УДК 004.932

RECOGNITION RECIPES WITH DEEP MACHINE LEARNING

Ulyanov Sergey[1], Filipyev Andrey[2], Koshelev Kirill[3]

[1]*Doctor of Science in Physics and Mathematics, professor;*
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: ulyanovsv@mail.ru.

[2]*PhD Student;*
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: avfilipev@gmail.com.

[3]*PhD Student;*
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: kirill_koshelev18@rambler.ru.

This article aims to reveal that deep machine learning algorithms can be applied in a variety of commercial companies in order to improve developing intelligent systems. The major task which would be discussed in the application of convolutional neural networks for recognizing recipes of products and providing the possibility of maintenance decision making in business processes. Besides algorithms, the problems of real projects like gathering and preprocessing data would be considered and possible solutions suggested.

Keywords: Deep Learning, Intelligence Systems, Convolutional Neural Networks, Image Recognition, Decision Making Systems, Artificial Intelligence.

КОНТРОЛЬ КАЧЕСТВА ПРОДУКЦИИ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ГЛУБОКОГО МАШИННОГО ОБУЧЕНИЯ

Ульянов Сергей Викторович[1], Филипьев Андрей Владимирович[2], Кошелев Кирилл Викторович[3]

[1]*Доктор физико-математических наук, профессор;*
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: ulyanovsv@mail.ru.

[2]*Аспирант;*
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: avfilipev@gmail.com.

[3]*Аспирант;*
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: kirill_koshelev18@rambler.ru.

Основная цель данной работы – продемонстрировать эффективность применения алгоритмов глубокого машинного обучения в деятельности различных коммерческих компаний. Создание интеллектуальных систем для поддержки различных коммерческих проектов – довольно актуальная задача на сегодняшний день. В работе описывается применение сверточных нейросетевых моделей для решения задачи распознавания различных продуктов, использующихся при приготовлении пиццы. Обсуждается возможность использования таких моделей для поддержки принятия решений в бизнес-процессах. Также в работе рассматриваются важнейшие этапы построения интеллектуальных систем – сбор данных и их предварительная обработка.

Ключевые слова: глубокое машинное обучение, интеллектуальные системы, сверточные нейронные сети, распознавание образов, системы поддержки принятия решений, искусственный интеллект.

Introduction

This article aims to reveal that deep machine learning algorithms can be used in a variety of commercial companies whose direct goals are not bounded to developing the science sphere. This research could show the importance of using modern technologies inside the companies because of the synergy of two opposite directions the science and the business could help each other in reaching their archives. And last years reveal that these spheres are becoming more and more bounded. The major task which would be discussed below is using convolutional neural networks in order to recognize recipes of products and provide the possibility of maintenance decision making in business processes.

Dodo Pizza operates in a very competitive market. Only after a few years of developing, it became a leader in the Russian market and opened its stores in more than twelve other countries. One of the key reasons for becoming a leader in the way of automatization most of the business processes by developing its own software. It lets to gather all data about every order that was generated by the information system. Developing software for automatization business processes gives an advantage to the company for fast-growing in most countries' markets. Developing data-based features can improve the research of market preferences and personalize offers for clients. Independently of advancement data analytic direction there are research projects that are aimed to investigate the possibility of intelligent systems help to optimize and improve production processes of products.

Developing a company in different markets makes to optimize operating processes in order to make the business profitable. Heads of various directions have to examine opportunities to create an environment where every person who is participating in interactions between the company's service and customers can get the best experience. Employees should have good working conditions and customers should get the products that they actually want. The real challenge for the company was to open several points of sales in the competitive market of the USA in order to find a profitable business model for overloaded foodservice environments.

People in this country have a large selection of different food companies and one of the unique customer experience features is the possibility to customize lots of their orders. In a business of selling pizza providing service which lets to add or remove any ingredient is an ordinary feature. But this overloads employees with the responsibility of controlling cooking lots of recipes.

The first results of a launching customization service have shown that recipe identification of baked products and matching them with order numbers takes too much time. This leads to the fact that there is a chance to make mistakes during the cooking process and provide the foodservice worse than customers expect. Testing new for the company features with a small amount of point of sales is a good way to find invisible problems of operating processes before scaling it for the whole chain. Managing controlling the cooking process on the whole chain of point of sales may require a large investment with unpredictable results.

In order to examine an opportunity and expediency of intelligent system's application for decision making, the company has decided to develop the AI module which can help to recognize the recipe of baked products and suggest controller the order number. For the company, it may be the first positive experience of investment in developing complicated algorithms and its integration into its own Dodo Information System, which automates most business processes. Successful results of this kind of experiment are the only way to start a commercial business work together with science because heads of companies should think first of justified investments in order to be competitive.

1. Method

The purpose of computer vision is the processing of images obtained from various sensors, the selection of images and their subsequent classification. Today, high results in pattern recognition are obtained using convolutional neural networks (CNN). With a sufficiently large size, CNN has a small number of configurable parameters and trained quite quickly, which allows them to be called the most universal and effective neural network models for solving computer vision problems.

The combined use of various sensors and their combinations with stereo vision technology allows for a more qualitative and complete construction of the "world scene" of the robot (most often, machine learning and computer vision algorithms are used in robotics). We also pursued the goal of creating such a computer vision system that can be transferred to a robotic platform to solve various problems, thereby improving its interaction with the environment. The stereo vision technology, which to some extent repeats the features of the development of natural vision, allows the on-board system to receive information not only about the color and brightness of the object but also about the distance to it, about its geometric shape, about obstacles to the object, which plays an extremely important role in the tasks of a mobile robot. Intellectualization of the control system, in particular, the use of a neural network approach in the recognition system, significantly reduced the negative impact of external factors on the quality of recognition (recognition error when changing the angle of the object, changing lighting, software sensitivity, etc.).

Convolutional neural network architecture

The CNN is a multi-layer sensor network, it represents a further development of the multi-layer perceptron, however, unlike the latter, the convolutional network has a much smaller number of weights (the principle of sharing weights). Quite often the CNN model is divided into two main parts: the part responsible for the selection of features, and the part with which classification is performed [1].

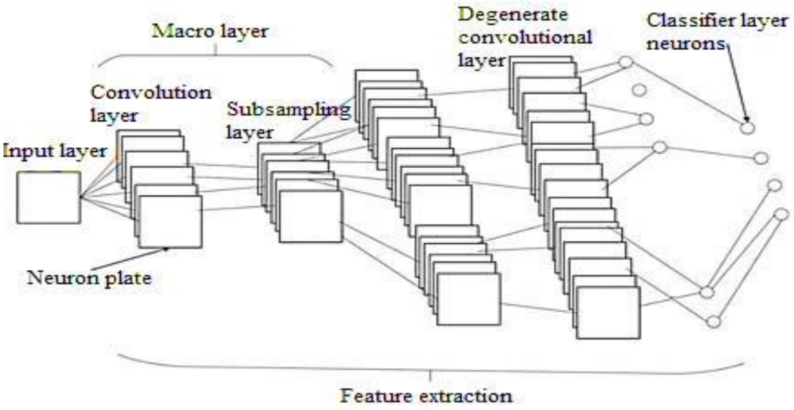

Fig. 1. General structure of the convolutional neural network

Features of CNN structure. The selection of features of an object occurs using layers of convolution and subsampling. Each layer is a set of plates of neurons, which are also called feature maps. The convolution network structure includes two types of layers: convolution layers and subsampling layers. Convolution and subsampling layers are combined into macro layers (a macro layer is a convolution layer followed by a subsampling layer [2]). A set of several architecture-similar macro layers is called a feature separator in a sensory neural network. Each neuron of the convolution layer and the subsample layer is associated with a receptive field (RF). RF is a certain square area that includes neurons capable of transmitting signals to a neuron that has a given RF.

Figure 2 shows a convolutional layer diagram, and also shows the process of its interaction with the previous layer.

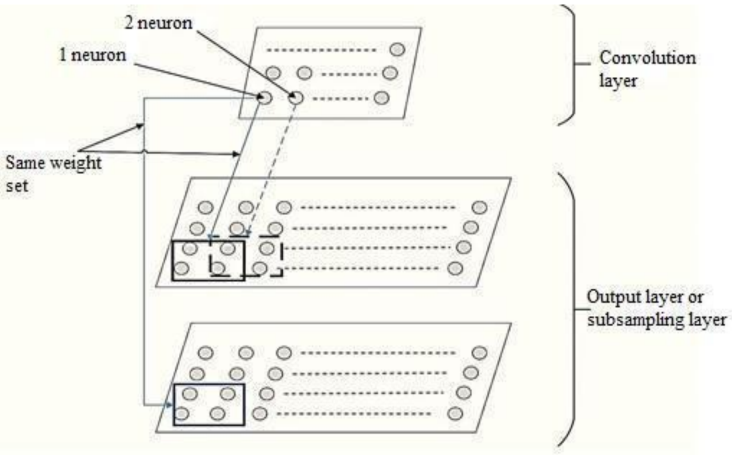

Fig. 2. Scheme of the interaction of the convolutional layer with the previous layer

Figure 2 makes it clear that RF of each neuron of the convolutional layer is immediately associated with two feature maps of the previous layer (as an example parallel processing of each image channel in RGB color format can be given). A key feature of convolutional layers is that bonds formed within the same feature map have the same set of weights. These are the so-called bound weights. Using the associated weights, certain features are selected in an arbitrary position on the feature map. The connection of the convolutional layer card with several cards of the previous layer provides the opportunity to equally interpret differently presented information [1].

Figure 2 shows that RF of the neurons intersect (the step is subject to adjustment). Reducing the step of applying the RF increases the number of neurons in the map of the next layer. Features of an object with the help of RF and associated weights are extracted [3]. If K_C is the number of neurons that make up the RF of the *n*-th neuron of the convolutional layer, *Kernel[k]* is the convolution core, *b* is the displacement of the *n*-th neuron (*b* and *Kernel [k]* retain their values for the entire map of the convolutional layer), *x[n + k]* is the input for the *n*-th neuron of the convolutional layer (k = 0..K_C-1), then the convolution operation can be displayed by the Eq. (1) [1]:

$$p = b + \sum_{k=0}^{K_c-1} Kernel_k * x_{n+k} \tag{1}$$

The weighted sum *p* is supplied to the input of the activation function – the response of the neuron is determined [1]. The output of the neuron has the following form (2):

$$y = f(p) \tag{2}$$

Each neuron of the convolutional layer is a detector of a certain feature that was isolated during training. The interaction of convolution with the activation function of a neuron allows us to assess the degree of presence of a particular trait in the current RF of this neuron. The convolution of an input element with general customizable parameters is an analog of passing an image on a map through some filter [3].

Figure 3 shows the interaction pattern of the subsampling layer with the previous convolutional layer. The main task of the subsampling layer is to reduce the scale of the processed display obtained using the previous convolutional layer. Each map of a subsampling layer is associated with only one map of the previous convolutional layer.

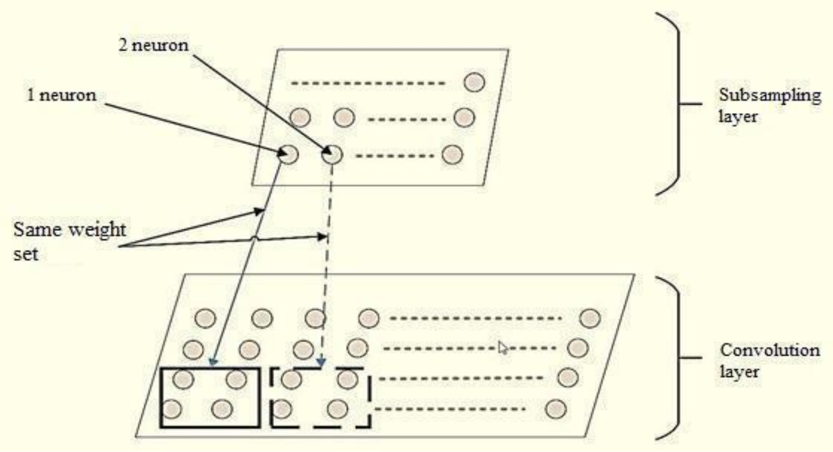

Fig. 3. The scheme of interaction of the subsampling layer with the previous layer

It is important to note that RFs of the neurons of the subsampling layer do not intersect. Configurable parameters are common to all neurons of each plate. The number of these parameters is equal to two; it does not depend on the number of elements included in the RF of these neurons.

Since RFs of the neurons do not intersect, the convolution *p* for the *n*-th neuron of the subsampling layer is defined as follows:

$$p = b + u * \sum_{k=0}^{K_s - 1} * x_{n * K_s + k} \qquad (3)$$

In Eq. (3) K_S is the total number of neurons included in the RP of the *n*-th neuron of the subsampling layer [2].

The second part of the CNN is a feature classifier. The classifier, as a rule, is a single-layer or two-layer perceptron. The number of neurons in the classifier layer usually corresponds to the number of classes to which the input image belongs. There are no associated weights in the classifier. The weighted sum *p* for the neuron of the classifier layer can be defined as

$$p = b_n + \sum_{k=1}^{K} x_k * w_{n,k} \, , \qquad (4)$$

In Eq. (4) b_n is the offset, different for each neuron, *x[k]* is the input element, *w[n, k]* are the custom parameters of the *n*-th neuron (unique to each neuron), *K* is the input size for the classifier layer [1].

Remark: there are many works [4-9] that are devoted to the creation and training of the CNN. In this work the recognition system based on stereo vision technology uses the classical CNN architecture, which includes convolution and averaging layers. Network training was done with a teacher. In relation to the recognition problem, a teacher is the number of a class that is encoded in a vector. This vector is equal to the size of the output layer of the neural network. This is the desired result corresponding to this input pattern. The actual response is obtained as a result of the reaction of the neural network with the current parameters on the input pattern. Error signal - the difference between the desired signal and the current response of the neural network. It is on the basis of the error signal that the tunable parameters of the neural network are corrected [2]. A significant minus of this training scheme is the great difficulty in creating training samples, however with a small number of classes the negative influence of this factor can be neglected.

The error function depends on the system settings being configured. For such a function, one can construct a multidimensional error surface in the coordinates of free parameters. In this case the real surface of the error is averaged over all possible examples, which are presented in the form of input-output pairs. To improve system performance over time, the error value should shift to the minimum. This minimum can be both local and global [2]. The most common and reliable methods for achieving a local or global minimum on the error surface are local optimization methods [10-12].

The main and most important stage in the implementation of the recognition system based on CNN is the stage of formation of the training sample. The design, creation and training of CNN were carried out using the *TensorFlow* library. The structure of the CNN is set using this library directly in the program code, which imposes questions on the choice of the optimal structure.

At the CNN input the images are received in the matrix rather than in vector form (which is necessary to save information about the topology). Input image size 256x256 pixels, format – RGB. The first convolutional layer contains 256 4x4 feature maps (each with its own convolution kernel), i.e. each convolutional neuron is connected to a square 4x4 image. The next convolutional layer has a similar architecture.

It is known that convolution layers and subsampling layers are responsible for highlighting certain attributes of various objects (borders, simple colors, and curves) in images. Deepening into the network (going through the following layers of convolution and subsampling) allows you to define less abstract (most characteristic of any class of objects) features. The following dependency can be traced - an increase in the number of macro layers (“Convolution-Subsampling”) makes it possible to find more and more complex features of certain objects in the image.

The next step after convolution is averaging (subsampling). This operation reduces the dimension of feature maps obtained from the previous convolution layer. This method is based on the fact that neighboring pixels are very slightly different from each other (the so-called "pixel correlation"). The averaging operation significantly reduces the dependence of the recognition result on the scale of the input image, and also significantly reduces the computational load. A fully connected output layer contains the probability that the object in the analyzed image belongs to a certain class. Convolutional neural networks are currently one of the best application tools for solving recognition and classification problems.

2. Research goals

The research task aims to check the possibility of deep learning algorithms to recognize recipes by the pictures of products. In case these algorithms show enough performance of recognition the company will have tools for integrating computational intelligence modules into business processes and unload employees. It is necessary that the inference of a machine learning algorithm has a high statistical percentage of right recognitions, otherwise, the resulting error of a human factor adds the error of the algorithms. The whole empirical experiment splits into three stages:

1. Gathering data;
2. Developing operating processes structure that allows integrating artificial intelligence system;
3. Developing and testing a module with a deep machine learning algorithm.

Gathering data might be a not obvious task for companies that just investigate opportunities for using complicated intelligence algorithms. On the one hand, it may seem that there are lots of sources of visual data for fitting machine learning models. However, every goal requires lots of individual marked data for solving certain problems. In this experiment, we faced the same problem, in the beginning, it appearances that there are lots of open data sources and the company has its own sets of images for fitting deep learning models. Nevertheless, when you start investigating available data lack of information emerges.

In order to manage this, the special data gathering was organized and company employees started taking pictures of baked products and marking them with information about order number, product name, etc. Set of one thousand images required near about one month and online communication between IT specialists, product owners and employees in the sale points. Awareness of the costs of this communication may help to understand the real amount of investments in intelligent systems and that the real innovation projects right now are more complicated than it sounds. That is why it is worth to have specialists which can estimate the real cost-effectiveness and profitability of the project.

Figure 4 illustrates the developed scheme of taking the order and cooking it with an integrated intelligent system.

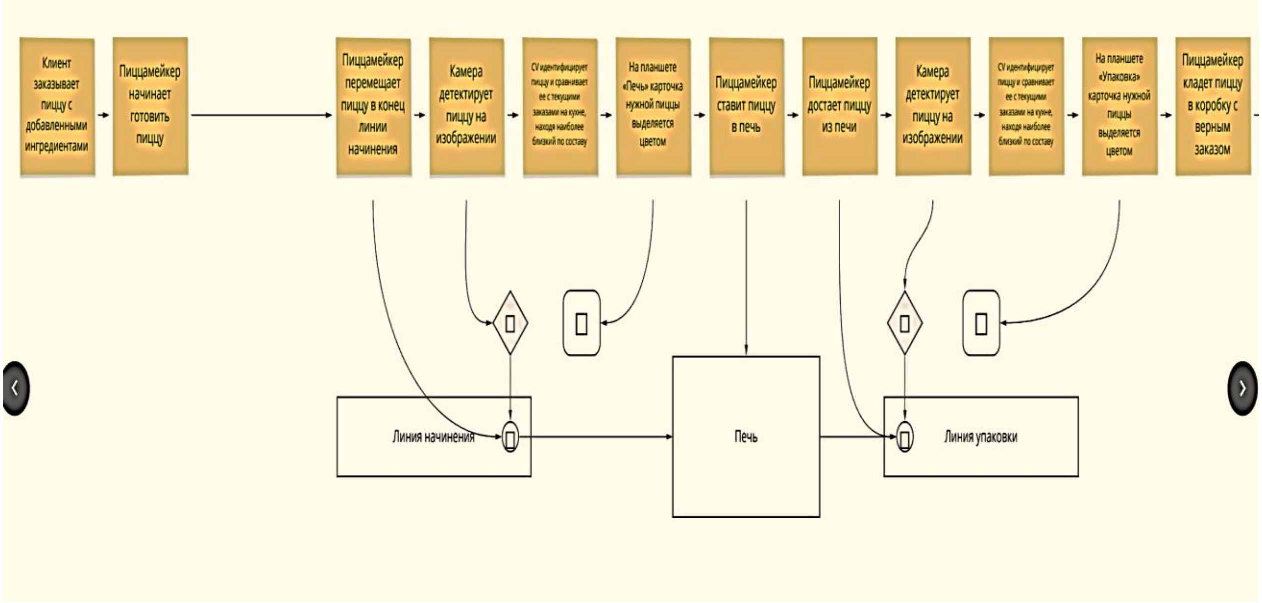
Клиент заказывает пиццу с добавленными ингредиентами
Пиццамейкер начинает готовить пиццу
Пиццамейкер перемещает пиццу в конец линии начинения
Камера детектирует пиццу на изображении
CV идентифицирует пиццу и сравнивает ее с текущими заказами на кухне, находя наиболее близкий по составу
На планшете «Печь» карточка нужной пиццы выделяется цветом
Пиццамейкер ставит пиццу в печь
Пиццамейкер достает пиццу из печи
Камера детектирует пиццу на изображении
CV идентифицирует пиццу и сравнивает ее с текущими заказами на кухне, находя наиболее близкий по составу
На планшете «Упаковка» карточка нужной пиццы выделяется цветом
Пиццамейкер кладет пиццу в коробку с верным заказом
Линия начинения
Печь
Линия упаковки

Fig. 4. The operating process of taking orders with an integrated computational intelligence module

This scheme was developed in order to create a theoretical basis for possible changes in operating processes and learn how artificial intelligence might be integrated into the real sale points. One of the theories was that recognition pizza recipe before the oven could help the identification of ingredients because the melted cheese would interfere with the detection of other products. But in this case, there is a problem of revealing the results to an employee who packs the products and marks them with the order number. Pictures of products before and after the oven were gathering during the first stage in order to check the performance of recognition.

At the beginning of the experiment, the main metrics of machine learning model performance were defined:

1. Custom accuracy.
2. Precision
3. Recall
4. F1 score

Custom accuracy shows the percentage of eight recognized recipes. It calculates like:

$$accuracy = \frac{\sum r}{n} \qquad (5)$$

In Eq. (5) r is the number of right recognitions and n is the number of pictures. If every ingredient is recognized right r equals 1, else 0.

These metrics have a good interpretation property for classification problems and it is possible to map them to business goals. Important note that the recipe a product is considered as recognized if only all ingredients are correctly identified.

3. Experiment

Good solutions with deep learning algorithms require lots of various data in order to provide all possible options of recognition conditions: different shape results of baked products, camera angle to the subject, lightning level, location of the ingredients on the product, etc. In the situation of gathered data limits, it is necessary to generate additional pictures from the available dataset. Augmentation techniques, such as rotation, translation, zoom, flips, shear, mirror, and color perturbation, allows increasing initial dataset and as the result improve machine learning model performance [13]. In order to use the high performance of neural networks in classification problems with limited image datasets, it is worth to use pre-trained deep learning models.

The idea is using pre-trained networks and adds classification layers that are trained on available images. To solve the task of recipe recognition the following models were used:

1. VGG16 [14]
2. Xception [15]
3. ResNeXt [16]

The problem of the lack of data can be solved with different open-source frameworks for augmentation and in this experiment, the *Albumentations* library was used in order to generate additional images for classification layers fitting [17], because it works fast and has a simple program interface. Base features that helped to create new images are horizontal and vertical blur the input image, resizing and randomly applying affine transforms (scaling and rotating).

In model fitting, all layers were first frozen and classifier layers were added, then after several epochs, some of the last layers of the base model were defrosted and the fitting process continued.

3. Results

The metrics of validation and test stages are illustrated bellow with a type of data set splitting, model size, average model fitting time and main experiment scores.

Table 1. The result metrics of experiment

Pre-trained model	Dataset split (validation/test)	Model size (Mb)	Average model fitting time (sec.)	Accuracy	F1 score	Precision	Recall
VGG16	val.	118	400	0.576	0.834	0.940	0.751
Xception	val.	150	188	0.859	0.949	0.948	0.950
Xception	test	150	188	0.515	0.802	0.755	0.854
Resnext	val.	221	210	0.915	0.970	0.755	0.945
Resnext	test	221	210	0.828	0.966	0.974	0.958

VGG has a good level of recognition of sauces, but *Xception* and *Resnext* has not. However, there is no certainty that they are correctly compared in the context of available data. It may be worth waiting for gathering new images and re-training all three models for a more correct comparison capability.

At the moment, the best results in quality show *Resnext*, but it is the largest in size (221 MB). *Xception* is worse, but faster and easier, and it seems to have overfitted quite a bit, you can see it on quality of validation and test accuracy on new data dropped from 85% to 51%.

There is also a possibility that *Resnext* also overfitted, because of accuracy also decreased when testing on new data, and it does not determine the specific ingredients. It seems that model inference shows the usual ingredient recipe when the type of pizza is identified. In order to test this hypothesis, we need to collect more data that would consist of different types of pizza in one recipe, for example, the combination of the various halves of pizza.

Conclusion

At the moment, the best results were shown by a model based on *ResNext* architecture with added layers: global average pooling, a fully connected layer (512 neurons) with *Rectified Linear Unit* (*ReLU*) activation function, a dropout layer (probability equals 0.5) and a fully connected layer with sigmoid activation for the final determination of classes.

In a sample consisting of photos of pizzas after the oven, this model achieved quality in the proportion of correct answers in 86%. Correctly defined composition of ingredients is considered to be only one that is defined to the accuracy of each ingredient, i.e. incorrectly defined composition is considered to be one in which at least one ingredient is defined incorrectly.

The main mistakes were in identifying some hard-to-distinguish ingredients that either blend in color with the cheese or are coated with it completely. Difficulties in visually determining the presence of ingredients on some pizzas have arisen and people. These ingredients include chicken was similar in color to cheese, covered with cheese and ham were on some pizzas completely covered with cheese.

To see if the quality would improve if the ingredients could be seen better, training was done on photos of pizzas before the oven. As a result, the proportion of correct answers increased to 92% - which leads to the conclusion that better identification of ingredients without melted cheese can increase the quality of the definition of the model.

The problem of misidentifying some other ingredients like chorizo salami, jalapeno, pickles, etc. is related to their low representation in the data set. Most likely, the quality will improve when the model is fitted on more complete data.

This research revealed that artificial intelligence modules have the ability to be applied in the various business processes in order to improve the quality of provided services. The final decision about integrating modern technologies is possible after comparing the results of the impact of the human recognition time of a customized recipe on the entire process of production of products and service quality with artificial intelligence.

References

1. Nemkov R. M. Development of neural network algorithms for invariant pattern recognition: dissertation of the candidate of technical sciences: 05.13.18. Stavropol. — 2015. — P. 162.
2. Haykin S. Nejronnije seti: polnij kurs [Neural networks: full course]. — Moscow, 2008. — Pp. 113, 281-330.
3. Gonzalez R. Tsifrovaya obrabotka izobrazhenij [Digital image processing]. — Moscow, 2012. — P. 1104.
4. Ranzato M. A., Jarrett K., Kavukcuoglu K., LeCun Y. What is the best multi-stage architecture for object recognition? In ICCV, 2009.
5. Richard S. Sutton. Obuchenije s podkreplenijem [Reinforcement training]. — Moscow, 2011. — P. 399.
6. Ackley D. H., Hinton G. E., Sejnowski T. J. A learning algorithm for Boltzmann machines. Cognitive Science. — 1985. — Vol. 9. — Pp. 147-169.
7. Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning. — 2009. — Vol. 2. — Issue 1. — Pp. 1-127.
8. Ciresan D., Meier U., Schmidhuber J. Multicolumn Deep Neural Networks for Image Classification [R]. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR '12, pp.3642-3649, Washington, DC, USA, 2012. — IEEE ComputerSociety.
9. Ranzato M., Huang F., Boureau Y., LeCun Y. Unsupervised learning of invariant feature hierarchies with applications to object recognition [R]. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR’07), IEEE Press, 2007.
10. Gill F. Prakticheskaya optimizatsiya [Practical optimization]. — Moscow, 1985. — P. 509.

11. Izmajlov A. F. Chislennije metody optimizatsii [Numerical optimization methods]. — Moscow, 2005. — P. 304.

12. Tarkhov D. A. Nejrosetevije modeli i algoritmy [Neural network models and algorithms]. — Moscow, 2014. — P. 352.

13. Lemley J., Bazrafkan S., Corcoran P. Smart augmentation learning an optimal data augmentation strategy. — IEEE Access, 2017. — Vol. 5. — Pp. 5858-5869.

14. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. — arXiv:1409.1556.

15. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. — arXiv:1610.02357.

16. Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He Aggregated Residual Transformations for Deep Neural Networks. — arXiv:1611.05431.

17. Buslaev A., Parinov A., Khvedchenya E., Iglovikov V., Kalinin A. Albumentations: fast and flexible image augmentations. — arXiv:1809.06839.