Synchronized Conceptual Representations in Generative Learning

Serge Dolgikh [0000-0001-5929-8954]

Dept. of Information Technology,
National Aviation University

Abstract. In this work we examined latent representations of image data with a collective of generative neural network models. A convolutional autoencoder with steep redundancy reduction was used to create low-dimensional latent representations of a dataset of geometrical shapes. Individual models trained in entirely unsupervised process with minimization of generative error were then exposed to a process of synchronization of symbolic concepts associated with characteristic density structures in the latent representations. It was demonstrated that conceptual representations with good decoupling of concepts can be produced with generative models of limited depth; and that a simple process can lead to synchronization of symbolic concepts between learning individuals and a possibility of communication with sharing semantic information about the observed environment. The results demonstrate the potential of latent conceptual frameworks emergent in unsupervised generative learning as a natural platform for abstract conceptualization and communication.

Keywords: Unsupervised Learning, Concept Learning, Representations.

1 Introduction

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

In experimental studies, an array of interesting results was reported such as the “cat experiment”, spontaneous emergence of concept sensitivity on a single neuron level in unsupervised deep learning with images [6]. Disentangled representations were produced and discussed [7] with a deep variational autoencoder and different types of data pointing at the possibility of a general nature of the effect. Concept-associated structure was observed in latent representations of raw Internet traffic and aerial surveillance images [8,9] and other results [10,11].

These results demonstrated that structure that emerges in the latent representations created by models of generative learning in the process of unsupervised self-learning with minimization of generative error can be used as a foundation for learning methods and processes based on compression or distillation of characteristic patterns in the observable environment in an entirely unsupervised process, based on ability to compress and restore the observations into an informative low-dimensional representation. Interestingly, these observations in unsupervised learning of artificial systems were paralleled very recently by several results in biologic sensory networks [12,13] that demonstrated commonality of low-dimensional representations in processing sensory information by mammals, including humans.

Based on these findings, we undertook an investigation of the process of generative learning with models of unsupervised self-learning of limited complexity, to understand its role in forming the conceptual basis for perception of the environment and the ability to share semantic information about observations between learners.

2 Models and Data

We used a convolutional autoencoder model with strong dimensionality reduction to a compact three-dimensional latent representation, with a dataset of greyscale images of geometric shapes as described in this section.

2.1 Convolutional Autoencoder Model

A convolutional autoencoder model had the encoding stage with convolution-pooling layers followed by several layers of dimensionality reduction with the encoding layer of size 3 (i.e., a three-dimensional latent representation). The dimensionality of the latent layer was chosen based on the recent results pointing at low dimensionality of representations of visual sensory data in the human sensory cortex [12] as well as for convenience of evaluation and observation of the resulting latent representation.

The decoding / generative stage was fully symmetrical to the encoder. The diagram of the model is given in Figure 1.

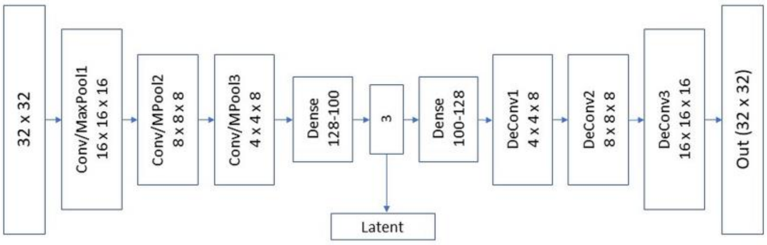

Fig. 1. Convolutional autoencoder with dimensionality reduction.

Overall, the model had 21 layers and 30,916 trainable parameters. The models were implemented in Keras / Tensorflow [14] and trained for minimization of generative error with binary cross-entropy cost.

2.2 Data

The dataset consisted of greyscale images of geometric shapes: circles, triangles and greyscale backgrounds of size (32×32), varying in size and contrast of the fore/background. The training dataset consisted of 200 images of each type: circles with background; triangles with background; and empty backgrounds, each type varying in size and contrast, for 600 in total.

To imitate real-world observations, test images were not prepared beforehand but generated in each test with varying size and contrast.

2.3 Training

A success of generative training was verified by the validation cost and the ability of the learner to regenerate a random subset of images of the types represented in the training dataset. The majority of learning models were successful, though a spread in the generative quality was observed.

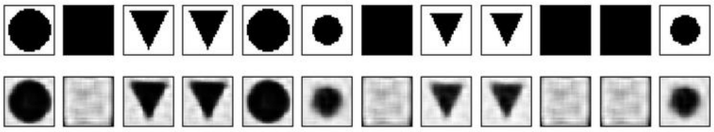

Fig. 2. Generative performance of trained models (top: input; bottom: regeneration).

3 Results

3.1 Generative Learning

For the experiments, models in the better segment of training metrics were selected, based on the value of the final validation cost in unsupervised learning and generative ability.

In black-box experiments with selected individual models it was confirmed that the latent regions containing the representations of different types of images were compact and continuous. With a subset of images *S* of the same type *C*, for example, circles in the observable space, generated image of the mean of the encoded representations of *S* in the latent representation R was an image of the same type:

$$g(S) = G(\bar{S}) \in C \qquad (1)$$

where $G: R \rightarrow O$, generative transformation from the latent representation to the observable space.

3.2 Conceptual Representations

Visualizations of the distribution of images in the latent representation created by models confirmed compact, well-defined and separated distribution of latent regions associated with different types of images, as illustrated in Figure 3: circles (green); triangles (red); background (magenta).

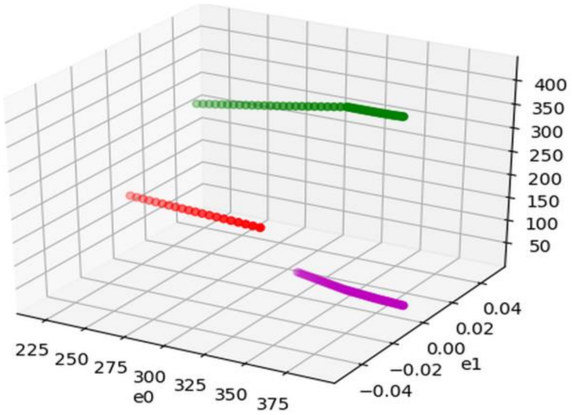

Fig. 3. Latent representation of image classes.

Linear, filament shape of latent concept clusters confirmed independent results of the previous section. Similar patterns of compact and detached, or "decoupled" distributions of characteristic types of observable data were seen with all models that were successful in generative learning pointing at the general nature of this effect [7,8].

Individual models learned to classify observations X presented to them to internal concept classes, or symbolic concepts { T_k } by identifying characteristic cluster K_i in the latent representation that contained the encoded stimulus:

$$r(X) = E(X) \in K_i \rightarrow X \in T_i \tag{2}$$

$E: O \rightarrow R$, the encoding transformation from the observable space to latent representation performed by a learned model.

In this sense, the individual concepts T are simply symbolic labels associated by each learner to a characteristic pattern in the observable data identified by a distinct latent cluster. In this work, the characteristic clusters were identified with a density clustering method, MeanShift [15] was used among numerous options, applied to the latent image of a subset of the observable data. However, a number of alternative strategies are equally possible as will be discussed in a future study.

It can be noted that trained learners were able not only to classify an observation as one of the learned symbolic concepts, but also, importantly, to interpret symbolic information as a representative instance or prototype of the concept. This ability clearly differentiates generative models from standard methods of supervised classification where such a task would not be meaningful.

Indeed, with an observation X its internal concept can be obtained from the association of the latent image of X to one of the learned latent clusters (2); then a representative instance can be associated to T_i by a number of strategies: as the generated image of a mean of known representatives of the concept (1); as the image of a characteristic point in the latent region of the concept, such as center of cluster identified with density clustering. Then, the association $X \rightarrow P_i$, the observable prototype of the symbolic concept associated with X can be established as:

$$P(X) = G(t_i) = G(t(T_i)) \tag{3}$$

with *t(T)* being the representative instance strategy. The strategy can be static, implemented as a dictionary of pairs { (T_i, t_i) }, or dynamic where representative instance is calculated on each invocation or periodically, for example if the set of representative points of T_i is updated. In the experiments we used cluster centers identified by density clustering method as representative instances, a type of static association until clusters are recalculated.

In the experiments with learned models, they were given a sequence of symbols associated with internal concepts and generated individual images, or perceptions, of offered symbols. Evidently, such an ability could potentially provide a basis for sharing perceptions between learners in a collective of learning individuals. However, at this stage of the experiment it was not yet possible, because internal symbolic representations of concepts were specific to each individual and had no semantic meaning for other learners.

3.3 Synchronized Representations

In the following set of experiments, a straightforward process was used to synchronize symbolic representations of individual learners by exposing them to the same, shared sequence of observations.

Individual trained models were shown an image X_i of a known type accompanied with a pre-defined symbol S_i. For example, circles were associated with *"c"*, triangles with *"t"* and empty backgrounds with one or more background symbols, *"u1", "u2",* … The synchronization set thus consisted of pairs (X_i, S_i) of synchronized stimuli and symbols and was consistent, i.e. did not have contradictory associations.

Upon observation, the learners were instructed to update their internal symbols representing the observed concept to the common, or synchronized ones as:

$$T(E(X_i)) = S_i \tag{4}$$

As a result of the synchronization process the learners received a common set of symbolic labels that could still be interpreted individually by each learner. Synchronization of individual symbolic representations then enabled the possibility of communicating information about observations between individual models, whereby an individual that was not presented with an observation, but given only a symbolic label of a known concept was capable of creating a perception or experience of it, as in (3) and (4):

$$I_{ind}(S) = G\big(t(T_i)\big) = P_{ind}(T_i) \tag{5}$$

where I_{ind} an individual perception of the symbolic communication *S* as the individual prototype of the internal concept associated with the observation denoted by *S*. The result is illustrated in Figure 4 with two different learners given the same information as a sequence of synchronized symbols: *t, c, t, c, u, t ….*

It can be noted that though the symbols are synchronized between the learners, each one was creating its own, individual perception of the transmitted symbolic information. For example, the interpretation of the background image *("u")* by learners was quite different, whereas other types of images with a strong population in the training set were interpreted similarly.

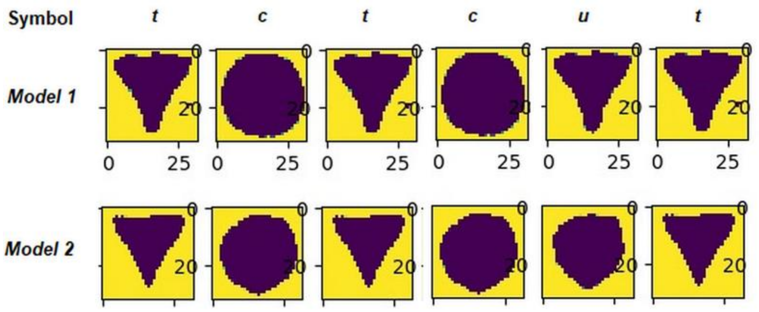

Fig. 4. Synchronized observations and perceptions.

The possibility of consistent interpretation of symbolic information via process of synchronization of individual latent concept frameworks can thus provide a basis for communication between individuals that use similar generative models in unsupervised observation and learning from the environment.

4 Discussion

The experiments in this study demonstrated and confirmed that decoupled representations can emerge in generative models of limited size and complexity, providing arguments for a general nature of the effect. The complexity of studied generative models was well within the range of primitive biologic organisms, comparable to that of jellyfish, snails and leeches, in the range of 10,000 – 30,000 neurons [16,17], providing independent evidence for the hypothesis that conceptual behaviors can be more common in biologic systems.

The capacity to associate behaviors with general concepts as opposed to each specific instance of observation would offer a clear and strong evolutionary advantage in massive reduction of required memory and complexity of processing and for that reason is likely to be selected in an evolutionary process.

Another essential finding of the study is that relatively simple social behaviors can result in synchronization of symbolic conceptual models of the environment in a collective of learning individuals with similar learning architectures, with latent representations emergent in unsupervised generative learning providing a basis for development of communication and sharing of semantic information about the observed environment.

There is no need to emphasize the evolutionary advantage of the ability to communicate information about observations between individuals in a collective. Immediate parallels can be drawn between many forms of social behavior and synchronization of individual latent conceptual frameworks, though the complexity and diversity of these processes and behaviors cannot be underestimated and would certainly require further study.

References

1. Hinton, G., Osindero, S., Teh Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006).
2. Fischer, A., Igel, C.: Training restricted Boltzmann machines: an introduction. Pattern Recognition 47, 25–39 (2014).
3. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning 2(1), 1–127 (2009).
4. Coates, A., Lee, H., Ng, A.Y.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of 14th International Conference on Artificial Intelligence and Statistics 15, 215–223 (2011).
5. Ranzato, M.A., Boureau Y.-L., Chopra, S., LeCun, Y.: A unified energy-based framework for unsupervised learning. In: Proceedings of 11^{th} International Conference on Artificial Intelligence and Statistics 2, 371–379 (2007).
6. Le, Q.V., Ransato, M. A., Monga, R. et al. Building high level features using large scale unsupervised learning. arXiv 1112.6209 (2012).
7. Higgins, I., Matthey, L., Glorot, X., Pal, A. et al.: Early visual concept learning with unsupervised deep learning. arXiv 1606.05579 (2016).
8. Dolgikh, S.: Categorized representations and general learning. In: Proceedings of 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions 1095, 93–100 (2019).
9. Prystavka, P., Cholyshkina, O., Dolgikh, S., Karpenko, D.: Automated object recognition system based on convolutional autoencoder. In: 10th International Conference on Advanced Computer Information Technologies (ACIT-2020), Deggendorf, Germany, 830-833 (2020).
10. Shi, J., Xu, J., Yao, Y., and Xu, B.: Concept learning through deep reinforcement learning with memory augmented neural networks. Neural Networks 110, 47–54 (2019).
11. Rodriguez, R.C., Alaniz, S., and Akata, Z.: Modeling conceptual understanding in image reference games. In: Advances in Neural Information Processing Systems (Vancouver), 13155–13165 (2019).
12. Yoshida, T., Ohki, K.: Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. Nature Communications 11, 872 (2020).
13. Bao, X., Gjorgiea, E., Shanahan, L.K. et al.: Grid-like neural representations support olfactory navigation of a two-dimensional odor space. Neuron 102(5), 1066–1075 (2019).
14. Keras: Python deep learning library. https://keras.io/, last accessed: 2020/11/21.
15. Fukunaga, K., Hostetler, L.D.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory 21(1), 32–40 (1975).
16. Garm, A., Poussart, Y., Parkefelt, L., Ekström, P., Nilsson, D-E.: The ring nerve of the box jellyfish Tripedalia cystophora. Cell and Tissue Research 329 (1), 147–157 (2007).
17. Roth G, Dicke U.: Evolution of the brain and intelligence. Trends in Cognitive Science 9 (5), 250 (2005).