Characteristics of Categorized Latent Representations in Unsupervised Generative Learning

S Dolgikh

Department of Network Engineering, Solana Networks, 305 Moodie Dr., Ottawa, Canada

E-mail: serged.7@gmail.com

Abstract. In this work the effect of spontaneous categorization in the latent representations of unsupervised generative neural network models was investigated and verified in experiments with real world aerial image data. Distributions of images of several terrain classes were compared in the input data space and in the latent representations created by unsupervised autoencoder neural network models in the process of unsupervised training with minimization of generative error. The results demonstrated significantly improved correlation of density structures in the latent representations of generative models with the concept classes than in the unprocessed data, leading to the conclusion that unsupervised training with minimization of generative error combined with the constraint of strong redundancy reduction can lead to emergence of structured representations correlated with concepts with significant representation in the input data. The observed effect can be used for effective learning with minimal training data in the environments with severe deficit of labels.

1. Introduction

The study of unsupervised representations with the purpose to identify and extract informative parameters in general data has a long history. Unsupervised hierarchical representations created with models like Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBN) different types of autoencoder models [1-3] proved to be effective in informative feature extraction and improving the accuracy of subsequent supervised training [4]. The deep relationship between training of intelligent models and the statistical principles such as minimization of free energy was studied in [5,6] leading to understanding that methods commonly used in training of artificial learning systems such as gradient descent in deep neural networks generally produce configurations compatible with the principles of minimization of free energy and variational Bayesian inference.

Results pointing to spontaneous high-level concept sensitivity in unsupervised generative neural network models were obtained in a number of works. Google Lab team [7] observed an interesting effect of spontaneous formation of concept-sensitive neurons, activated by images in higher-level categories with a large, deep sparse autoencoder model trained in entirely unsupervised mode without any exposure to ground truth with very large arrays of images obtained from YouTube videos.

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 [8]. 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 [9], 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 [10] suggest that certain neural network models 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 the information by characteristics of similarity in the representations, thereby 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 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. The aim of this work was to validate the results on unsupervised categorization obtained in earlier studies [8] by answering the essential questions:

1. Do the unsupervised representations obtained with generative unsupervised training have better concept-clustering structure compared to the original data?
2. Is the association between most represented concepts in the original data and the density structure in the representation stable and significant, with respect to change of the model and data?

To address these questions, a number of experiments were designed and executed with a dataset of genuine images, as described in the following sections.

2. Materials and Methods

Among different types of generative models, neural networks have a high potential due to their versatility and universal approximation power that makes them suitable for data of virtually unlimited types and complexity [11] as was demonstrated in a number of results including cited above.

The models used in this work produced two stages of latent 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 dimension (64, 64). The resulting representation was used as input to the second stage autoencoder with physical dimensionality reduction to three dimensions, based on principal component analysis of the first stage representation. For a detailed description of the model and the dataset used in the study refer to [12].

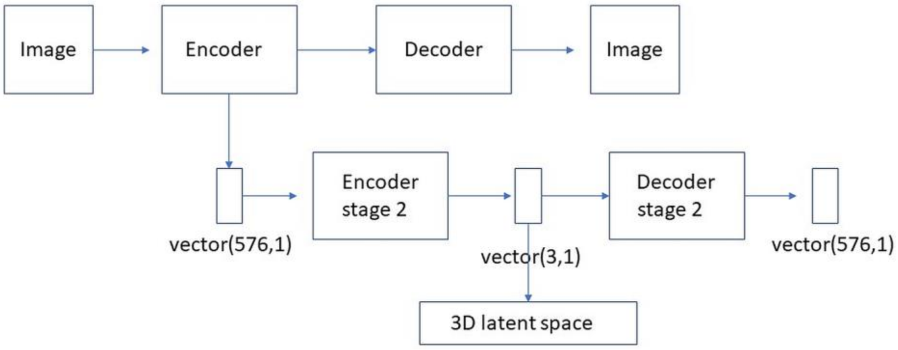

Figure 1. Two stage convolutional autoencoder model

The dataset consisted of real, unprocessed raw aerial images manually labeled with terrain classes. In the experiments, six out of ten classes in the dataset were used in the experiments in unsupervised learning and self-learning; however, the rest of data was still used as uncategorized background to verify the resolution of concept classifiers. The labeled classes used in self-learning were the following: fields; forest; water; roads; large construction structures and vehicles (classes 1 – 6, Table 2).

In the process of unsupervised training the models have achieved significant improvement in all metrics such as: cost function, cross-categorical accuracy and a number of post-training metrics of generative quality (see Section 3.3 for details) indicating that learning models have indeed learned and retained some essential information about the input distributions despite strong compression in the latent representation.

An important aspect in investigating data distributions in the input data and latent representation was identification of data density structure. To this end, density clusters were calculated with MeanShift density clustering method [13] that does not depend on labeled data and can be used in an entirely unsupervised mode; a comparison of the resulting cluster structure was then performed between the input distribution and the latent representation.

To address the objectives of the study, distribution parameters were compared for distributions of concept data samples in the input space and in the latent representation of pre-trained unsupervised models.

The association of higher-level concepts and the obtained density structure was tested by applying landscape learning approach introduced in [8], with the selected concepts in the input data and representation. Given that the method uses extremely small positive sample of the concept, the results of the learning experiments show how closely the density structure is related to the concepts of interest.

In conclusion, to demonstrate that non-trivial learning is taking place with the generative models used in this work, a random dataset was used to compare the outcomes of training and learning between the genuine and randomly generated data.

3. Results

3.1. Unsupervised Categorization

In this section we compare the characteristics of concept distributions in the original input data space with the latent representation created by trained generative models. Some measurements in this section were made with unsupervised methods that required no labeled data. The measured properties were:

1. Structure, the total number of clusters identified with density clustering;
2. Concentration, the fraction of the sample that is found in large clusters with the size over 2% of the dataset.
3. Characteristic size and density of the concept clusters; relative to the overall dimension of the dataset and average density, respectively.
4. Separation, a measure of overlapping of concept distributions calculated as the number of concepts with representation over 30% of the concept sample in the same unsupervised density cluster. The results are presented in Table 1.

Table 1. Distribution parameters in the latent representation

Sample	Dimensionality	Structure	Concentration	Average concept size / density	Separation
Input dataset	576	210	57%	0.44 / 65	3-6
Representation	3	46	73%	0.18 / 340*	1-2*

* Compact concept clusters

In the latent representation, two essentially different types of concept distributions were observed. One was compact and dense, as illustrated in Figure 2, referred to as a "compact cluster". A different type of concept distributions was observed for classes with smaller relative area in the image, such as classes 5 - 7 (construction structures, vehicles, etc.). It is believed that such a difference can be caused by the structure of the model for example, size and depth. These questions will be explored in more detail in another work.

Another interesting conclusion that can be drawn from the analysis of compact concept distributions is the connected and smooth shape of the concept manifolds in the latent representation space. It is a clear confirmation of the manifold assumption [14] that is used widely in unsupervised and semi-supervised machine learning.

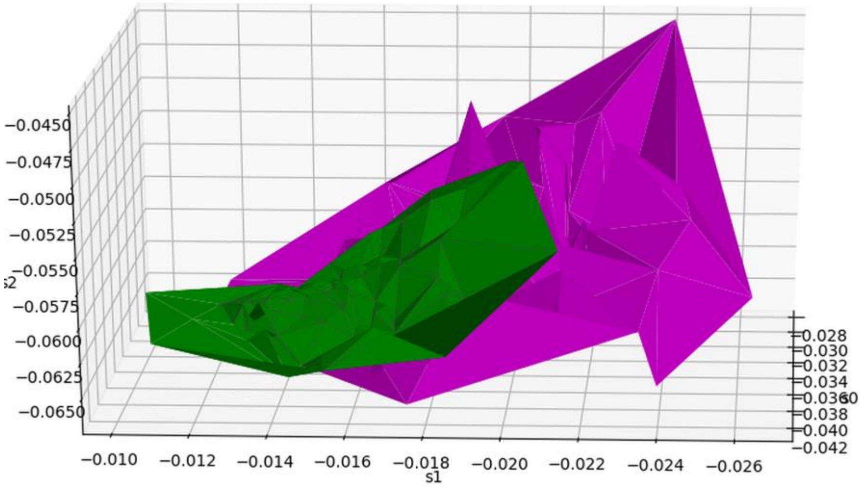

Figure 2. Compact concept distribution in the latent representation

Overall, as can be concluded from the results of this section, latent representations of generative models in the experiments demonstrated more structured and concentrated character than the unprocessed samples in the input data space.

3.2. Minimal Sample Learning

In this section, the method of signal, or minimal sample learning with very small sets of labeled data [8] was applied, comparatively, in the original data and unsupervised representations of generative neural network models.

The method is based on generating samples for training of concept classifiers from the density distribution in the unsupervised latent representation following unsupervised training and for that reason, comparing classification outcomes in these cases allows to estimate categorization capacity of the models, that is, how effectively they were able to associate similar data to compact regions in the latent representation space for more effective learning, and how closely these unsupervised structures reflected distributions of known higher-level concepts.

The method of minimal sample learning is based on a direct association between training of concept classifiers and the unsupervised density structure or landscape, of native density clusters. For this reason, the success or otherwise, of the classifier to learn concepts with a minimal truth sample, that can be as small as a single positive sample of the concept being learned, allows to draw a conclusion on how closely the unsupervised density structure obtained in unsupervised generative learning was correlated with the higher-level concepts in the input data.

In the experiments in this section, classification results of concept classifiers trained with minimal concept samples were measured in 100 tests with randomly selected 100 samples of in- and out-of-concept classes, 20,000 predictions in total. Signal accuracy of classifiers trained with a minimal sample was measured as recall and false positive rate representing errors of two types, and the combined accuracy measure taking into account errors of both types (F1-score).

The null hypothesis in this experiment would be represented by one of the following outcomes:

1) A failure of the representation classifier to learn, i.e. a strongly biased prediction to acceptance, or rejection;
2) The accuracy of the representation classifier on the level of a random prediction i.e. for a binary classifier, (½, ½) or F1-score of 0.5.

whereas successful learning of representation classifiers for different concepts would support the hypothesis of a correlation between unsupervised landscape in the latent representation that emerges in unsupervised generative training, and higher-level concepts in the input data. The results of these experiments representing the accuracy of concept classifiers trained with 1-3 positive samples of classes in the original data space (middle column) and unsupervised representation (left column) are given in Table 2.

Table 2. Self-learning accuracy in the input vs latent space

Class	Description	Signal Accuracy, Representation (mean)	Signal Accuracy, Input (mean)	Signal Accuracy, Representation (best)
Class 1	fields	0.70 / 0.33	0.95 / 0.76	0.77 / 0.34
Class 2	forest	0.87 / 0.38	1.00 / 0.75	0.89 / 0.32
Class 3	water	0.93 / 0.24	1.00 / 0.75	0.95 / 0.25
Class 4	roads	0.52 / 0.42	1.00 / 0.76	0.51 / 0.38
Class 5	construction	0.80 / 0.42	1.00 / 0.76	0.86 / 0.39
Class 6	vehicles	0.65 / 0.36	1.00 / 0.78	0.71 / 0.38

As can be seen from these results, while classifiers in the representation space for all concepts were able to achieve learning accuracy better, and in most cases significantly better than random, those in the input space were not able to converge to a meaningful resolution and remained nearly 100% biased for acceptance and in some recorded cases, for rejection. In over 100 tests across most concepts, the null hypothesis has not been observed, and classifiers trained with unsupervised density structure in the latent representation obtained in unsupervised generative learning were able to predict the concept successfully, with better than random accuracy.

The results in this section indicate that the density structure in the latent representations that emerges in unsupervised generative learning appears to be essential for successful learning of the concept with minimal data that can only be the case if the emergent density structure is correlated with the common concepts in the input data. It is worth noting that the accuracy results in the input space were observed in the entire spectrum of the bandwidth parameter of the density clustering method as confirmed by a grid search in the entire meaningful range of the parameter, and therefore cannot be attributed to specific choice of the parameter.

3.3. A Negative Case: Learning with Random Data

In the previous sections generative unsupervised models have shown interesting results in unsupervised concept learning with real-world image data. But a question can be raised, do these results show a genuine effect of unsupervised learning or perhaps, an artifact of the particular selection of the model and training process, for example, models overfitting data in training?

To address this question, a negative case experiment was designed based on the observation that in the latter case one should observe similar results with training of different datasets including ones with random data, while in the former, only genuine data would produce meaningful results. To this end, a random data array with the same size and parameter range as the Stage 1 encoded representation (Figure 1) was generated and used to train the second stage encoder with dimensionality reduction.

The comparative results of training and generative metrics for the genuine and random datasets are shown in Table 3. In addition to commonly used training metrics such as Mean Squared Error (MSE) and cross-categorical accuracy, two metrics of generative ability of trained models were used: correlation coefficient of the input sample and its image generated by the model (in the input data space); and a ratio of the average norm of the generative error by the norm of the input sample. The value of the correlation coefficient that is close to 1 indicates a high degree of correlation, whereas a low ratio of generative error to the input shows that the model has learned to regenerate the input distribution from the latent representation successfully.

Table 3. Training and generative performance, genuine vs. random data

Dataset	Description	Training metrics: MSE, cross-categorical accuracy	Correlation, input / generated output	Norm ratio, generative error to input
Genuine image data	Phase 1 encoding of real images, 576	up to x100 improvement	0.8 – 0.9	0.1 – 0.2
Random data	Randomly generated data array, 576	not changed significantly	0.1 – 0.15	~ 0.5

The analysis of training and generative results for genuine and random datasets in Table 3 shows that while unsupervised training was successful for genuine image data, it was not so with the random dataset. Unlike models trained with genuine image data where a range of generative performance was observed, none of the models trained with the random dataset showed successful ability to regenerate input samples.

In our view, the results of this experiment clearly demonstrated that not every data can be categorized successfully via generative self-learning and the effect of concept-correlated representations observed in the previous sections is likely to be genuine.

4. Discussion

The objective of this study was to address common questions about unsupervised concept learning, such as generality, effectiveness and reproducibility of results. While strong support for correlation between unsupervised structure in the representations and concepts in the observable data was demonstrated in several results [7-10], questions about generality of these results were raised.

By using a model of lower complexity in this work we first attempted to address the problem of generality. Indeed, observation of the effect of unsupervised categorization with clear correlation with higher-level concepts in the real-world image data shows that specific and complex design, while very likely, essential for superior performance, is not the cause of the observed effect.

Secondly, it was demonstrated (Section 3.1) that strong redundancy reduction combined with generative learning leads to emergence of a structured latent representation with a strong correlation to higher-level concepts. All categorization parameters for concepts of the compact type were significantly better than in the input data, despite strong (almost 200-fold) dimensionality reduction. This conclusion agrees with the earlier results on unsupervised categorization and strengthens the argument for a general nature of this effect.

The results of minimal sample learning in Section 3.2 support this conclusion as well. Clearly, if unsupervised structure in compressed latent representations had no significant correlation with higher-level concepts, there would be no reason to expect any improvement in classification performance comparatively to classifiers trained with raw input data. Yet, as the results show, for most concepts, classifiers trained in the latent representation were successful in learning with minimal samples of concept data.

The experimental results presented in Sections 3.1 - 3.3 provide new convincing empirical arguments for the effect of spontaneous categorization in unsupervised representations of models with self-encoding and regeneration. Theoretical approaches to explanation of this effect were outlined in [5-6].

In our view the presented results answer the questions and achieve the objectives set out for this work, demonstrating consistent and clear effect of correlation between unsupervised the structure in the latent representations of generative neural network models that emerges in unsupervised generative self-learning and common higher-level concepts in the input data.

Conclusion

The methods of unsupervised learning are receiving more attention and with increasing number of applications in the problems and environments where application of conventional supervised methods is limited for example, due absence or deficit of labeled data due to significantly reduced requirements for labeled data needed for successful learning. For example, in supervised machine learning the performance of the model in real world very often directly relates to the accuracy and size of labeled training data. Producing such data in some problems is associated with significant challenges as obtaining large amounts of accurate truth labeled samples may be expensive and/or technically challenging.

Unsupervised models bypass this challenge by identifying principal patterns in the data without the need for labeled samples, via constraints imposed on the model in training, such as generative quality and redundancy reduction. This ability can serve as a foundation for more flexible environment-driven learning with close resemblance to learning process of biologic systems and humans [15].

Thus, generality and flexibility allows the models based on the principles of generative unsupervised learning to be applied in a broad range of fields and applications, that deal with large amounts of unlabeled data such as navigation, different types and aspects of flight information, communications, positioning and others providing an effective and intelligent approach and a set of tools in complex data analysis with effective applications in multiple technology domains including public aviation and aerospace.

References

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

[2] Fischer A and Igel C 2014 Training restricted Boltzmann machines: an introduction. Pattern Recognition. 47: p 25-39.

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

[4] Coates A, Lee H and Ng A Y 2011 An analysis of single-layer networks in unsupervised feature learning. In: International Conference on AI and Statistics (AISTATS)

[5] Friston K 2012 A free energy principle for biological systems. Entropy. 14: p 2100-2121.

[6] Tishby N, Pereira F C and Bialek W 2000 The information bottleneck method. Arxiv:physics/0004057

[7] Le Q V, Ransato M A, Monga R, Devin M, Chen K, Corrado J S, Dean J and Ng A Y 2012 Building high-level features using large scale unsupervised learning. Arxiv:1112.6209

[8] Dolgikh S 2019 Categorized representations and general learning. In: 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions – ICSCCW-2019. 1095: p 93-100

[9] Higgins I, Matthey L, Glorot X, Pal A, Uria B, Blundell C, Mohamed S and Lerchner A 2016 Early visual concept learning with unsupervised deep learning. ArXiv:1606.05579

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

[11] Hornik K, Stinchcombe M and White H 1989 Multilayer feedforward neural networks are universal approximators. Neural Networks. 2(5): p 359-366

[12] Prystavka P, Cholyshkina O, Dolgikh S and Karpenko D 2020 Automated object recognition system based on aerial photography. In: 10th International Conference on Advanced Computer Information Technologies – ACIT-2020

[13] Fukunaga K and Hostetler L D 1975 The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions in Information Theory. 21(1): p 32-40

[14] Zhou X and Belkin M 2014 Semi-supervised learning. Academic Press Library in Signal Processing, Elsevier. 1: p 1239-1269.

[15] Hassabis D, Kumaran D, Summerfield C and Botvinick M 2017 Neuroscience-inspired artificial intelligence. Neuron. 95(2): p 245-258