Achieving Statistical Dependence of the CNN Response on the Input Data Distortion for OCR Problem*

I. M. Janiszewski[I], V. V. Arlazarov[II,III,IV], D. G. Slugin[I,II]

[I]Federal Research Center Computer Science and Control of Russian Academy of Sciences, Moscow, Russia
[II]Smart Engines Service LLC, Moscow, Russia
[III]Institute for Information Transmission Problems of Russian Academy of Sciences, Moscow, Russia
[IV]Moscow Institute of Physics and Technology (State University), Moscow, Russia

Abstract. The paper proposes an approach to training a convolutional neural network using information on the level of distortion of input data. The learning process is modified with an additional layer, which is subsequently deleted, so the architecture of the original network does not change. OCR of data based on the MNIST dataset distorted with Gaussian blur using LeNet5 architecture network is considered. This approach does not have quality loss of the network and has a significant error-free zone in responses on the test data which is absent in the traditional approach to training. The responses are statistically dependent on the level of input image's distortions and there is a presence of a strong relationship between them.

Keywords: Convolutional neural networks, pattern recognition, machine learning, distortion, Gaussian blur, OCR, MNIST

DOI 10.14357/20718632190409

Introduction

Nowadays the mobile devices market, in particular smartphones, grows explosively. Typical modern devices already have enough high-productive processors and high-quality cameras, which allows to solve many problems using neural networks directly on the device, without cloud technologies [1, 2]. One of these tasks is document recognition in a video stream, part of which is optical character recognition (OCR), that is typically performed using a classifying neural network. Convolutional neural networks are actively used in the task of character classification and recognition, as they provide a suitable balance between the complexity of the network (which affects the performance) and the recognition quality. The source data for such networks are symbols selected from the video stream. It is worth noting that the images of symbols are usually seriously different from the "ideal" in the original document due to the difference in lighting when shooting and glares from light sources, spatial distortions of the document on the frames, noise of the camera matrix, blurring the image at slow shutter speeds, shaking the device, and so on [3]. To improve the neural network classification, it is necessary not only to have a sufficient data set for training, but also to find a suita-

*The reported study was partially funded by RFBR according to the research projects 17-29-07093 and 17-29-03263.

ble augmentation model that can simulate various types of distortions and to supply the original set with missing examples [4, 5]. The result of the classifying neural network is a set of estimates of the symbol images, besides to the absolute values as a criterion for belonging to the class, these estimates can be interpreted as sort of "confidence" of the network in the classification. A certain threshold for estimates can be chosen empirically, below which the responses of the neural network are considered doubtful and require an additional analysis. Aside from the classification quality of the trained network (the percentage of correct answers on the test dataset), for practical use in real-world tasks, we would like to have the following features:

a. The incorrect classifications of the network have a low confidence. It helps to avoid the most of incorrect classifications by choosing a suitable threshold.

b. Monotonic dependence between confidence and the level of distortion of the original image. This is useful, for example, to integrate recognition results from many frames of video stream containing the same document, as it allows to make some assessment of the original image quality [6].

The paper proposes a method of learning convolutional neural network using the level of distortion in the construction of the extended network model. The original neural network has the same quality and obtains these properties.

1. Method Description

1.1. Problem Statement

First we consider an initial model M of the multi-layered convolutional network, which receives objects x_i, $i \in 1, \dots, N$, from K classes ($K > 1$). The network response is a vector $s_i = M(x_i)$, $s_i = \left\{s_i^{(1)}, \dots, s_i^{(K)}\right\}$

$$s_i^{(k)} = \frac{\exp\left(z_i^{(k)}\right)}{\sum_{l=1}^{K} \exp\left(z_i^{(l)}\right)}, \quad (1)$$

where $z_i = \left\{z_i^{(1)}, \dots, z_i^{(K)}\right\}$ is an output of the last rectified linear unit layer

$$ReLU\langle X\rangle(v) = \min\left(X, \max(0, v)\right) \quad (2)$$

of the model M. An additional point to emphasize is that it is possible to predict the result of classification as follows $c_i = \underset{k}{\operatorname{argmax}}\, s_i^k$.

The scheme of the model M is illustrated in Fig.1a.

Further, assume that each object x_i has some property with a value $q_i \in [a, b] \subset R^1$. Note that different objects can have the same property values. In general, our purpose is to construct and train a neural network with the following feature. Let's consider two property values $q_1, q_2 \in [a, b]$ such that $q_1 < q_2$. If this condition is fulfilled, then it is to be expected the fulfillment of (3)

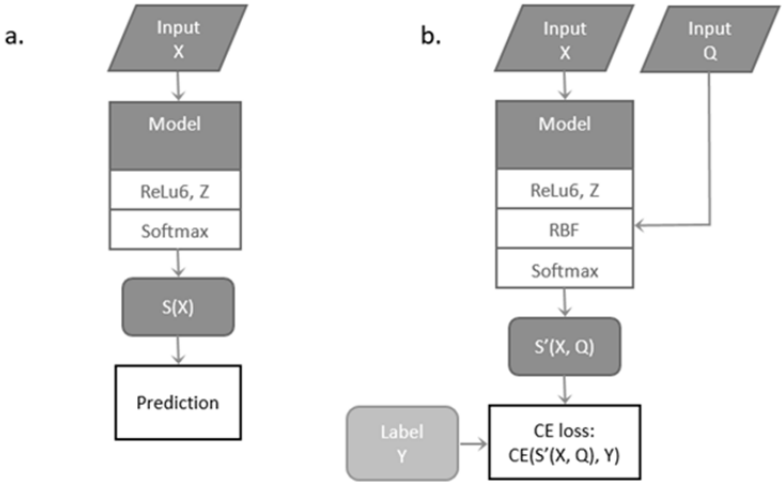

Fig. 1. Networks M and M' architecture

$$Q_\tau\left(s_i^{(c_i)}\middle|q_1\right) > Q_\tau\left(s_j^{(c_j)}\middle|q_2\right), \quad (3)$$

where $\tau \in (0,1)$ is a prespecified quantile, $Q_\tau\left(s_i^{(c_i)}\middle|q\right) = \inf\left\{s: P\left(s_i^{(c_i)} < s\middle|q_i = q\right) \geq \tau\right\}$.

1.2. Construction of Network

We proceed now to the construction of new model M', part of which will be the original model M. The model M' is schematically represented in the Fig.1b. A distinguishing feature of the model M' is that the response of the model M' is a vector $s_i' = M'(x_i, q_i)$. Additionally, we have used an intermediate non-trainable layer with radial basis function (RBF). Let us consider in detail, step by step, the construction of M'.

Step1. Divide the interval $[a, b]$ into n subintervals $a_0 = a < a_1 < \cdots < a_n = b$.

Step2. Select n values $p_1 > \cdots > p_n$, where $p_r \in [\frac{1}{K}; 1.0), r = 1, \ldots, n$. We call p_r "target value" for each x_i with distortion $q_i \in [a_{r-1}; a_r)$. This value can be interpreted as target confidence score for object x_i.

Step 3. Using logit function p

$$p(\zeta) = \frac{\exp(\zeta)}{\exp(\zeta)+(K-1)} \quad (4)$$

and selected p_r it is possible to obtain ζ_r with inverse logit function

$$\zeta_r = \ln\big(p_r \cdot (K-1)\big) - \ln(1-p_r) \quad (5)$$

It is easy to see that the expression (4) can be derived from (1) under the assumption that $z_i^{(k)} = 0, k \neq c_i$. If response $z_i^{(c_i)}$ tends to ζ_r the confidence score $s_i^{(c_i)}$will be close to target estimate p_r in case that other $z^{(k)}$ are around zero.

Step 4. Define function $\mu(q)$ by the following way

$$\mu(q) = \sum_{r=1}^{n} \zeta_r \cdot I_{[a_{r-1}, a_r)}(q) \quad (6)$$

where $I_{[a_{r-1}, a_r)}(q)$ is the indicator function for the corresponding interval.

Step 5. Introduce element-wise transformation on RBF layer where Gaussian function [7] is used as a radial-basis function:

$$G(z_i; A, \mu(q_i), \sigma)^{(k)} = A \cdot \exp\left(-\frac{\left(z_i^{(k)} - \mu(q_i)\right)^2}{2\sigma^2}\right) \quad (7)$$

Where A is a height of peak ($A > 0$), ζ_i is a center of peak and σ is a standard deviation ($\sigma > 0$).

Step 6. Categorical Cross Entropy Loss is used as a loss function (8) where $y_i^{(k)} = 1$ for correct class k and 0 otherwise.

$$L = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^{(k)} \cdot \log\left(s'^{(k)}_i\right) \quad (8)$$

Thus, the constructed model M' in learning process approximates the value of the vector $\{z_i\}$, corresponding to the correct class, to the center of the peak, as a result the confidence score of the model M falls into the neighborhood of the target estimate for the corresponding q_i. The remaining values of $\{z_i\}$ are initially small and according to the domain of definition do not go beyond the neighborhood of zero because of the monotonic decrease of the Gaussian function from the center of the peak. When the training is completed, the RBF layer is removed from the model M' and the network of the original M architecture is obtained.

2. Data Model

Well-known dataset of handwritten numbers MNIST [8] is used as a source data. As a model of data distortion, the Gaussian blur is considered [9]. Let's put the property $q = \sigma$ where σ is a standard deviation - parameter of Gaussian blur. The samples of image's distortion in shown in Fig. 2. The numbers of the training and testing datasets were increased using this augmentation in 10 times uniformly in interval $q \in [0; 4]$. As a result, size of train dataset is equal 600 000, the test is 100 000. It should be noted that for small values of q the distortion is practically absent as the weights of neighboring pixels are rather small (see Fig. 3). Because of that only samples with $q \geq 0.5$ was used for training.

3. Experiment Results

Consider the convolutional neural network like LeNet5 [10] with training on the selected data model. It consists of a combination of the following layers: a convolution, full connection and subsampling. A linear rectifier ReLU is used as an activation function, on the last layer there is a rectifier with a limit of the maximum value ReLU6. The visualization of the network architecture is shown in Fig. 4.

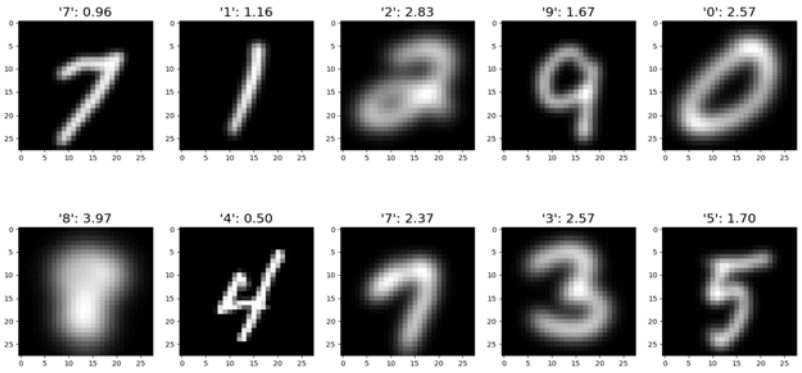

Fig. 2. Samples of data augmentation depending on distortion parameter

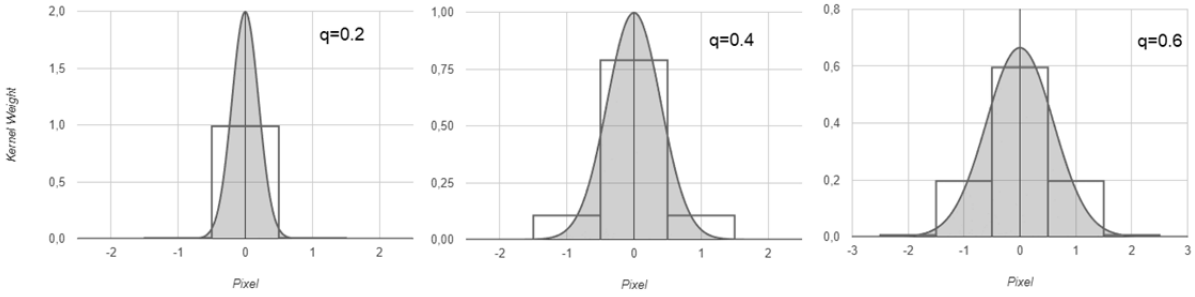

Fig. 3. Weights of neighboring pixels dependng on distortion parameter

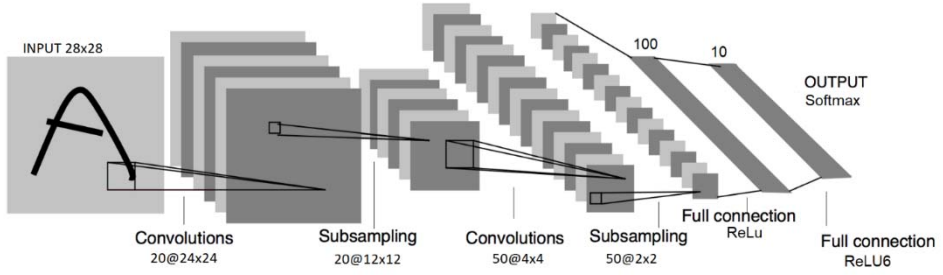

Fig.4. LeNet5 convolution network

The training parameters were the following:

- number of classes $K = 10$;
- number of trainable weights 46 680;
- distortion $q \in [0; 4]$;
- $p_{max} = 0.6$, $p_{min} = 0.3$;
- peak $A = 10$;
- deviation $\sigma = 0.7$.

The training was performed using Keras with a limit of 200 epochs. The results on the test dataset are summarized in Table 1, where M1 is the result of the network trained as is and M2 – as part of the M' model.

Table 1. Experiment's results

Network	Quality, %	Error-free rate, %	Spearman's correlation coefficient
M1	99.12	0	-0.11
M2	99.12	20.21	-0.87

The qualities of the trained networks M1 and M2 are equal. Consider the distribution of confidences for correct and incorrect classifications depending on the level of distortion in the form of a correlation field. Fig. 5 presents the field for the M1 network, Fig. 6 – for M2.

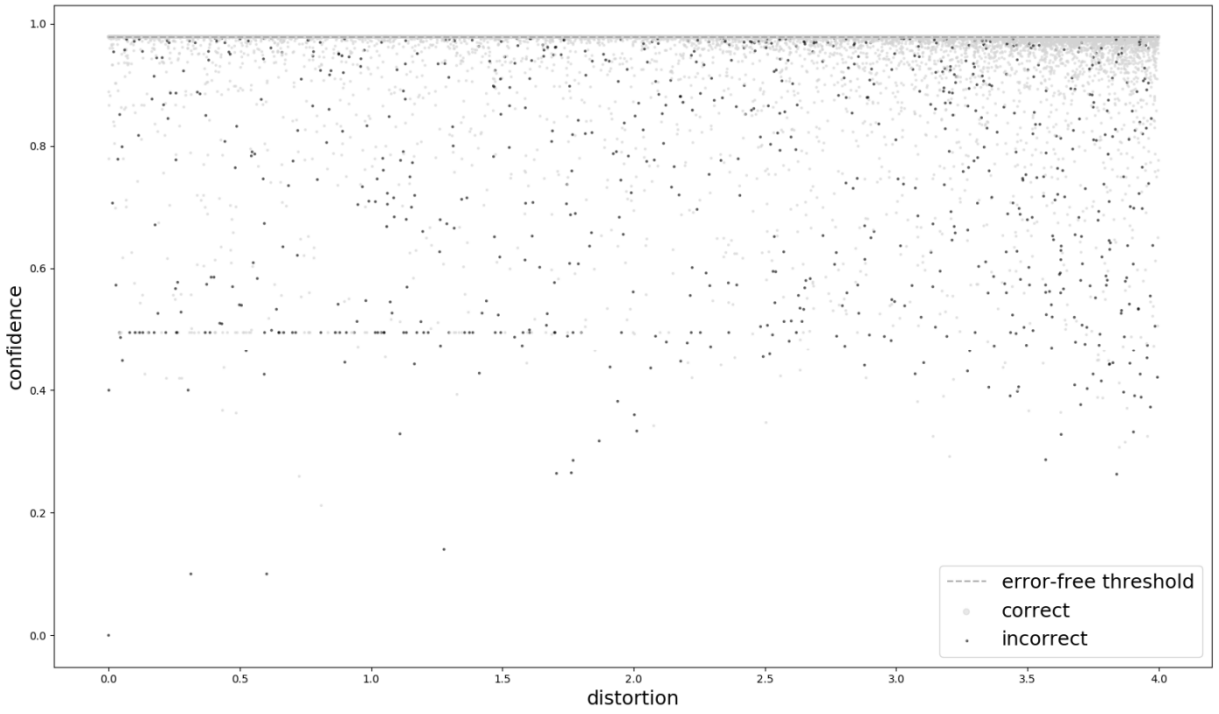

Fig. 5. Correlation field for M1

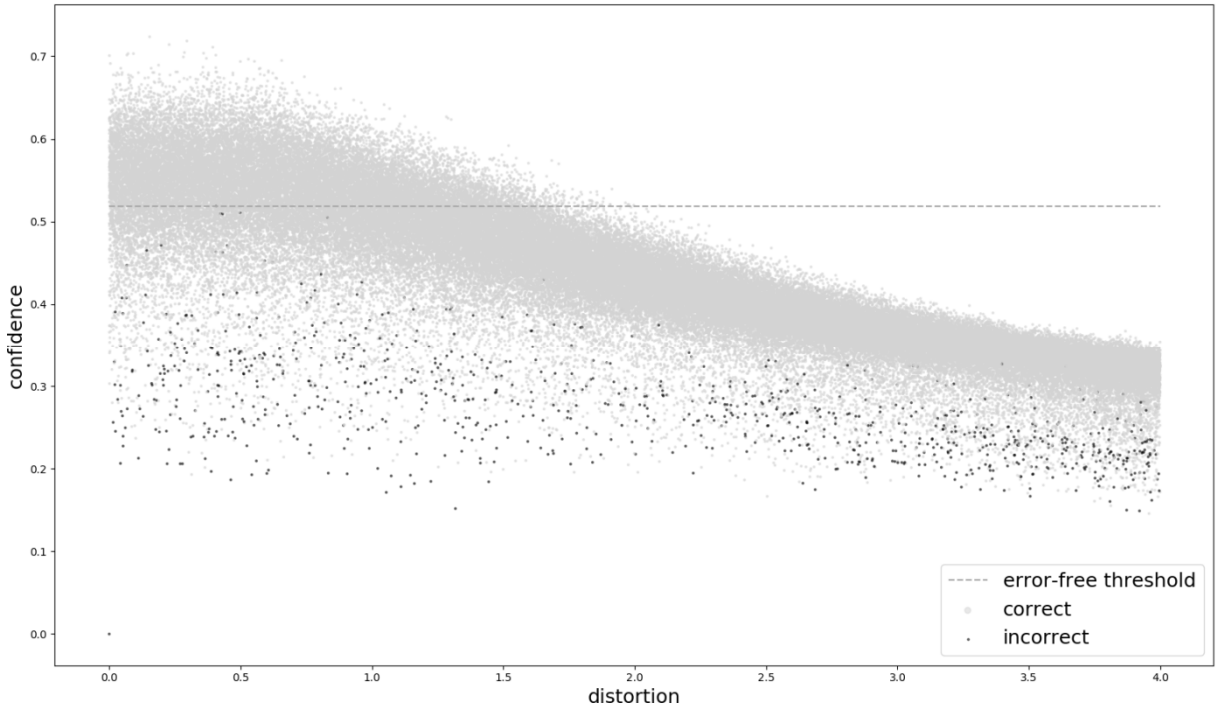

Fig. 6. Correlation field for M2

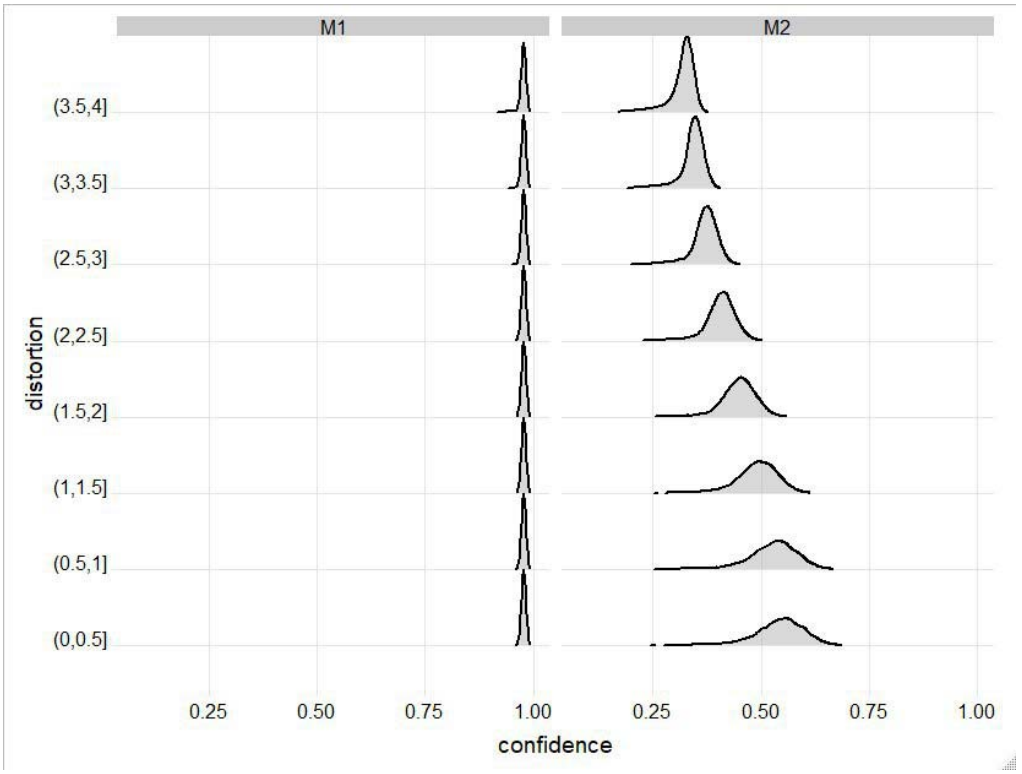

Fig. 7. Sample distributions

For test dataset the error-free threshold — the maximum value of the confidence score with incorrect classifications and the error-free rate — the percentage of correct classifications with confidence above the error-free threshold were calculated. These values characterize the level of trust in the network responses and show the presence or absence of incorrect classifications with an abnormally high level of confidence. The results from Table 1 tell us that M2 has a significant level of error-free rate while M1 does not have it at all. In contrast to M1, a correlation between the level of distortion and the confidence score is visually observed for the network M2. For verification of this assertion, we calculate the Spearman's correlation coefficient [11] for each network. The correlation coefficient for M1 is small. In the case of M2, coefficient is roughly as -1 that gives evidence of a strong inverse monotonic relationship between distortion and confidence.

To keep matters clear, consider in Fig. 7 a graphical representation of the sample distributions of points, forming a correlation field.

4. Regression Analysis

The method of analysis outlined in the last section shows only the presence of a statistical relationship between the level of distortion and network confidence. It is necessary to apply regression analysis to determine the form of dependence. Detailed studies have revealed that the normality of residuals requirement is not fulfilled for test data. Therefore, regression in mean isn't suitable. Therefore, we use quantile regression analysis [12] to verify a linear dependence of the network response on the level of distortion in terms of quantiles. This type of regression refers to nonparametric research methods, that is, which do not require a priori knowledge about the distribution of random variables. A remarkable feature of quantile regression is the major robustness of quantile methods versus classical least squares estimation.

The τ-quantile linear regression model may be expressed formally as follows:

$$Q_\tau(s) = \boldsymbol{q}^{\mathrm{T}}\beta(\tau) \qquad (9)$$

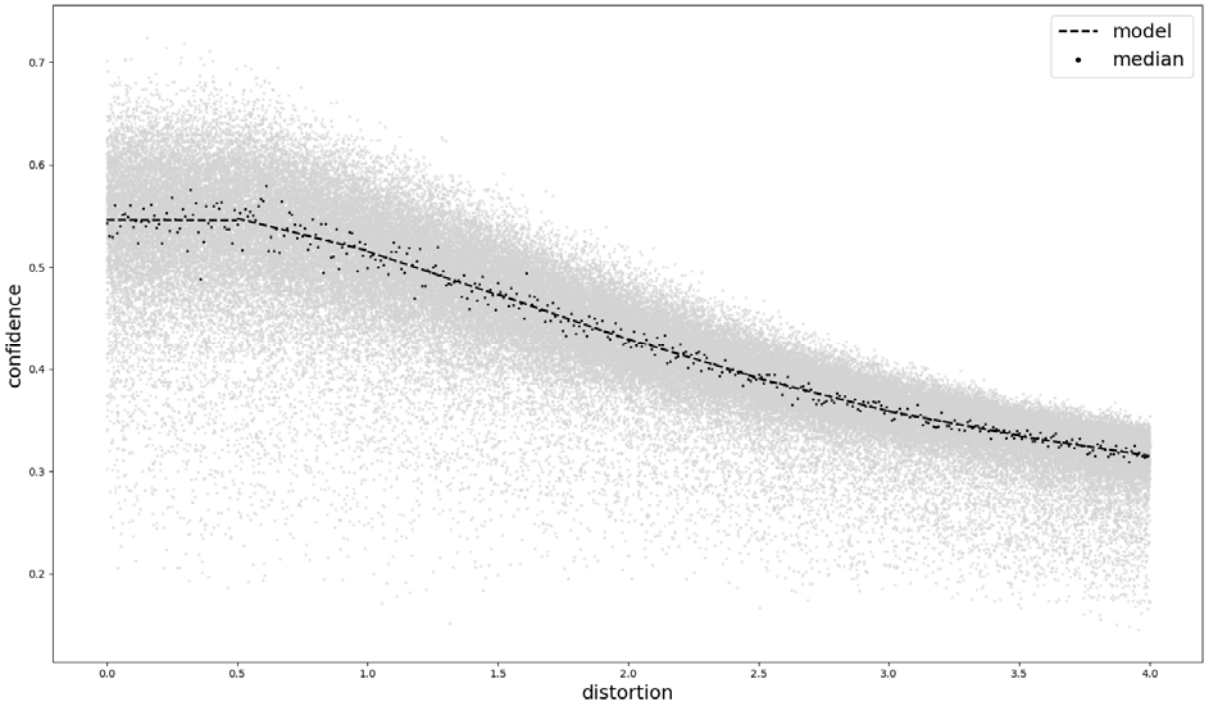

Fig. 8. Interval model of median regression

where $\boldsymbol{q}$ is a d-dimensional vector of covariates (including 1 as first element) and $\beta(p) = [\beta_0(\tau), \beta_1(\tau), \dots, \beta_d(\tau)]^{\mathrm{T}}$ is a vector of coefficients. Evaluation can be carried out using the so called "check-function"

$$\rho_\tau(u) = u\big(\tau - I(u < 0)\big) = \begin{cases} u\tau & \text{if } u \geq 0, \\ u(\tau - 1) & \text{if } u < 0, \end{cases} \quad (10)$$

where I represents the indicator function of an event. It allows us to use techniques of linear programming to find a solution. Given a set of observations $\{(\mathrm{q_i}, \mathrm{s_i}),\ \mathrm{i} = 1, \dots, N\}$, the sample regression quantile can be obtained by minimizing

$$\sum_{i=1}^{N} \rho_\tau\big(s_i - \boldsymbol{q}_i{}^{\mathrm{T}}\mathbf{b}\big) \quad (11)$$

We restrict the discussion to the case of $\tau = 0.5$, i.e., the median regression, and $d = 2$. This means that at this distortion level, the network responses are equally likely to be greater or less than the value of the regression model. It is possible to build the models for different quantiles and then calculate the confidence interval of network responses for a given probability.

The studies demonstrated that there is no global linear regression model. Nevertheless, we shall divide the entire interval of values $q \in [0; 4]$ into some sub-intervals and work out models for each of them. Fig. 8 shows quantile regression lines for each interval, as well as separately calculated real median values.

Obviously, it is necessary to justify the using of quantile regression model by applying a corresponding goodness-of-fit test. Let $\hat{\beta}$ be the τ-th quantile estimate of model (9). Let $r_i = s_i - \boldsymbol{q}_i{}^{\mathrm{T}}\hat{\beta}$ denote the residuals. Let us consider

$$R_N(\boldsymbol{t}) = N^{-1/2} \sum_{j=1}^{N} \psi\big(r_j\big)\boldsymbol{q}_i I(\boldsymbol{q}_i \leq \boldsymbol{t}) \quad (12)$$

where $\psi(r) = \tau I(r > 0) + (\tau - 1)I(r < 0)$. We can define the test statistic [13]

$$T_n = \max_{\|a\|} N^{-1} \sum_{i=1}^{N} \Big(a^{\mathrm{T}} R_N(\boldsymbol{q}_i)\Big)^2 \quad (13)$$

as the largest eigenvalue of $N^{-1} \sum_{j=1}^{N} R_N(\boldsymbol{q}_i) R_N^{\mathrm{T}}(\boldsymbol{q}_i)$.

We have been carried out computational estimation of quantile functions using the package Qtools [14]. As results in Table 2 indicate, the hypothesis of linearity of medians will not be rejected with significance level not less than 0.13.

Conclusion

The paper proposes a method of convolutional networks training using additional data of the training sample such as level of distortion of input data.

Table 2. Calculated p-values depending on interval

Interval	[0;0.5]	[0.5;1.0]	[1.0;1.5]	[1.5;2.0]	[2.0;2.5]	[2.5;3.0]	[3.0;3.5]	[3.5;4.0]
p-value	0.34	0.93	0.90	0.13	1.0	0.77	1.0	1.0

The experiment shows that the network trained by this way does not loss in quality, save its original architecture and obtains some additional properties such as the statistically dependence between network responses and level of distortion, as well as a significant error-free rate. In addition, the regression model of the dependence of confidence on the level of distortion was built to show a strong linear correlation in terms of medians. Nowadays many convolutional networks use only synthetic training data [15], which allows to create complex augmentation models and apply their parameters in training. The proposed method is promising in terms of building reliable OCR systems with a high level of confidence in responses that is very important for OCR of identity documents.

References

1. K. B. Bulatov, V. V. Arlazarov, T. S. Chernov, O. A. Slavin and D. P. Nikolaev, “Smart IDReader: Document Recognition in Video Stream”, ICDAR2017, IEEE Computer Society, ISSN 2379-2140, ISBN 978-15-38635-86-5, pp. 39-44, 2017. DOI: 10.1109/ICDAR.2017.347
2. M. M. Luqman, P. Gomez-Kramer, and J.-M. Ogier, “Mobile Phone Camera-Based Video Scanning of Paper Documents”. Cham: Springer International Publishing, 2014, pp. 164–178.
3. V. V. Arlazarov, A. Zhukovsky, V. Krivtsov, D. Nikolaev, and D. Polevoy, “Analysis of using stationary and mobilesmall-scale digital cameras for documents recognition,” Information Technologies and Computing Systems (3), 71–81 (2014). (in Russian)
4. A. V. Gayer, A. V. Sheshkus and Y. S. Chernyshova, “Effective real-time augmentation of training dataset for the neural networks learning”, ICMV 2018, 11041 ed., SPIE, vol. 11041, 2019. DOI: 10.1117/12.2522969
5. L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning”, CoRR abs/1712.04621 (2017).
6. K. B. Bulatov, A. E. Lynchenko and V. E. Krivtsov, “Optimal frame-by-frame result combination strategy for OCR in video stream”, ICMV 2017, 10696 ed., SPIE, Apr. 2018, vol. 10696, 758 pp., ISBN 978-15-10619-41-8, 106961Z, 2018, DOI: 10.1117/12.2310139
7. Buhmann, Martin Dietrich, “Radial basis functions : theory and implementations”. Cambridge University Press. ISBN 978-0511040207 (2003).
8. The MNIST database of handwritten digits. URL: http://yann.lecun.com/exdb/mnist
9. Shapiro, L. G. & Stockman, G. C: "Computer Vision", page 137, 150. Prentice Hall, 2001
10. LeCun, Yann; Léon Bottou; Yoshua Bengio; Patrick Haffner (1998). "Gradient-based learning applied to document recognition" (PDF). Proceedings of the IEEE. 86 (11): 2278–2324. DOI:10.1109/5.726791
11. Daniel, Wayne W. (1990). “Spearman rank correlation coefficient”. Applied Nonparametric Statistics (2nd ed.). Boston: PWS-Kent. pp. 358–365. ISBN 978-0-534-91976-4
12. Koenker and G. Bassett, Jr. “Regression Quantiles” Econometrica, Vol.46 No1 (January, 1978)
13. He XM, Zhu LX. “A lack-of-fit test for quantile regression”. Journal of the American Statistical Association (2003);98(1):1013-1022. DOI: 0.1198/016214503000000963
14. Geraci M. “Qtools: A collection of models and tools for quantile inference”. The R Journal (2016), 8(2), 117-138. DOI:10.32614/RJ-2016-037
15. Yulia S. Chernyshova, Alexander V. Gayer, and Alexander V. Sheshkus "Generation method of synthetic training data for mobile OCR system", Proc. SPIE 10696, Tenth International Conference on Machine Vision (ICMV 2017). DOI: 10.1117/12.2310119

Janiszewski I.M. Senior researcher at Institute for Systems Analysis, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia. PhD. Number of publications: 22. Research interests: machine learning, algorithms, statistical modeling, applied probability, pattern recognition, stochastic models. E-mail: yanishevsky@isa.ru

Slugin D.G. Researcher at Institute for Systems Analysis, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia. Number of publications: 11. Research interests: machine learning, image processing, pattern recognition. E-mail: slugindm@gmail.com

Arlazarov V.V. Lead researcher at IITP RAS, Moscow, Russia. Lecturer at MIPT, Moscow, Russia. PhD. Number of publications: 70. Research interests: artificial intelligence, machine learning, recognition systems, information technology, queueing theory. E-mail: vva777@gmail.com