REDUCING OVERCONFIDENCE IN NEURAL NETWORKS BY DYNAMIC VARIATION OF RECOGNIZER RELEVANCE

Konstantin Bulatov
Dmitry Polevoy
Institute for Systems Analysis of Russian Academy of Sciences (ISA RAS)
Prospekt 60-letiya Oktyabrya, 9, Moscow, 117312, Russia
National University of Science and Technology "MISIS" (MISIS)
Leninskiy prospekt 4, Moscow, 119049, Russia
E-mail: hpbuko@gmail.com

KEYWORDS

Optical character recognition, combination of recognizers, recognition post-processing

ABSTRACT

Contemporary recognition systems use various methods of symbol recognition and post-processing methods designed for enhancing the quality of text recognition. For some recognition problems it may be difficult to create an adequate dataset for training symbol recognizers, so several symbol recognizers are used to ensure better performance. In this paper the concept of recognizer relevance is introduced as a way of analysing the recognizer output. A method is described using this concept, allowing to use external information about the input samples in order to balance the contributions of the recognizer and the post-processing subsystem.

INTRODUCTION

For any recognition algorithm based on machine learning the training dataset is always an approximation of the universal set of test cases. The goal of the recognizer training, whether it is an artificial neural network (Haykin, 1998), or support vector machine (Hastie et al., 2001), or another technique, is almost always a minimization of some error function on a training dataset. One of the most important problems which has to be dealt with during the recognizer training is local minima of this error function.

A number of techniques has been developed to avoid this problem. In the scope of artificial neural networks, such methods as stochastic gradient descent (Moulines and Bach, 2011; Zeiler and Fergus, 2013) have proven quite effective when dealing with machine learning problems, partly because they remedy the convergence of the error function to local minima.

The synthetic extension of the training dataset is also proposed (Nikolaev et al., 2014). This technique can be used in order to approximate the universal set of test cases more accurately.

Engineering approaches to avoid the local minima after the recognizer training (i.e. at the operational stage) include the combination of recognizers (Tulyakov et al., 2008; Slavin, 2004) and various techniques of recognition post-processing (Sholomov et al., 2005; Arlazarov V.L. et al., 2014).

Machine learning methods are commonly used in such a way that the classification result given the input sample is represented as a vector of probability estimations for each class. In the case of Artificial Neural Networks (ANNs) this is achieved using the softmax output unit, which can provide the probability estimates for classes consistent with the input and training data (Denker and Le Cun, 1991).

Yet recognition systems using ANNs for character recognition and post-processing techniques based on the probabilistic language models (such as Hidden Markov Models (Bouchaffra et al., 1997) or Weighted Finite-State Transducers (Llobet et al., 2010), or other methods of statistical post-processing (Sholomov et al., 2005, Arlazarov V.L. et al., 2014) often suffer from the inability to optimally calibrate the character recognizer output. Using single-vector recognizer output model a problem of recognizer overconfidence occurs. That means that even if the symbol recognition result is false the distribution of the estimations provided by the recognizer can be strongly shifted to the probability estimation of the incorrect class. In that case the post-processing techniques have less chance to correct the recognition result using the language model of the recognized text.

In this paper we propose a model of recognizer relevance and a way of looking at the recognizer output not as a vector of class probabilities, but as a vector-function of the recognizer relevance argument. This model allows us to vary the classification information output of the symbol recognizer. By knowing the properties of the recognizer behaviour given the different nature of the input data we can decrease the recognizer relevance value for some characters if we want the post-processing algorithm to pay more attention for these characters.

RELEVANCE MODEL

Character recognition can be formally considered as a problem of classification of an input image to one of the K classes. A classifier (recognizer) is an agent that

Proceedings 29th European Conference on Modelling and Simulation ©ECMS Valeri M. Mladenov, Petia Georgieva, Grisha Spasov, Galidiya Petrova (Editors)
ISBN: 978-0-9932440-0-1 / ISBN: 978-0-9932440-1-8 (CD)

assigns a result vector $p=(p_1,...,p_K)$ to an input image x. We assume that $p_1,...,p_K$ are real numbers in range $[0,1]$ with $\sum_{i=1}^{K} p_i = 1$. The classification result is an index of a class and it corresponds to an index of the maximal item of a vector p.

Let's introduce the relevance as an abstract external argument of the recognizer, represented as a nonnegative real number $\tau \in [0,+\infty]$. Using this term we will now consider another model for the recognition output: instead of the result vector p the classifier now assigns a vector-function $p(\tau)=\left(p_1(\tau),...,p_K(\tau)\right)$ to an input image x. The relevance parameter τ represents the amount of information which will be provided to a system by the recognizer. The value $p(0)$ will correspond to a vector with zero classification information $\left(1/K,...,1/K\right)$. The value $p(+\infty)$ will correspond to a vector with maximum information. The relevance τ can also be regarded as parameter proportional to the inverse entropy of the classifier output.

The relevance model can be defined for different recognizers, but perhaps the most natural definition comes in terms of ANNs with softmax output unit. The softmax output unit of the ANN takes a vector of values $A=(a_1,...,a_K)$ as an input and performs the softmax transformation:

$$p_i = e^{a_i} \Big/ \sum_{j=1}^{K} e^{a_j}$$

Using the measurable or known in advance properties of the classes distribution, represented by positive "gain" values $g_1,...,g_K$, the softmax unit can be used to yield a probability distribution more suited to the desired recognition system (Denker and Le Cun, 1991). So, more generally:

$$p_i = e^{g_i a_i} \Big/ \sum_{j=1}^{K} e^{g_j a_j}$$

The relevance parameter τ can be introduced to the softmax output unit as follows:

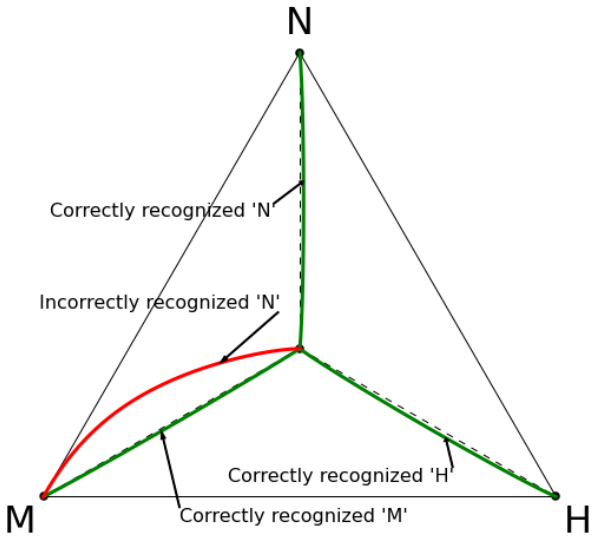
$$p_i(\tau) = e^{g_i a_i \tau} \Big/ \sum_{j=1}^{K} e^{g_j a_j \tau}$$

Note that all $p_i(\tau)$ are real number in range $[0,1]$, with fixed τ all $p_i(\tau)$ sum up to 1, and with $\tau = 0$ all $p_i(\tau)$ equal to $1/K$. If the vector A had only one maximal value a_i then with τ approaching $+\infty$, $p_i(\tau)$ approaches 1, where i is class index corresponding to the recognition result. If the vector A had L identical maximal values then all their corresponding values of $p(\tau)$ approach the value $1/L$.

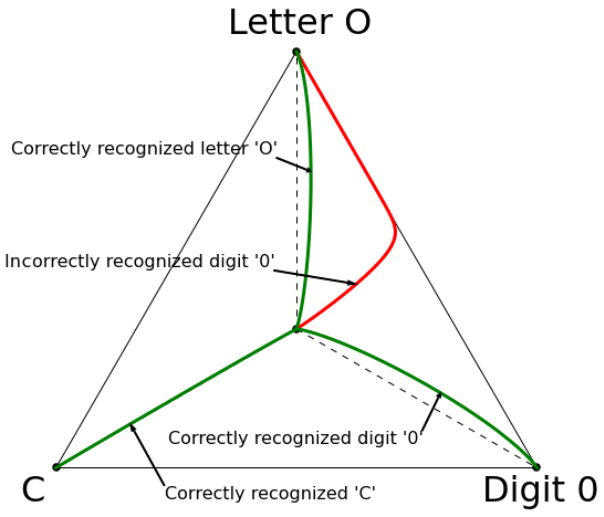
This parameterization can be seen as a 'gamma-correction' of the output pseudo-probability estimates.

Recognizer output vectors can be seen as points in K-dimensional space belonging to the $(K-1)$-dimensional simplex with it's center in $\left(1/K,...,1/K\right)$ and vertices in $(1,0,...,0)$, $(0,1,0,...,0)$, etc. The vector-function $p(\tau)$ can be represented as a curve connecting the center of the simplex ($\tau = 0$) with the center of its $(L-1)$-dimensional facet ($\tau = +\infty$) where L is the number of identical maximal values in the vector A passed as an input to the softmax unit.

Fig. 1 and 2 show the sample curves corresponding to the recognizer outputs on several input samples.



Figures 1: Sample Recognizer Output Vector-Functions, Projection on the 'M'-'N'-'H' Facet



Figures 2: Sample Recognizer Output Vector-Functions, Projection on the 'C'-'O'-'0' Facet

The analysed recognizer is a convolutional ANN trained to recognize projectively distorted symbols of the OCR-B font. The alphabet consists of 37 symbols (decimal digits, Latin letters and one special character). Fig. 1 shows the orthogonal projection of the recognizer output vector-function curves to the 2-dimensional facet of the simplex containing the characters 'M', 'N' and 'H'. Fig. 2 shows the projection to the 2-dimensional facet containing the characters 'C', 'O' (letter) and '0' (digit).

As shown in Fig. 1, one of the samples of the letter 'N' was incorrectly recognized as 'M'. The curve corresponding to the recognizer output connects the center of the simplex with its 0-dimensional facet, corresponding to the answer 'M' (that means that there was only one maximal value of the softmax unit input vector). However, it can be observed that the curve is significantly drawn to the direction of the 'N' vertex. More so in Fig. 2 - letter 'O' and digit '0' in the OCR-B font is poorly distinguishable, so even the correctly recognized samples show curves which are slightly drawn to the direction of the similar symbol vertex. The curve of the Incorrectly recognized sample of the digit '0' is very noticeably drawn to the 'correct' vertex, though in the end it still approaches the vertex corresponding to the letter 'O'.

In the original model of the recognizer outputs only one point on the curve is taken as a classification result. Using the whole vector-function of the recognition result we can vary the amount of information obtained from the symbol recognizer, thus allowing more space to other methods, such as recognition post-processing techniques.

APPLICATION OF THE MODEL

We performed our experiments for the system designed for machine-readable zone (MRZ) recognition on images obtained from a camera of a mobile device. These images often suffer from bad lighting conditions and various distortions, such as motion blur and defocus (Arlazarov V.V. et al., 2014).

Recognition errors were more frequently observed on poorly-focused character images. Since the recognition results even in those cases often have shifted class estimations distribution, the post-processing method based on the MRZ language model sometimes fails to correct those errors.

Prior to the recognition we calculate the quality estimation for each character image. First, we estimate the gradient of the image I using derivative images in four directions: vertical, horizontal and two diagonal:

$$G^{V}_{i,j}(I) = I_{i+1,j} - I_{i,j}$$

$$G^{H}_{i,j}(I) = I_{i,j+1} - I_{i,j}$$

$$G^{D1}_{i,j}(I) = \left(I_{i+1,j+1} - I_{i,j}\right)\Big/\sqrt{2}$$

$$G^{D2}_{i,j}(I) = \left(I_{i,j+1} - I_{i+1,j}\right)\Big/\sqrt{2}$$

The quality estimation of an image I is then calculated as a minimal 0.95-quantile of these derivatives:

$$Q(I) = \min\begin{Bmatrix} q\left(G^{V}(I)\right), q\left(G^{H}(I)\right), \\ q\left(G^{D1}(I)\right), q\left(G^{D2}(I)\right) \end{Bmatrix}$$

Here q(G) is a 0.95-quantile of the derivative image, i.e. a minimal g such that 95% of the values $G_{i,j}$ are less than or equal to g.

The estimation $Q(I)$ can be used as a measure of image quality containing the information about contrast, defocus and motion blur. This estimation is proportional to the image contrast, and on defocused images all derivative values will be lower. On images with motion blur only derivative in some directions will be lower, that is why the final estimation $Q(I)$ is calculated as a minimum of four derivative quantiles.

Using the calculated quality estimations of the character images, we assigned the recognizer relevance parameter for each character cell as a linear function of $Q(I)$:

$$\tau(I) = a \cdot Q(I) + b$$

Here a and b are selected in such way that the character image with maximal quality estimation would be assigned with $\tau(I) = 1$ and a hypothetical character image with the lowest tolerable quality estimation $Q(I) = Q_0$ would be assigned with relevance $\tau(I) = \tau_0$.

Using this pre-evaluation of the character image quality in order to adjust a recognizer relevance score we make the post-processing algorithm use less information from the recognition result of the images with poor quality, relying more on the language model instead.

This method was tested on a reference dataset of 3000 unique images of the machine-readable zones, containing in total ~260'000 characters. The recognition precision of the system with the fixed relevance parameter ($\tau = 1$) is 99.30%. The variance of the relevance parameter depending on the quality estimation of the character image (with $\tau \in [\tau_0 = 0.6,\ 1]$) helped to achieve the precision of 99.67%.

CONCLUSION

In this paper a new model of recognizer output is described. Using this model the resulting probability distribution of classes can be more accurately analysed. With the help of external information about the nature of the input samples this models allows us to balance the contributions of the symbol recognizer and the post-processing subsystem to the final recognition result.

This method was successfully applied to the existing system for machine-readable zones recognition on images captured from a mobile device camera. Using the described method better recognition precision is achieved at a reference dataset.

FUTURE WORK

The combination of several symbol recognizers can also be improved using the model of recognizer output as a vector-function of the relevance argument. In this case the relevance can be regarded as an easily controlled parameter which allows to balance the contribution of the recognizers in combination.

The analysis of the curvature of the recognizer output vector-function can also be used to estimate the self-confidence (Arlazarov et al., 2013) of the symbol recognizer in order to create a reliable rejection criteria for the recognition system.

ACKNOWLEDGEMENTS

This work is supported by Russian Foundation for Basic Research (projects 13-07-12172, 14-07-00730).

REFERENCES

Arlazarov V.L.; A. Marchenko; D. Sholomov. 2014. “Cumulative contexts for the problem of recognition”. *Proceedings of Institute for Systems Analysis RAS,* Vol. 64, No. 4, 64-72 (In Russian).

Arlazarov V.V.; K.B. Bulatov; S.M. Karpenko. 2013. “Methods for recognition reliability estimation for the problem of embossed symbols recognition”. *Proceedings of Institute for Systems Analysis RAS,* Vol. 63, No. 3, 117-122 (In Russian).

Arlazarov, V.V.; A.E. Zhukovsky; V.E. Krivtsov; D.P. Nikolaev; D.V. Polevoy. 2014. “Analysis of compact stationary and mobile digital cameras usage for document recognition”. *Information technologies and computing systems*, No. 3, 71-78 (In Russian).

Bouchaffra, D.; V. Govindaraju; S.N. Srihari. 1997. “Postprocessing of Recognized Strings Using Nonstationary Markovian Models”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, V. 21, No. 10, 990-999.

Denker, J.S.; and Y. Le Cun. 1991. “Transforming neural-net output levels to probability distributions”. *Advances in Neural Information Processing Systems* 3, 853-859.

Hastie, T.; R. Tibshirani; J. Friedman. 2001. *The Elements of Statistical Learning*. Springer New York Inc., N.Y., USA.

Haykin, S. 1998. *Neural Networks - A Comprehensive Foundation.* 2nd, Prentice Hall PTR, Upper Saddle River, N.J., USA

Llobet, R.; Cerdan-Navarro J.-R.; Perez-Cortes J.; Arlandis J. 2010. “OCR Post-processing Using Weighted Finite-State Transducers”. *Pattern Recognition (ICPR),* 2021-2024.

Moulines, E.; and F.R. Bach. 2011. “Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning”. *Advances in Neural Information Processing Systems* 24, 451-459.

Nikolaev, D.P.; D. Polevoy; N. Tarasova. 2014. “Training dataset synthesis for the problem of text recognition in 3D space”. *Information technologies and computing systems,* Vol. 3, 82-88 (In Russian).

Sholomov, D.L.; V.V. Postnikov; A.A. Marchenko; A.V. Uskov. 2005. “OCR result post-processing using partially defined syntax”. *Proceedings of Institute for Systems Analysis RAS, Intellectual information technologies - Concepts and Instruments,* Vol. 16, Moscow, KomKniga, 146-165 (In Russian).

Slavin, O.A. 2004. “Combined recognition methods for printed and hand-printed symbols”. *Proceedings of Institute for Systems Analysis RAS, Document processing - Concepts and instruments,* 151-174 (In Russian).

Tulyakov, S.; S. Jaeger; V. Govindaraju; D. Doermann. 2008. “Review of Classifier Combination Methods”. *Machine Learning in Document Analysis and Recognition - Studies in Computational Intelligence,* Vol. 90, 361-386.

Zeiler, M.D.; and R. Fergus. 2013. “Stochastic Pooling for Regularization of Deep Convolutional Neural Networks”. *CoRR,* abs/1301.3557.

AUTHOR BIOGRAPHIES

KONSTANTIN B. BULATOV was born in Petrozavodsk, Russia, went to the National University of Science and Technology "MISIS" in Moscow to study Applied Mathematics, obtained his Specialist degree in 2013 and entered a PhD program there. In 2014 he became a junior researcher at Institute for Systems Analysis of Russian Academy of Sciences. His research interests are optical document recognition and computer vision.
His e-mail address is: hpbuko@gmail.com.

DMITRY V. POLEVOY was born in 1981. He went to the Moscow Institute of Physics and Technology (MIPT) in 1998 and obtained his degree of Master of applied mathematics and physics in 2004. In 2007 he obtained his PhD in tabular documents recognition. In 2009 he became a senior researcher at Institute for Systems Analysis of Russian Academy of Sciences.
His e-mail address is: dvpsun@gmail.com.