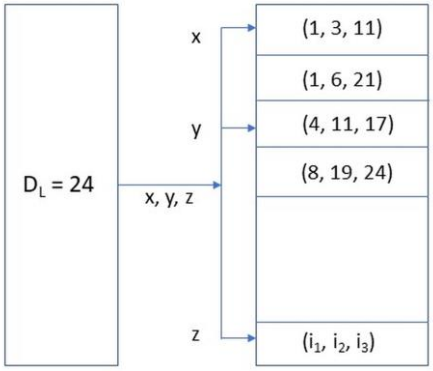

Figure 2. Stacked structure in a sparse latent space.

B. Data

Datasets of grayscale images of basic geometric shapes: circles, triangles and grey-scale backgrounds of size 64×64 were used to model simple yet realistic (i.e., minimally realistic) visual environments. While images represented simple shapes, the intent was for the characteristic patterns in the datasets to have certain realistic meaning for some primitive learning system, for example, different types of shapes associated with sources of food versus predators and general background.The first dataset, Shapes-1, consisted of 600 grayscale images of circles, triangles and grayscale backgrounds with two representative samples per shape with variation in the size and contrast of fore / background.

The second dataset, Shapes-2, contained 1,000 of grayscale images of circles, tri-angles and backgrounds with variation of the size in the range 0.3 – 1.0 of the image size, with variation of contrast of fore- vs. background for each size.

To model more complex realistic visual environments, a dataset of handwritten digits (MNIST dataset, [18]) was used as well.

C. Training

A success of generative training was measured by the change in the validation cost function over the process of unsupervised training and the ability of trained models to generate a subset of images of the types represented in the training dataset (Fig. 3).

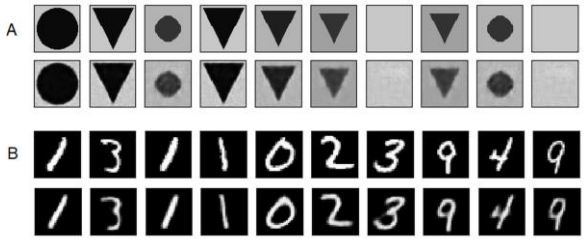

Figure 3. Generative performance of trained models (top: input; bottom: generation by a trained mod-el) A: geometric shapes (flat model); B: handwritten digits (sparse model).

A majority of learning models were successful in generative learning:

- Geometrical shapes datasets, flat models: ~ 80%
- Handwritten digits dataset, sparse models: 60 – 70%

though a spread in the generative quality was observed among individual models in the ensemble.

D. Encoding and Generation

A trained generative model can perform two essential transformations of data as a result of an unsupervised training process that does not use labeled samples of pre-known concepts. The encoding transformation *E*, realized by the encoding stage of the model (tensor *E*(Input, *L*), Fig. 1), transforms a sample *x* in the observable space *O* to its encoded position *y* in the latent representation space *R*. The generative transformation *G* operates in the opposite direction, from the latent representation to the observable space and is realized by the generating stage of the model (tensor *G*(*L*, Out), Fig.1):

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

Straightforwardly but essential for subsequent analysis, the process of unsupervised training allows to decouple the encoding and generative stages of the model, so that not only an encoded image of an actual observation $E(x)$ but any position *y* in the latent representation space *R* can produce an observable image via generative trans-formation (1).

The transformations are also essentially independent: once the training phase completes, the parameters of the encoding and generative transformations are fixed and contained in the corresponding tensors, so that no information about the encoding stage is needed to generate an observable image of a latent position (1) and vice versa, no knowledge of generative parameters is needed for encoding.

III. Synchronized Conceptual Representations

A. Conceptual Representations

In the first phase of the process of production of conceptual representations, the ability of models to create structured low-dimensional latent representations correlated with characteristic patterns in the observable data was verified. It is supported by a number of earlier results [11-14] with data of different types and origin.

Models of both types, flat and sparse, produced structured representations correlated with characteristic types of images in the training sets. Methods of density clustering [19] were used to identify density features in the latent representations

of trained models in an entirely unsupervised process. With the latent density structure identified in a clustering process, correlation of the produced structure with characteristic content of the data in the training sets was established by propagating positions of the identified structural features, for example, centers of density clusters, to the observable space with generative transformation (1). With all successfully trained models in the study, density structure obtained with this method showed a clear correlation with characteristic types of images in the datasets (Figure 4).

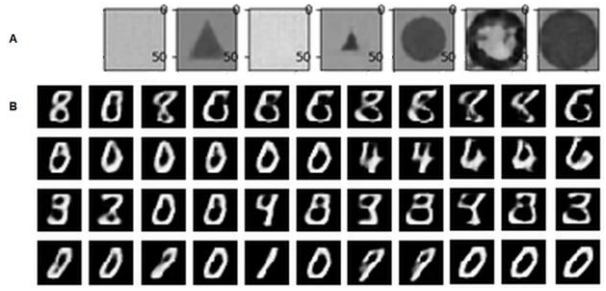

Figure 4. Observable images of latent density clusters. A: set Shapes-2 (flat model); B: handwritten digits (sparse model).

The ability of generative models to produce structured latent representations can be used to identify latent regions associated with internal or "natural" concepts in the observed data. Structured geometry of latent representations allows trained models to associate observations x in the observable space (that is, images) to internal or "natural" concepts $T = \{ T_k \}$ via the relationship of containment of the encoded position of x in a characteristic region R_k in the latent representation as:

$$r(x) = E(x) \in R_i \rightarrow x \in T_i; T(x) = T_i \qquad (2)$$

where $T(x)$: natural concept associated with an observable sample x.

As has been demonstrated earlier [16], characteristic concept regions can be identified by unsupervised methods such as density clustering or novelty/similarity based, signifying that association of sensory data to concepts (2) can be obtained in an entirely unsupervised process that does not require labeled concept samples or any essential prior knowledge about the training dataset and for this reason has to be defined only by the internal characteristics of the data and generative architecture.

Several strategies can be applied for this purpose as has been demonstrated in earlier studies [8,13], including entirely unsupervised methods that require no prior concept knowledge or semi-supervised ones, using small sets of concept samples. In this work, as an illustration of possible approaches, two different methods were used, though no attempt to optimize the performance was made. The first one is an application of a density clustering method, such as MeanShift and similar [19,20] to a representative sample of the dataset encoded to the latent space to produce a set of latent density clusters. Such methods are in the essence, unsupervised and do not require prior knowledge about conceptual content in the data.

The second method is based on a similarity relationship between latent samples. In the first iteration, there is a single set of samples S_1 defined by some similarity relationship, for example, "a triangle". A geometry-based binary classifier such as Nearest Neighbor [21] for the concept associated with the samples can be obtained with a) the encoded set $P_1 = E(S_1)$ representing in-class training subset; and b) a subset of the encoded general sample g, $E(g)$ at the maximum distance from the center of P_1. The process is repeated iteratively for each new next concept, with positive samples of known concepts used as negative ones for the novel concept (refer to the Appendix for details). An internal concept $T_k \in T$ can then be interpreted as: 1) the latent region associated via encoding and generative transformation (1) with a certain characteristic pattern in the observable data, for example, a type of geometrical shape or digit; 2) a symbolic token or index q_k associated with conceptual regions by individual learning models, allowing to distinguish between them. As a result of this phase, a structure of concept features such as density clusters is produced, with an ability of a trained model to associate an observable sample to its internal or natural concept.

B. Sparse Conceptual Representations

Sparse representations are similar to the flat ones described in the preceding section, with an extension that clustering is performed independently in the low-dimensional subspaces or "slices" of the full M-dimensional latent space (Fig.2).

The slices with highest populations of activations can be identified with a representative sample of input images encoded to the latent space; inputs are placed in a given slice if their most significant activations match the slice's index (i.e. the tuple of latent coordinates, such as (i, j, k), i, j, k = 1 .. M) and achieve a minimal activation threshold. Then, density clustering can be applied independently in each slice on its population. The resulting structure of density clusters indexed by slices represents a complete generative structure or "landscape" in the latent space with a natural unique index of (slice, cluster). A subset of a generative landscape in the form of observable images generated from density clusters in several 3-dimensional slices of a 24-dimensional latent space is shown in Figure 4.

C. Symbolic Representations

An essential observation used in the rest of this work is that in addition to the ability to associate observable inputs to native concepts (2), generative models are also capable of interpreting symbols associated with concepts as representative instances or "prototypes" [22]. Concept prototypes can be defined in both latent (representation) and observable spaces. A representative latent instance t_k of a concept T_k associated with a latent region R_k can be determined by several methods:

- As a set of known samples or a function such as the mean of a set of known samples of the concept in R_k.
- As a characteristic position in the latent region of the concept, for example, the center of an associated density cluster, the mean position of a cluster and so on.
- Calculated from geometric parameters of the concept regions R_k, if it is known with sufficient detail, such as geometric center; and others.

Then, a representative observable prototype P_k of the concept T_k can be obtained as:

$$P_k = G(t_k) = G(t(R_k)) \qquad (3)$$

with $t(T) = t(R_k)$ being the prototype-producing strategy that associates representative instances to a latent concept region as discussed earlier. In this work two different strategies were used: cluster centers identified with density clustering; and the mean position of concept samples identified by similarity relationship. Based on these observations it can be concluded that the unique index of a concept T_k or any symbol S_k uniquely associated with it can be interpreted as an observable image (prototype) of the concept as: $S_k \rightarrow T_k \rightarrow t_k \rightarrow P_k$ (3).

The ability to interpret symbolic tokens as observable prototypes clearly distinguishes generative models from conventional methods of supervised learning, where such a task would not be meaningful. However, in this phase symbolic tokens of concepts are produced independently by each learner, for example, as an index of the identified density cluster and have no semantic meaning for other learners in the ensemble. For example, an internal index of a density structure in a sparse latent space of a generative model described in Section III.B can be associated with a different type of image, or not be valid at all for a different model. To enable the exchange of information about sensory observations in a collective, a process of synchronization of individual symbolic concept frameworks (concept maps) is needed.

D. Detection and Interpretation

The ability to exchange symbolic information about sensory observations is based on several necessary conditions:

1. Structural consistency of latent representations, that can be defined as a set of latent regions associated with characteristic patterns in the observed data between the learners in the ensemble. For visual data similar to that used in the study, it is supported by the results in this study and a number of previous results [10,14,16].
2. An ability to associate a symbol to a specific latent position or region. As discussed in Section III.A this ability naturally exists in models of generative learning in which symbols can be associated for example, with internal indices of latent features.
3. An ability to produce an observable image of a latent position or region. This ability is also natural in generative models as a representative instance (prototype) function discussed in Section III.A and can be considered as "interpretation" of a latent position or a symbol associated with such.
4. Finally, an ability to produce symbolic responses to observations that can be characterized as "detection".

The processes of interpretation and detection are illustrated in the Figure 5.

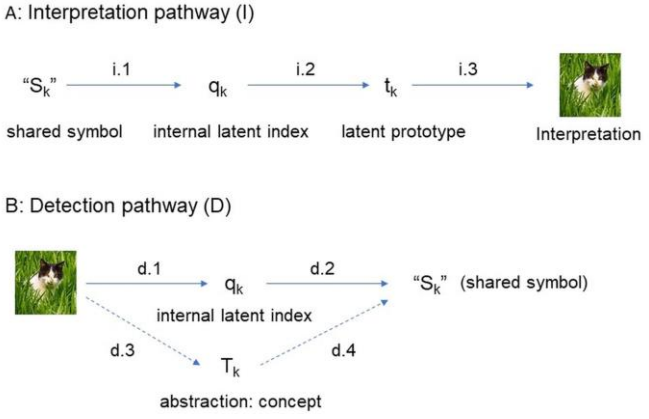

Figure 5. Interpretation and detection in generative learning

The interpretation sequence (A, Fig. 5) is more straightforward and we begin with it. Suppose through some process of synchronization, internal indices or tokens of features in the latent structure (conceptual maps or landscape, as discussed in Sections III.A, III.B) are associated with a set of shared symbols, $\{ S_k \}$. As a result of this process, a shared symbol can be interpreted by all or most individuals in the collective of learners via an ability to translate it to an internal token associated with a latent structure learned in generative training. Then a reception of a shared symbol can be translated to a token of a certain latent structure in the latent landscape via (1) and through the segment i.3, Fig.5 produce an observable interpretation of the communicated symbol. Importantly, interpretations of the same symbol by different individuals in the ensemble do not have to be identical and likely would not be, however based on the assumption of consistency, it can be expected that they would represent the same or similar types of observable patterns. In the detection pathway (B, Fig.5), a direct connection from an observation to a shared symbol is possible through segments d.1, d.2 in the diagram, i.e., as:

Observation, $x \rightarrow$ latent image, $y = E(x) \rightarrow$ latent feature token, $q_k(y) \rightarrow$ shared symbol, S_k, that can be communicated.

This strategy works well for simpler types of data with a small number of latent features, such as geometrical shapes datasets in this work. In this scenario, the likely outcome of the synchronization process is that all or most latent features would be associated with a shared symbol and detection can be implemented by a direct translation latent feature to symbol ("direct" strategy). However, with more complex data such as handwritten digits, it can be observed (Fig.3) that latent positions of characteristic observable patterns, such as a certain digit can be spread across multiple, and possibly large number of latent features. It can be less likely in this case that in the synchronization process all landscape features are associated with symbols and translation of an observation to a shared symbol via direct strategy may not be successful in all cases. A possibility to improve effectiveness of detection is provided by another function of generative models, that of generalization. Suppose, a more experienced learner in the collective has acquired, via certain process, an ability to classify sensory observations into general classes: $K = \{ K_j \}$, for example, interpret multiple individual versions of a digit "0" as being in the same general class. Approaches in empirical, environment driven learning based on conceptual structure discussed earlier have been proposed [23] but will not be discussed here due to limitations of this study. Then, a different detection pathway ("abstraction" or "general") would be possible as well (Fig.5 B, d3, d4):

Observation, $x \rightarrow$ general class, $K_j(x) \rightarrow$ shared symbol, S_j associated with class K_j.

The detection strategy above would provide a more confident detection of essential sensory patterns because it could incorporate cases where the natural feature associated with an observation, $q_k(x)$ has not been associated with a shared symbol in the synchronization process. The pathways of detection and interpretation described in this section, along with a process of symbolic synchronization of conceptual maps of individual learners in an ensemble provide a basis for sharing of sensory observations in a collective.

E. Synchronized Representations

In this section a basic implementation of the process of synchronization of conceptual maps between individuals in

the ensemble based on shared observation is described. It is intended to demonstrate a possibility of such a process and its possible role in the emergence of communication and other collective intelligence behaviors but optimization for effectiveness and performance was not intended at this stage. The intent of the experiment was to examine the ability of individual learners in a collective to synchronize their conceptual models of the sensory data via a simple process of group exposure to a limited subset of sensory inputs with production of a shared symbol ("orchestration") and thereupon produce consistent interpretations of symbolic information.

In the first, synchronization phase of the experiment, individual generative models were trained independently in an unsupervised generative process and conceptual models or maps were produced as described in Sections III.A, III.B as maps of density clusters indexed by a unique internal index. For flat models (datasets Shapes-1, Shapes-2) the index was an integer number associated with identified density clusters, ordered by population; for sparse models used with the dataset of handwritten digits, the index was an integer tuple (slice, cluster id) as discussed in Section III.B.

For an individual model, a valid value of the index uniquely identifies corresponding latent feature in the conceptual map; but as commented earlier a given value of an index has semantical meaning only for a given individual and has no information value for other individuals in the group. A group of prepared models with individual conceptual maps was then shown a sequence of images of the types present in the training set, along with "shared" symbols S = { S_k }; for example, circles were associated with "c", triangles with "t" and so on. The synchronization set thus consisted of pairs (X_i, S_i) of sensory images and shared symbols and was consistent, that is, did not contain contradictory associations. Upon each observation, learners were instructed to associate the symbol associated with the shown image to the latent prototype of the image obtained with (3).

The process created an association or "dictionary" allowing translation of internal indices q_k to and from shared concept symbols S_k, and the models in the group are considered "synchronized" (Figure 6). For models with an ability to generalize and classify observations to general concepts as discussed in Section III.D an additional synchronization sequence associating the identified concept of the synchronization image with the shared symbol. Such a strategy can improve the effectiveness of the detection due to incorporation of multiple latent features into a single general concept ("general" strategy, Section III.D).

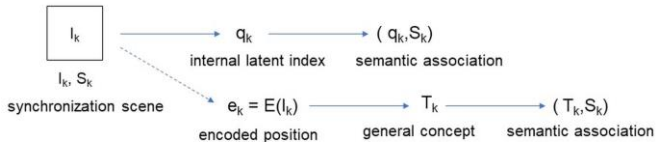

Figure 6. Synchronization process. Top: direct; bottom: with generalized concept classifier.

In the verification stage of the experiment, individual learners synchronized via the described process were presented with a sequence of shared symbols such as: "t, c, t, c, b, t …" and instructed to produce interpretations of symbols as observable proto-types or interpretations, associated with them. Finally, as discussed in Section III.D, the detection sequence modeling production of symbolic communications of sensory observations was examined as well. For flat models with the data of lower complexity, the sequence based on direct association of an observation to latent feature was tested (Fig.5, B segments d1, d2). For sparse models with more complex handwritten digits images, the pathway based on generalization (Fig.5, B segments d3, d4) was used. If the assumptions of the study were correct and learning and synchronization processes were successful, interpretations of shared symbols by individual models would be consistent, supporting the ability to share information about sensory observations in the collective via exchange of shared symbols.

IV. Results

A. Conceptual Representations

In experiments with individual models, it was confirmed that the latent regions containing the representations of different types of images were connected and continuous. For a subset of images S of the same type, for example, circles in the observable space, generated image of the mean of the encoded representations of S in the latent space R was of the same type, indicating a connected topology of concept regions. This observation was confirmed by detailed investigation of the topology of the latent space [16] demonstrating a structure of well-defined connected concept regions separated by boundary surfaces of lower dimensionality.

An analysis of the structure of latent representations of individual models trained with the same dataset confirmed the assumption of structural consistency. Generative training produced latent representations with consistent structure that was confirmed in the experiments with different instances of trained models (Table 2, sparse model, MNIST dataset).

Model instance	**Sparse-1**	**Sparse-2**	**Sparse-3**
Size	474	396	485
Recognition	0.973	0.975	0.971
All digits represented	True	True	True
Highest representation	0, 7, 3	0, 7, 1	4, 7, 0
Lowest representation	4, 6, 9	2, 5, 6	2, 8, 6

Size: the number of non-empty density clusters
Recognition: the fraction of the landscape associated with recognizable digits
Representation: slices with highest / lowest population

Table 2. Consistency of latent landscape.

The findings in this section were consistent with results of other studies [14,16] indicating consistency of latent structure in the representations of generative models trained with similar data.

B. Symbolic Representations

Methods described in Sections III.A, III.B were used to produce conceptual maps following generative training with datasets Shapes-1, Shapes-2 (flat models) and MNIST (sparse models).

With the Shapes-1 dataset, a density clustering method (MeanShift) was used with a general sample of images encoded to the latent space, producing a set of density clusters ranked by the size (i.e., the population) of the cluster. The index of the cluster was used as the unique concept token q_k.

Concept prototypes were defined as the center positions of the identified density clusters.

With the Shapes-2 dataset, similarity-based kNN classifiers were produced for concepts as described in Section III.B. The accuracy of concept identification is shown in Table A1 (Appendix). Concept prototypes were associated with geometrical means of the representative latent instances of the concepts.

With the MNIST dataset, a sparse latent landscape was produced indexed by a two-dimensional integer tuple as described in Section III.B. Concept prototypes were defined as the center positions of the identified density clusters.

As a result of this process, each trained model had an individual concept map produced with the ability to associate sensory inputs to internal concept tokens, and produce an observable prototype for a given concept token.

For interpretation pathway, with flat models only direct interpretation sequence based on association of shared symbol – latent feature, (S_k, q_k) was examined. A general classification ability, as discussed in Section III.E is likely to improve the effectiveness of interpretation; it was illustrated with one of the groups of sparse MNIST data models.

With the detection pathway, the effectiveness of symbolic response was measured as follows:

- Shapes-1 models (flat): direct strategy (d.1, d.2, Fig.5 B)
- Shapes-2 models (flat): direct strategy (d.1, d.2, Fig.5 B)
- MNIST models (sparse): direct and general strategies (Fig.5 B)

The effectiveness of interpretation was measured in a group of three independently trained models as: the rate of consistent interpretation I_c (all models produced consistent and correct interpretation of a test image; the rate of a partial agreement I_p (majority of models produced consistent and correct interpretation) and the rate of inconsistent interpretation, I_f.

The effectiveness of detection was measured in a group of two independently trained models as: the mean rate of a symbolic response to a sensory stimulus, D_r; and the mean rate of a correct response to a sensory stimulus D_c, between the models.

C. Synchronized Representations

The description the groups of generative models with synchronization and detection strategies used in the synchronization experiment are provided in Table 3. Each group contained several independently trained models with the same unsupervised dataset. Due to more complex nature of the MNIST data, a larger synchronization sample was used (up to 10 synchronization images per digit).

Group	Shapes,1	Shapes,2	Mnist,1	Mnist,2
Size	3	3	3	3
Type	flat	flat	sparse	sparse
Detection	direct	direct	direct	general
Prototype	density	mean	density	density
Sample*	10	10-15	30-100	30-100

* Size of synchronization sample

Table 3. Synchronization and detection strategies.

The results of the experiment measuring the effectiveness of interpretation and detection strategies as described in Section IV.C are presented in Table 4. With MNIST (digits) data, three digits were synchronized: 0, 1, 3.

Group	Shapes, 1	Shapes, 2	Mnist, 1	Mnist, 2
Interpretation, I_c	0.92	0.88	0.77	0.77
Interpretation, I_p	0.96	0.93	0.98	0.98
Interpretation, I_f	0.04	0.07	0.02	0.02
Detection, D_r	0.97	0.96	0.17–0.59 (1)	0.74–0.96 (1)
Detection, D_c	0.93	0.99	0.97	0.92

(1) Digits, "0" (lowest) to "1" (highest)

Table 4. Synchronization and detection, results

Comparing the effectiveness of detection strategies for models MNIST it can be concluded that generalization strategy allowed to significantly improve detection rate compared to direct detection via "tagged" latent features. The cause of failed interpretation i.e., "disagreement" between individual models has been identified as a variation in the generative structure between the latent position of the synchronization image and the center of the associated latent cluster from which an observable prototype was produced. By employing more sophisticated prototyping strategies it should be possible to improve these results.

Examples of the output of the interpretation experiment for models in groups Shapes-2 and MNIST are shown in Figure 6.

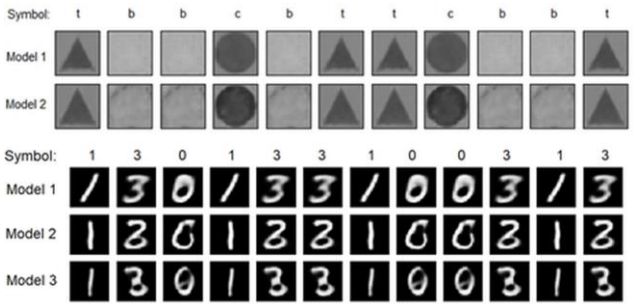

Figure 7. Interpretation of symbolic information following synchronization. Models: Shapes-2 (top), MNIST (bottom).

It can be noted that though the symbols were synchronized between the learners, each one was producing its own, individual interpretation of the transmitted symbolic information. For example, the interpretation of the background image ("b") by individual models in Figure 6 was noticeably different, differences in the individual prototypes of the types of images in the MNIST dataset (Fig.6, bottom) can be observed as well. In some observed cases, individual models produced different internal indexing of the concept clusters: (0, 1, 2) vs. (0, 2, 1) for background, circles, triangles, models Shapes-1. Nevertheless, the synchronization process associated correct shared symbols to internal indices, allowing consistent interpretations of the shared information by individual models.

A simple structure of latent representations of the geometric shapes with a small number of latent features data makes synchronization process straightforward and effective in these cases. It is indicative of the generality of the method that it works with certain effectiveness with significantly more

complex and realistic images of handwritten digits with significantly more variable content of characteristic patterns.

V. Discussion

A possibility to use structured generative representations of sensory data, including of more complex types such as real-world images with significant variation of content for successful environment-driven learning of concepts can be instrumental for a number of reasons.

First, it offers a direction for studying and modeling natural learning, methods and strategies based on direct observation of the sensory environment; with minimal confident samples obtained incrementally in empirical interactions with the environment; in a flexible process based on empirical trials, not dependent on availability of known concept data upfront, before learning process can begin. It can be hoped that models and systems designed on these principles can be more effective in the environments where confident knowledge of the domain is not available, as well as contribute to investigation of emergence and evolution of intelligent functions and behaviors.

Secondly, it can provide essential insights into the origins of higher-level concepts and concept-based intelligence. According to the results presented in this and a number of other studies, concept prototypes can emerge in generative processing of sensory data as native structures in generative representations related to and correlated with characteristic common patterns in the sensory inputs. Essential conditions for emergence of such structured representations appear to be generative accuracy, that is, encoding sufficient information about the observed distributions in the latent space, and redundancy reduction [10,14].

The line of investigation based on unsupervised structure emergent in representations of successful generative models can provide a solution to the conceptual "chicken and egg" puzzle: if true instances of higher-level concepts are needed to analyze and determine their representations, how can they be defined and what is their origin? Methods of analysis of unsupervised generative representations dis-cussed here allow to associate origins of general higher-level concepts in the sensory data with characteristic latent structures that can be determined with entirely unsupervised methods and without any prior knowledge of external concepts.

As the results in Sections III.A, III.B appear to suggest, concepts in this process may not emerge as a single broad class with subsequent specialization (hierarchical stratification). Rather, concept-associated features such as density clusters can be spread across the components of a complex latent space, such as stacked low-dimensional slices observed in this work. Some concepts can be associated with relatively small number of latent structures in the same low-dimensional slice (i.e., produced by a constant group of latent neurons). Other concepts can be distributed between different slices (Fig. 4), encoded by variable groups of neurons. Generalization of multiple prototypes such as concept-associated clusters into a single concept class can happen in a process of empirical testing as described in Section III.D.

For a brief illustration of this point, let us consider a concrete example. Suppose we have a single positive instance of a concept of interest, for example in the context of the work, an image a digit "2". There can be different latent regions associated with different variations of representations of the digit written by different individuals in the dataset. Further, suppose an early iteration of a classifier produced a positive prediction for a different version of the same digit, located in a different cluster, and slice. It is possible for example, due to relative proximity of the latent positions of the samples in the full latent space.

A positive prediction would cause an empirical test of the identified input sample. If the test confirms similarity of the outcomes (for example, similar amount of useful substance obtained in the trial), the sample can be recognized as another true positive representative of the concept, and the model of its distribution in the latent space can be updated with a possibility of improvement in the accuracy the concept classifier. The process driven by empirical interaction with the environment can continue until confident recognition of the concept is achieved. In this process, generality of concepts emerges as a synthesis of characteristic latent features associated with common patterns in the sensory data and empirical trials, from the lower, "flat" levels of latent structure up, rather than in a hierarchical model, top down.

Another observation pointed to the character of latent encoding and landscape learning as essentially geometrical in nature. While proximity-type classifiers such as kNN were capable of successfully leaning concept classes in an iterative process with minimal empirical samples (Table 4, Section 3.4.3) classifiers of several other types including neural network models such as perceptron [24] and SVM [25] were unable to interpret information encoded in the latent positions and produced strongly overfitted classifiers incapable of successful learning. Investigation of geometrical and topological properties of generative representations could provide further in-sights into learning processes based on generative latent structure.

Interestingly, recent results in experimental neuroscience have demonstrated commonality of informative low dimensional representations with small number of participating neurons in processing of sensory information of different types, including visual, audio, olfactory [26,27] in neural systems of animals and humans. The complexity of neural network models in this work, in the order of 10^4 - 10^5 neural parameters, was similar to that of some primitive biologic organisms, comparable to that of jellyfish, snails and leeches [28,29] indicating that an ability to construct simple conceptual models of sensory environments could be within capacity of such organisms, although it does not prove the ability to communicate that essentially depends on developing more complex interpretation and synchronization behaviors. Effectiveness of informative low-dimensional representations in interpretation of sensory environments thus provides an interesting and exciting connection between learning processes of artificial and biological systems.

Harnessing the informative structure of latent landscape in the initial phase of learning when confident data can be very scarce allows "kick start" an iterative and incremental learning process based on empirical interactions with the environment. Such a process of learning with minimal samples resembles learning of natural bio-logical systems [30] and it can be hoped that further investigation along the directions outlined in this work may produce models that are flexible, adaptive and effective in learning via direct interaction with the sensory environment.

VI. Conclusions

The results presented in this work demonstrated a possibility of consistent interpretation of symbolic information via the process of synchronization of individual latent concept maps, that can provide a basis for communication of essential information about observations between individuals with similar architecture of processing sensory information in unsupervised generative learning with the sensory inputs from the environment.

It can be noted that despite some similarities, the process of symbolic synchronization is not closely related to supervised learning. First, synchronization sets of observations can be very small, down to single instances per concept. Secondly, and more importantly, synchronization is not teaching or training the models to interpret sensory inputs as individual learners are capable of interpreting sensory inputs immediately upon production of conceptual maps based on unsupervised landscape created in unsupervised generative learning; the purpose of synchronization process is rather to allow sharing of symbolic information about observations of the environment by introducing shared communication symbols that can be translated by individual learners in a collective to and from their private conceptual maps. This is possible in a collective of learners with similar generative architecture that produces consistently structured latent representations via a process of orchestrated observation as described and examined in this work.

Experiments in the study demonstrated that relatively simple social behaviors based on simultaneous observation can result in synchronization of symbolic conceptual maps of the environment in a collective, with latent representations emergent in unsupervised generative learning providing a basis for a shared semantic framework associated with principal natural concepts in the sensory data. Such frameworks, as shown by the results in Section VI.C can serve as a natural foundation of the capacity of communication and sharing of semantic information about the observed environment.

It was shown that these abilities can be well within learning ability of artificial and biological learning systems even of limited complexity. For these reasons, in the authors view, further investigation of conceptual representations in generative learning and their role in intelligent functions and behaviors in both artificial and biological systems merits attention of the research community.

Appendix: Similarity-based concept identification

This method can be seen as a type of a novelty detection approach, with the extension of harnessing unsupervised latent structure that is produced in unsupervised generative learning. The method is bases on producing latent nearest neighbor classifier (or another geometry-based classifier) with a small set of samples of interest identified by a relationship of similarity.

In the first iteration, there is a single set of samples S_1 defined by some similarity relationship that can be acquired in empirical trials, for example, "a food source". A binary classifier for the concept associated with the samples can be obtained with 1) the encoded set $P_1 = E(S_1)$ representing in-class training samples; and 2) a subset of the encoded general sample g, E(g) at the maximum distance from the center of P_1, negative out-of-class samples. An observable prototype of the concept can be obtained as an image generated from the position of the latent prototype produced with P_1, for example, mean(P_1).

The process is then repeated iteratively for the next concept, with positive samples of known concepts used as negative ones for the new concept. Despite simplicity of the method, structured latent representations with strong correlation of latent regions to principal native concepts allowed to obtain reasonable accuracy in identification of principal concepts with very small similarity sets, as illustrated in Table 1A.

Samples per concept	2	3	5
Sensitivity / false positives*	0.89 / 0.09	0.96 / 0.04	0.99 / 0.01

*An average of 3 independently trained models

Table 1A. Similarity-based classification, 3 concepts.

In a real environment, the accuracy of the classifier can be improved iteratively by adding concept observations verified in an empirical test to the training set for the concept classifiers, with the potential to further improve the accuracy in identification of principal native concepts.

Acknowledgment

The author is grateful to Prof. Pylyp Prystavka and colleagues at the Department of Information Technology, NAU for reviewing this research, fruitful discussions and helpful advice.

References

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

[2] R. Salakhutdinov, G.E. Hinton. "Deep Boltzmann machines". In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 448–455, 2009.

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

[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, 215–223, 2011.

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

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

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

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

[9] I. Higgins, I., L. Matthey, L., X. Glorot et al. "Early visual concept learning with unsupervised deep learning". *arXiv* 1606.05579, 2016.

[10] L. Gondara. "Medical image denoising using convolutional denoising autoencoders". In *16th IEEE*

International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, pp. 241–246, 2016.
[11] Dolgikh. "Spontaneous concept learning with deep autoencoder". International Journal of Computational Intelligence Systems 12 (1), pp. 1–12, 2018.
[12] J. Shi, J. Xu, Y. Yao, B. Xu. "Concept learning through deep reinforcement learning with memory augmented neural networks". *Neural Networks* 110, pp. 47–54, 2019.
[13] M.Elleuch, M. Kherallah. "Off-line handwritten Arabic text recognition using convolutional DL networks". *International Journal of Computer Information Systems and Industrial Management Applications*, 12, pp. 104–112, 2020.
[14] A. Banino, C. Barry, D. Kumaran. "Vector-based navigation using grid-like representations in artificial agents". *Nature*, 557, pp. 429–433, 2018.
[15] S. Dolgikh. Topology of conceptual representations in unsupervised generative models. In *Proceedings of 26th International Conference Information Society and University Studies (IVUS)*, Kaunas, Lithuania pp. 150–157, 2021.
[16] W. Luo, J. Li, Y. Yang et al. "Convolutional sparse autoencoders for image classification". *IEEE Transactions on Neural Networks and Learning Systems* 29 (7), pp. 3289–3294, 2018.
[17] Keras: Python deep learning library. Online: https://keras.io/, last accessed: 2021/08/2021.
[18] Y. LeCun. "The MNIST database of handwritten digits". Courant Institute, NYU Corinna Cortes, Google Labs, New York Christopher J.C. Burges, Microsoft Research, Redmond, 2008.
[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), pp. 32–40, 1975.
[20] M. Ester, H.-P. Kriegel, J. Sander, X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise". In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231, 1996.
[21] N.S. Altman. "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician* 46 (3), pp.175–185, 1992.
[22] E.H. Rosch. "Natural categories". *Cognitive Psychology*, 4, pp. 328–350, 1973.
[23] S. Dolgikh. "Categorized representations and general learning". In *Proceedings of the 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions (ICSCCW-2019)*, 1095 pp. 93–100, 2019.
[24] Liou, D.-R., Liou, J.-W., Liou, C.-Y. *Learning behaviors of perceptron*. iConcept Press ISBN 978-1-477554-73-9, 2013.
[25] B. Schölkopf, A.J. Smola. *Learning with Kernels*. Cambridge, MIT Press ISBN 0-262-19475-9, 2002.
[26] T. Yoshida, K. Ohki. "Natural images are reliably represented by sparse and variable populations of neurons in visual cortex". *Nature Communications*, 11, pp. 872, 2020.
[27] X. Bao, Gjorgiea, L.K. Shanahan et al. "Grid-like neural representations support olfactory navigation of a two-dimensional odor space". *Neuron* 102 (5), pp. 1066–1075, 2019.
[28] A. Garm, Y. Poussart, L. Parkefelt, P. Ekström, D.-E. Nilsson. "The ring nerve of the box jellyfish Tripedalia cystophora". *Cell and Tissue Research* 329 (1), pp. 147–157, 2007.
[29] G. Roth, U. Dicke. "Evolution of the brain and intelligence". *Trends in Cognitive Science* 9 (5), pp. 250 (2005).
[30] D. Hassabis, D. Kumaran, C. Summerfiel, M. Botvinick. "Neuroscience inspired Artificial Intelligence". *Neuron*, 95 (2), pp. 245–258, 2017.

Author Biographies

Serge Dolgikh holds the degrees of Distinction M.Sc. in theoretical and mathematical physics from the National Nuclear Research University (MEPhI) Moscow, and M.Sc. in telecommunications engineering, Coventry University. Serge has a number of publications in Theoretical Physics, Information Theory research and technology applications and worked on industry projects with leading network technology providers as an engineer and project manager. His current research interests include on the areas of Unsupervised Learning and Self-learning Systems as well as international research and development projects with the Department of Information Technology, National Aviation University, Kyiv Ukraine, universities and business organizations in the European Union and Canada.