МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
КОРАБЛЕБУДУВАННЯ ІМЕНІ АДМІРАЛА МАКАРОВА
МИКОЛАЇВСЬКА ОБЛАСНА ДЕРЖАВНА АДМІНІСТРАЦІЯ
ДЕПАРТАМЕНТ ОСВІТИ І НАУКИ
МИКОЛАЇВСЬКОЇ ОБЛАСНОЇ ДЕРЖАВНОЇ АДМІНІСТРАЦІЇ

ІННОВАЦІЇ В СУДНОБУДУВАННІ ТА ОКЕАНОТЕХНІЦІ

XVI Міжнародна науково-технічна конференція

МАТЕРІАЛИ

25–26 вересня 2025 рік

Національний університет кораблебудування
імені адмірала Макарова
просп. Героїв України, 9

ВИДАВНИЦТВО
НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ
КОРАБЛЕБУДУВАННЯ
ІМ. АДМІРАЛА МАКАРОВА
2025

УДК 004.032.26:004.89

DYNAMIC VALIDATION AND RELEVANCE MAINTENANCE OF NEURAL CORRELATES UNDER CONTINUAL LEARNING

Verbytskyi O.S.
M.Sc. Student in Computer Science
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine
soomoalex@gmail.com
ORCID: 0009-0000-6848-8880
Gaidaienko O.V.
Ph.D. in Technology, Associate Professor,
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine
oksana.gaidaienko@nuos.edu.ua
ORCID: 0000-0002-6614-5443

Abstract. Deep neural networks trained in static environments suffer from model degradation when faced with evolving data distributions (concept drift). This study proposes a framework for dynamic validation and targeted maintenance of concept-specific neurons, addressing the continual learning challenge. Building on L0-optimization for neuron localization, we introduce a "Concept Adequacy Score" to detect knowledge decay. Upon detecting drift, a localized update mechanism fine-tunes only the minimal neuron set responsible for the concept, avoiding costly full-network retraining. Experiments simulating a continual learning scenario show that our approach maintains high classification accuracy over time, reducing computational costs by an order of magnitude compared to full retraining, thereby enabling efficient and adaptive AI systems.

Keywords: continual learning, concept drift, explainable AI, model degradation, localized fine-tuning, neural correlates.

Introduction. Deep Neural Networks (DNNs) have demonstrated outstanding performance, yet their static nature makes them vulnerable to degradation in dynamic environments. A model trained on a specific data distribution rapidly loses its relevance when that distribution changes – a phenomenon known as "concept drift" [1]. Previous work demonstrated a method for precisely localizing minimal sets of neurons responsible for specific concepts (e.g., "cat"), which was a step toward solving the "black box" problem.

However, localized knowledge is not permanent on its own. The correlations learned by neurons can become outdated. The traditional approach to this problem—full network retraining on new data – is computationally expensive and inefficient, often leading to the "catastrophic forgetting" of previously acquired knowledge [2].

The goal of this research is to create a framework for the regular assessment of the adequacy of knowledge encoded in conceptual neurons and their "targeted" updating without full model retraining. We propose:

1. A validation module, based on an adequacy metric, to track the relevance of neural correlates over time.

2. A localized update mechanism that, upon detecting drift, fine-tunes only the identified group of neurons, preserving the rest of the network.

This work bridges Explainable AI (XAI) with Continual Learning, offering a path toward creating adaptive and resource-efficient intelligent systems [3].

Materials and Methods. Problem Formulation – the task is to adapt a model M with parameters Θ to a changing data stream D_t. Initially, using L0-optimization on data D_0 [4], a set of concept-specific neurons S_c is identified. The problem is to maintain the relevance of S_c over time with minimal computational cost. We formalize this task through two components:

1. Drift Detection. The relevance of the set S_c on new data D_t is measured by the **Concept Adequacy Score (A)**:

$$A(S_c, D_t) = \mathbb{E}_{x \in D_t}\left[P\left(c \middle| M(x, S_c)\right)\right]$$

Drift is considered detected if $A(S_c, D_t)$ falls below a threshold τ. This metric serves as the foundation of the validation module.

2. Localized Adaptation. When drift is detected, an update is initiated. Instead of minimizing the loss over all parameters Θ, we solve the following optimization problem:

$$\min_{\Theta(S_c)} \mathbb{E}_{(x,y) \in D_t}\left[\mathcal{L}\left(M\left(x; \Theta_{\text{frozen}}, \Theta(S_c)\right), y\right)\right]$$

Here, $\Theta(S_c)$ are the parameters directly associated with the neurons in S_c, and Θ_{frozen} are all other parameters of the network. This process constitutes the "targeted fine-tuning" mechanism.

Algorithm Implementation. The dynamic adaptation algorithm is presented in Fig. 1.

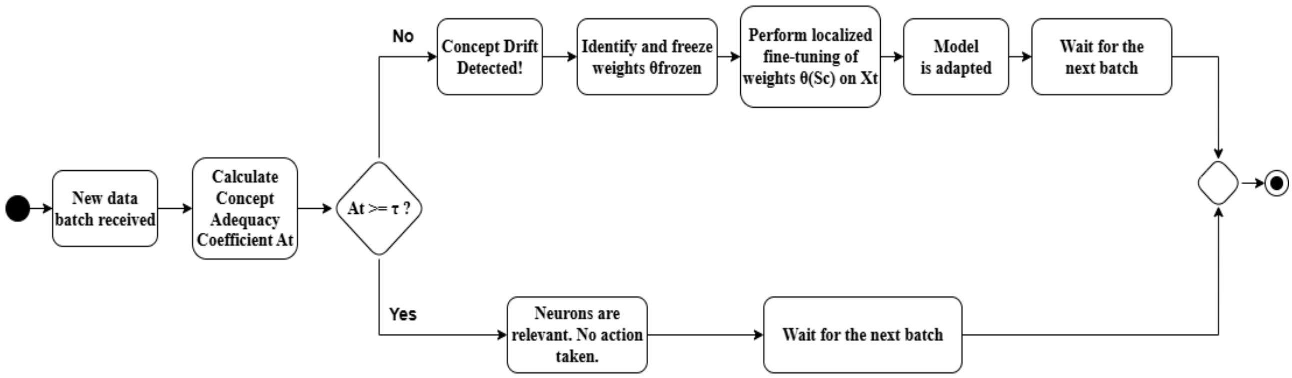

Fig. 1: Flowchart of the dynamic validation and adaptation algorithm

Validation Module. The module continuously tracks the distribution of activations and confidence scores for key neurons on new data, implementing the statistical test described above. Alternatively, for critical stages, verification can involve a human-in-the-loop.

Targeted Fine-Tuning Mechanism. After the validation module is triggered, the system freezes the weights of all layers except for those containing the neurons in S_c and possibly one adjacent layer. Then, a few epochs of gradient descent are performed on the new data batch to update only this small subset of weights.

Experimental Validation. Experimental Setup. We simulate a continual learning scenario. **Model:** VGG16, pretrained on ImageNet.

– **Stage 1 (Initial Training).** The model is trained on images of a single class, for example, "Persian cats." A set of neurons S_{cat} is identified.
– **Stage 2 (Drift 1 – New Subclasses).** Images of other cat breeds (e.g., Sphynx, Maine Coon) are gradually introduced into the data stream.
– **Stage 3 (Drift 2 – New Classes).** Images of other, similar objects (e.g., small dogs, foxes) are added to the stream.

Compared Approaches

– **Static Model.** No updates are performed.
– **Full Retraining.** The model is completely retrained on each new dataset.
– **Proposed Framework (Localized Update).** The model is updated using our algorithm.

Metrics: classification accuracy over time, computational cost per update (measured in GFLOPS – billion floating-point operations per second).

Results and Discussion. The experimental results are summarized in Table 1.

Table 1 – Comparative Performance Results

Approach	Accuracy (Stage 1)	Accuracy (Stage 2)	Accuracy (Stage 3)	Update Cost (GFLOPS)
Static Model	98.1%	75.4%	51.2%	0
Full Retraining	98.1%	97.8%	96.5%	~15.5 / stage
Our Framework	98.1%	96.2%	93.1%	~0.8 / stage

Fig. 2. Accuracy dependency on time (stages) for the three approaches.

Key Findings

Adaptation Efficiency. The static model degrades rapidly. Our framework successfully adapts to concept drift, maintaining a high level of accuracy (a drop of < 4% compared to full retraining).

Resource Efficiency. The proposed method reduces the computational cost of updates by over **90%** compared to full retraining, making it suitable for real-time systems.

Conclusions. This study offers a solution to the problem of model degradation in dynamic environments by introducing a framework for the dynamic validation and localized updating of concept-specific neurons. The main contributions of this work are: formalization of Drift at the Neuron Level – we introduced the "Concept Adequacy Score" metric to quantify the relevance of neural correlates, efficient Adaptation Algorithm – we developed a targeted fine-tuning mechanism that allows the model to adapt to new data with minimal computational costs, avoiding full retraining. The proposed approach is a significant step toward creating more reliable, adaptive, and interpretable AI systems capable of operating in real-world, ever-changing conditions.

References

[1] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 1–37.

[2] Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(12), 2935-2947.

[3] Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. *Advances in Neural Information Processing Systems, 30.*

[4] Louizos, C., Welling, M., & Kingma, D. P. (2018). Learning sparse neural networks through L0 regularization. *International Conference on Learning Representations (ICLR)*.

УДК 004.658.2:004.853

AUTOMATED RULE GENERATION FOR OPTIMAL DATA SELECTION IN ROBUST NEURAL NETWORK TRAINING

Verbytskyi O.S.
M.Sc. Student in Computer Science
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine
soomoalex@gmail.com
ORCID: 0009-0000-6848-8880
Gaidaienko O.V.
Ph.D. in Technology, Associate Professor,
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine
oksana.gaidaienko@nuos.edu.ua
ORCID: 0000-0002-6614-5443

Abstract. This study tackles the bottleneck of manual data curation by proposing a framework for automated data selection. We train a meta-model to quantify the "utility" of each data point using key metrics like entropy and representativeness. This model then generates human-readable rules to filter noisy datasets, creating optimized subsets for training. The goal is to simultaneously improve the final accuracy, robustness, and convergence speed of neural networks.

Keywords: data selection, data curation, meta-learning, active learning, curriculum learning, robust models, neural networks.

Introduction. Deep Neural Networks (DNNs) have achieved state-of-the-art performance across numerous domains, but their success is intrinsically tied to the data they are trained on. The principle of "garbage in, garbage out" is particularly salient in this context, where noisy, redundant, or uninformative data can lead to suboptimal models with poor generalization capabilities [1]. This paper introduces a framework to automate the generation of data selection rules for building more robust and efficient neural networks. We propose a system that moves beyond simple data cleaning and instead intelligently filters a large dataset to find the most "useful" samples for training. Our core contribution is a meta-model that learns to predict the value of each data point and subsequently synthesizes this knowledge into explicit filtering rules [2]. By training on a smaller, more information-rich subset of data, we aim to achieve three primary goals:

1. Maximize Model Quality. Improve the final accuracy and generalization of the target model.

СЕКЦІЯ 7. ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА УПРАВЛІННЯ ПРОЕКТАМИ