Unsupervised Generative Learning and Native Explanatory Frameworks

Serge Dolgikh
National Aviation University / Lubomyra Huzara Ave.,1
Kyiv Ukraine

Abstract

A framework of native concepts emergent in unsupervised generative training under the constraints of redundancy reduction and generative accuracy was observed and investigated with an unsupervised generative neural network model and real-world image data. Characteristics of concept distributions in the latent representations were measured and possibility of effective learning with the identified density structure demonstrated. We discuss the potential of using frameworks of native information clusters in the effective latent representations of learning models as a natural platform for explanation of learning processes in machine and biological systems based on the relations and criteria of native similarity that form in the process of generative learning under certain constraints. This approach can be general, intuitive and independent of specific model architectures and types of data.

1 Introduction

A number of results reported the effect of spontaneous categorization, that is, higher-level concept correlation in the latent representations of unsupervised generative models emergent in entirely unsupervised training with the constraints of redundancy reduction and minimization of generative error. This effect was observed independently in several studies with different models and data types (Le et al., 2012). In (Higgins et al., 2016; Dolgikh, 2019) the structure of the concept-sensitive regions in the representations was investigated in some detail, whereas (Friston, 2012; Ransato et al., 2007) proposed theoretical arguments in explanation of learning processes in machine and biological systems.

These results suggest that unsupervised categorization, that is, association and grouping of information by certain criteria of similarity established in the process of generative unsupervised learning can have general character and applicable to a wide range of generative models under certain constraints. The question we wanted to raise and investigate here is whether such native information structures, related to the inner patterns of similarity in the data of general type and origin, can be used as an explanatory framework for the learning processes in the unsupervised systems with possible extension and / or application to supervised cases as well.

2 Methods and Data

The model used to produce compacted (i.e. redundancy-reduced) latent representations of real-world image data was a stacked autoencoder neural networks (Bengio, 2009) comprising the convolutional and dimensionality reduction stages. The overall compression factor measured as the ratio of the dimensionality of inputs to that of the latent representation was over 1,000 (from RGB images with resolution 64×64 pixels to 3-dimensional latent representation). The diagram of the model is shown in Fig.1.

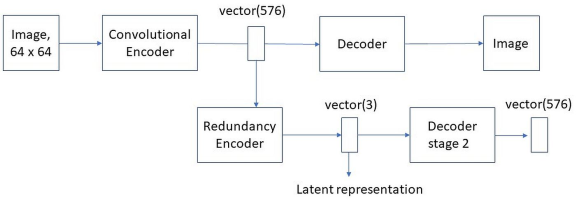

Figure 1: Stacked covolutional autoencoder with redundancy reduction

Convolutional stage was used to acquire higher scale features and consisted of a sequence of convolutional-pooling layers producing an invariant latent representation with dimension 576. The compression, i.e. dimensionality reduction stage was a flat autoencoder of depth up to 10 layers with the resulting latent representation with dimension 3. The dimensionality of the latent representation was selected based on PCA-analysis of the Stage 1 representation that indicated three principal components with variation of over 90%.

2.1 Data

The dataset contained real unprocessed images of terrain obtained in live aerial surveillance. The size of the dataset was approximately 5,000 images, of which approximately 1,100 were manually classified into ten classes, including: built-up area; wooded area; fields; water; roads; construction structures, such as dams and bridges; vehicles and equipment (small and large) and several others. Labeled samples were not used in unsupervised training of the models, but in the analysis of distributions of the higher-level concepts in the latent representations created by trained models.

2.2 Training

The models were trained in unsupervised autoencoder mode to achieve good reproduction of inputs measured by a cost function such as Mean Squared Error (MSE). Several criteria of effectiveness of unsupervised training were used, such as monitoring the cost function and cross-categorical accuracy during the training with both showing significant improvement following the training.

Additionally, generative performance of trained models was measured by calculating correlation coefficient of the input to regenerated output and the mean deviation of the input sample from the generated output to the mean norm of the input sample, in the range of 0.1 - 0.15.

It allowed to conclude that with a strong reduction of dimensionality, latent representations created by the models retained significant information about the original distribution.

2.3 Representation Analsyis

A trained unsupervised model can perform encoding and generative transformations from the observable data space to the latent representation, and from the latent representation to the observable space, respectively as the output of the encoder and generator submodels.

In the latent representation of a trained model, the emergent density structure can be identified by applying a density-based clustering method such as MeanShift and many variations (Fukunaga and Hostetler, 1975). It allows to identify density clusters of the encoded samples in the representation space without external concept labels in completely unsupervised mode. For example, the associated density cluster for a sample X in the input data space can be calculated as:

$$K_{nat}(X) = cluster\_model.predict(X) \quad (1)$$

where $cluster\_model$ is a density-based clustering method trained with a general data sample in the latent representation.

K_{nat} above can be seen as the native concept of the sample identified by the model based on similarity relationships established in unsupervised generative training, and the set of clusters $K = \{K_{nat}\}$ perhaps with sufficient representation above certain minimal threshold defines the native concept framework of the dataset.

Distributions of native and external concepts can be investigated in the latent representations by unsupervised (i.e. without pre-labeled samples) and supervised methods. Parameters such as the number, population, denstity and geometrical characteristics of native clusters can be measured without labeled data and provide insights into the native similarity relationships established by the model to successfully regenerate the original distribution from a compact representation.

Distributions of explicit concepts can be studied by transforming a labeled sample of given concept in the observable space to the latent represenation and measuring the parameters of distribution such as size, density, topological and geometrical characteristics of the concept region in the representation space, as well as associations of the concept data to the native concepts in the matrix form:

$$Y_{ie,jn} = card(C_{ie} \in K_{jn})/card((C_{ie}) \quad (2)$$

where $C_{ie},\ K_{jn}$, explicit concept class and native cluster, respectively. Clearly, the cross-concept population matrix Y indicates the degree of correlation between the explicit and native concepts in the dataset.

A common practice in machine learning is to train a supervised classifier with a labeled sample of explicit concepts transformed to the latent representation of some pre-trained unsupervised model. The explicit class of a sample X in the observable space can then be predicted as:

$$C_{ex}(X) = classifier.predict(encode(X)) \quad (3)$$

C_{ex} and K_{nat} in (1), (3) then represent respectively, the externally known class of an observable sample and its native or implicit concept identified from the density distribution in the latent space created by the model in the process of unsupervised generative learning.

3 Results

The results in this section were obtained with several instances of models trained as explained earlier, in the top 20% range of the learning metrics. Whereas all models succeded in learning by the identified criteria, a distribution of learning performance was observed, not unlike among learning individuals. In most cases the results were consistent among the model instances.

3.1 Native Concept Framework

The methods were applied as described to the dataset and models producing the following general characteristics of the unsupervised data clusters (native concept framework):

- Number of unsupervised density clusters: 102
- Number of significant clusters (with population of 2% and above of the test sample): 18
- Concentration (the fraction of the test sample in significant clusters): 0.744
- Relative size of cluster, range: 0.031 - 0.055[1]
- Relative density of cluster, range: 380 - 1400[2]

The measurements showed that the most populated clusters were indeed compact and dense, with a well defined region in the latent space as illustrated in Fig.2.

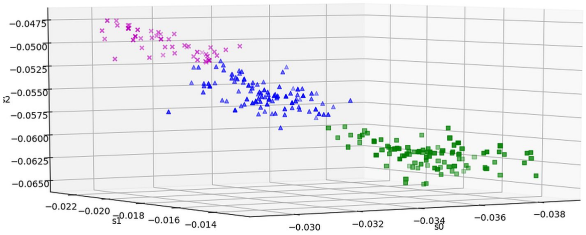

Figure 2: Native clusters in the latent representation of a generative model

Based on identified density distributions in the latent representation, a binary native concept classifier can be created by training a common classifier such as Nearest Neighbour (Altman, 1992) and training it with the dataset of positive in-cluster samples and negative samples in different clusters (Dolgikh, 2019). A trained the classifier can predict the native concept of a sample in the observable space as in (3), except for training of native classifiers no data labeled with external concepts is needed. In the tests, native classifiers for most clusters showed good accuracy in both selectivity and specificity, with F1-score above 95% for all larger clusters.

3.2 Native vs External Concepts

With a test set of samples labeled with external higher-level concepts as described in Section 2.1, it is possible to measure the external-to-native and vice versa concept cross-population matrix $Y_{e,i}$ as defined in (2). Whereas a detailed report of these results would be of considerable length and will be attempted elsewhere, it is sufficient to note that for several external concepts with a strong representation in the dataset, the association between the native clusters and higher-level concepts was clear as illustrated in Table 1.

Ext. concept	Clusters	Max.Population
woods	1, 11	0.38
fields	0, 1, 16	0.67
water	1, 2, 6	0.51
roads	0, 1	0.43
vehicles	0, 1, 2, 6	0.27

Table 1: External vs native concepts, cross-population.

From the example above, an association between concepts ”woods”, ”fields” and ”water’ and native clusters ”0” (field) and 1 (woods-water) can be established.

An interesting and quite intriguing, were it not for the external vs native concept cross-correlation results above, observation was made with native classifiers discussed in the previoius section. When applied to their associated external concepts, these classifiers trained without any (literally) labeled data and based entirely on the density structure developed in unsupervised training, showed better than random, and in some cases, good classification results in both sensitivity and specificity. For example, native classifiers achieved F1-score of 0.75

[1] Relative to the size of the test sample, larger clusters

[2] Relative to the uniform density of the test sample, larger clusters

- 0.77 and specificity above 80% for background-type concepts "fields", "woods" and "water", based only on the density structure in the latent representation and without any labeled samples of the concepts!

These results were obtained with randomly selected positive and negative samples of the concept in multiple independent test runs, excluding statistical fluctuations due to selection of test samples and in our view, provide a strong argument in favour of external-to-native concept correlation. They may also offer a direction toward explainability of learning in deep neural network models via association of learned concepts and native information structure that emerges in the representations of generative models during training.

3.3 Explanatory Framework

As demonstrated in the previous sections, native concept frameworks that emerge in unsupervised generative training, can be used as a basis for explanation of the learning process in unsupervised generative models. Unlike approaches that attempt to establish rules defining the association between the input sample and target outcome (Goebel et al., 2018), native concept frameworks are based on the relations of similarity, or proximity in the latent coordinates that are developed by the learning model during training and emerge from common patterns of similarity in the observable data. Arguments can be made (Dolgikh, 2020) that such information structures can be of general character, that is, not specific to a particular data or model design, and applicable to unsupervised, and supervised cases under certain constraints imposed by the learning process.

As was shown, such an explanatory framework can include a detailed description of the information structure in the emergent latent representations such as the set of identified density clusters and native classifiers capable of associating input samples with a native concept. In more complex models, the methods and instruments of analysis of effective latent representations would need to be developed to advance in this direction.

In the next step, a subset of native concepts can be probed and tagged with external concept labels via a sequence of empirical trials with labeled samples. It can happen via a common training process with massive labeled datasets, but an alternative, environment driven process with empirical trials based on the experience as and when it becomes available is equally possible, resulting in assocation of external concept labels or tags with the native clusters. Such a process, driven by the interaction with the environment, gradual and iterative is more reminiscent of learning of biological systems (D.Hassabis et al., 2017).

The resulting explanation can be represented as a set of similarity, or proximity decisions for an input sample X based on a framework of labeled native information nodes, represented by native clusters labeled with associated external concepts:

$$X \rightarrow \{(K_1,\, p_1), (K_2,\, p_2), \ldots\} \tag{4}$$

Unlike rule-based approaches, an explanatory framework of this type would be more geometrical in nature, based not on hard rules but rather on more flexible proximity relationships within the effective internal model of the environment constructed in the process of generative learning.

4 Discussion

Frameworks of native concepts that form in generative learning discussed in this work can serve as a platform for explaining learning processes in artificial learning systems though not necessarily from a strict rule-to-outcome viewpoint. Such models can be interrogated about what was learned, how it was learned and why a certain decision was taken on any specific input. Let's attempt to illustrate with an imagined dialog based on presented results:

1. What did you learn from this data?

- I learned that the observable distribution can be modeled with F effective parameters (informative features) producing K characteristic patterns (native clusters) with the following characteristics ...

2. How did you learn it?

- By creating a latent representation of the observable distribution under the constraints of redundancy reduction and generative accuracy, that allowed to model the observable data with the distribution in the latent representation. It means that the combination of the latent coordinates and the latent distribution was effective in retaining the essential information content of the observable data under the constraints imposed in training.

3. Why did you classify this sample X as class C?

- Because in the effective latent coordinates it looked similar to one (or more) of the known concepts, as shown by its relative position with respect to identified native concepts (clusters): ...

4. Can you show a sample of the learned native concept in the observable space?

- Yes. A representative sample of a native concept can be obtained by forward-propagating points or regions in the latent representation through the generative submodel.

5. Can you identify the formal rules that associate X with C?

- Not sure. The answer depends on the definition of an invariant representation of the input that is not dependent on the specific method or format of observation (such as in this case, color, resolution, size, format and so on of the images in the dataset). The problem of developing formal rules that explain learning of machine systems is tied to a definition of representations that are invariant to the specifics of the observation process, a task that is far from trivial in its own right. Generative models were in many cases successful in bypassing this obstacle by creating effective latent representations that incorporate the essential informative features in the observable data.

References

N.S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Y. Bengio. 2009. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127.

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

S. Dolgikh. 2019. Categorized representations and general learning. In *Proceedings of the 10th International Conference on Theory and Application on Soft Computing*, pages 93–100.

S. Dolgikh. 2020. Why good generative models categorise. *International Journal of Modern Education and Computer Science*. In press.

K. Friston. 2012. A free energy principle for biological systems. *Entropy*, 14(1):2100–2121.

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

R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, and S. Stumpf. 2018. Explainable AI: the new 42? In *Cross-domain Conference for Machine Learning and Knowledge Extraction CD-MAKE 2018*, pages 295–303.

I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner. 2016. Early visual concept learning with unsupervised deep learning. *Computing Research Repository*, arXiv:1606.05579.

Q.V. Le, M.A. Ransato, R.Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, and A.Y. Ng. 2012. Building high-level features using large scale unsupervised learning. *Computing Research Repository*, arXiv:1112.6209.

M.A. Ransato, Y.L. Boureau, S. Chopra, and Y. LeCun. 2007. A unified energy-based framework for unsupervised learning. In *Proceedings of the Conference on AI and Statistics (AI-Stats)*.