UDC 004.93(045)
DOI:10.18372/1990-5548.66.15225

[1]V. M. Sineglazov,
[2]A. T. Kot

TRAINING DATA SAMPLING FOR CONVENTIONAL NEURAL NETWORKS CONFIGURING

[1]Aviation Computer-Integrated Complexes Department, Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine
[2]Technical Cybernetic Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine
E-mails: [1]svm@nau.edu.ua ORCID 0000-0002-3297-9060, [2]anatoly.kot@gmail.com

Abstract**—***The problem of generating training data for setting up the convolutional neural networks is considered, which is of great importance in the construction of intelligent medical diagnostic systems, where due to the lack of elements of the training sample, it is proposed to use the approaches of artificial data multiplication based on the initial training sample of a fixed size for the image processing (the results of the ultrasound, CT and MRI). It shows that the increase of the training sample resulted in less informative and poor quality elements, which can introduce extra errors in the goal achievement. To eliminate this situation the algorithm for assessing the quality of a sample element with the subsequent removal of uninformative elements is proposed.*

Index Terms—Convolution neural networks; artificial reproduction of data; uninformative data; intelligent medical systems.

I. INTRODUCTION

Neural networks have been actively developing in the last decade. They demonstrate excellent quality of solving classification problems, forecasting, etc. versus other machine learning algorithms. Neural networks contributed actively to image processing which became widely used throughout different applications. A huge dataset (up to several million images) should be used for neural network training which in turn is extremely computationally-demanding.

Convolutional neural networks (CNNs) are efficient image processing tools. Currently, they are actively used in intelligent medical diagnostic systems for processing the results of the ultrasound, CT, and MRI images. The CNN training issue consists of the importance to create a training dataset, which, on the one hand, must have the optimum size, and, on the other hand, be informative in order to train the convolutional neural network to mark the mean image parameters. It is a very important to create the training data set of the minimum size, such as the usage of a large available dataset for training neural networks leads to a significant increase in computational and memory resources and can also cause excessive models.

The existing sampling techniques suffer from many shortcomings, making them either very slow processed, or resource-demanded, or undefined quality criteria for the sample to be formed. These efforts aim to create the improving samples quality algorithm, due to the introduction of quality criteria, reducing computing resources use, as well as time reduction for the formation process.

II. DATA SAMPLING METHODS BASED ON EXISTING DATA

During training data sampling, the fact of the insufficient available initial data should be considered. Basing on a few data, there is no way to distinguish a correct pattern and a false one. Such false patterns, arising as a result of lack of data, are called background regularities. Some types of retraining consist of background regularity learning. An example of background regularity is the correlation between an image class and a color of one specific pixel. An important special case of the problem is missing data of a certain type. Part of the variables often has the same values or a very narrow range of values. Semantically, an unreasonable disbalance of the amount of different data types in the dataset leads to an unsubstantiated overstated impact on the result of some data and an underestimated impact or total disregard of other data, and as a result, causes a suboptimal decision. The problem goes back to the uneven representation of different classes when some classes of images are represented by substantially smaller data volumes than others.

An alternative approach to solving the problem of the limited amount of training data is to implement direct artificial data propagation based on the initial training sample of a fixed size. New data can be obtained both by modifying the original data and by generating new random values that have any properties characteristic of the data in the original dataset.

Increased by means of transformations of the initial data training dataset is most often used in image recognition, therefore, these methods are focused primarily on image processing. Particularly frequently, when generating the images, such transformations are used as rotation by a certain random angle, compression and stretching vertically and horizontally, tilt, mirror reflection, cropping, displacement, and many others [1] – [3]. This group of methods also includes the noise of the initial data, as well as various morphing transformations, similar to the described in this work [4], where new data are generated by “crossing” the initial data with each other.

The solution to the problem of diagnosing the liver fibrosis stages (F0, F1, F2, F3, F4) based on the results of processing ultrasound studies using the ResNet 101 convolutional network has been highlighted in the article [5]. The results of the network operation and the original defective sampling for class F2, and the increased sampling by the rotation method and scaling and blurring methods of some pixels. were compared. As a result, the methods of scaling and blurring some pixels showed the best results.

Taking a subsample that stood out from the initial sample as a kind of a prototype, there are various data sampling methods [6], [7], subdivided into the types by the prototype process approach:

- prototype construction methods;
- prototype selection methods;
- mixed methods which combine prototype construction and prototype selection methods;

Prototype selection methods [8] – [12] are based on the sampling from the original sample using specific approaches. Such methods can sequentially add the instances from the original selection to the prototype [8], form the prototype from the set of the initial sample elements and sequentially remove instances from the prototype [9]. Some defined criteria are used while adding or removing the elements:

- noise filtering [10] – removal of noisy elements, the marking of which does not match the adjacent ones;
- condensation [11] – inverse to noise filtering approach when adding to the prototype the instances carrying the new information (new markup);
- stochastic search [12] – walk through the possible prototype combinations and obtaining the best upon the specified criterion.

The effective usage of the methods on a little data is the common disadvantage of these methods.

The number of iterations for forming a prototype increases nonlinearly during the linear growth of the initial sample, which leads to the nonlinear growth of the used computing resources and/or time.

The artificial synthesis approaches are used under the conditions of poor initial sampling in terms of the element number, or insufficient elements with certain characteristics. Thus, the initial sample is expanded due to the synthesis of its new elements. The main drawback of this approach is the "quality" of artificial elements, which results in additional costs for verification.

III. Problem Statement

Let the initial sample $<X, Y>$ be given with a volume of S instances characterized by a set of values N of input (descriptive) features X and one output (target) feature Y. Then the task of forming a training sample $<x, y>$ from the original sample $<X, Y>$ can be represented as a search for such a minimal subset $<X, Y>$ for which the value of a given quality functional $\overline{I}(<x, y>)$ will have a maximum value. In this case, the quality functional $\overline{I}(<x, y>)$ should reflect the requirements regarding the topological and statistical representativeness of the formed subsample $<x, y>$ relative to $<X, Y>$.

IV. Algorithm for Forming a Training Sample

There are two stages of the classification problem in medical diagnostics: the first is the task of image segmentation obtained using CT, MRI or ultrasound; the second is the selection of informative characteristics on the segmented image with the further task of the classification.

The quality of decision-making is influenced, on one hand, by the qualitative and quantitative composition of the data sampling and, on other hand, by the qualitative and quantitative composition of the space of informative characteristics.

The following characteristics of a training sample are introduced in [13]: the confidence-building measures in data sampling, the feature space, the representativeness of a sample, in the sample size, the confidence-building experts' measure in a sample, in informative or informational value, the confidence-building experts' measure in the composition of features, in the dimension (number) of informative characteristics. Solving a multicriteria problem of finding the elements of the data sampling is supposed the usage of those criteria, which is computationally expensive.

The less sophisticated approach is the usage of the search algorithm based on the alternately discarding elements of an expanded (implementation of direct artificial data reproduction based on an initial data sampling of a fixed volume) sample monitoring such criteria as Jaccard similarity index, sensitivity, specificity, accuracy, dice similarity ratio and Matthew's correlation coefficient for solving the segmentation problem and the accuracy criterion for solving the classification problem. Despite the apparent simplicity of this approach, it is time-consuming.

It is advised to use an evolutionary approach (genetic algorithm [14] – [17]) to reduce the number of enumerated combinations, which includes the following stages.

1. Initialization: set the initial sample <*X*, *Y*> with a volume of *S* instances (corresponds to the volume of data obtained as a result of ultrasound, CT or MRI studies), as well as the maximum allowable volume S_{F} sample <*x*, *y*> formed by the initial sample and generating images; designate the values of the quality criteria of the initial and extended samples, determined based on the total error in solving segmentation problems, highlighting essential characteristics and decision support, respectively $\bar{I}$ and $\bar{I}^* \le \bar{I}$. Set the size of the decision population *H*, the maximum number of iterations *T*, the probability of mutation P_{m}, and the acceptable quality criterion of the result.

2. Using the genetic algorithm the chromosome is formed only from genes related only to generated images, which are added to *S* images obtained as a result of ultrasound, CT or MRI studies. The initial solutions population establishment is carried out as follows. We represent the *k*th solution h^k as a binary combination of *S* bits, the *S* bit of which is $\mathrm{h}^{k}_{\mathrm{s}}$ determines the inclusion of the *S* instance of the added sample in the solution (if $h_s^k = 0$, the *S* instance is not included in the *k*th solution, otherwise the case when $h_s^k = 1$, the *s*th instance is included in the *k*th solution). Let randomly generate *H* binary combinations for $k = 1, 2, ..., H$; $s = 1, 2, ..., S_F - S$.

2.1 Set the probabilities of including instances in the K solution:

2.2 Moving from the digits with high probabilities of including instances in the k-th solution to the digits with lower probabilities, set no more than S_F digits equal to one with the highest probabilities, but not less than 0.5, and set the remaining digits equal to zero.

3. *The end of the search check.* Form an appropriate sample for each *k*th decision of the population, for which to *evaluate*. If more than *T* iterations are performed or among the set of solutions there is such a solution with number *k*, for which stop the search and return as a result the sample with the highest value of the quality criterion.

4. *Cross selection process.* Considering the maximised fitness function, to form parental pairs based on the quality criterion presented in step 1, thus ensuring the probability for evaluating the probability of a decision to be allowed to cross.

5. *Crossing*. Implement the crossing of selected solutions to produce the new solutions based on a single crossing over as described in [18].

6. *Mutation*. Implement the mutation operator for each of the available solutions, according to the algorithm described in [18]. Exclude the solutions encountered earlier in the previous method cycles from the current population. Proceed to step 3.

V. CONCLUSIONS

The data sampling algorithm for convolutional neural networks with an insufficient volume of the initial sample based on the use of image generation methods and a genetic algorithm has been developed.

REFERENCES

[1] L. Yaeger, R. Lyon, and B. Webb, *Effective Training of a Neural Network Character Classifier for Word Recognition*, NIPS, 1996.

[2] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition," *Neural Computation*. vol. 22(12), 2010.

[3] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best Practices for Convolu- tional Neural Networks Applied to Visual Document Analysis," *Int'l Conf. Document Analysis and Recognition*. 2003.

[4] S. Kachalin. "Increasing the stability of training large neural networks by supplementing small training samples of parent examples, synthesized biometric descendant examples*," Proceedings of the scientific and technical conference of the cluster of Penza enterprises that ensure the security of information technologies*. Penza. vol. 9, 2014, pp. 32–35.

[5] O. I. Chumachenko and A. T. Kot. "Formation of a Learning Set for the Task of Image Processing," Electronics and Control Systems, N 3(65), Kyiv, NAU: Osvita Ukrainy, pp. 9–17, 2020. https://doi.org/10.18372/1990-5548. 65.14978

[6] N. Jankowski and M. Grochowski, "Comparison of instance selection algorithms I. Algorithms survey," *Artificial Intelligence and Soft Computing: 7th International Conference ICAISC-2004*, Zakopane, 7–11 June, 2004: proceedings. Berlin: Springer, 2004, pp. 598–603. – (Lecture Notes in Computer Science, vol. 3070).

[7] T. Reinartz, "A unifying view on instance selection," *Data Mining and Knowledge Discovery*, no 6, pp. 191–210, 2002.

[8] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, no 6, pp. 37–66, 1991.

[9] G. Gates, "The reduced nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 18, no 3, 1972, pp. 431–433.

[10] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 14, 1968, pp. 515–516.

[11] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, Cybernetic*s, vol. 2, no 3, 1972, pp. 408–421.

[12] D. R. Wilson and T. R. Martinez, "Reduction techniques for instancebased learning algorithms," *Machine Learning*, vol. 38, no 3, pp. 257–286, 2000.

[13] Ghosh Ashish & Dehuri Satchidananda, "Evolutionary Algorithms for Multi-Criterion Optimization: A Survey," *International Journal of Computing & Information Sciences*, 2, 2004.

[14] C. A. C. Coello, "Evolutionary multi-objective optimization: a historical view of the field," *Comput. Intell. Mag. IEEE* 1 (1), 28–36, 2006. https://doi.org/10.1109/MCI.2006.1597059

[15] K. Deb, "Multi-Objective Optimization Using Evolutionary Algorithms," vol. 16, John Wiley & Sons, 2001.

[16] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang, "Multi-objective evolutionary algorithms: a survey of the state of the art," *Swarm Evol. Comput*, 1 (1), 32–49, 2011. https://doi.org/10.1016/j.swevo.2011.03.001.

[17] B. Li, J. Li, K. Tang, and X. Yao, "Many-objective evolutionary algorithms: a survey," *ACM Comput. Surv.* 48 (1), pp. 1–35, 2015. https://doi.org/10.1145/2792984

[18] Michael Z. Zgurovsky, Victor M. Sineglazov, Olena I. Chumachenko, *Artificial Intelligence Systems Based on Hybrid Neural Networks*, Springer, 2020, 390 p. [Electronic resource]. Access mode: https://link.springer.com/book/10.1007/978-3-030-48453-8. Customer can order it via https://www.springer.com/gp/book/9783030484521.

Received October 21, 2020

Sineglazov Victor. orcid.org/0000-0002-3297-9060. Doctor of Engineering Science. Professor. Head of the Department.
Aviation Computer-Integrated Complexes Department, Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine.
Education: Kyiv Polytechnic Institute, Kyiv, Ukraine, (1973).
Research area: Air Navigation, Air Traffic Control, Identification of Complex Systems, Wind/Solar power plant.
Publications: more than 660 papers.
E-mail: svm@nau.edu.ua

Kot Ananatoliy. Post-graduate student.
Technical Cybernetic Department, National Technical University of Ukraine “Ihor Sikorsky Kyiv Polytechnic Institute,” Kyiv, Ukraine.
Education: National Technical University of Ukraine “Ihor Sikorsky Kyiv Polytechnic Institute,” Kyiv, Ukraine, (2017).
Research area: artificial Intelligence.
Publications: 8.
E-mail: anatoly.kot@gmail.com

В. М. Синєглазов, А. Т. Кот. Формування навчальної вибірки для налаштування згорткових нейронних мереж

Розглянуто задачу формування навчальної вибірки для налаштування згорткових мереж, що має велике значення при побудові інтелектуальних медичних систем діагностики в яких для обробки зображень використовуються результати УЗД, КТ та МРТ. У зв’язку з нестачею елементів навчальної вибірки запропоновано використовувати підходи штучного розмноження даних на основі вихідної навчальної вибірки фіксованого обсягу. Показано, що в результаті такого збільшення обсягу навчальної вибірки в неї можуть потрапити малоінформативні і поганої якості елементи, які можуть внести додаткові похибки у розв’язання поставленої задачі. Для усунення такої ситуації в роботі запропоновано алгоритм оцінки якості елемента вибірки з подальшим видаленням малоінформативних елементів.

Ключові слова: згорткові нейронні мережі; штучне розмноження даних; малоінформативні дані; інтелектуальні медичні системи.

Синєглазов Віктор Михайлович. orcid.org/0000-0002-3297-9060.
Доктор технічних наук. Професор. Завідувач кафедрою.
Кафедра авіаційних комп’ютерно-інтегрованих комплексів, Факультет аеронавігації електроніки і телекомунікацій, Національний авіаційний університет, Київ, Україна.
Освіта: Київський політехнічний інститут, Київ, Україна, (1973).

Напрям наукової діяльності: аеронавігація, управління повітряним рухом, ідентифікація складних систем, вітроенергетичні установки.
Кількість публікацій: більше 660 наукових робіт.
E-mail: svm@nau.edu.ua

Кот Анатолій Тарасович. Аспірант.
Кафедра технічної кібернетики, Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського», Київ, Україна.
Освіта: Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського», Київ, Україна, (2017).
Напрям наукової діяльності: штучний інтелект.
Кількість публікацій: 8.
E-mail: anatoly.kot@gmail.com

В. М. Синеглазов, А. Т. Кот. Формирование обучающей выборки для настройки сверточных нейронных сетей

Рассмотрена задача формирования обучающей выборки для настройки сверточных сетей, что имеет большое значение при построении интеллектуальных медицинских систем диагностики в которых для обработки изображений используются результаты УЗИ, КТ и МРТ. В связи с нехваткой элементов обучающей выборки предложено использовать подходы искусственного размножения данных на основе исходной обучающей выборки фиксированного объема. Показано, что в результате такого увеличения объема обучающей выборки в нее могут попасть малоинформативные и плохого качества элементы, которые могут внести дополнительные погрешности в решение поставленной задачи. Для устранения такой ситуации в работе предложен алгоритм оценки качества элемента выборки с последующим удалением малоинформативных элементов.

Ключевые слова: сверточные нейронные сети; искусственное размножение данных; малоинформативные данные; интеллектуальные медицинские системы.

Синеглазов Виктор Михайлович. orcid.org/0000-0002-3297-9060.
Доктор технических наук. Профессор. Заведующий кафедрой.
Кафедра авиационных компьютерно-интегрированных комплексов, Факультет аэронавигации электроники и телекоммуникаций, Национальный авиационный университет, Киев, Украина.
Образование: Киевский политехнический институт, Киев, Украина, (1973).
Направление научной деятельности: аэронавигация, управление воздушным движением, идентификация сложных систем, ветроэнергетические установки.
Количество публикаций: более 660 научных работ.
E-mail: svm@nau.edu.ua

Кот Анатолий Тарасович. Аспирант.
Кафедра технической кибернетики, Национальный технический университет Украины «Киевский политехнический институт им. Игоря Сикорского», Киев, Украина.
Образование: Национальный технический университет Украины «Киевский политехнический институт им. Игоря Сикорского», Киев, Украина, (2017).
Направление научной деятельности: искусственный интеллект.
Количество публикаций: 8.
E-mail: anatoly.kot@gmail.com