DEVELOPMENT OF THE METHOD OF UNSUPERVISED TRAINING OF CONVOLUTIONAL NEURAL NETWORKS BASED ON NEURAL GAS MODIFICATION

Viacheslav Moskalenko
PhD, Associate Professor
Department of Computer Science
Sumy State University
2 Rymskoho-Korsakova str., Sumy, Ukraine, 40007
v.moskalenko@cs.sumdu.edu.ua

Abstract: Technologies for computer analysis of visual information based on convolutional neural networks have been widely used, but there is still a shortage of working algorithms for continuous unsupervised training and re-training of neural networks in real time, limiting the effectiveness of their functioning under conditions of nonstationarity and a priori uncertainty. In addition, the back propagation method for learning multi-layer neural networks requires significant computational resources and the amount of marked learning data, which makes it difficult to implement them in autonomous systems with limited resources. One approach to reducing the computational complexity of deep machine learning and overfitting is use of the neural gas principles to implement learning in the process of direct information propagation and sparse coding to increase the compactness and informativeness of feature representation.

The paper considers the use of sparse coding neural gas for learning ten layers of the VGG-16 neural network on selective data from the ImageNet database. At the same time, it is suggested that the evaluation of the effectiveness of the feature extractor learning be carried out according to the results of so-called information-extreme machine learning with the teacher of the output classifier. Information-extreme learning is based on the principles of population optimization methods for binary coding of observations and the construction of radial-basic decision rules optimal in the information criterion in the binary Hamming space.

According to the results of physical modeling, it is shown that learning without a teacher ensures the accuracy of decision rules to 96.4 %, which is inferior to the accuracy of learning with the teacher, which is equal to 98.7 %. However, the absence of an error in the training algorithm for the backward propagation of the error causes the prospect of further research towards the development of meta-optimization algorithms to refine the feature extractor's filters and parameters of the unsupervised training algorithm.

Keywords: neural gas, convolutional neural network, sparse coding, information criterion, classifier.

1. Introduction

Technologies of computer analysis of visual information have received wide application in robotic systems and infocommunication services for various purposes. However, modern methods of intelligent data processing require significant computational resources and volumes of training dataset, which makes it difficult to implement them in autonomous systems with limited resources. In addition, there is a shortage of working algorithms for continuous (self) learning of deep data models in real time, limiting the effectiveness of their functioning under conditions of nonstationarity and a priori uncertainty.

One of the approaches to reducing the computational complexity of deep machine learning is the so-called Transfer Learning [1, 2], where accumulated knowledge in the analysis of visual images can be adapted to a new task. At the same time, the task of continuous machine unsupervised and additional training of the hierarchical feature representation on new observations in an unknown environment can be solved by using autoencoders or a Restricted Boltzmann Machine, but these models are fully connected and therefore resource-intensive [2]. It is proposed in [3, 4] to use the neural gas principles for learning the vocabulary of basic vectors used for sparse coding of the feature description of observations. At the same time, the rapid convergence of this algorithm for unsupervised learning of attributes and the reduction of the overfitting effect of decision rules with a limited volume of the training sample is proved. However, until now it has not been considered the use of this approach for the unsupervised learning of the hierarchic feature extractor, for example, based on the convolutional neural network.

The most popular method of classification analysis of the feature description of observations is the method of support vectors, but this method requires many manual adjustments to regularize the model [1]. A promising approach is the use of binary coding of observations and radial-basis functions in the binary Hamming space on the basis of comparison operations and "Excluding OR", since these operations are the most computationally simple [5, 6]. At the same time, the use of information efficiency criteria and population-based algorithms for searching for optimal parameters of functioning makes it possible to realize automatic regularization of the data model.

It is proposed to use the principles of neural gas and sparse coding for the training of the hierarchical extractor of visual features using the example of a multi-layered neural network VGG-16 [7, 8]. At the same time, the efficiency evaluation of the extractor is supposed to be carried out based on the results of learning the information-extreme classifier with binary coding of observations.

2. Methods

A model of the well-known convolutional network VGG-16 [7] is given, the configuration of ten layers of which is shown in **Fig. 1.**

A training sample is provided from the ImageNet image database [8], along which a sample of image patches $\{x_i^{(j)} \mid i=\overline{1,N}, j=\overline{1,n}\}$ is generated, where N, n – the number of recognition features (N=9) and patches, respectively.

It is necessary in the process of machine learning without a teacher to determine the weight coefficients of the first convolutional layer in the form of M basis vectors

$$C=\{c_{m,i} \mid m=\overline{1,M}, i=\overline{1,N}\}.$$

On the basis of the obtained feature maps, it is necessary to form a new sample of patches of the next layer and continue the layered unsupervised learning to the tenth layer inclusive.

It is necessary to compare the recognition efficiency of a sample of images by the maximum value of an information criterion when using an unsupervised and a supervised trained extractor

$$\overline{E}^{*}=\frac{1}{K}\sum_{k=1}^{K}\max_{\{s\}}E_k, \quad (1)$$

where E_k – the information criterion of the effectiveness of learning decision rules to recognize X_k^o class implementations; K – the number of recognition classes; {s} – many steps of machine learning.

An important step in image analysis is their previous normalization in order to eliminate the linear correlation of the observational components. Zero-phase Component Analysis (ZCA) is one of the most common pre-normalization methods [7].

Fig. 1. The configuration of the first ten layers of neural network VGG-16

Learning of the convolutional layer filters is proposed to be carried out on the basis of the modification of the neural gas proposed in [3, 4]. This algorithm has the following basic steps:

1) initialization of the dictionary of basis vectors C by random values from the uniform distribution;

2) choice of the t-th input observation x from the set X, which has a volume t_{max};

3) normalization of basis vectors C in unit vectors;

4) calculation of the proximity and learning speed coefficients:

$$\lambda_t := \lambda_o(\lambda_{final}/\lambda_o)^{t/t_{max}},$$

$$\alpha_t := \alpha_o(\alpha_{final}/\alpha_o)^{t/t_{max}},$$

where $\lambda_o, \lambda_{final}$ – the initial and final value of the coefficient λ_t; α_o, α_{final} – the initial and final value of the coefficient α_t;

5) initialization of the set of indices of those columns $\{x_{c,i}\}$ that have already been used during t-iterations, $U = \varnothing$;

6) initialization of the minimized remainder $\varepsilon^U := x$;

7) initialization of the time matrix $R := (r_1,...,r_l,...,r_M) = C$ orthonormal in accordance with C^U;

8) initialization of the counter of steps to refine the remainders

$$h := 1, h = \overline{1, K-1};$$

9) calculation of the similarity degree of the vector r_{l_k} to ε^U and their ordering in ascending order

$$-(r_{l_0}^T\varepsilon^U)^2 \le ... \le -(r_{l_k}^T\varepsilon^U)^2 \le ... \le -(r_{l_{M-h-1}}^T\varepsilon^U)^2;$$

10) initialization of the step counter for the refinement of the basis vectors C, $k = 0$, $k = \overline{0, M-h-1}$;

11) updating at the k-th step of the basis vectors using the orthogonality principles for the subspace given in C^U and Oja's rule [4]

$$c_{l_k} := c_{l_k} + \Delta_{lk}, \; r_{l_k} := r_{l_k} + \Delta_{lk},$$

where

12) normalization r_{l_k} by reducing to a unit vector;

13) if $k < 0$, $M-h-1$ – the transition to step 11;

14) determination of the basis of the winner by the formula

$$l_{win} := \arg\max_{l, l \notin U}(r_l^T\varepsilon_i^U)^2;$$

15) updating the matrix R and the current remainder ε_i^U by the formulas:

$$r_l := r_l - (r_{l_{win}}^T r_l)r_{l_{win}}, \; \varepsilon_i^U := \varepsilon_i^U - (r_{l_{win}}^T\varepsilon_i^U)r_{l_{win}},$$

where $r_{l_{win}}$ – the column of the matrix R, has the maximum overlap with the current remainder ε_j^U, the index of which has not yet been added to U;

16) updating the matrix of selected basis vectors

$$U = U \cup l_{win};$$

17) if h <K–1, then the transition to step 11;

18) if $t < t_{max}$, then the transition to step 2, else – the end of processing.

When constructing decision rules in the process of information-extreme learning, binary coding of each feature is used. Coding is done by comparing the value of the i-th attributes with the corresponding lower $A_{B,l,i}$ and upper $A_{T,l,i}$ limits of the field of control tolerances.

The formation of a binary learning matrix

$$\{b_{k,i}^{(j)} \,|\, i = \overline{1, L\cdot N}; \; j = \overline{1, n_k}; \; k = \overline{1, K}\},$$

where N – the number of features of a classifier, n_k – the number of X_k^o class vectors and K – the number of recognition classes, is performed according to the rule

$$b_{k,(l-1)*N+i}^{(j)} = \begin{cases} 1, \text{ if } A_{B,l,i} \le y_{k,i}^{(j)} \le A_{U,l,i}; \\ 0, \text{ else.} \end{cases}$$

Calculations of the values of the coordinates of the binary etalon (averaged) vector b_k, with respect to which the class containers are constructed in the radial basis, are carried out according to the rule

$$b_{k,i} = \begin{cases} 1, \text{ if } \frac{1}{n_k}\sum_{j=1}^{n_k} b_{k,i}^{(j)} > \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} b_{k,i}^{(j)}; \\ 0, \text{ if else;} \end{cases} \quad i = \overline{1, N\cdot L},$$

where n – the total volume of the labeled vectors of the training sample.

As a criterion for the effectiveness of machine learning of the classifier, a modification of the Kullback's information measure [5] is considered:

$$J_k = \frac{1-(\alpha_k+\beta_k)}{\log_2(2+\varepsilon)-\log_2\varepsilon}\cdot\log_2\left[\frac{2-(\alpha_k+\beta_k)+\varepsilon}{(\alpha_k+\beta_k)+\varepsilon}\right], \qquad (2)$$

where α_k, β_k – the estimates of the errors of the first and second kind, which define the admissible domain of the function (2) in the form of inequalities $\alpha_k \ge 0{,}5$ and $\beta_k \ge 0{,}5$; ε – a small positive sign to avoid uncertainty when dividing by zero, is, as a rule, a number from the range $[10^{-4}...10^{-2}]$.

To optimize the parameters of the field of control tolerances, it is proposed to use the Particle Swarm Optimization (PSO) algorithm, which is characterized by simplicity of implementation and interpretation [9, 10]. Optimization of the radii of class containers can be carried out by the method of sequential direct search with a specified step.

3. Results

The proposed unsupervised machine learning algorithm of the multi-layered convolutional neural network VGG-16 is used to synthesize the extractor of the feature description and classifier of objects from the alphabet $\{X_k^o \,|\, k = \overline{1, K}\}$, where K=21. The volume of the training and test samples for each class are $n_k = 100$ images. To learn the extractor, a sample of 3×3 patches with the volume n=1,000,000 from ImageNet images is randomly generated.

The classifier is trained on the basis of the formed feature presentation. The graphs of the change in the maxima of the information criterion (1) during the process of optimizing the field of the control tolerances for the value of the features formed by the 10 layers of the extractor learned without a teacher by the modified neural gas method are shown in **Fig. 2, *a***. The step counter k corresponds to the number of migrations of the swarm agents.

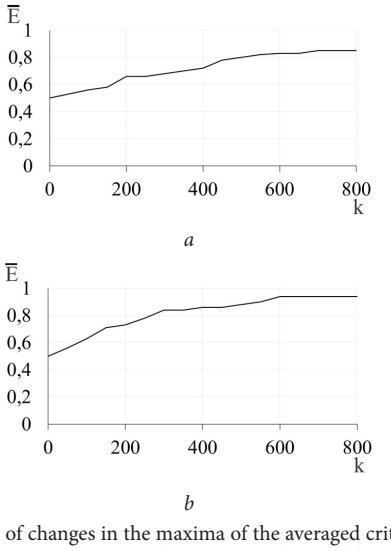

Fig. 2. Graph of changes in the maxima of the averaged criterion (1) in the optimization process $A_{B,l,i}$ and $A_{T,l,i}$ using a learned feature extractor: *a* – without a teacher; *b* – with a teacher

The analysis of **Fig. 2, *a*** shows that during the 800 iterations of the swarm search algorithm, the global optimum of the criterion function (1) is found. In this case, the value of criterion (1) for the extractor learned with a teacher is equal to =0.84, which corresponds to a sampling accuracy of A=0.987.

Thus, the proposed algorithm of unsupervised machine learning of a feature extractor makes it possible to synthesize decision rules with acceptable accuracy for many practical applications.

4. Discussion

Based on the results of physical modeling (**Fig. 2, *a***), the effectiveness of unsupervised training of a hierarchical feature extractor constructed on the first ten layers of the neural network VGG-16 by using a modification of the neural gas algorithm aimed to sparse coding of observations is proved.

To compare the effectiveness of unsupervised learning with the traditional approach in **Fig. 2, *b*** shows the graph of the change in the maxima of the information criterion (1) during the process of optimizing the field of control tolerances for the feature representation of the extractor formed by 10 layers, supervised trained on the full ImageNet sample by the stochastic gradient descent method [8]. The analysis of **Fig. 2**, ***a*** shows that the use of supervised learning on the basis of the stochastic gradient descent algorithm makes it possible to obtain for the alphabet of the 21 recognition class the accuracy of the decision rules equal to 98.7 %, while the proposed algorithm for the unsupervised training of feature extractor is somewhat inferior and provides an accuracy equal to 96.4 %. However, learning without a teacher allows to carry out additional training from time to time without long procedures for backward propagation of the error.

A promising direction of further research is the development of meta-optimization algorithms for clarifying the filters of feature extractor and parameters of the algorithm of its unsupervised learning.

Acknowledgments

The work is supported in the framework of the research work of ДР № 0117U003934 "Intellectual autonomous on-board system of an unmanned aerial vehicle for identification of objects on the ground" (Ukraine).

References

1. Huang, Z., Pan, Z., Lei, B. (2017). Transfer Learning with Deep Convolutional Neural Network for SAR Target Classification with Limited Labeled Data. Remote Sensing, 9 (9), 907. doi: 10.3390/rs9090907
2. Masci, J., Meier, U., Ciresan, D., Schmidhuber, J.; Honkela, T., Duch, W., Girolami, M. A., Kaski, S. (Eds.) (2011). Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. Artificial Neural Networks and Machine Learning – ICANN 2011, 52–59. doi: 10.1007/978-3-642-21735-7_7
3. Labusch, K., Barth, E., Martinetz, Th. (2008). Learning Data Representations with Sparse Coding Neural Gas. 16th European Symposium on Artificial Neural Networks – ICANN 2008, 233–238. doi: 10.1007/978-3-540-87536-9_81
4. Labusch, K., Barth, E., Martinetz, T. (2009). Sparse Coding Neural Gas: Learning of overcomplete data representations. Neurocomputing, 72 (7-9), 1547–1555. doi: 10.1016/j.neucom.2008.11.027
5. Dovbysh, A. S., Moskalenko, V. V., Rizhova, A. S. (2016). Information-Extreme Method for Classification of Observations with Categorical Attributes. Cybernetics and Systems Analysis, 52 (2), 224–231. doi: 10.1007/s10559-016-9818-1
6. Dovbysh, A. S., Moskalenko, V. V., Rizhova, A. S. (2016). Learning Decision Making Support System for Control of Nonstationary Technological Process. Journal of Automation and Information Sciences, 48 (6), 39–48. doi: 10.1615/jautomatinfscien.v48.i6.40
7. Ng, H.-W., Dung Nguyen, V., Vonikakis, V., Winkler, S. (2015). Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning. 17th International Conference On Multimodal Interaction (ICMI'15). Seattle, 443–449. doi: 10.1145/2818346.2830593.
8. Simonyan, K., Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations (ICLR2015), 1–14.
9. Parpinelli, R. (Ed.) (2012). Theory and New Applications of Swarm Intelligence. London: InTech, 204. doi: 10.5772/1405
10. Moskalenko, V., Pimonenko, S. (2016). Optimizing the parameters of functioning of the system of management of data center it infrastructure. Eastern-European Journal of Enterprise Technologies, 5 (2 (83)), 21–29. doi: 10.15587/1729-4061.2016.79231