INFORMATION TECHNOLOGY

UDC: 004.896 (045)
DOI 10.18372/2306-1472.80.14274

Serge Dolgikh

SPONTANEOUS CATEGORIZATION AND SELF-LEARNING WITH DEEP AUTOENCODER MODELS

National Aviation University, 1, Lubomyr Husar ave, Kyiv, 03058, Ukraine
E-mail: serged.7@gmail.com

Abstract

In this studythe author investigates information processing in deep autoencoder models. It is demonstrated that unsupervised training of autoencoders of certain class can result in emergence of compact and structured internal representations of the input data space that can be correlated with higher level categories. The authors propose and demonstrate the practical possibility to detect and measure this emergent information structure by applying unsupervised density clustering in the activation space of the focal hidden layer of the model. Based on the findings of the studya new approach to training neural network models is proposed that is based on the emergent in unsupervised training information landscape, that is iterative, driven by the environment, requires minimal supervision and with interesting similarities to learning of biologic systems. In conclusion, a discussion of theoretical foundations of spontaneous categorization in self-learning systemsis provided.

Keywords: artificial intelligence; machine learning; neural networks; unsupervised learning

1. Introduction

Over the recent years, the domain of biology-motivated machine learning has seen very fast, one can even say exploding growth. A number of significant advances have been made, bringing efficiency and confidence in learning of machine systems and specifically, deep neural networks, in certain areas of application such as image recognition, time series analysis, games and others to that of human abilities or even surpassing them.

2. Related Work

In a breakthrough in self-learning with training method based entirely on self-play reinforced learning with no human supervision, DeepMind team developed Zero Go machine player that achieved superior performance among both machine and human players while learning entirely on its own through self-play with no supervised training (Silver, Shrittwieser *et al*., [1]). Iterative, progressive and self-reinforcing unsupervised learning can prove an important step toward general learning directly from the environment with minimal external supervision.

Interesting results in unsupervised training with deep autoencoder neural networks were reported by Le, Ranzato *et al*., [2]. Training an experimental deep neural network in unsupervised mode with a very large array of images they observed emergence of concept sensitive neurons – those activated by images of certain abstract category such as a human or animal face.

While accuracy of recognition reported in the study was not yet at a confident level, these results open new possibilities in studying spontaneous emergence of concept associated structures in the information landscape of deep neural networks.

An in-depth review of essential up-to-date developments in biology-motivated machine learning with applications of advances and findings in neuroscience to machine intelligence can be found in Hassabis, Kumaran *et al.* [3], notably in application to spontaneous learning and continual learning models , probabilistic and deep generative learning, progressive learning and conceptual representation, while essential concepts, results, promises and challenges in application of deep neural networks in artificial intelligence were investigated and summarized in great scope and detail by Bengio [4].

Applications of clustering techniques novelty

detection are numerous and well known, such as OLINDDA method by Spinosa *et al*. [5] for novelty and concept drift detection in data streams, Fanizzi *et al*. on concept clustering [6], applications of self-organizing neural networks in novelty detection [7], density-based clustering [8], and deep autoencoder models for anomalydetection [9], see Pimentel et al. for a comprehensive review of the field [10].

While impressive progress has been made in adapting AI systems and specifically, neural networks to a wide and growing by day array of tasks and applications often with outstanding success, one cannot help pointing out some areas where advance has been slower. First, the achieved success is often limited to a specific application, skill or problem area, with limited capacity for more general and environment motivated self-learning.

Secondly, the process of training machine intelligence systems with fixed categories and massive amounts of truth data may not always be efficient or practical in a dynamic and fluent information environment, where the emergence of new concepts and/or obsolescence of others would require frequent retraining of the learning system; nor is it reminiscent of learning processes of biologic systems. As pointed out by Hassabis *et al*., *"human cognition is distinguished by its capacity to rapidly learn about new concepts from only a handful of examples*" that is, it tends to be iterative, adaptive to the environment and based on trials and errors with limited ground truth data, while achieving gradually high levels of confidence in recognition of newly learned concepts.

The motivation for this study is to approach both of these challenges from the direction suggested by the earlier studies, that is, by exploring the link between unsupervised training of certain deep neural network models and emergence of concept sensitive structures in their inner layers. Should such a link beestablished, could it be used as a foundation for novel approaches to training of machine intelligence systems that can learn with minimal supervision?

The structure of the paper is as follows: in Section 2 we describe the model, data and methods used in the study. Section 3 contains the results of simulation experiments and evaluation of the emergent encoded structure. In Section 4 we discuss theoretical aspects of unsupervised categorization ability of autoencoder models. Finally, Section 5 contains a discussion of the results, possible applications, and further directions of research.

3. Model and Methods

The model in this study contains several essential components with a deep autoencoder neural network in its core. Autoencoder models were studied extensively in applications in unsupervised learning and were chosen in this study for the following reasons:

1. Being a universal approximator [11], feedforward neural networks can have virtually unlimited versatility and are suitable to most complex data types such as images and video [12], hyper-spectral image streams [13] and other;
2. The effect of spontaneous emergence of higher-level concept sensitive structure in deep neural network and autoencoder models was reported in the earlier studies [2, 14 15];
3. Neural networks are widely present in biologic systems that are also highly successful in self-learning with minimal ground truth data [3, 15, 16];
4. Finally, a comparison of higher-level concept correlation of deep autoencoder model and PCA encoded spaces appeared to indicate somewhat stronger relation for autoencoder model [15].

Based on these arguments we expect that deep autoencoder models would be a good starting point for a study into spontaneous categorization by higher level concepts and learning based on the emergent unsupervised information landscape.

3.1. Model

The model used in the study to produce a transformed representation of input data space is a deep autoencoder neural network of near-symmetrical layout, with significant compression in the central layer as illustrated in Fig. 1. A compression factor, that is, the ratio of the size of the input to the central layer of the model up to 10 was used, see [15] for the complete graph of the model.

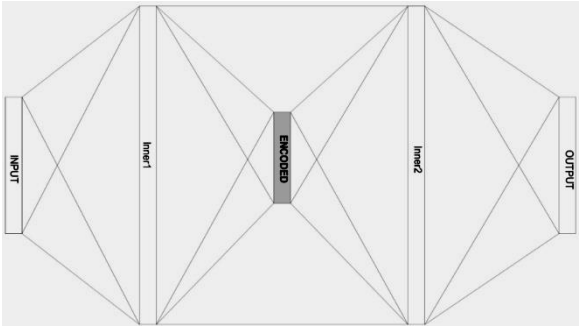

Fig. 1. Model layout

Depending on the size of hidden layers, models in the study had up to 8,000 parameters as described in Table 1.

Table 1

Model Parameters

Layer	Size	Range	Activation	Shape	Cost
Input	F=22	[0 .. 1]		(n, F)	
Inner 1,2	M= 10..50	Any	Leaky Relu	(n,M)	
Encoded	N= 3..10	Any	Leaky Relu	(n,N)	
Out	O = F = 22	[0 .. 1]	Sigmoid	(n, F)	MSE[1]

The models were implemented in Python with Keras [17]. Common software packages such as: sklearn-kit, matplotlib and others were used as well.

3.2. Data

The data in the study represents Internet sessions recorded in two different networks by geographic location and source [18]. In a sense, the task of Internet application identification can be compared to a recording of sound in a busy shopping mall, with the task to classify conversations in by some characteristics of the speaker, such as gender, age, occupation, etc (Table 2).

Table 2

Input Data Parameters

Type	No	Description
General	6	Total duration, total data size (per direction), number of packets (per direction), data protocol
Packet size	8 - 12	min, max, mean, standard deviation, entropy of packet size, per direction
Packet timing	8 - 12	min, max, mean, standard deviation, entropy of packet inter-arrival time, per direction

Each data sample represents an instance of Internet session such as a voice call, web browsing session, instant messaging session, file download, *etc.* and is defined by 22-30 parameters derived from temporal and volume statistics of data packets in the session [19].

Being a live recording in a core Internet network, the data has a wide representation of conversation patterns, with over 4,000 distinct applications represented in the dataset.For this reason, we believe it is well suited to test the validity of the developed approach with data of significant diversity and variation.

3.3. Components

Along with autoencoder model described above, for experiments and measurements we use components that were introduced and described in [15].

1. The autoencoder model is trained in unsupervised mode to match the output to the input X with Mean Squared Error loss function[1].

$$model.train(input=X, output=X, \ldots). \quad (1)$$

A trained model performs "encoding transformation" from the input data space X to its representation in "Encoded" (Fig. 1) layer of the model y as:

$$y = encoder.predict(X), \quad (2)$$

where *encoder* is a sub-model mapping the input to activation of the encoding layer.

2. To classify input samples to categories {C}, a classifier is trained with ground truth labeled set (X, L) in the encoded space of the model:

$$classifier.fit(encoder.predict(X), L). \quad (3)$$

Together, the encoder and classifier can predict the class *C* of an input sample *S* as:

$$C = classifier.predict(encoder.predict(S)). \quad (4)$$

Here we used a geometry-based classifier such as nearest neighbor.

3. In the unsupervised training phase one can apply density-based clustering method that doesn't require fitting with labeled samples, such as MeanShift [20]. It is fitted on a subset of data in the encoded space of the model to learn and visualize its structure as:

$$structurer.fit(encoder.predict(Y),\ldots), \quad (5)$$

where *Y* is the structuring sample, a significant subset of the input dataset.

Note that while unsupervised *structurer* cannot predict the higher-level category of the input sample that is, its class *C,* it can predict its implicit cluster *Cl* as one of the clusters identified in the structuring phase (5) as:

$$Cl = structurer.predict(encoder.predict(S)) \quad (6)$$

[1] Additional optimization terms such as L2 and sparse were used in some experiments as well.

C and *Cl* thus signify the distinction between the externally known higher-level category of the sample and its internal concept ("implicit knowledge") derived in unsupervised training of the model and clustering in its encoded space.

3.4. Training, Visualization and Classification

The models are first trained in an unsupervised autoencoder mode to achieve good reproduction of inputs. Two measures of the quality of reproduction i. e., the average deviation of the output of the model from the input were used:

1) costfunction, MSE, had starting value in the range of 0.25 dropping to 0.001-0.002 after 100 epochs of self-supervised training; and

2) accuracy, measured as the match of *softmax(input), softmax(output)*thus, a measure related to covariance of input and output. Measured in this way, accuracy has increased after 100 epochs from ~ 1% to, on average, 95%. Both cost and accuracy were measured on the validation sample, separate from the one used in training.

The structure in encoded space that emerges as a result of unsupervised training, also referred to as "unsupervised landscape", can be measured and observed by the following methods:

1. By applying an unsupervised clustering method in encoded space and identifying clusters populated by samples of a given application category;
2. By applying multi-dimensional histogram analysis.
3. By measuring the parameters of the distribution of application category samples in encoded space.
4. By plotting and direct observation and measurement of application category samples in encoded space, with common plotting instruments such as [21].

Classification accuracy can be measured with labeled data by obtaining prediction as in (4) that can be compared with ground truth. We use accuracy metrics as commonly defined: *classification accuracy* or *recall* as True Positive samples (class) / Total samples (class); and *false positive rate* as False Positive (class) / Total samples (not in class).

It's worth noting that *a priori*, there's no expectation of correlation between accuracy in unsupervised training vs. classification accuracy with labeled data. Where needed for clarity, they are referred to as "training accuracy" vs "classification accuracy" in the rest of the study.

3.5. Landscape-Based Learning

Based on results pointing to possible correlation of the emergent information structure in the encoded space with higher-level categories an attemptwas made to illustrate the possibility of using this structure in training unsupervised machine systems to learn and recognize new higher-level concepts.

The method is based on developing a set of "concept markers" in encoded space over a series of learning iterations that aim to identify clusters or structures relevant to the concept being learned. Concept markers are built with small number of truth samples in trial and error iterations and artificial or "synthetic" markers derived from structures identified in the clustering phase following unsupervised training. In each learning iteration, the set of concept markers is updated based on real world inputs and classifier is retrained with the updated set of markers iteratively improving category prediction.

4. Results

4.1. Shape and Structure

In all tests we observe that unsupervised training of models results in compact and structured representation of the input data space.

For each application category C, that is a distinct Internet application, the following parameters were measured:

1. Dispersion, or the relative volume of the category sample to general sample:

$$(C) = Vol\ ((cat\_s(C)))/Vol\ (gen\_s),$$

where cat_sC) and gen_s are the category and generic samples, respectively.

2. Resolution, as the ratio of the number of visually identifiable features in the application category sample to the total number of identified clusters:

$$Res(C) = Count(C) / Count(gen\_s).$$

3. Size, the typical size of an individual feature in the category sample relative to the mean size of generic sample.

4. Density, of category features calculated as the ratio of the number of points in the category cluster to its volume:

$$Dn(C) = Count(cat\_s(C)) / Vol(cat\_s(C)).$$

5. Accuracy, defined as the accuracy of classification for in- and out-of-class samples:

$$Acc(C) = (Recall(C), FPR(C)),$$

where Recall, FPR are recall and false positive rate of the category classifier.

The results of measurements of the emergent unsupervised structure for a subset of common Internet applications were summarized and analyzed. Not surprisingly, the measurements show that the categories with the most expressed structure in encoded space, that is, a small number of compact and dense clusters (DNS, NTP and Telnet) produced the highest accuracy of classification. On the opposite end of the categorization spectrum we observed applications with greater variability of content and behavior, such as streaming, BitTorrent, and Web protocol (HTTPS). Technically speaking, the latter should not be considered as a distinct application as it can carry many applications different in content and behavior, so the higher number of associated clusters and sparsity of the category space in this case is hardly surprising.

4.2. Classification

In the previous section it was observed and concluded that unsupervised training of models produced compact and structured encoded space, however there was no indication if or how it is correlated with higher-level concepts. The results in this section support the hypothesis that such correlation indeed exists.

It is possible to monitor training of a neural network model with a callback. A simple callback was implemented to record classification accuracy of the same labeled sample during unsupervised training of a model after each 10-th epoch of training. In this example, classification accuracy across all categories improved by over 5%:

epoch: 0 accuracy: 0.865
epoch: 20 accuracy: 0.908
epoch: 40 accuracy: 0.910
epoch: 60 accuracy: 0.914
epoch: 80 accuracy: 0.916
epoch: 100 accuracy: 0.917.

In the recorded experiments classification accuracy increased as a result of unsupervised training in all cases, with mean of 4.6% and the range 2.8 – 7.7%. In our view this is a strong indication that the emergent structure in the encoded space is correlated with higher-level features in the input data.

If the emergent unsupervised structure had no significant correlation with higher-level concepts, positive correlation between unsupervised training and classification accuracy of concepts would be difficult to explain.

4.3. Visualization in Encoded Space

Categorization can be defined as a characteristic of the encoding transformation whereby samples of same higher-level categories are likely to be transformed to distinct regions in the encoded space of the model. This spontaneous clustering by higher-level concept can be observed with the models directly by visualizing category labeled samples in encoded space.

In Fig. 2, samples of Internet applications, including: DNS requests (green), Escale Newton game (magenta) and MSN messenger (red) are plotted in the encoded space the model with non-categorized data of other categories (bottom plot, 10,000 samples, gray).

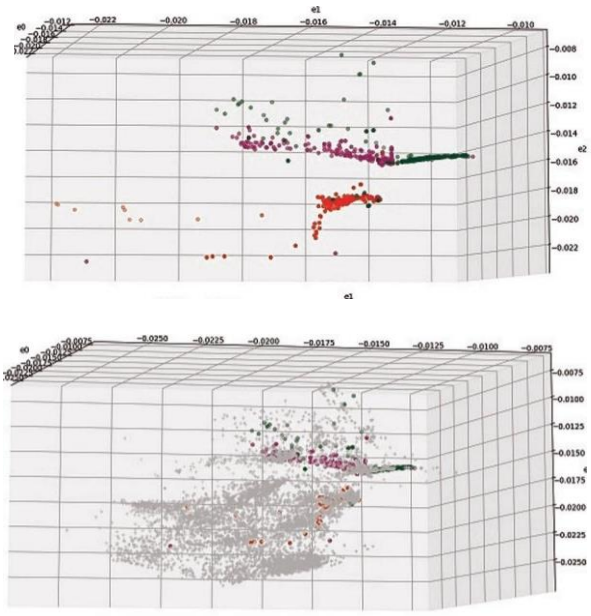

Fig. 2. Categorized sample in encoded space

This visualization demonstrates that application category samples are indeed transformed into distinct regions in encoded space, though categorization parameters such as shape, size, density and others, may vary significantly across applications.

4.4. Landscape-Based Learning

Based on the results presented in the previous sections pointing to association between spontaneous structure in autoencoder models and higher-level concepts in the input data the authors attempted to develop a method that would harness this emergent structure for more efficient learning.

The approach is based on detecting and measuring the unsupervised spontaneous structure in

encoded space and using it, along with small streams of trusted data obtained in trial and error iterations to construct a set of markers in the encoded space that would identify the concept-associated regions well enough for a confident classification. We used a simple form of the method whereby synthetic markers were generated randomly within a small sphere around cluster centers and believe that refining it may further improve the performance.

The learning process involves several stages:

- unsupervised phase: spontaneous structure is detected by structuring method and "synthetic" markers calculated from identified clusters;
- "encounter", that registers the first labeled samples of the new concept allowing to identify clusters in the encoded space associated with the concept and build first iteration of concept markers from identified concept clusters and labeled samples;
- trial and error iterations: the set of concept markers is updated based on the outcomes of trials with small streams of labeled data and classifier retrained on the updated set;
- reinforcement and permanent learning: check and maintain classification performance achieved in the learning phase.

Iterative landscape-based learning was applied to samples of several Internet applications, followed by verification of accuracy of classification as presented in Table 3.

Table 3

Landscape-based Learning, Accuracy

Applica-tion	LM, Start	LM, Final	DNN	kNN
DNS	89.5/ 19.3	90.5 / 6.9	92.0 / 8.3	92.5 / 7.3
NTP	66.8 / 14.4	99.4 / 13.2	100 / 11.5	99.5 / 13.0
Telnet	97.8 / 25.1	97.9 / 12.1	96.8 / 9.2	97.9 / 11.6
XBox	76.6 / 39.2	78.8 / 13.8	83.9 / 7.6	87.8 / 10.2
Messenger	98.5 / 37.8	91.7 / 6.6	88.9 / 8.2	89.9 / 5.0
Email	87.9 / 46.8	85.7 / 17.9	90.4 / 19.4	92.0 / 20.6
Streaming	78.2 / 23.1	84.0 / 14.7	98.7 / 1.8	91.0 / 6.9

How, and why does iterative learning work? Visualizing training samples generated by the model in learning iterations may give an answer to this question empirically.

In Fig. 3, training samples of one application category (network time protocol, NTP), both genuine and generated artificial ones, were visualized in several learning iterations, with the background of the category sample in encoded space of the learning model.

The diagram illustrates how over a period of learning iterations the set of training data points spreads over the category space of the concept being learned coincidentally with progressive improvement in the accuracy of the category classifier as learning process proceeds.

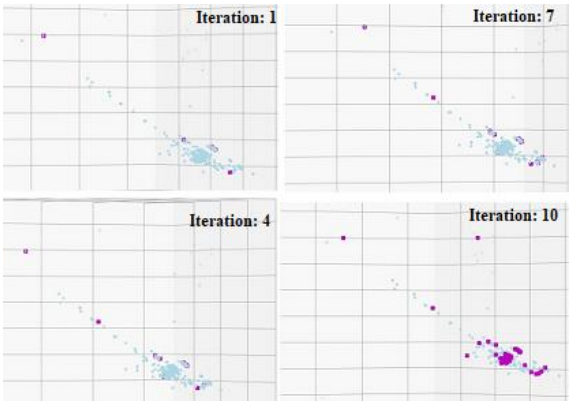

Fig. 3. Training samples in learning iterations

5. Conclusions

The results reported in this section can be summarized as follows:

1. Unsupervised training of deep autoencoder models studied here results in compact and structured representation of the input data space. This conclusion can be reached from shape and structure analysis.
2. Some models can achieve high classification accuracy being trained with very small amounts of truth data pointing to possible correlation between unsupervised spontaneous structure and higher-level categories in the input data.
3. Accuracy in classification is correlated with unsupervised training and improves considerably over the course of training, supporting the argument for correlation between the emergent unsupervised structure and higher-level categories.
4. Visualization analysis directly supports the hypothesis that spontaneous structure emergent in unsupervised training reflects higher-level categories in the input data.
5. A method of landscape-based concept learning based on unsupervised spontaneous structure with

iterative learning process and very light requirement for ground truth data was proposed, with good classification performance.

6. Theoretical Approaches to Categorization

A question can be asked: what theoretical reasons stand behind categorizing ability of models investigated in this study? Let's start with some definitions.

We shall define a learning model M as having a "good generalization", or GM, of certain set of higher-level concepts C = { C_k} if: 1) it is finite; and 2) is constant, that is, does not require refitting; and 3) the mean measure of error in prediction of category in C is below certain maximum margin ε.

We shall define a transformation of the input data space *I* to a certain data space *E* as "categorizing transformation" T_c if:

1) for each category C_k and category space A_k in *I*, $T_c(A_k)$, that is, the encoded representation of A_k in *E*, is a continuous region in the encoded space, or a finite set of such regions; and

2) that encoded representations of category spaces { E_k } in *E* overlap with measure not exceeding certain maximum overlap factor η, i.e.:

$$M(U_{i,j}(E_i \cap E_j)) / M(U_i E_i) < \eta, \qquad (7)$$

where M – the measure of volume in *E*.

If at least one categorizing transformation exists, the encoded representation of input data space *E*can also be referred to as the category manifold, M_c.

It is easy to see that for regular feed-forward neural networks trained with marked category samples, the condition of good generalization is equivalent to the existence of at least one categorizing transformation with overlap factor related to margin of error.

Corollary: A Generalizing Model for data I and category space C exists if and only if there exists a categorizing transformation T_c.

Proof:

If categorizing transformation exists, then by theorem of universal approximation [14], it can be approximated with any precision by a certain neural network N_E. Then the mapping from category manifold M_c to the category space can be approximated by another finite subnetwork N_c with margin of error not exceeding certain factor derived from the overlap factor *F* of T_c. The combined network of N_E and N_c then satisfies both conditions of finiteness and maximum error of a GM.

The reverse statement is straightforward, as transformation of *I* into *C* by a GM itself satisfies the conditions of categorizing transformation[2] so the model itself is a categorizing transformation, with *E*=*C*.

To illustrate this statement, let's consider the case of random data I_R. For any given set of samples S it possible to construct a model that could fit it to given categories with expected precision. However, the next batch of random data would require refitting of the model and possibly, increasing its size and so on. So, the conditions of constancy and finiteness cannot be met in this case and good generalization is not possible.

The result above applies if category space C is known *a priori*, and the existence of a model with good generalization can be seen as an essential property of the input data that can be "packed" into given categories with controlled error and for as long as the nature of data *I* does not change significantly. But there are cases where concepts to which input data can be classified aren't known, or significant volumes of representative ground truth data aren't available to train models as in traditional machine learning approaches.

One example of models that could deal with such cases naturally are autoencoder models. They do not require massive (or in fact, any) supervision in training and were used extensively [22, 23] for pre-processing, dimensionality reduction and feature selection. "Templating" that is, grouping similar samples into category clusters is a successful strategy for deep learning and self-learning [4], But does it explain categorizing ability of autoencoder models observed in this and other related studies?

It would be tempting to try to extend the results above to autoencoder systems as well, for example by defining "natural categorization", similar to definition of categorizing transformation with given externally category space. However, on this path right from the start we encounter certain challenges.

First, it appears that a trivial autoencoding network can successfully reproduce any input, including random, by passing it along to the output, if the size of the hidden layer matches or exceeds the input: $w_{ij} = \delta_{ij}$, where *w* – the weights in the hidden layer, δ – Kronecker's symbol (and biases set to 0).

So, for any non-trivial categorization effect we need to impose the condition of compression in the hidden layer: $H < I$, or even $H << I$, where *H* and *I*

[2] It is assumed that category classes have no overlap

are the sizes of the hidden and input layers, respectively [3].

We can then define "natural" categorizing transformation, with no externally defined category space, as:

$$E, T_E: I \rightarrow E, \text{ such as } M_E = T_E(I) = U M_k \quad (8)$$

where M_E is the implicit category manifold, M_k – a continuous region in E; and dim(E) << dim(I); and T_E is continuous in I; and the overlap condition for M_k (7) is met.

With this definition, it's quite straightforward to outline the proof that the data that is naturally categorizable (that is, at least one categorizing transformation exists) can be encoded to a lower dimension with good reproduction (that is, there exists at least one encoding model with good reproduction accuracy).

Really, from definition of categorizing transformation, both T: I → E, and T^{-1}: E → I must be continuous, and therefore can be approximated with a neural network model, say N_E and N_I. By combining N_E and N_I output to input we can obtain an autoencoding model N_{EI}: I → I, with limited error and good accuracy of reproduction.

However, the reverse case, that is, sufficiency of the existence of a good encoding network with compression for theexistence of a natural categorizing transformation appears to be more challenging, as several different cases need to be carefully considered and will be addressed in the future studies.

7. Discussion

7.1. Advantages of Landscape-based Learning

The proposed method offers a number of essential advantages over common machine learning methods, particularly in early learning of novel, previously not known concepts in areas where significant prior knowledge is not available:

(1) It is environment driven and iterative: the learning process can be triggered by an encounter with a single instance of concept, and proceeds in an iterative manner as and when training data become available, without dependence on massive amounts of ground truth data upfront.

(2) It is effective from the start: providing better than random classification accuracy from the start of the learning process and over the entire interim learning phase.

(3) It is lightweight: the method requires minimal amounts of ground truth data and model resources and in this way, is significantly more efficient than common methods. For example, with ten learning iterations, a classifier of a landscape-based model has only about 100 of three-dimensional data points, as opposed to thousands of weights and biases if a dedicated to concept neural network classifier was used.

(4) It is flexible: learned categories can be easily added and / or "forgotten" without any negative impact on other learned concepts, nor does it require a retraining of the encoding model as with most common methods.

7.2. Applications

Iterative environment driven learning of new concepts based on the unsupervised landscape can provide insights to building machine intelligence models that could learn directly from the environment and acquire new higher-level concepts without massive supervised training. Such approaches would be useful as an alternative to commonly used methods in cases and areas where the concept being learned is new and large amounts of ground truth data are not yet available.

Passive self-learning models similar to those studied here can be used in a range of research and technology applications to monitor and identify categories in data that were not known or detected previously such area surveillance image categories, network traffic patterns, operational and security events and situations, demographic and social data classes, and others.

7.3. Further Work

The results of the study need to be confirmed with data of different types and nature, that is the intent of future studies.

Acknowledgments

I'd like to thank colleagues at Solana Network Engineering and Kyiv National Aviation University for reviewing this article and valuable discussions and suggestions.

[3] Or sparsity condition [5, 26]; in complex models such as in image recognition, we mean an effective subnetwork rather than a single physical layer and an effective factor of compression imposed by the sparsity condition.

References

[1] Silver D. Shrittwieser J. Simonyan K. Antonoglou I. Huang A. et al (2017), Mastering the game of Go without human knowledge, *Nature*, volume 550, 354–359.

[2] Le Q.V. Ranzato M.A. Monga R. Devin M. Chen K. et al. (2012), Building high-level features using large scale unsupervised learning, *arxXiv:1112.6209*.

[3] Hassabis D. Kumaran D. Summerfield C. Botvinick M. (2017), Neuroscience inspired Artificial Intelligence, *Neuron* 95, 245-258.

[4] Bengio Y. (2009), Learning deep architectures for AI, *Foundations and Trends in Machine Learning* Vol. 2 no. 1, 1–127.

[5] Spinosa E., de Carvalho A.C.P.L.F.., Gama J. (2007), OLINDDA: a cluster-based approach for detecting novelty and concept drift in data streams, 2007 ACM Symposium on Applied Computing (SAC), Seoul.

[6] Fanizzi N., d'Amato C., Esposito F. (2008) Conceptual clustering and its application to concept drift and novelty detection. The Semantic Web: Research and Applications, ESWC

[7] Albertini M.K., de Mello R.F. (2007), A self-organizing neural network approach to novelty detection, ACM symposium on Applied computing, Seoul, 462-466.

[8] Cassisi C., Ferro A., Guigno R., Pigola G., Pulvirenti A. (2013), Enhancing density-based clustering: parameter reduction and outlier detection, *Information Systems*, Vol. 38 no.3, 317-330.

[9] Japkowicz N., Myers C., Gluck M. (1995), A novelty detection approach to classification, 14th international joint conference on Artificial intelligence Montreal, Vol.1, 518-523.

[10] Pimentel M, Clifton D, Clifton L, TarassenkoL(2014). A review of novelty detection, *Signal Processing*, Vol. 99, 215-249.

[11] Hornik K, Stinchcombe M., White H. (1989), Multilayer feedforward neural networks are universal approximators, *Neural Networks*, Vol.2 no 5, 359-366.

[12] He K., Zhang X., Ren S., Sun J. (2015), Deep residual learning for image recognition, *arXiv:1512.03385*.

[13] H. Peterson, D. Gustaffson, D. Bergstrom (2016), Hyperspectral image analysis using deep learning, 6th Int. Conf. Image Proc. Theory, Tools and Appl. (IPTA) Oulu U.S.A.

[14] A. Banino, C. Barry, D. Kumaran (2018), Vector-based navigation using grid-like representations in artificial agents, *Nature*vol. 557 pp. 429–433.

[15] S. Dolgikh (2018), Spontaneous concept learning with deep autoencoder, *Int. J. Comp. Intel. Syst*. vol. 12 no. 1 pp. 1-12.

[16] Getting P. A. (1989), Emerging principles governing the operation of neural networks, *Annual Review of Neuroscience*, Vol.12, 185-204.

[17] Keras: Python deep learning library. Available at: https://keras.io/

[18] Waikato Internet Traffic Storage (WITS) passive datasets.Available at:https://wand.net.nz/wits

[19] Wright C., Monrose F., Masson G. M., "HMM Profiles for Network Traffic Classification", ACM DMSEC, pp. 9-15, 2004.

[20] Comaniciu D. and Meer P. (2002), Mean Shift: A robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 603-619.

[21] Matplotlib. Available at: https://matplotlib.org

[22] Ng A. (2011), Sparse autoencoder CS294A lecture notes, 1-19.Available at: http://ailab.chonbuk.ac.kr/seminar_board/pds1_files/sparseAutoencoder.pdf.

[23] A. Gogna, A. Majumdar (2019), Discriminative Autoencoder for Feature Extraction: Application to Character Recognition" Neural Processing Letters, vol. 49 no. 3 pp 1723–1735.

С.М. Долгих

Спонтанна категоризація та самонавчання з глибокими моделями автокодування

Національний авіаційний університет, просп. Любомира Гузара, 1, Київ, 03058, Україна
E-mail: serged.7@gmail.com

У цій роботі автор досліджував обробку інформації у моделях глибокого автокодування. Було продемонстровано що непідконтрольне навчання автокодерам певного класу може призвести до появи компактного та структурованого внутрішнього представлення простору вхідних даних, що може бути співвіднесено з категоріями вищого рівня. Була запропонована і продемонстрована практична можливість виявити та виміряти цю формується інформаційну структуру шляхом

застосування непідконтрольного кластеризації щільності в просторі активації фокусного прихованого шару моделі. На основі отриманих висновків запропонований новий підхід до навчання моделей нейронних мереж, що базується на структурах виникаючих у в неконтрольованом інформаційному середовищі навчання, який є ітеративним, керованим навколишнім середовищем, вимагає мінімального нагляду та з подібністю до вивчення біологічних систем і також дає хороші результати класифікації при навчанні нових концепцій вищого рівня навіть при мінімальному наборі маркованих даних. На закінчення надається обговорення теоретичних основ спонтанної категоризації в системах самонавчання.

Ключові слова: штучний інтелект; машинне навчання; нейронні мережі; непідконтрольне навчання

С.Н. Долгих

Спонтанная категоризация и самообучение в глубоких самокодирующих моделях

Национальный авиационный университет, просп. Любомира Гузара, 1, Киев, 03058, Украина
E-mail: serged.7@gmail.com

В этой работе автор исследовал обработку информации в глубоких моделях автоэнкодеров. Было продемонстрировано, что неконтролируемая подготовка автоэнкодеров определенного класса может привести к появлению компактного и структурированного внутреннего представления пространства входных данных, которое можно соотнести с категориями более высокого уровня. Была предложена и продемонстрирована практическую возможность обнаружить и измерить эту возникающую информационную структуру, применяя кластеризацию неконтролируемой плотности в пространстве активации фокусного скрытого слоя модели. Основываясь на выводах, предложен новый подход к обучению моделей нейронных сетей, основанный на категоризованных представлениях возникающих в неконтролируемой информационной среде обучения, который является итеративным, управляемым средой, требует минимальных маркированных данных и с интригующим сходством с процессами обучения биологических систем. В заключение дано обсуждение теоретических основ спонтанной категоризации в самообучающихся системах.

Ключевые слова: искусственный интеллект; машинное обучение; нейронные сети; некотролируемое обучение

Serge Dolgikh
Senior Project Engineer, Solana Networks, 301 Moodie Drive, Ottawa, K2H 9R4, Canada
Education: Master of Science, Coventry University (United Kingdom), Master of Science, National Nuclear Research University
E-mail: serged.7@gmail.com