MSC 62B15 DOI: 10.14529/mmp250209

THE IMPACT OF DATASET SIZE ON THE RELIABILITY OF MODEL TESTING AND RANKING

A.V. Chuiko[1]***, V.V. Arlazarov***[1,2]***, S.A. Usilin***[1,2]
[1]Federal Research Center "Computer Science and Control" RAS, Moscow, Russian Federation
[2]LLC "Smart Engines Service", Moscow, Russian Federation
E-mail: a.chuyko@smartengines.com, vva@smartengines.com, usilin@smartengines.com

Machine learning is widely applied across diverse domains, with research teams continually developing new recognition models that compete on open datasets. In some tasks, accuracy surpasses 99%, and the differences between top-performing models are often marginal, measured in hundredths of a percent. These minimal differences, combined with the varying size of the benchmark datasets, raise questions about the reliability of model evaluation and ranking. This paper introduces a method for determining the necessary dataset size to ensure robust hypothesis testing for model performance. It also examines the statistical significance of accuracy rankings in recent studies on MNIST, CIFAR-10, and CIFAR-100 datasets.

Keywords: dataset size; object recognition; statistical significance; model evaluation; recognition quality assessment.

Introduction

Recognition models play a crucial role in the modern world, being applied across various domains such as document recognition [1], forensics [2], mobile number recognition [3], and text recognition in road scenes [4]. Therefore, it is imperative to reliably assess their capabilities. The ability of a model to perform a given task is a critical measure of its quality. In the scientific community, benchmarks – comprising datasets and associated quality metrics – are commonly used to evaluate model performance. However, benchmarks often represent tasks beyond merely predicting corresponding dataset objects. We interpret the benchmark's task more broadly as making predictions for objects "similar" to those in the comprising datasets. Therefore, benchmark metrics estimate model performance but do not fully reflect its predictive abilities in broader contexts.

Commonly used benchmarks in computer vision include MNIST [5], CIFAR-10 [6], and CIFAR-100 [6], with accuracy as the standard metric for evaluating overall prediction accuracy across all classes.

MNIST consists of 60 000 training and 10 000 testing 28×28 grayscale images of handwritten digits across 10 classes. It has become a standard for evaluating machine learning models, including neural networks. Top models achieve accuracy exceeding 99%, with recent studies reporting results from 99,75% to 99,87% [7–12].

CIFAR-10 and CIFAR-100 are more complex datasets, each with 50 000 training and 10 000 testing images. CIFAR-10 consists of 32×32 color images across 10 classes (e.g., airplanes, cars, animals), while CIFAR-100 has images labeled into 100 classes (and 20 superclasses). These datasets are used to evaluate classification algorithms in more challenging scenarios. The highest accuracy on CIFAR-10 reaches 99,612% [13], and only 96,08% on CIFAR-100 [14].

In the field of machine learning, datasets such as MNIST, CIFAR-10, and CIFAR-100 have historically served as important benchmarks for evaluating the performance of various models. These datasets have played a pivotal role in the development of recognition algorithms by providing researchers with a common platform to test and compare their algorithms. However, as models have achieved near-perfect accuracy on these benchmarks, their use for comparing models has become less effective. Marginal improvements in accuracy on these datasets no longer correspond to significant practical advancements, raising questions about the relevance of continued competition on them. For some models, the difference in accuracy on the aforementioned datasets is on the order of 0,1%. In this work, we demonstrate that such advantages in accuracy between one model and another are often not statistically significant when considering the broader task at hand for datasets such as MNIST, CIFAR-10, and CIFAR-100.

This paper explores the relationship between test set accuracy and task performance. We present a method to determine the necessary dataset size for valid hypothesis testing of whether a model belongs to a high-accuracy class and provide a formula for assessing the statistical significance of accuracy differences between models. This analysis is applied to models validated on MNIST, CIFAR-10, and CIFAR-100. Scripts for calculating the main formulas in this work are available at https://github.com/AlexanderChuiko/Impact-of-Dataset-Size.

1. Mathematical Model

Thus, let us consider a feature space X, whose points represent recognizable objects, and a response space Y. We assume that a probability measure π and a loss function $L : Y \times Y \to \{0, 1\}$ are defined on $X \times Y$:

$$L : (r, y) = \begin{cases} 0, & r = y; \\ 1, & r \neq y. \end{cases} \tag{1}$$

The model m is a mapping from X to Y. We define the loss of the model on a pair $(x, y) \in X \times Y$ as the expression $L(m(x), y)$, and we assume that the function $L(m(x), y)$ is measurable. Let us introduce the accuracy of the model m as the inverse of the expected value of its loss:

$$A_m = 1 - E_\pi[L(m(x), y) \mid (x, y) \in X \times Y]. \tag{2}$$

Denoting $g_m(x, y) = 1 - L(m(x), y)$, we obtain

$$A_m = E_\pi[g_m(x, y) \mid (x, y) \in X \times Y]. \tag{3}$$

The empirical accuracy A_m^t, which is used to evaluate the performance of the model m, is computed based on the test sample $O_t \subset X \times Y$:

$$A_m^t = \sum_{j \in O_t} g_m(j) / \left|O_t\right|, \tag{4}$$

where $|O_t|$ denotes the size of the test sample. Note that the empirical accuracy A_m^t coincides with the quality metric Accuracy. We will further assume that the random variables $\{g_m\}$ are mutually independent for all considered models.

2. Assessment of Required Dataset Size and Statistical Significance of Ranking

This section addresses two key questions regarding model accuracy in practical applications. The first question concerns classifying a model as either high-quality or low-quality, based on whether the distribution of g_m for model m falls within one of two classes: $A_m \geq p_0$ or $A_m \leq p_1 < p_0$. The second question examines the "advantage" between two models m_1 and m_2, specifically whether the difference $A_{m1} - A_{m2}$ is positive. We begin with the first question.

2.1. Model Classification: High-Quality vs Low-Quality

High-quality models satisfy $A_m \geqslant p_0$, while low-quality models meet $A_m \leqslant p_1 < p_0$, where p_0 and p_1 are fixed thresholds. Models with $A_m \in (p_1, p_0)$ are considered unclassifiable, but this is not crucial for our analysis.

The goal is to classify any model with $A_m \notin (p_1, p_0)$ while maintaining a Type I error rate α and a Type II error rate β. When testing the hypothesis $H_0 : A_m \geqslant p_0$ against the alternative $H_1 : A_m \leqslant p_1 < p_0$, the most difficult cases occur when $A_m = p_0$ and $A_m = p_1$, as the distributions of g_m are closest in these scenarios.

We first consider testing the simple hypothesis $\hat{H}_0 : A_m = p_0$ against the simple alternative $\hat{H}_1 : A_m = p_1$. Let us construct a likelihood ratio criterion. The quantity A_m^t represents the sample mean of indicators of the model's correct predictions on the test sample. Therefore, under the assumption that the null hypothesis H_0 is true, the quantity $|O_t| \cdot A_m^t$ follows a binomial distribution $Bin(|O_t|, p_0)$. Similarly, under the assumption that the alternative hypothesis H_1 is true, this quantity follows a binomial distribution $Bin(|O_t|, p_1)$. Assuming that the sample size $|O_t|$ is sufficiently large, we approximate the distributions $Bin(|O_t|, p_0)$ and $Bin(|O_t|, p_1)$ by a normal distribution:

$$\begin{array}{l} |O_t|A_m^t \sim \mathcal{N}(|O_t|p_0,\ |O_t|p_0(1-p_0)),\ \text{under } H_0; \\ |O_t|A_m^t \sim \mathcal{N}(|O_t|p_1,\ |O_t|p_1(1-p_1)),\ \text{under } H_1. \end{array} \tag{5}$$

The likelihood ratio for A_m^t after simplification will take the form:

$$\Lambda(A_m^t) = \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}} \exp\left(-\frac{(A_m^t - p_0)^2}{2\frac{p_0(1-p_0)}{|O_t|}} + \frac{(A_m^t - p_1)^2}{2\frac{p_1(1-p_1)}{|O_t|}} \right). \tag{6}$$

It can be shown that this is an increasing function of A_m^t for $0 \leqslant A_m^t \leqslant 1$ and $p_1 < p_0$. Then, a criterion of the form $\Lambda(A_m^t) \geqslant \lambda_0$ is equivalent to a criterion of the form $A_m^t \geqslant \lambda_1$, given the corresponding values of λ_0 and λ_1. Given that $\frac{\sqrt{|O_t|}(A_m^t - p_0)}{\sqrt{p_0(1-p_0)}} \sim \mathcal{N}(0,1)$ under the true hypothesis H_0, for the Type I error rate α, we obtain the likelihood ratio criterion:

$$\begin{array}{ll} H_0: & A_m^t \geqslant p_0 + z_\alpha \sqrt{p_0(1-p_0)/|O_t|}, \\ H_1: & \text{otherwise}, \end{array} \tag{7}$$

where z_α is the quantile of level α of the standard normal distribution.

We compute the required size of the test sample $|O_t|$ such that the Type II error is no greater than β:

$$P_{H_1}\left(A_m^t \geqslant p_0 + z_\alpha \sqrt{p_0(1-p_0)/|O_t|} \right) \leqslant \beta, \tag{8}$$

Transforming the expression, we obtain:

$$P_{H_1}\left(\frac{\sqrt{|O_t|}\,(A_m^t - p_1)}{\sqrt{p_1(1-p_1)}} \geqslant \frac{\sqrt{|O_t|}\left(p_0 - p_1 + z_\alpha\sqrt{p_0(1-p_0)/|O_t|}\right)}{\sqrt{p_1(1-p_1)}}\right) \leqslant \beta. \tag{9}$$

Given that $\dfrac{\sqrt{|O_t|}\,(A_m^t - p_1)}{\sqrt{p_1(1-p_1)}} \sim \mathcal{N}(0,1)$ under the true hypothesis H_1, we obtain:

$$\frac{\sqrt{|O_t|}\left(p_0 - p_1 + z_\alpha\sqrt{p_0(1-p_0)/|O_t|}\right)}{\sqrt{p_1(1-p_1)}} \geqslant z_{1-\beta}, \tag{10}$$

from which we can express:

$$|O_t| \geqslant \left(\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}}{p_0 - p_1}\right)^2. \tag{11}$$

By the Neyman–Pearson lemma, the obtained criterion is the most powerful. Therefore, the size of the test sample, which is obtained by rounding up the right-hand side of expression 11, is the minimal size required to construct a criterion with Type I and Type II errors at levels α and β, respectively.

It is evident that for $A_m \geqslant p$, for any fixed threshold c, it holds that $P_{A_m}\left(A_m^t \geqslant\ c\right) \geqslant P_p\left(A_m^t \geqslant\ c\right)$. Therefore, when applying the criterion 7 to test the hypothesis $H_0 : A_m \geqslant p_0$ against the alternative $H_1 : A_m \leqslant p_1 < p_0$, the Type I and Type II errors will also be at levels α and β, respectively.

2.2. Hypothesis of Advantage

Consider two models m_1 and m_2 with accuracies A_{m1}^t and A_{m2}^t, where $A_{m1}^t > A_{m2}^t$ on the test set. We investigate the relationship between A_{m1} and A_{m2} using the statistic:

$$St\left(A_{m1}^t, A_{m2}^t\right) = \sqrt{2|O_t|}\frac{A_{m1}^t - A_{m2}^t}{\sqrt{A_{m1}^t + A_{m2}^t}\sqrt{1 - A_{m1}^t + 1 - A_{m2}^t}}, \tag{12}$$

which follows an asymptotic normal distribution as $|O_t| \to \infty$ and $A_{m1} = A_{m2}$. To test the null hypothesis $H_0 : A_{m1} \leqslant A_{m2}$ against the alternative $H_1 : A_{m1} > A_{m2}$, we define the P-value p_{St} as:

$$p_{St}\left(A_{m1}^t, A_{m2}^t\right) = 1 - F_{\mathcal{N}}\left(St\left(A_{m1}^t, A_{m2}^t\right)\right), \tag{13}$$

where $F_{\mathcal{N}}$ is the cumulative distribution function of the standard normal distribution. Rejecting H_0 at $p_{St} \leqslant \alpha$ controls the Type I error rate at α. By inverting (13) with respect to A_{m2}^t, we derive the range of A_{m2}^t values for which $p_{St} \leqslant \alpha$ at a fixed A_{m1}^t:

$$A_{m2}^t \in \left[0, \frac{2|O_t|A_{m1}^t - z_\alpha^2 A_{m1}^t + z_\alpha^2 - \sqrt{D}}{2|O_t| + z_\alpha^2}\right], \tag{14}$$

$$D = \left(2|O_t|A_{m1}^t - z_\alpha^2 A_{m1}^t + z_\alpha^2\right)^2 - A_{m1}^t\left(2|O_t| + z_\alpha^2\right)\left(2|O_t|A_{m1}^t - 2z_\alpha^2 + z_\alpha^2 A_{m1}^t\right). \tag{15}$$

Thus, if A^t_{m2} falls within the range (14), we can claim that model $m1$ has a statistically significant accuracy advantage over model $m2$ at the Type I error rate α.

We can also determine the required test set size $|O_t|$ such that the difference $A^t_{m1} > A^t_{m2}$ indicates a statistically significant advantage:

$$|O_t| \geqslant z^2_\alpha \frac{(A^t_{m1} + A^t_{m2})\,(1 - A^t_{m1} + 1 - A^t_{m2})}{2\,(A^t_{m1} - A^t_{m2})^2}. \tag{16}$$

This formula specifies the necessary test set size to assert that model m_1 has a statistically significant advantage over model m_2 based on their observed accuracy values, $A^t_{m1} > A^t_{m2}$.

3. Computational Experiments

We calculate the discriminability boundary using formula (14) for models evaluated on the MNIST [7–12, 15], CIFAR-10 [10, 13, 16–19], and CIFAR-100 [10, 14, 20–23] datasets. Tables 1–3 list the models, their accuracy metrics, and the highest accuracy values for which each model demonstrates a statistically significant advantage.

Table 1

Upper bound of the range (14) for models on MNIST

Model	Accuracy	Upper bound of the range (14) at Type I error rate of 0,05
[11]	0,9987	0,99772
[12]	0,9984	0,99733
[7]	0,9982	0,99707
[9]	0,9979	0,99669
[8]	0,9977	0,99644
[10]	0,9975	0,99620
[15]	0,9859	0,98302

Table 2

Upper bound of the range (14) for models on CIFAR-10

Model	Accuracy	Upper bound of the range (14) at Type I error rate of 0,05
[13]	0,99612	0,99453
[16]	0,995	0,99322
[17]	0,995	0,99322
[10]	0,9949	0,99310
[18]	0,9905	0,98811
[19]	0,984	0,98095

If a model's accuracy exceeds the value in the third column, the claim that the model in the second column has greater accuracy than the first is not statistically significant. From this comparison, we make the following conclusions:

1. On the MNIST dataset, the top model does not have a statistically significant advantage over its three closest competitors. The second-best model [12] has a significant advantage

only over the last model listed [15]. Thus, leading models on MNIST do not exhibit statistical superiority over one another. It is plausible that the 2016 model [9] with an accuracy of 0,9979, may outperform the 2021 model [11] with an accuracy of 0,9987, on a larger dataset.

2. A similar trend is observed on the CIFAR-10 dataset, where the top model [13] shows no statistically significant advantage over models [16], [17], and [10].

3. On CIFAR-100, model rankings are more robust. The top-1 model has a statistically significant advantage over the top-2 model, the top-2 over the top-3, and the top-3 over the top-5. Therefore, the current test set size for CIFAR-100 still supports statistically significant ranking by accuracy.

Table 3

Upper bound of the range (14) for models on CIFAR-100

Model	Accuracy	Upper bound of the range (14) at Type I error rate of 0,05
[14]	0,9608	0,95616
[20]	0,9510	0,94586
[10]	0,9495	0,94428
[21]	0,942	0,93644
[22]	0,9409	0,93529
[23]	0,9395	0,93383

To improve the statistical significance of rankings on MNIST, larger test sets are required. For example, to distinguish a model with 0,9987 accuracy from the one with 0.9979 accuracy (such as models [9] and [11]), at least 14 349 test instances are necessary (using formula (16) with $\alpha = 0,05$). For models like [11] and [12], an even larger test set – 87 053 instances – is required.

Conclusion

This study examines the statistical significance of evaluating and ranking recognition models, particularly as top models approach near-perfect accuracy on classical benchmarks, making it difficult to distinguish between them.

We formalized accuracy as the expected value of the correct response indicator, treating it as the sample mean. This approach enables conclusions to be drawn about model performance. Using the likelihood ratio test, we derived a formula to determine the required test set size to reliably classify a model as high-quality. This formula is useful in practical applications for designing datasets to ensure that models meet or exceed critical accuracy thresholds. For comparing two models, we introduced a method, using formula (14), to identify when one model has a statistically significant accuracy advantage over another.

Our computational experiments applied formula (14) to models evaluated on MNIST, CIFAR-10, and CIFAR-100 datasets. The results show that, for MNIST and CIFAR-10, the top models do not have statistically significant advantages over their nearest competitors. However, for CIFAR-100, the highest-ranked models demonstrate clear statistically significant advantages. From these observations, we can conclude that conducting competitions on classical datasets is irrelevant when the achieved accuracy on them is already sufficiently high. However, MNIST, CIFAR-10, and CIFAR-100 retain their significance for individual research.

References

1. Arlazarov V.L, Slavin O.A. Issues of Recognition and Verification of Text Documents. *Intelligent Systems AND Technologies*, 2023, no. 3, pp. 55–61. DOI: 10.14357/20718632230306
2. Kunina I.A., Sher A.V, Nikolaev D.P. Screen Recapture Detection Based on Color-Texture Analysis of Document Boundary Regions. *Computer Optics*, 2023, vol. 47, no. 4, pp. 650–657. DOI: 10.18287/2412-6179-CO-1237
3. Gayer A.V. Context-Independent Fast Text Detection Method for Recognizing Phone Numbers. *Proceedings of ISA RAS*, 2024, vol. 74, no. 3, pp. 39–47. DOI: 10.14357/20790279240305
4. Maksimova T.R., Bulatov K.B. Reducing Errors and Computational Load in Road Scene Text Recognition. *Intelligent Systems AND Technologies*, 2024, no. 3, pp. 1–15. DOI: 10.14357/20718632240301
5. Deng Li. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 2012, vol. 29, no. 6, pp. 141–142. DOI: 10.1109/MSP.2012.2211477
6. Krizhevsky A., Hinton G. *Learning Multiple Layers of Features From Tiny Images*. Toronto, University of Toronto, 2009.
7. Kowsari K., Heidarysafa M., Brown D.E., Meimandi K.J., Barnes L.E. RMDL: Random Multimodel Deep Learning for Classification. *Proceedings of the 2nd International Conference on Information System and Data Mining*, New York, 2018, pp. 19–28. DOI: 10.1145/3206098.3206111
8. Ciregan D., Meier U., Schmidhuber J. Multi-Column Deep Neural Networks for Image Classification. *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 2012, pp. 3642–3649. DOI: 10.1109/CVPR.2012.6248110
9. Romanuke V. Training Data Expansion and Boosting of Convolutional Neural Networks for Reducing the MNIST Dataset Error Rate. *Research Bulletin of the National Technical University of Ukraine "Kyiv Politechnic Institute"*, 2016, no. 6, pp. 29–34. DOI: 10.20535/1810-0546.2016.6.84115
10. Gesmundo A., Dean J. An Evolutionary Approach to Dynamic Introduction of Tasks in Large-Scale Multitask Learning Systems. *arXiv: Machine Learning*, 2022. Avialable at: https://arxiv.org/abs/2205.12755. DOI: 10.48550/arXiv.2205.12755
11. Byerly A., Kalganova T., Dear I. No Routing Needed Between Capsules. *Neurocomputing*, 2021, vol. 463, pp. 545–553. DOI: 10.1016/j.neucom.2021.08.064
12. Hirata D., Takahashi N. Ensemble Learning in CNN Augmented with Fully Connected Subnetworks. *IEICE Transactions on Information and Systems*, 2023, vol. 106, no. 7, pp. 1258–1261. DOI: 10.1587/transinf.2022EDL8098
13. Bruno A., Moroni D., Martinelli M. Efficient Adaptive Ensembling for Image Classification. *arXiv: Computer Vision and Pattern Recognition*, 2022. Avialable at: https://arxiv.org/abs/2206.07394. DOI: /10.48550/arXiv.2206.07394
14. Foret P., Kleiner A., Mobahi H., Neyshabur B. Sharpness-Aware Minimization for Efficiently Improving Generalization. *arXiv: Machine Learning*, 2020. Avialable at: https://arxiv.org/abs/2010.01412. DOI: 10.48550/arXiv.2010.01412
15. Gehring J., Auli M., Grangier D., Yarats D., Dauphin Y.N. Convolutional Sequence to Sequence Learning. *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1243–1252. DOI: 10.48550/arXiv.1705.03122

16. Dosovitskiy A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv: Computer Vision and Pattern Recognition*, 2020. Avialable at: https://arxiv.org/abs/2010.11929. DOI: 10.48550/arXiv.2010.11929

17. Oquab M., Darcet T., Moutakanni T., Huy Vo, et al. Dinov2: Learning Robust Visual Features without Supervision. *arXiv: Computer Vision and Pattern Recognition*, 2023. Avialable at: https://arxiv.org/abs/2304.07193. DOI: 10.48550/arXiv.2304.07193

18. Kabir H.M. Reduction of Class Activation Uncertainty with Background Information. *arXiv: Computer Vision and Pattern Recognition*, 2023. Avialable at: https://arxiv.org/abs/2305.03238. DOI: 10.48550/arXiv.2305.03238

19. Zhichao Lu, Sreekumar G., Goodman E., Banzhaf W., Deb K., Boddeti V.N. Neural Architecture Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, vol. 43, no. 9, pp. 2971–2989. DOI: 10.1109/TPAMI.2021.3052758

20. Ridnik T., Sharir G., Ben-Cohen A., Ben-Baruch E., Noy A. Ml-Decoder: Scalable and Versatile Classification Head. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 32–41. DOI: 10.1109/WACV56688.2023.00012

21. Ridnik T., Ben-Baruch E., Noy A., Zelnik-Manor L. Imagenet-21k Pretraining for the Masses. *arXiv: Computer Vision and Pattern Recognition*, 2021. Avialable at: https://arxiv.org/abs/2104.10972. DOI: 10.48550/arXiv.2104.10972

22. Haiping Wu, Bin Xiao, Codella N., Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang. CVT: Introducing Convolutions to Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31. DOI: 10.1109/ICCV48922.2021.00009

23. Ching-Hsun Tseng, Liu-Hsueh Cheng, Shin-Jye Lee, Xiaojun Zeng. Perturbed Gradients Updating Within Unit Space for Deep Learning. *IEEE International Joint Conference on Neural Networks*, 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9892245

Received December 24, 2024

УДК 519.248 **DOI: 10.14529/mmp250209**

ОЦЕНКА РЕЛЕВАНТНОСТИ ТЕСТИРОВАНИЯ И РАНЖИРОВАНИЯ МОДЕЛЕЙ В ЗАВИСИМОСТИ ОТ ОБЪЕМА ДАТАСЕТА

А.В. Чуйко[1]***, В.В. Арлазаров***[1,2]***, С.А. Усилин***[1,2]
[1]Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Российская Федерация
[2]ООО «Смарт Энджинс Сервис», г. Москва, Российская Федерация

Методы машинного обучения все чаще используются в различных областях жизнедеятельности. Ежегодно множество научных коллективов разрабатывают новые

распознающие модели, соревнуясь при этом в показателях качества на открытых датасетах. В некоторых задачах показатели точности давно превысили 99%, при этом лучшие в таблице ранжирования модели зачастую отличаются между собой на сотые доли процентов. Принимая в расчет объемы датасетов, резонным становится вопрос о релевантности оценки качества и достоверности ранжирования различных распознающих моделей. В работе описан метод расчета необходимого объема датасета для возможности корректной проверки гипотезы о точности модели, а также представлен анализ статистической значимости ранжирования по точности некоторых современных работ на датасетах MNIST, CIFAR-10 и CIFAR-100.

Ключевые слова: объем датасета; распознавание объектов; статистическая значимость; оценка качества модели; оценка качества распознавания.

Литература

1. Арлазаров, В.Л. Вопросы распознавания и верификации текстовых документов / В.Л. Арлазаров, О.А. Славин // Информационные технологии и вычислительные системы. – 2023. – № 3. – С. 55–61.
2. Kunina, I.A. Screen Recapture Detection Based on Color-Texture Analysis of Document Boundary Regions / I.A. Kunina, A.V. Sher, D.P. Nikolaev // Computer Optics. – 2023. – V. 47, № 4. – P. 650–657.
3. Гайер, А.В. Контекстно-независимый метод быстрой детекции текста для распознавания номеров телефонов / А.В. Гайер // Труды института системного анализа РАН. – 2024. – Т. 74, № 3. – С. 39–47.
4. Максимов, Т.Р. Снижение ошибки и вычислительной нагрузки в распознавании текста дорожной сцены / Т.Р. Максимов, К.Б. Булатов // Информационные технологии и вычислительные системы. – 2024. – № 3. – С. 1–15.
5. Deng Li. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web] / Li Deng // IEEE Signal Processing Magazine. – 2012. – V. 29, № 6. – P. 141–142.
6. Krizhevsky, A. Learning Multiple Layers of Features From Tiny Images / A. Krizhevsky, G. Hinton. – Toronto: University of Toronto, 2009.
7. Kowsari, K. RMDL: Random Multimodel Deep Learning for Classification / K. Kowsari, M. Heidarysafa, D.E. Brown, K.J. Meimandi, L.E. Barnes // Proceedings of the 2nd International Conference on Information System and Data Mining. – New York, 2018. – P. 19–28.
8. Ciregan, D. Multi-Column Deep Neural Networks for Image Classification / D. Ciregan, U. Meier, J. Schmidhuber // IEEE Conference on Computer Vision and Pattern Recognition. – Providence, 2012. – P. 3642–3649.
9. Romanuke, V. Training Data Expansion and Boosting of Convolutional Neural Networks for Reducing the MNIST Dataset Error Rate / V. Romanuke // Research Bulletin of the National Technical University of Ukraine “Kyiv Politechnic Institute”. – 2016. – № 6. – P. 29–34.
10. Gesmundo, A. An Evolutionary Approach to Dynamic Introduction of Tasks in Large-Scale Multitask Learning Systems / A. Gesmundo, J. Dean // arXiv: Machine Learning. – 2022. – URL: https://arxiv.org/abs/2205.12755.
11. Byerly, A. No Routing Needed Between Capsules / A. Byerly, T. Kalganova, I. Dear // Neurocomputing. – 2021. – V. 463. – P. 545–553.
12. Hirata, D. Ensemble Learning in CNN Augmented with Fully Connected Subnetworks / D. Hirata, N. Takahashi // IEICE Transactions on Information and Systems. – 2023. – V. 106, №. 7. – P. 1258–1261.

13. Bruno, A. Efficient Adaptive Ensembling for Image Classification / A. Bruno, D. Moroni, M. Martinelli // arXiv: Computer Vision and Pattern Recognition. – 2022. – URL: https://arxiv.org/abs/2206.07394
14. Foret, P. Sharpness-Aware Minimization for Efficiently Improving Generalization / P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur // arXiv: Machine Learning. – 2020. – URL: https://arxiv.org/abs/2010.01412
15. Gehring, J. Convolutional Sequence to Sequence Learning / J. Gehring, M. Auli, D. Grangier, D. Yarats, Y.N. Dauphin // Proceedings of the 34th International Conference on Machine Learning. – 2017. – P. 1243–1252.
16. Dosovitskiy, A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale / A. Dosovitskiy // arXiv: Computer Vision and Pattern Recognition. – 2020. – URL: https://arxiv.org/abs/2010.11929
17. Oquab, M. Dinov2: Learning Robust Visual Features without Supervision / M. Oquab, T. Darcet, T. Moutakanni, Vo Huy, et al. // arXiv: Computer Vision and Pattern Recognition. – 2023. – URL: https://arxiv.org/abs/2304.07193
18. Kabir, H.M. Reduction of Class Activation Uncertainty with Background Information / H.M. Kabir // arXiv: Computer Vision and Pattern Recognition. – 2023. – URL: https://arxiv.org/abs/2305.03238
19. Zhichao Lu. Neural Architecture Transfer / Lu Zhichao, G. Sreekumar, E. Goodman, W. Banzhaf, K. Deb, V.N. Boddeti // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2021. – V. 43, № 9. – P. 2971–2989.
20. Ridni, T. Ml-Decoder: Scalable and Versatile Classification Head / T. Ridnik, G. Sharir, A. Ben-Cohen, E. Ben-Baruch, A. Noy // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. – 2023. – P. 32–41.
21. Ridnik, T. Imagenet-21k Pretraining for the Masses / T. Ridnik, E. Ben-Baruch, A. Noy, L. Zelnik-Manor // arXiv: Computer Vision and Pattern Recognition. – 2021. – URL: https://arxiv.org/abs/2104.10972
22. Haiping Wu. CVT: Introducing Convolutions to Vision Transformers / Wu Haiping, Bin Xiao, N. Codella, Liu Mengchen, Dai Xiyang, Lu Yuan, Lei Zhang // Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2021. – P. 22–31.
23. Ching-Hsun Tseng. Perturbed Gradients Updating Within Unit Space for Deep Learning / Tseng Ching-Hsun, Liu-Hsueh Cheng, Shin-Jye Lee, Xiaojun Zeng // IEEE International Joint Conference on Neural Networks. – 2022. – P. 1–8.

Александр Васильевич Чуйко, аспирант, федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация), a.chuyko@smartengines.com.

Владимир Викторович Арлазаров, доктор технических наук, генеральный директор, ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация); заведующий отделом, федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация), vva@smartengines.com.

Сергей Александрович Усилин, кандидат технических наук, исполнительный директор, ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация); старший научный сотрудник, федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация), usilin@smartengines.com.

Поступила в редакцию 24 декабря 2024 г.