ISSN 2519-2310 (Online)

КОМП'ЮТЕРНІ НАУКИ ТА КІБЕРБЕЗПЕКА, 2024, № 2 (26)

COMPUTER SCIENCE AND CYBERSECURITY, 2024, Iss. 2 (26)

DOI: https://doi.org/10.26565/2519-2310-2024-2-02
УДК 004.032.26

PROBLEMS OF THE NEURAL NETWORKS OUTPUT DATA QUALITY ASSESSMENT

Yurii Halaichuk[1], PhD student, department of theoretical and applied systems engineering, e-mail: yurii.halaichuk@student.karazin.ua, ORCID: https://orcid.org/0009-0004-1048-9425
Maryna Miroshnyk[1], Doctor of Technology, professor, professor of the department of theoretical and applied systems engineering, e-mail: m.miroshnyk@karazin.ua, ORCID: https://orcid.org/0000000222312529
[1]*V. N. Karazin Kharkiv National University, Svobody Square, 4, Kharkiv, 61022, Ukraine*
Manuscript was received October 20, 2024; Received after review November 23, 2024; Accepted December 13, 2024

Abstract: Today, artificial intelligence, particularly neural networks, is increasingly being used in software in a variety of industries, from mission-critical applications such as healthcare and the military to commerce and entertainment. One of the main stages of development and implementation of such software is the stage of quality control. To prevent fatal errors and to survive in a highly competitive environment, the software needs proper testing taking into account the peculiarities inherent in the data obtained as a result of the neural network. This article presents the relevance of using artificial intelligence systems in general and neural networks in particular and analyzes the main challenges that arise when assessing the quality of such networks. The author compares the properties of the output data of the artificial intelligence systems of the previous generation and the latest neural networks, highlights the key differences of the latter, such as the potential infinity of the input data sets and their relative unpredictability, the dependence of the results on the network training stage, and the subjective nature of the evaluation of such results. Based on the analysis, the author formulates a set of problems that can be solved using mathematical algorithms and methods. The main part of an article contains a general overview of existing solutions, with an emphasis on such algorithms and methods as calculating accuracy and loss, finding the F-score, interpretation methods and imitation modeling. As a result of the research, the author comes to the conclusion that, despite a sufficient number of existing solutions that can be used to solve the highlighted problems, they still have to be improved to increase the accuracy of neural network evaluation, as one hundred percent accuracy in evaluating data obtained as a result of the operation of neural networks has not yet been achieved.

Keywords: *artificial intelligence, expert systems, F-score, loss function, neural networks, quality assessment*

In cites: Halaichuk Y., Miroshnyk M. (2024). Problems of the neural networks output data quality assessment. *Computer Science and Cybersecurity.* 2(26): 19–24. https://doi.org/10.26565/2519-2310-2024-2-02

1. Introduction

According to Statista.com, the global artificial intelligence (AI) market is expected to reach nearly $126 billion by 2025, up from just $10.1 billion in 2016. There are various information systems and software that use neural networks algorithms to process data and generate results. Examples of such software include services for finding similar images, navigation devices for finding the most optimal route, dating sites, text generators that use machine learning, computer games, and many others.

Evaluation of neural networks is a very relevant and timely topic given the growing integration of AI into various aspects of our lives. Neural networks are becoming an important tool in decision-making, creative endeavors and problem-solving. As society increasingly relies on the results of neural networks, understanding how to evaluate their performance becomes critical. Currently, quite effective methods, models, and techniques for assessing software quality are widely used, including such software quality indicators as functionality, reliability, usability, efficiency, mobility, interactivity, etc. However, neural networks have peculiarities that should be considered separately.

2. Setting of the problem and the aim of the article

One of the peculiarities of neural networks is that unlike other algorithms, where the decision-making logic is often directly coded and is transparent, as in the case of expert and scoring systems, neural networks operate as complex, interconnected layers of nodes whose internal workings are hidden from data assessors.

One of the most important challenges in assessing neural networks is their ability to generate an infinite variety of outputs. Unlike traditional algorithms that produce a fixed set of numerical or categorical results, neural networks can generate diverse outputs such as text, images, music, or even complex behaviors in games. For instance, large language models (LLMs) can produce countless variations of a story or response to a prompt, each differing in tone, style, or content. This infinite output set complicates assessment because it is impossible to evaluate every possible result.

Another significant challenge is the variability of neural network outputs based on their training stage. Neural networks are iterative learners, meaning their performance evolves as they are exposed to more data or undergo additional training. Consequently, the same input can produce different outputs depending on when the model is evaluated. For example, a partially trained image recognition model might misclassify a cat as a dog, while a fully trained version of the same model correctly identifies the cat. This variability makes it difficult to assess the quality of a neural network consistently. The dependence on the training stage introduces uncertainty in evaluation. A model that performs well during one phase of training might degrade or behave unpredictably later due to overfitting, underfitting, or changes in the training data. This is particularly problematic for applications requiring stable outputs, such as medical diagnosis or autonomous driving, where inconsistent results could have serious consequences.

The next complex problem is the lack of clearly defined criteria for assessing the quality of creative or non-numerical outputs, such as text, images, or game behaviors. In creative tasks such as art, literature, or music, where subjectivity plays a significant role, the understanding of a single correct result becomes unclear. This subjectivity introduces a significant human factor into the assessment process. Human evaluators often disagree on what is a correct output, leading to inconsistent and biased evaluations. For instance, one tester might approve a generated image for its aesthetics, while another critiques it for lacking realism. This raises profound questions about how we define success and failure in the context of neural network-generated content and requires a shift to more nuance-oriented evaluation criteria.

To understand the complexity of testing neural networks in relation to knowledge-based systems, such as expert systems, the following comparison of their output data can be made:

Table 1. Comparison of output data properties of expert systems and neural networks

Output data properties	Expert systems	Neural networks
Potential quantity	Determined by the size of the knowledge base or is a range of numbers	Endless
Uniqueness	All possible sets of output data or possible combinations are known	Each set of output data can be unique and unpredictable
Dependence on input data	The output data depends on the input data according to a known algorithm	The output data depends on the input data according to the logic used by the algorithm at its current learning stage
Complexity of the assessment	Data accuracy can be assessed objectively	Assessment of data accuracy is partly subjective

Thus, we can identify the following key issues that arise when evaluating the performance of neural networks:

- *the potential number of output results (including results that are not a set of numbers) can be infinite;*
- *output data depend on the algorithm training - they can change with the same input data, which makes the assessment of the quality of such an algorithm dependent on its learning stage;*
- *there are no clearly defined criteria for the quality of such output data as, for example, creative text, images, and game behavior, which leads to a large share of human factors in assessment of such data.*

3. Existing solutions

For example, to test expert systems and obtain accuracy which is close to 100% it is sufficient to use the method of boundary values and the method of equivalent classes, which allow covering the entire range of possible outcomes.

However, the difficulties associated with evaluating the performance of neural networks, emphasize the need for a more systematic approach. Mathematical methods make it possible to quantify and analyze the intricacies of neural networks, offering a more objective and reproducible assessment.

3.1. The most essential approach to obtaining quality metrics of neural networks performance is the calculation of accuracy and loss function values.

Accuracy represents the proportion of correctly classified or predicted instances out of the total number of samples in the test dataset. It offers a straightforward and easily interpretable measure of how well the neural network generalizes to unseen data. A high accuracy score suggests that the model is making correct predictions for a significant portion of the input, showing confidence in its ability to perform well in real-world applications. For tasks like image classification, where the goal is

to assign a single correct label to each image, accuracy serves as a natural and intuitive benchmark. A 95% accuracy, for instance, directly translates to the model correctly identifying 95 out of every 100 images.

$A = \frac{N(correct)}{N}$, where N is the amount of test samples (1)

The loss function quantifies the discrepancy between the network's predictions and the actual ground truth labels. It provides a more granular and continuous measure of the model's performance, reflecting the degree of error in its predictions. Unlike accuracy, which focuses on discrete correctness, the loss function captures the confidence and correctness of each individual prediction. Different types of loss functions are used depending on the specific task and the neural network architecture.

The general rule of application of obtained accuracy and loss values to assess the quality of neural network output is described in Table 2:

Table 2. Relation of Accuracy / Loss values to overall model performance

	Low Loss	High Loss
High Accuracy	The model correctly predicts most instances with high confidence.	The model makes correct predictions with low confidence or incorrect predictions have large errors.
Low Accuracy	The model consistently predicts values close to the true values but does not cross a classification threshold correctly.	The model makes many incorrect predictions with significant errors.

3.2. The F1 score method provides a single, robust metric for evaluating classification models, especially in situations with imbalanced classes. By considering both the accuracy of positive predictions and the ability to identify all positive instances, it offers a more insightful assessment than accuracy alone, allowing to build more reliable and effective classification systems.

The algorithm assigns a score weight to all the output data and emphasizes the following criteria:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}, \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}, \quad (3)$$

where:

- *True Positives* is the number of data samples correctly predicted as "positive";
- *False Positives* is the number of data samples wrongly predicted as "positive";
- *True Negatives* is the number of data samples correctly predicted as "negative";
- *False Negatives* is the number of data samples wrongly predicted as "negative".

Then F1 score is being calculated:

$$F1\ score = 2 * \frac{Precision * Recall}{(Precision + Recall)}, \quad (4)$$

which will provide the percentage values of positive results for the investigated model. A high F1 score indicates that the model has both high precision and high recall, signifying a well-balanced classifier.

3.3. Another approach is to use model interpretation methods to explain the logic which was used during the prediction of output results. The most widely used methods are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which are based on game theory and local surrogate models respectively and allow to conduct black-box testing of neural network algorithms.

LIME focuses on explaining individual predictions by approximating the complex black-box model with a more simple, interpretable model (like a linear model) to analyze the specific data point. It is relatively fast and easy to understand, but its explanations can be unstable depending on the perturbation and sampling methods, and it only provides a local view without guarantees of global consistency.

SHAP provides a more theoretical approach using game theory, specifically Shapley values. It aims to quantify the contribution of each feature to a specific prediction by considering all possible combinations. SHAP offers stronger theoretical guarantees and more consistent explanations, but at the cost of execution speed and simplicity.

3.4. The capabilities of imitational modeling allow testing of neural networks in a similar way it is done with static data using the combination of boundary values and equivalent classes methods. Such methods as Synthetic Data Generation, Environment Simulation, Adversarial Attack Simulation, Monte Carlo Simulation can perform testing covering many non-standard cases, however, they require the creation of new data sets for testing and are more appropriate for using the development flow and unit testing of neural networks.

The algorithms and methods described above can help to improve the quality of neural network output data evaluation and require further research to improve the accuracy of that evaluation

4. Conclusions

To solve the problems associated with evaluating neural networks, the integration of mathematical methods is a crucial component. Quantitative metrics, algorithms, and mathematical models provide a structured framework for objective assessment of the quality and performance of neural networks.

Despite a sufficient number of existing methods and algorithms that can be used to perform tasks of evaluation of the quality of neural networks, they still need further improvement, as they do not yet allow for 100% accuracy in forming an assessment of the output data of AI algorithms.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Giesecke K., Horel E., (2020) Significance Tests for Neural Networks by Enguerrand Horel. *Journal of Machine Learning Research.* July 20, 2020. 1-29. https://jmlr.csail.mit.edu/papers/volume21/19-264/
2. Saurabh Ranjan Srivastava., (2016) F1 Score Analysis of Search Engines. *Skit Research Journal*, 2(6)
3. Ponakala R., Dailey M., (2019). Testing Deep Neural Networks for Classification Tasks Through Adversarial Perturbations on Test Datasets. December, 2019. Asian Institute of Technology, School of Engineering and Technology, Thailand. https://doi.org/10.31237/osf.io/r7wcn

4. Christian Terwiesch, (2023) Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023. URL: https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch/
5. ISO/IEC/IEEE 29119-1:2013. Software and Systems Engineering. Software Testing Part 1: Concepts and Definitions. Geneva: International Organization for Standardization, 2013
6. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30, с. 4765-4774. DOI: https://doi.org/10.48550/arXiv.1705.07874
7. Russell S., Norvig P. Artificial Intelligence: A Modern Approach: підручник, 4th ed, Boston: Pearson, 2020, 1152 с.
8. Zhang J. M., Harman M., Ma L., Liu Y. Machine Learning Testing: Survey, Landscapes and Horizons // IEEE Transactions on Software Engineering. 2020, Vol. 48 No. 1, с. 1-36. DOI: https://doi.org/10.1109/TSE.2019.2962027
9. Hossain E. Machine Learning Crash Course for Engineers / Eklas Hossain., 2024. - 453 с.

ПРОБЛЕМИ ОЦІНЮВАННЯ ВИХІДНИХ ДАНИХ НЕЙРОННИХ МЕРЕЖ

Юрій Галайчук[1], аспірант кафедри теоретичної та прикладної системотехніки;
e-mail: yurii.halaichuk@student.karazin.ua; ORCID: https://orcid.org/0009-0004-1048-9425
Марина Мірошник[1], доктор технічних наук, професор; професор кафедри теоретичної та прикладної системотехники; e-mail: m.miroshnyk@karazin.ua;
ORCID: https://orcid.org/0000000222312529
[1]*Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків, 61022, Україна*
Рукопис надійшов 20 жовтня 2024 р. Отримано після рецензування 23 листопада 2024 р.
Прийнято 13 грудня 2024 р.

Анотація: Станом на сьогодні штучний інтелект, зокрема нейронні мережі, все більше використовується у програмному забезпеченні в різних галузях, від критично важливих, таких як охорона здоров’я та військова справа, до комерції та сфери розваг. Одним із основних етапів розробки та впровадження такого програмного забезпечення є етап контролю якості. Для запобігання фатальних помилок та втримання у надто конкурентному середовищі програмне забезпечення потребує належного тестування з урахуванням особливостей даних, що отримані у результаті роботи нейронної мережі. У статті наводиться актуальність використання нейронних мереж та аналізуються основні виклики, що виникають при оцінюванні якості таких мереж. Автор робить порівняння властивостей вихідних даних систем штучного інтелекту минулого покоління та новітніх нейронних мереж, виділяє ключові відмінності останніх, такі як потенційна нескінченність наборів вихідних даних, залежність результатів від етапу навчання мережі та суб’єктивну природу оцінювання таких результатів. Основний зміст складає узагальнений огляд існуючих рішень, з акцентом на такі алгоритми та методи, як розрахунок функції втрат, знаходження F-score, методи інтерпретації та імітаційне моделювання. У результаті дослідження автор приходить до висновку, що, не зважаючи на достатню кількість наявних рішень, вони все ще потребують вдосконалення для підвищення точності оцінювання нейронних мереж, адже стовідсоткова точність оцінювання даних все ще не досягнута.

Ключові слова: *F-score, експертні системи, нейронні мережі, оцінювання якості, функція втрат, штучний інтелект*

Конфлікт інтересів: автори повідомляють про відсутність конфлікту інтересів.